Essays on Predictability & Excess Profitability of Quantitative Methods: Modelling Implied Volatility, Technical Trading, Data Snooping and Market Efficiency

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy
by
Ioannis Psaradellis

October 2017

ACKNOWLEDGEMENTS

I would like to record my deepest gratitude to my advisors, Mr Jason Laws, Dr Athanasios Pantelous and Dr Georgios Sermpinis, for their three years of extraordinary guidance, helpful comments and suggestions and unconditional support. Without them, this dissertation could not have been finished. I wish to express my appreciation to my dearest parents and brother. Their understanding and gracious support during these years gives me the biggest encouragement to move on. I would also like to thank all my close friends and colleagues, for their support in tough times and the joyful memories the have offered me. Finally, I would like to acknowledge the gracious support for this work provided by the EPSRC and ESRC Centre for Doctoral Training on Quantification and Management of Risk & Uncertainty in Complex Systems & Environments (EP/L015927/1).

ABSTRACT

Essays on Predictability & Excess Profitability of Quantitative Methods: Modelling Implied Volatility, Technical Trading, Data Snooping and Market Efficiency

Ioannis Psaradellis[1]

The first essay concentrates on the modelling and trading of three daily market implied volatility indices issued on the Chicago Board Options Exchange (CBOE) using evolving combinations of prominent autoregressive and emerging heuristics models, with the aim of introducing an algorithm that provides a better approximation of the most popular U.S. volatility indices than those that have already been presented in the literature and determining whether there is the ability to produce profitable trading strategies out-of-sample. A heterogeneous autoregressive process (HAR) is combined with a genetic algorithm–support vector regression (GASVR) model in two hybrid algorithms. The algorithms' statistical performances are benchmarked against the best forecasters on the VIX, VXN and VXD volatility indices. The trading performances of the forecasts are evaluated through a trading simulation based on VIX and VXN futures contracts, as well as on the VXZ exchange traded note based on the S&P 500 VIX mid-term futures index. Our findings indicate the existence of strong nonlinearities in all indices examined, while the GASVR algorithm improves the statistical significance of the HAR processes. The trading performances of the hybrid models reveal the possibility of economically significant profits.

This second essay investigates the debatable success of technical trading rules, through the years, on the trending energy market of crude oil. In particular, the large universe of 7846 trading rules proposed by Sullivan et al., (1999), divided into five families (filter rules, moving averages, support and resistance rules, channel breakouts, and on-balance volume averages), is applied to the daily prices of West Texas Intermediate (WTI) light, sweet crude oil futures as well as the United States Oil (USO) fund, from 2006 onwards. We employ the *k*-familywise error rate (*k*-FWER) and false discovery rate (FDR) techniques proposed by Romano and Wolf

[1] B.Sc., M.Sc., Athens University of Economics & Business.

(2007) and Bajgrowicz and Scaillet (2012) respectively, accounting for data snooping in order to identify significantly profitable trading strategies. Our findings explain that there is no persistent nature in rules performance, contrary to the in-sample outstanding results, although tiny profits can be achieved in some periods. Overall, our results seem to be in favour of the adaptive market hypothesis.

The third essay examines technical trading rules performance on the statistical arbitrage investment strategy, pairs trading, using daily data over the period 1990-2016 for 15 commodity, equity and “famous” currency pairs. Adopting the false discovery rate test of Barras et al., (2010) to control for data snooping bias and exercising 18,412 technical trading rules, we find evidence of significant predictability and excess profitability, especially for commodity spreads, where the best performing strategy generates an annualized mean excess return of 17.6%. In addition, we perform an out-of-sample analysis to cross-validate our results in different subperiods. We find that whilst the profitability of rules based on technical analysis exhibits a downward trend over the sample, the opportunities for pairs trading remains has increased in certain cases.

Table of Contents

CHAPTER 1
INTRODUCTION

In recent years, rapid improvements in technology and statistics are reshaping the financial markets. All major market players (J.P. Morgan, Goldman Sachs, CITADEL) have raised the need for hiring computer and data science experts to integrate new technology frameworks in their business activities. Firms such as McKinsey (2017) and the Boston Consulting Group (2015) have already predicted that the interaction between new quantitative techniques for analysing big data and machine learning will dominate the fields of financial risk management, corporate finance and asset management within the next decade. J. P. Morgan's (2017) quantitative investing and derivatives strategy team reports that machine learning, artificial intelligence and computerized trading have become crucial to the future functioning of markets, while analysts, portfolio managers, traders and chief investment officers will have to familiarize themselves with these techniques.

Big data and machine learning strategies are already eroding some of the advantage of fundamental analysts, equity long-short managers and macro investors, as well as systematic strategies increasingly adopt machine learning tools and methods. As data sets get larger and more complex, investors are sometimes forced to use sophisticated data analytics. The tools used for these tasks include machine learning (drawn from traditional statistics) or deep learning (inspired by the functioning of the human brain). Understanding however the economics behind the data and the signals triggered is always more important than developing complex techniques, as certain data may have no value and more complex techniques do not always produce better forecasts.

In terms of statistical inference, the continuous growth of datasets, make hypothesis testing even more complex. For instance, portfolio managers' aim is to assess datasets for their ability to generate alpha, the so-called 'alpha content'. Searching for 'alpha content' however is gradually dependent upon the cost of the data, the amount of processing required and how well-used the dataset is already. In such way, the data snooping issue has nowadays become even more urgent because of the severe usage of big datasets, leading to promising results sometimes even by pure luck. Despite that, data replication quite often involves the cost of incorrectly discovering a profitable strategy yielding alpha. Classical statistical inference focusing on single hypothesis testing for each rule, without paying attention to the performance of the rest of the strategies, usually leads to false rejections or the so called, Type I error due to extensive specification search. Recently, Harvey (2017) raises this issue as the *p*-hacking phenomenon (i.e. frequent falsely significant *p*-values) based on numerous significant variables and explains that new, adjusted thresholds in multiple hypothesis testing reflecting genuine significance for an investment strategy should be defined. In his recent presidential address at the American Financial Association annual meeting (2017), he remarks that economic plausibility should also be a part of the statistical inference for computing such adjusted thresholds in order to control for data mining issues in big data analytics. In similar manner, Cochrane (2011) raises a concern about the discovery of a "zoo" of new significant factors in asset pricing literature over the last years and the need of adjusted discount theories. Thus, new multiple hypothesis frameworks developed to minimize such occurrences, while performing statistical inference are more than necessary nowadays. Although, there is an ongoing research towards this direction over the last years (Sullivan et al., 1999; Lo et al., 2000; Hansen, 2005; Hsu et al., 2010; Neely and Weller, 2011; Brajgowicz and Scaillet, 2012), enhanced frameworks achieving a good balance between Type I and Type II errors still should be reviewed.

This thesis studies expected returns using new technologies introduced by machine learning and data science. In particular, we investigate the predictability and excess profitability of financial markets using up-to-date powerful techniques pooled from machine learning and statistics, which accounting for data snooping effects. The motivation is to study the financial market predictability and to yield new insights in the empirical dynamics of asset pricing using statistical inference. This will guide us to revisit existing methods and to identify new advancements. For this purpose, we rebuilt trading strategies and systems, mainly in the second part, commonly utilized by portfolio managers (i.e., technical analysis in its algorithmic-trading version) to grasp to what extent markets are predictable and define the main factors explaining such predictability. We also extensively evaluate the out-of-sample performance of financial markets using the above techniques to understand the rationale underpinning predictability, as well as examining whether our findings reconcile with the Efficient

Market Hypothesis (EMH) or more recent formulations revisiting market efficiency such as the Adaptive Market Hypothesis (ADH) of Lo (2004). Another objective based on the above simulations and theories is to try to explain why do signals rise and disappear through time, which will provide us with a valuable awareness of the nature of short term market predictability. In fact, a trading strategy rarely reserves a uniform performance for more than a few years or even months. Finally, we also try to investigate the new horizons data science offers in the field of algorithmic trading, triggering a scientific revolution. Indeed, the analytical power of statistics in analysing big and complex data patterns and correlations can lead us to new scientific theories.

The first part of the thesis exploits the use of machine learning and artificial intelligence techniques to evaluate the predictability of volatility and the exercise of trading strategies based on the employed technique's forecasts. Volatility trading has been of growing interest not only from the quantitative financial analyst community but also from institutional and high-net-worth clients due to their characterization as alternative risk-premium strategies, (Sepp, 2016), as well as the availability of volatility-related indices and instruments (see CBOE) for trading purposes over the recent years. When it comes to machine learning we mostly focus on supervised learning and deep learning specifications even though various iterations exist (i.e., supervised, unsupervised, deep and reinforcement learning). The aim of supervised learning is primarily to determine the relationship between two datasets and to use one dataset to predict the other. The aim of deep learning is to employ artificial intelligence techniques, such as multi-layered neural networks, to estimate a trend, while it encourages the algorithms to explore and identify the most profitable trading strategies.

We employ an empirical investigation, which concentrates on the modelling and trading of three daily market implied volatility indices issued on the Chicago Board Options Exchange (CBOE), namely the VIX, VXN and VXD, using evolving combinations of prominent autoregressive and emerging heuristics models. The aim is to develop an algorithm that provides a better approximation of the most popular U.S. volatility indices than those that have already been presented in the literature and to determine its ability to compose a profitable trading strategy. We combine the heterogeneous autoregressive process (HAR) of Corsi (2009), which captures long memory, with a genetic algorithm–support vector regression (GASVR) model in two hybrid algorithms. We benchmark the algorithms' statistical performances against the best forecasters on the VIX, VXN and VXD volatility indices. Moreover, we evaluate the trading performances of the forecasts through a trading simulation based on VIX and VXN futures contracts, as well as on the VXZ exchange traded note based on the S&P 500 VIX mid-term futures index. The main findings indicate the existence of strong nonlinearities in all indices examined, while the GASVR algorithm improves the statistical significance of the HAR processes. In terms of out-of-sample excess profitability of the hybrid models, we reveal the possibility of economically significant profits. The proposed methodology and the empirical evidence of this study have already been published in the International Journal of Forecasting (see, Psaradellis and Sermpinis, 2016).

The second part deals more extensively with the issue of data snooping when a great number of trading signals occur, based on trading strategies employed in professional trading desks. We assess the rationale underpinning the emergence of trading signals and how the existence of significant ones can be consistent with current economic theories. We revisit the strategies of technical trading (i.e., technical indicators and oscillators) and statistical arbitrage built to be delta neutral (i.e., pairs trading). The notion is to shed new light in the world of professional trading, as trading desks do in practice to maximize their profits, while in the meantime to exercise emerging approaches controlling for data snooping effects in both back testing and out-of-sample environments. This will allow us to measure the significance of those profits yielded by some of the most puzzling principles in the field of asset pricing, momentum and mean reversion.

We perform two empirical applications towards the above direction, which also constitute the third and fourth chapter of this thesis. In the third chapte we investigate the debatable success of technical trading rules, through the years, on the trending energy market of crude oil. In particular, we revisit the large universe of 7846 trading rules proposed by Sullivan et al., (1999), divided into five families (filter rules, moving averages, support and resistance rules, channel breakouts, and on-balance volume averages) and applied to the daily prices of West Texas Intermediate (WTI) light, sweet crude oil futures as well as the United States Oil (USO) fund, from 2006 onwards. We employ the *k*-familywise error rate (*k*-FWER) and false discovery rate

(FDR) techniques proposed by Romano and Wolf (2007) and Barras et al., (2010) respectively, accounting for data snooping in order to identify significantly profitable trading strategies. We conclude that there is no persistent nature in rules performance, contrary to the in-sample outstanding results, although tiny profits can be achieved in some periods. In overall, our results seem to be in favour of the adaptive market hypothesis.

In the fourth chapter, we examine technical trading rules performance when using a statistical arbitrage investment strategy, pairs trading, using daily data over the period 1990-2016 for 15 commodity, equity and "famous" currency pairs. Adopting again the false discovery rate test of Barras et al., (2010) to control for data snooping bias and exercising 18,412 technical trading rules this time, we find evidence of significant predictability and excess profitability, especially for commodity spreads, where the best performing strategy generates an annualized mean excess return of 17.6%. In addition, we perform an out-of-sample analysis to cross-validate our results in different subperiods. We find that whilst the profitability of rules based on technical analysis exhibits a downward trend over the sample, the opportunities for pairs trading remains has increased in certain cases.

The remainder of the thesis is structured as follows. In Chapter 2, we provide the methodology and empirical evidence of our first application on modelling and trading the U.S. implied volatility indices with hybrid models based on autoregressive and machine learning techniques. Chapter 3 describes the application of Sullivan's et al., (1999) technical trading rules universe on the crude oil market as well as the detailed methodology and performance of the powerful *k*-FWER and FDR specifications for controlling data snooping on the generated returns. Chapter 4 presents our proposed universe of 18,412 technical trading rules, the construction of the pairs considered, as well as the former's predictability and excess profitability after accounting for data snooping bias through the FDR approach. Finally, Chapter 5 presents the concluding remarks of the thesis.

CHAPTER 2
MODELLING & TRADING THE U.S. IMPLIED VOLATILITY INDICES: EVIDENCE FROM THE VIX, VXN AND VXD INDICES.

1. Introduction

The Chicago Board Options Exchange (CBOE) implied volatility index (VIX), the so-called "investor fear gauge" (Whaley, 2000), has been widely used as a key measure of risk by both academics and practitioners, since it relies on the market expectations of volatility implied by the supply and demand of the S&P 500 index options. Its popularity as a hedging instrument for investors encouraged the CBOE to calculate several other volatility indices as well, measuring the expectations conveyed by option prices traded in other markets; for example, the Nasdaq-100 volatility index (VXN) and the Dow Jones Industrial Average volatility index (VXD). In particular, the VIX, VXN and VXD are forward-looking indicators that represent expected future market volatility over the next 30 calendar days. They are all characterized by sharp increases during periods of uncertainty and turmoil in the options market (Whaley, 2009). This specific feature of the volatility indices makes them very popular tools for decision makers and financial analysts, because they reveal whether or not the most liquid markets have reached an extreme level of sentiment. Thus, being able to predict these specific volatility indices accurately is of great importance not only for derivative markets but for the hedge fund industry in general. This paper concentrates on modelling the VIX, VXN and VXD using evolving combinations of prominent autoregressive and emerging heuristic techniques, which are distinguished by their forecasting potential.

Examining the empirical evidence on modelling the term structure of implied volatility, we find a considerable degree of variation in the literature. Gonzales Miranda and Burgess (1997) and Malliaris and Salchenberger (1996) apply non-parametric techniques successfully to the modelling of the Black-Scholes implied volatility of the S&P 100 ATM call options and Ibex35 index options, respectively. They find that Neural Networks (NN's) are able to express some characteristics of the data better than traditional models. Dumas et al. (1998) use a deterministic function (DVF) to capture the dynamic S&P 500 options’ implied volatility, employing the asset prices, moneyness ratio and expiration date of options as inputs. The model that they examine does not show a significant stability across the implied volatility surface relative to a fully stochastic one. Following a similar methodology by Malliaris and Salchenberger (1996), Refenes and Holt (2001) take a step forward by not only applying a multi-layer perceptron network (MLP) for forecasting the implied volatility of Ibex35 options but also using the Durbin-Watson test on (NN) residuals for purposes of misspecification analysis. Gonçalves and Guidolin (2006) express these dynamics using a vector autoregressive (VAR) technique. They also assess the economic significance of the VAR's forecasts by constructing a variety of trading and hedging strategies. Ahn et al., (2012) follow a different and unique approach by forecasting the directional movements of the implied volatility of the KOPSI 200 options precisely as a function of Greeks using an artificial NN and a sliding window technique.

Other researchers such as Blair et al. (2001a, b), Fleming et al. (1995) and Harvey and Whaley (1992) have conducted noteworthy research on the predictability of the VXO implied volatility of the S&P 100 index. The first approach is an economic variables model under the Black–Scholes assumptions. The last three methodologies demonstrate that the movements of the VXO are explained by a first-order autocorrelation model that incorporates mean reversion and an ARCH model that consolidates leverage effects, index returns and VIX observations. Similarly, Brooks and Oozeer (2002) also use a macroeconomic variables model to forecast and trade the implied volatility derived from at-the-money options on Treasury bond futures of LIFFE.

There are numerous papers in the literature that have investigated the dynamics of the VIX, such as for pricing implied volatility derivatives or for predicting the directional movements of the S&P 500 index (see e.g. Dotsis et al., 2007). However, only a limited number of studies in the literature have addressed the question of forecasting the dynamics of the implied volatility indices directly. Ahoniemi (2006) uses a hybrid ARIMA-GARCH model to produce point forecasts for the VIX index, while Konstantinidi et al., (2008) examine the predictive ability of a mixture of methodologies, such as an economic variables model, a vector autoregressive (VAR) model and an autoregressive fractionally integrated moving average (ARFIMA) model for producing point and interval forecasts of several U.S. and European implied volatility indices. Both studies indicate that the ARFIMA explains the U.S. volatility indices better, and apply out-of-sample forecasts for trading purposes. Clements and Fuller's (2012) study focuses on the implementation of a long volatility hedge for an equity index, based on semi-parametric forecasts that capture increases in the VIX. Fernandes et al., (2014) use the heterogeneous autoregressive (HAR) process of Corsi (2009) to model the VIX, while considering numerous macro-finance variables from the U.S. economy. The rationale behind the use of the HAR is the long memory which characterizes the implied and realized volatility of options (Bandi & Perron, 2006; Corsi, 2009; Koopman et al., 2005). They also develop a semi-parametric HAR model that includes a neural network (NN) term for capturing any nonlinearities of unknown form that define the index. Their stimulus lies in the fact that some macro-finance variables (e.g., the USD index) do not seem to have statistically significant effects on the VIX if one controls for nonlinear dependence; conversely, their effect on the index is significant in a linear structure.

This study employs a heterogeneous autoregressive process (Corsi, 2009; Muller et al., 1997) to predict the VIX, VXN and VXD and combines it with one of the most promising heuristic techniques, a hybrid genetic algorithm-support vector regression (GASVR) model. GASVR is a promising, fully adaptive heuristic algorithm, that is free from the data snooping effect and parameterization bias, and has had only a small number of applications in the field of forecasting (Dunis et al., 2013: Pai et al., 2006; Sermpinis et al., 2014; Yuang, 2012). This is the first application of the GASVR to the modelling of option volatilities.

Financial series (particularly tradable series such as the ones under study) are vulnerable to both behavioural (Froot et. al., 1992) and exogenous factors such as political decisions (Frisman, 2001). These factors are impossible to captured with mathematical models, and include noise to time-series estimations. Linear models (like those that dominate the relevant literature) on the other hand, will be only partially successful at capturing the relevant underlying trend. They seem to unable to help traders to generate profitable series, and have low forecasting power and volatile behaviour through time (LeBaron, 2000; Qi & Wu 2006). For instance, the HAR process is one of the most dominant approaches to modelling and forecasting the implied volatility in a linear form, based on three past volatility components (daily, weekly and monthly). However, when considering our proposed semiparametric approach, we find that the daily component of the HAR specification is no longer statistically significant. This indicates that the series under study exhibits nonlinear characteristics. By combining the best linear performers for forecasting the U.S. volatility indices with one of the most up-to-date and promising non-linear heuristic approaches, this research aims to create a superior hybrid forecaster that will surpass the statistical and trading performance of the models presented previously in the relevant literature. More specifically, a HAR process (the most promising linear model according to Fernandes et al., 2014) is developed and combined with a GASVR model in two hybrid models. The forecasting performance of the hybrid models will indicate whether there are non-linear elements that HAR is unable to capture and whether the evolutionary concept of GASVR can actually mimic the market dynamics and be capable of producing profitable forecasts. Their performances are benchmarked against a hybrid non-linear heuristic model (incorporating a HAR process and a recurrent neural network (RNN)), a simple HAR process, an ARFIMA model and a hybrid ARFIMA algorithm. In this study, we do not consider macro-finance variables models because their forecasting performances for predicting the VIX, VXN and VXD, are poor relative to those of stochastic processes (see Fernandes et al., 2014; Konstantinidi et al., 2008)[2]. We verify the robustness of our proposed methodology, by examining two out-of-sample datasets. The first one covers between the Lehmann Brothers collapse (mid-September 2008) and the end of 2009, the period of financial crisis. The second one is a more recent period, starting from January 2013 until April 2014. The paper also performs a heuristic analysis based on the residuals obtained from the autoregressive models, with the aim of extracting any other unknown form of nonlinearity that is not captured by the residuals of HAR and ARFIMA specifications. The aim of the paper is not to map the series under study, as this would be impossible to achieve for any financial tradable series while using a mathematical model. Instead, this study aims to introduce an algorithm that approximates the examined indices better than those already presented in the literature.

[2] Experimentation has shown that the use of explanatory variables, such as the continuously compounded return on the S&P 500 index, the S&P 500 volume change and the continuously compounded return on the one-month crude oil futures contract, as inputs, does not improve the performance of our hybrid algorithm.

The forecasting performance of the models under study are examined using three different predictive ability tests: the superior predictive ability (SPA) and model confidence set (MCS) tests of Hansen (2005) and Hansen et al., (2011) respectively, and the Giacomini and White (2006) test. Finally, we perform an out-of-sample realistic trading simulation by employing VIX and VXN futures contracts acquired from the Chicago Board Options Exchange (CBOE) volatility futures market in order to check for possible abnormal profits. For investigating the excess profitability of VIX index, we also employ exchange traded notes (ETNs), specifically the iPath S&P 500 VIX mid-term futures index (VXZ). ETNs linked with volatility indices are strongly preferred by investors as a good diversification hedge, while they are available with tiny investor fee rates. Our results reveal that a HAR- GASVR residual hybrid model is the only algorithm that generates statistically significant excess profitability, when taking futures contracts into account. When considering trading performances for the VXZ ETN, in which the transaction costs are substantially lower, all HAR specifications are capable of yielding statistically significant profits. To the best of our knowledge, this is the first time that a HAR process has been employed for trading purposes other than modelling.

The remainder of the paper is structured as follows. Section 2 provides a detailed description of the implied volatility indices, the VIX and VXN futures contracts and the VXZ ETN. Section 3 presents a synopsis of the benchmark models, the semiparametric architectures applied and the combination methods implemented. The statistical forecasting and trading performances are discussed in Sections 4 and 5, respectively. Finally, the last section presents our conclusions.

2. Implied Volatility Indices and Related Financial Data

The VIX was introduced on the Chicago Board Options Exchange (CBOE) in 1993, while VXN and VXD were introduced a few years later. All three indices are settled on daily basis. VIX, VXN and VXD represent weighted indices that mixed together different types of stock index options from S&P 500, Nasdaq-100 and DJIA respectively. As has been mentioned, the indices portray the expected future market volatility over the next 30 calendar days. Hence, they are forward-looking illustrations of the level of volatility expected by the market in the short term. All indices apply the VIX algorithm (a model-free implied volatility estimator; see, Jiang &Tian, 2005) for the calculation of index values (see, Chicago Board Options Exchange, 2015). Thus, they do not depend on any particular option pricing structure such as the Black-Scholes model (Britten-Jones & Neuberger, 2000).

In this paper, we examine two periods for which we have daily closing prices of the VIX, VXN and VXD, from August 2002 to November 2009 and from January 2007 to April 2014, for the sake of robustness. The datasets were separated into in-sample and out-of-sample subsets (see Table 1), in which, the out-of-sample subset consists of approximately the last 14 months of each dataset. The dataset was obtained from the CBOE website. Furthermore, the descriptive statistics for the three series are presented in Table 2.

[Table 1]

[Table 2]

The three series under study are non-normal (see the Jarque-Bera *p*-values in levels at the 99% confidence level) and exhibit high levels of skewness and positive kurtosis. The series were therefore transformed into logarithms in order to overcome these issues. The summary statistics of the series (in logs) are also presented in Table 2, in which we observe that a slight skewness still remains.

The time series (in logs) were also tested for stationarity, a unit root and long memory through a variety of testing techniques (Table 3). including the augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) unit root tests. In addition, the KPPS test statistic for the null hypothesis of stationarity and the long memory rescaled variance test statistic (V/S) (see Giraitis et al., 2003) were employed to confirm that long memory models such as ARFIMA and HAR are appropriate for modelling our data. The number of lags for the KPSS test was selected using the quadratic spectral kernel with bandwidth choice (Andrews, 1991).

[Table 3]

Table 3 reports that the null hypothesis of a unit root is rejected at the 99% statistical level for the full sample according to the *p*-values of ADF and PP tests.

Likewise, the KPSS test cannot reject the null of stationarity at all the 10%, 5% and 1% significance levels for the full sample, which confirms the stationarity property[3]. The V/S test null hypothesis for short memory is rejected for both levels of significance, indicating that our sample is characterized by long memory.

In our trading simulation, we use VIX and VXN futures contracts[4] from the Chicago Futures Exchange (CFE), as well as the iPath S&P 500 VIX mid-term futures ETN (VXZ)[5]. The VIX and the VXN future contracts may trade up to nine near-term serial months and five months on the February quarterly cycle. The final settlement date is the Wednesday that is thirty days prior to the third Friday of the next month, when the standard S&P 500 and Nasdq-100 index options expire. The contract multiplier for each future is $1000. Our application examines seven different futures contracts traded in the second out-of-sample data set, which expire in 2013 and 2014[6]. We trade the contracts much closer to their expiration dates, when the futures price is almost equal to the spot price, in order to minimize the basis risk. Finally, we minimize the effect of noisy data by rolling from every future contract series to the following one, five days before each matures (see Dotsis et al., 2007). Table 4 presents the characteristics of the VIX and VXN futures contracts considered.

[Table 4]

On the other hand, VXZ, offers investors' a cheap alternative relative to the more expensive (in terms of transaction costs) futures contracts. In addition, trading an ETN does not require a margin account. The VXZ ETN is designed to offer exposure to the S&P 500 VIX mid-term futures index total return. This index provides access to a daily rolling long position in the fourth, fifth, sixth and seventh month VIX futures contracts. The investor fee rates for the VXZ ETN are 0.89% per annum. The VXZ ETN is the second biggest CBOE volatility index ETF in terms of total assets.

[3] The small size of our sample does not allow us to distinguish reliably between long and short memory processes (Lee & Schmidt, 1996).

[4] We do not take VXD futures contracts into account because they were delisted from the CBOE Futures Exchange in 2009.

[5] The futures contract and ETN specifications and settlement processes were retrieved from the CBOE and Barclays websites respectively.

[6] VXN futures and the VXZ ETN were developed after the period of the first dataset, unlike the VIX futures. Contracts with trading volumes of settlement prices that are less than five are excluded.

The biggest is the iPath S&P 500 short-term VIX features ETN (VXX), which seeks to replicate the daily rolling long position in the immediately first and second month VIX futures contracts. We choose the VXZ ETN because it is exposed to less contago effect arising from the volatility forward curve[7] and has lower basis risk compared to the VXX[8].

3. Forecasting Models

3.1. ARFIMA model

An ARFIMA (1, d, 1) model is employed as a benchmark for capturing the short and long memory properties of the implied volatility index. ARFIMA (1, d, 1) performs better than VAR models and other simple linear models based on economic variables for forecasting the U.S. implied volatility indices (Konstantinidi et al., 2008). A hybrid model based on the residuals of ARFIMA (1, d, 1) regressions and the GASVR algorithm is also explored. The intuition of the hybrid model is that VIX, VNX and VXD are most likely to follow a nonlinear pattern. The GASVR algorithm attempts to extract these non-linear elements from the residuals and to combine them with the ARFIMA forecasts in order to present a superior forecasting model.

The standard ARFIMA (*p, d, q*) process is given by

$$lriv_t = (1-L)^{-d}\{\rho(L)\}^{-1}\theta(L)\varepsilon_t, \quad \varepsilon_t \sim N(0,\sigma_\varepsilon^2) \tag{1}$$

where $lriv$ is the logarithm of the volatility index, $(1-L)^d$ is the fractional difference operator with a d order of fractional integration being required for stationarity, which is expressed in non-integer values; $\rho(L) = \left(1-\rho_1 L - \cdots - \rho_p L^p\right)$ and $\theta(L) = (1-\theta_1 L - \cdots - \theta_q L^q)$, are the lagged autoregressive and moving average polynomials, respectively, and ε_t is the Gaussian error term.

[7] Volatility ETFs are subject to a contango effect arising from the volatility forward curve, which is upward sloping, because they track VIX futures, not the VIX index itself.

[8] We have computed the basis risk for the two ETNs in the in-samples subperiods. VXZ demonstrates less basis risk than the VXX in both subperiods.

3.2. HAR model

Corsi (2009) proposes a heterogeneous autoregressive model for realised volatility, inspired by the Heterogeneous Market Hypothesis of Müller et al. (1993), which accepts the presence of heterogeneity across traders. Specifically, he focuses on the heterogeneity arising from different time horizons due to the divergent trading frequency of market agents. The notion is that there are three classes of market participants based on their trading frequencies. These agents are classified as short-term (e.g., intraday traders or speculators and hedge funds), characterized by higher trading rates, usually daily; medium-term (e.g., commercial banks), who perform a weekly rebalancing of their assets and long-term (e.g., pension funds, insurance companies), defined by lower frequency of transactions, usually on a monthly basis. This leads to three different types of volatility components (daily, weekly, monthly), which create an overall volatility cascade from low to high frequencies. At each level of the cascade, the underlying volatility component consists not only of its past observation but also the expectation of longer horizon partial volatilities. The proposed model is defined as an additive linear structure of first-order autoregressive partial volatilities able to capture the long-range dependence:

$$lriv_t = \beta_0 + \beta_{(d)} lriv_{t-1}^{(d)} + \beta_{(w)} lriv_{t-1}^{(w)} + \beta_{(m)} lriv_{t-1}^{(m)} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \quad (2)$$

where $lriv^{(h)} = \frac{1}{h}\sum_{j=1}^{h} lriv_{t-j+1}$ and $h = (1, 5, 22)'$ is an index vector that depicts the daily, weekly and monthly components of the volatility cascade. We use the HAR specification as a second benchmark for VIX, VXN and VXD modelling because of its excellent forecasting ability on implied and realised volatilities (see, Busch et al., 2011; McAleer and Medeiros, 2008, amongst others). In addition, we also employ the HAR structure to construct our semiparametric approaches involving NNs and the GASVR algorithm.

3.3 Neural network approach

Our third benchmark model is a semiparametric approach evolving the HAR process and a recurrent neural network (RNN). As was discussed above, several researchers have applied NNs successfully to the task of identifying patterns in implied or realized volatilities. Usually, NN specifications have at least three layers. The first layer is called the input layer, and the number of nodes corresponds to the number of explanatory variables. The last layer is called the output layer, and the number of nodes corresponds to the number of response variables. An intermediate layer of nodes, called the hidden layer, separates the input and output layers. The number of nodes in this layer controls the amount of complexity that the model is capable of fitting. In addition, the input and hidden layers each contain an extra node called the bias node. This node has a fixed value of one and has the same function as the intercept in traditional regression models. Normally, each node in a given layer is connected to all of the nodes in the next layer. The training of the network (which involves the adjustment of its weights such that the network maps the input values of the training data to the corresponding output values) starts with randomly chosen weights and proceeds by applying a learning algorithm called the backpropagation of errors (Shapiro, 2000). The iteration length is optimized by maximizing a fitness function in the test dataset.

The RNNs have activation feedback that embodies short-term memory. In other words, the RNN architecture can provide more-accurate outputs because the inputs are (potentially) taken from all previous values. While Tenti (1996) notes that RNNs need more connections and memory than standard back-propagation networks, these additional memory inputs allow RNNs to yield better results than simple MLPs. For more information on RNNs, see Sermpinis et al. (2012). A similar hybrid HAR process and a simple NN model (NNHARX) performed equally well for forecasting the VIX relative to different types of HAR processes (see Fernandes et al., (2014)).

Straightforward modelling of the implied volatility indices using only RNNs[9] does not seem sufficient. The hybrid HAR-RNN method is defined as follows:

$$lriv_t = \beta_0 + \beta_{(d)} lriv_{t-1}^{(d)} + \beta_{(w)} lriv_{t-1}^{(w)} + \beta_{(m)} lriv_{t-1}^{(m)}$$

$$+ \sum_{m=1}^{M} \frac{\lambda_m}{1+e^{-\alpha-\beta_{(m)} lriv_{t-1}^{(d)} - \beta_{(m)} lriv_{t-1}^{(w)} - \beta_{(m)} lriv_{t-1}^{(m)}}} + \varepsilon_t \qquad (3)$$

[9] We conduct NN experiments and a sensitivity analysis on a pool of autoregressive terms of VIX, VXN and VXD series. We find that a simple RNN approach performs poorly for both the in-sample and out-of-sample datasets. The problem is probably that a simple NN model cannot capture the long memory of implied volatilities efficiently, although it is very capable of capturing nonlinearities.

where $lriv_{t-1}^{(d)}, lriv_{t-1}^{(w)}$ and $lriv_{t-1}^{(m)}$ are the three volatility components of the HAR model, and $\sum_{m=1}^{M} \frac{\lambda_m}{1+e^{-\alpha-\beta_{(m)}lriv_{t-1}^{(d)}-\beta_{(m)}lriv_{t-1}^{(w)}-\beta_{(m)}lriv_{t-1}^{(m)}}}$ represents the transfer sigmoid function of the neural network. The neural network architecture is trained through the backpropagation method and the regularization parameter is optimized based on a cross-validation algorithm. The number of hidden units, M is set through a trial and error procedure on the in-sample dataset, which reveals the optimal results. In our simulation, the optimal number of hidden units is 3. For our NNs, we apply an objective fitness function that focuses on minimizing the mean squared error (MSE) of the network's outputs. After the networks have been optimized, the predictive value of each model is evaluated by applying it to the validation dataset (out-of-sample dataset).

3.4. HAR-GASVR framework

3.4.1. The GASVR

Support vector machines (SVMs) are nonlinear algorithms that are used to solve classification problems in supervised learning frameworks. SVM processes belong to the general category of kernel methods (Scholkopf & Smola, 2002). Their development involves, first, sound theory, then implementation and experiments, in contrast to the development of other heuristics that are purely atheoretic, such as NNs. Their main advantage is that, while they can generate nonlinear decision boundaries through linear classifiers, they still have a simple geometric interpretation. In addition, the solution to an SVM is global and unique; in other words, it does not suffer from multiple local minima such as the solutions of NNs occasionally do. Another advantage is that the practitioner can apply kernel functions to data such that their vector space is not fixed in terms of dimensions. SVMs can be used in regression problems by implementing the ε-sensitive loss function by Vapnik (1995). This function established SVRs as a robust technique for the construction of data-driven and nonlinear empirical regression models. Recently, SVR and its hybrid applications have become popular for time-series prediction and financial forecasting applications (see, among others, Dunis et al., 2013; Pai et al., 2006; Sermpinis et al. 2014; and Yuang, 2012). Finally, they also seem able to cope well with high-dimensional, noisy and complex feature problems (Suykens et al., 2002). A theoretical framework for SVRs is provided in Appendix A.

Although SVR has emerged as a highly effective technique for solving nonlinear regression problems, the design of such a model can be impeded by the complexity and sensitivity of parameter selection. The performances of SVRs depend on all parameters being set optimally. Numerous different approaches to this optimization have been presented in the literature, such as setting ε to a non-negative constant for the sake of convenience (Trafalis & Ince, 2000), using data-driven approaches (Cherkassky & Ma, 2004), applying cross-validation techniques (Cao, Chua, & Guan, 2003; Duan, Keerthi, & Poo, 2003), and controlling ε with v-SVR (Scholkopf, Bartlett, Smola, & Williamson, 1999).

In this study, SVR parametrization is conducted via a genetic algorithm (GA)[10]. The resulting algorithm (GASVR), searches genetically over a feature space and then provides a single optimized SVR forecast for each series under study. We perform this process using a simple GA in which each chromosome comprises feature genes that encode the best feature subsets, and parameter genes that encode the best choice of parameters. For our hybrid approach, we implement a radial basis function (RBF) v-SVR kernel, which is specified generally as

$$K(x_i, x) = \exp(-\gamma \|x_i - x\|^2), \gamma > 0 \tag{4}$$

where γ represents the variance of the kernel function. We optimize the γ parameter along with regulation parameters controlling the balance between learning the SVR and training errors (slack variables) with the GA as described in Appendices A and B. RBF kernels are the most common in similar SVR applications (see, Ince and Trafalis, 2006, 2008, amongst others), because they overcome overfitting efficiently and seem to excel in forecasting applications.

3.4.2. The HAR-GASVR

Following the approach described above, this study combines the HAR model with a genetically optimized v-SVR. In this hybrid model, the v-SVR parameters (C, v and γ) are optimized through a genetic algorithm. This HAR-type genetic support vector (HAR-GASVR) model is specified as follows:

[10] For a more detailed description of the GA algorithm see Appendix B.

$$lriv = \beta_0 + \beta_{(d)} lriv_{t-1}^{(d)} + \beta_{(w)} lriv_{t-1}^{(w)} + \beta_{(m)} lriv_{t-1}^{(m)}$$
$$+\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) K(lriv_i^{(h)}, lriv) + \varepsilon_t \quad (5)$$

where

$$K\left(lriv_i^{(h)}, lriv\right) = \exp\left(-\gamma \left\| lriv_i^{(h)} - lriv_0 \right\|^2\right), \gamma \quad (6)$$

is an RBF Kernel function that uses the index vectors of the three volatility components as inputs.

For the GA optimization, we set the crossover probability to 0.9. This setting enables our model to keep some population for the next generation, in hope of creating better new chromosomes from the good parts of the old ones. The mutation probability is set to 0.1 in order to prevent our algorithm from performing a random search, whereas the wheel roulette selection technique is applied to the selection step of the GA. Similar to NNs, our HAR-GASVR model requires the use of training and test subsets to validate the goodness of fit of each chromosome. The population of chromosomes is initialized in the training sub-period, and the optimal selection of chromosomes is achieved when their forecasts minimize the MSE in the test-sub period. Then, the optimized parameters and selected predictors of the best solution are used to train the SVR and produce the final optimized forecast, which is evaluated over the out-of-sample period.

We adjust the GA initial population to 100 chromosomes, and the maximum number of generations is set to 200. However, the algorithm may terminate the evolution earlier if the population is deemed to have converged. The population is deemed to have converged when the average fitness across the current population is less than 5% away from the best fitness of the current population. More specifically, when the average fitness is less than 5% away, the diversity of the population is very low, and more generations of evolution are unlikely to produce different and better individuals than the existing ones or those examined by the algorithm in previous generations.

3.5. Modelling the residuals

Adding to the previous models, we now proceed to a residual analysis of the implied volatility indices estimation approaches in order to express potential asymmetric effects that show up among the residuals. The GASVR regression method is applied to the residuals generated from our two linear benchmarks (ARFIMA and HAR). The idea behind these HAR and ARFIMA-type genetic support vector regression residual models (ARFIMA-GASVR(res) and HARGASVR(res), respectively) is to perform a heuristic analysis on the ARFIMA and HAR residuals and capture the nonlinear elements that are hidden in their noise. This specification should be able to forecast VIX, VXN and VXD more accurately than its linear counterparts.

We follow a two-step approach. The first step is to feed and train the GASVR algorithm using the series of residuals derived from the ARFIMA and HAR estimations, respectively. In the second step, the GASVR forecasted values are added to the ARFIMA and HAR forecasts. Again, following the GASVR methodology described above, the main goals are to genetically optimize the v-SVR parameters and to minimize the mean squared error (MSE) between the residuals and those that emerge from the SVR regression by employing the same fitness function. For this purpose, the optimization problem describing the v-SVR is transformed into

$$f(\varepsilon) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) \exp(-\gamma \|\varepsilon_i - \varepsilon\|^2) + b,\ 0 \leq \alpha_i, \alpha_i^* \leq \frac{C}{n}, \gamma > 0 \qquad (7)$$

where ε_i are the lagged values of residuals for each benchmark model, ε are the actual ones, while α_i and α_i^* represent langrage multipliers, helping to solve the above problem along with the kernel function (see Appendix A). In the absence of any formal theory for the selection of the inputs of a GASVR and based on the experiments during the in-sample period, we choose to feed our networks with the first five autoregressive lags of the VIX, VXN and VXD estimation residuals, representing a weekly time interval[11]. In addition, we keep the population and

[11] We experimented with various different numbers of lags in the in-sample periods (orders three to fifteen). In all cases, we obtained the best forecasting performance when using the first five autoregressive lags. The performance of GA-SV is highly sensitive to the selection of the inputs (see, Dunis et. al. 2013; Pai et. al. 2006; and Sermpinis et. al. 2014, amongst others).

generation levels and the crossover and mutation probabilities the same as in the previous approach.

4. Empirical findings

4.1. Statistical Performance

Tables 5 and 6 present the out-of-sample statistical performances[12] of each of the models inspected for the two periods considered. We report the root mean squared error (RMSE) and the mean absolute error (MAE) criteria for the statistical evaluation of our daily forecasts for the two out-of-sample periods, in which a lower output value for a model indicates that it has a better forecasting accuracy[13]. Apart from the models mentioned above, we also consider the predictive ability of a random walk without drift[14] and an AR(1) model[15].

[Table 5]

Concerning the statistical performance of the global financial crisis out-of-sample period, we observe that the HAR-GASVR(res) displays the best statistical results according to both measures for all of the implied volatility indices. In case of the VIX index in particular, the HAR-GASVR(res) and HAR-GASVR models outperform their competitors considerably. For instance, the former clearly outperforms the ARFIMA model in forecasting accuracy and achieves even better results than the HAR and HAR-RNN methods, which were established recently as the most accurate techniques for forecasting the VIX (Fernandes et al., 2014). The second-best predictive ability is achieved by the HAR-GASVR method, which presents a considerably better performance than either the HAR or HAR-RNN. Only in the case

[12] The in-sample statistical performances for both periods considered are available upon request.

[13] Daily forecasts are simply the forecasts for the rest days out-of-sample, compared with one-day-ahead forecasts which are generated through a recursive estimation of the models.

[14] Note that we also computed a random walk model with a drift, but found that incorporating the drift had a negative impact on the forecasting performance.

[15] In addition to the proposed models, we also explored forecast combinations of the best three and six models under study (ARFIMA, HAR, HAR-RNN, HAR-GASVR, ARFIMA-GASVR(res), HAR-GASVR(res)) with three different approaches: a simple average of the underlying forecasts, a Bayesian averaging method (Buckland et al., 1997), and a weighted average technique (Aiolfi and Timmermann, 2006). In all cases, the performances of the forecast combinations were inferior to that of our best model (HAR-GASVR(res)).

of VXN index does the HAR-GAVSR method perform equally with the above two processes. The results undoubtedly reveal the existence of nonlinearities and asymmetric effects on the implied volatility index, though from long memory and persistence. Strong evidence for this is provided by the recognition of HAR-GASVR(res) approach as the best forecasting model for the in-sample period. This shows that our proposed specifications have the ability to perform equally well, even in periods of turmoil.

[Table 6]

Concerning the out-of-sample statistical performance for the second period, our results display almost the same picture as in the first period, with our proposed forecast combinations being more accurate than the ARFIMA, HAR and HAR-RNN approaches. Specifically, the HAR-type approach with the GASVR error term again seems superior to the statistical measures employed for modelling the implied volatility indices. The hybrid HAR-GASVR model follows. HAR and HAR-RNN are next, presenting almost equally less precise out-of-sample results. However, the performance of the HAR-GASVR approach is equal to those of its previous HAR approaches in the case of the VXD index.

The above findings confirm the relative success of the HAR method bearing out the findings of Fernandes et al., (2014). Indeed, it is obvious that every HAR type specification outperforms the ARFIMA ones. This advantage might be attributable to the special ability of the HAR method to capture strong persistence in our dataset. A persistent nature really exists in VIX, VXN and VXD, which quantify the market expectations concerning the 22-trading-days ahead risk-neutral volatility. Furthermore, we find that there are also strong nonlinearities in the above indices, which makes our hybrid models perform better.

4.2. Robustness checks

We authenticate the results above by computing the unconditional Giacomini-White (2006) test for out-of-sample predictive ability testing and forecast selection, when the model may be misspecified. The null hypothesis of the test is the equivalence of forecasting accuracy between two forecasting models. The sign of the test statistic indicates which model has superior forecasting performance. A positive

GW test statistic indicates that the second model is more accurate than the first one, which produces larger losses, whereas a negative statistic specifies the opposite. We calculate the test in terms of the mean squared error loss function (MSE) for each forecast for both out-of-sample periods. Tables 7-9 display the *p*-values of the statistic under the null hypothesis that the performance of the model in the column is equivalent to that of the model in the row, for every index separately.

[Table 7]

[Table 8]

[Table 9]

It is obvious from the Table 7 that all of the HAR processes outperform the ARFIMA models at the 5% and 1% significance levels when forecasting the VIX index, according to MSE loss function. The HAR-GASVR (res) approach is superior to all the HAR processes. Similarly, only the HAR-GASVR and the HAR-GASVR(res) specifications produce significantly better forecasts than all of the competing models.

The picture seems to be much the same in Table 8 when applying the Giacomini–White test to the predictability of the VXN index. The results clearly show that the HARGASVR(res) model is again the best forecaster. However, the performance of our second proposed methodology, the HAR-GASVR model, is almost equal to that of the HAR-RNN approach for both periods. The rest of the specifications are inferior to the above ones, with the HAR methodology being superior to the ARFIMA models overall.

Table 9 indicates that the Giacomini–White test provides nearly the same information for the VXD index as for the VIX and VXN indices, with the HAR-GASVR(res) method being the most accurate approach for modelling VXD.

Table 10 now exhibits some descriptive results from Hansen's (2005) superior predictive ability (SPA) test and Hansen's et al.'s, (2011) model confidence set (MCS) procedure in order to allow an equal comparison of various methodologies considered under the mean squared error (MSE) and (MAE) criteria. The SPA test focuses on a comparison of the relative forecasting performances of multiple methodologies in a full set of models. The null hypothesis is that the benchmark forecast is not inferior to the best alternative one. Each model is used as the benchmark in turn each time we apply the SPA test, starting with the random walk. Low *p*-values indicate that the respective benchmark model is inferior to at least one alternative (rejecting the null), whereas high *p*-values specify the opposite.

The MCS procedure deduces the 'best' models from a full set of models under specified criteria and at a given level of confidence. Actually, it is a random data-dependent set of best forecasting models because a standard confidence interval covers the population parameter, while acknowledging the limitations of the data (Hansen et al., 2011). Hence, more-informative data can lead to only one best model, whilst less-informative data result in an MCS including several models because it is impossible to differentiate among the competing approaches. An equivalence test and an elimination rule are the key features of the MCS procedure. Low *p*-values indicate that it is unlikely that the model will belong to the set of the 'best' models. Therefore, *p*-values that exceed the usual levels of significance are preferable.

[Table 10]

The results of the SPA test indicate that most of examined models examined are inferior to at least one of the alternatives in almost all cases. This probably happens because the HAR-GASVR(res) model achieves the highest forecasting performance[16]. Only in the case of the VXN index do HAR processes seem to achieve the same performances during the global financial crisis period according to the MSE. In addition, HAR-GASVR and HAR-GASVR(res) yield the highest p-values for the VIX and VXN indices during the same period, which does not make them inferior to alternatives.

The MCS findings reveal the same picture. HAR-GASVR and the HAR-GASVR(res) are the only specifications that belong to the 'best' set for the VIX and VXN indices in the first out-of-sample period, whereas HAR-GASVR(res) is the only

[16] Applying the SPA test without considering the HAR-GASVR(res) approach, we find that all of the models are beaten by the second-best algorithm, the HAR-GASVR approach.

superior model for the rest of the cases considered[17]. This allows us to conclude that the data examined are indeed informative.

5. Economic Significance (Out-of-Sample Trading Simulation)

In this section, we apply a simple trading strategy to assess the economic significance of our models by employing the time series of VIX[18] and VXN futures and the VXZ ETN[19] for the second out-of-sample period. This is of great importance because statistical accuracy is not always synonymous with trading profitability. The trading strategy is executed separately for each of our forecasting models, and involves seven different futures contracts (see Table 4) and one ETN. The transaction costs are estimated at \$ 0.5 per transaction (see CBOE specifications) for future contracts and 0.89% per annum for the VXZ ETN (see Barclay's specifications).

We evaluate the trading efficiency of our forecasts and compare our results with those of previous studies by following a simple trading rule. The investor goes long (short) in the volatility futures and the ETN when the forecasted value of the implied volatility index is greater (smaller) than its current value.

The annualized Sharpe ratio (SR) and the annual Leland's (1999) alpha (A_p) are considered as performance measures. We calculate the Sharpe ratio and Leland's alpha using the continuously compounded annual U.S. Libor rate as the risk-free rate. Moreover, we bootstrap the 95% confidence intervals of the SRs and A_p values for each forecasting model in order to assess the statistical significance of the returns.

[17] The results remain the same whether we set the confidence level in our application to 10%, 5% or 1%, with the number of replications set to 10,000. Only when we exclude the HAR-GASVR(res) model and apply the procedure do we obtain a larger 'best' set.

[18] The regular VIX calculation uses the mid-point in the bid–ask spread of out-of-the money SPX options. The VIX futures settlement price is based on actually traded prices of SPX options. This difference can lead the VIX futures settlement price to diverge from the spot VIX, especially in some cases where the bid–ask spread in the SPX is very wide. However, Shu and Zhang (2011) have found that spot VIX and VIX futures generally react to information synchronously.

[19] It is also worth noting that the VIX and VXN futures and the VXZ ETN can be also applied as hedging tools on their respective indices. However, their efficiency is questionable (Alexander & Korovilas, 2013; Engle & Figlewski, 2015; Psychoyios & Skiadopoulos, 2006).

Leland's (1999) alpha is applied to tackle the existence of non-normality in the distribution of the returns found at the end of the trading strategies for each model[20]. It is specified as

$$A_p = E(r_P) - B_p\left[E(r_{mkt}) - r_f\right] - r_f, \tag{8}$$

where r_P is the return on trading strategy, r_f is the risk-free rate, r_{mkt} is the return on market portfolio, $B_p = \frac{cov(r_P, -(1+r_{mkt})^{-\gamma})}{cov(r_{mkt}, -(1+r_{mkt})^{-\gamma})}$, is a measure of risk similar to the CAPM's beta and $\gamma = \frac{\ln[E(1+r_{mkt})] - \ln(1+r_f)}{var\,[\ln(1+r_{mkt})]}$ is a risk aversion criterion (see Konstantinidi and Skiadopoulos, 2011). In addition, the continuously compounded annual return of the S&P 500 and Nasdaq 100 index are used as proxies for the benchmark market portfolio. The trading strategy presents an expected return over the risk adjusted degree, when $A_p > 0$. The trading performances of our models are presented in Table 11, while Appendix C presents the cumulative returns of the two best models over time.

[Table 11]

It is obvious from Table 11, that the *SR* and the A_p measures of the VIX and VXN futures' trading performances are statistically significant for the half of the cases examined, during the most recent out-of-sample period. Rejections of the null hypothesis of a zero value at the 5% significance level are indicated by an asterisk. On the other hand, all HAR specifications are capable of producing significant profits, when taking into account the performance of the VXZ ETN.

Specifically, our findings show that the HAR and HAR-GASVR(res) methods can produce significant profits for VIX and VXN future contracts, to some limited extent. However, the HAR specifications exhibit substantially larger gains when it comes to the trading simulation of the VXZ ETN. This is due to the very small investor fee rates for the volatility ETNs compared to the larger fees and margin requirements of futures contracts, as described earlier. The ARFIMA and ARFIMA-GASVR(res) models seem to produce losses for all products examined. The trading performances of the ARFIMA models seem to validate the conclusions of Konstantinidi and

[20] A statistical analysis shows that the distributions of the returns of the individual models are non-normal and far from Gaussian.

Skiadopoulos (2011) and Konstantinidi et al. (2008), who trade VIX volatility futures with the same model.

In summary, the HAR-GASVR(res) approach is to be found superior in terms of trading performances. It produces the largest gains for futures contracts and the ETN employed. In other words, it has a noteworthy prospect of achieving economically significant profits in the VIX and VXN volatility futures markets, which suggests that there is promise for the application of nonlinear methods, and specifically of the GASVR algorithm, even in trading strategies that involve future contracts and ETNs.

6. Conclusions

This paper examines the existence of nonlinearities in the evolution of the implied volatility. In particular, it provides evidence concerning the daily settlement of three market volatility indices, the VIX, VXN and VXD. Fernandes et al., (2014) recently showed that a HAR process seems to be very promising for forecasting the VIX, due to its long-range dependence and persistent nature. Two semiparametric methodologies are introduced as a combination of the HAR specification and one of the most promising heuristic techniques, a hybrid genetic algorithm–support vector regression (GASVR) model. The first semiparametric approach includes an extra optimization term in the HAR model. Specifically, the GASVR algorithm is fed the three volatility components (daily, weekly and monthly) of the HAR specification as inputs. The second specification performs a residual analysis for expressing the potential asymmetric effects that may be prevalent among the residuals. A heuristic regression between the residuals of HAR and its lagged values is applied to test for further persistence. The GASVR forecasted residuals are employed to develop the existing model. The performance of the proposed techniques is benchmarked with (1) an ARFIMA model, which predicts the US implied volatility indices well according to the literature (see Konstantinidi et al., 2008), (2) a semiparametric approach similar to our first, but using a recurrent neural network (RNN) instead of the GASVR algorithm, and (3) a semiparametric technique that is focused on the residual analysis of the ARFIMA model.

The HAR-GASVR(res) approach produces predictions that are more accurate than those of the other models by a significant margin. The second-best performance is achieved by the HAR-GASVR model. We authenticate the above results by applying the SPA test (Hansen, 2005), the MCS procedure (Hansen et al., 2011) and the Giacomini and White (2006) test. However, all of the HAR processes have better predictive abilities than the benchmark model. This justifies Fernandes et al.'s (2014) finding that this process cannot be beaten for forecasting the VIX, because of its persistent feature. The forecasting superiority of hybrid models confirms that the VIX, VXN and VXD indices exhibit nonlinear characteristics.

Finally, the economic significance of the forecasts is assessed by implementing trading strategies with VIX and VXN futures contracts, as well as an S&P 500 VIX midterm futures index ETN. A HAR process has been evaluated economically for the first time by using futures and ETNs. The results indicate that the HAR specifications, and particularly those optimized using the GASVR algorithm, are capable of producing statistically significant profits in normal conditions to some extent, when trading futures contracts. On the other hand, the ETN trading performance reports that HAR specifications can achieve much higher gains because of their lower investor fee rates.

Appendix A. SVR theoretical framework

Considering the training data $\{(x_1,y_1),\ (x_2,y_2),\dots,(x_n,y_n)\ \}$, where $x_i \in X \subseteq R$, $y_i \in Y \subseteq R,\ i = 1 \dots n$ and n is the total number of training samples, then the SVR function can be specified as

$$f(x) = w^T\varphi(x) + b \tag{A.1}$$

where w and b are the regression parameter vectors of the function and $\varphi(x)$ is the nonlinear function that maps the input data vector x into a feature space in which the training data exhibit linearity.

The ε-sensitive loss function L_ε finds that the predicted points lie within the tube created by two slack variables, $\xi_i\ \xi_i^*$:

$$L_\varepsilon(x_i) = \begin{cases} 0 \; if \; |y_i - f(x_i)| \le \varepsilon \\ |y_i - f(x_i)| - \varepsilon \;\; if \; other \end{cases}, \; \varepsilon \ge 0 \tag{A.2}$$

However, the lack of information on the noise in the training datasets makes the *a priori* *ε*-margin setting off *ε*-SVR a difficult task. In addition, the parameter ε takes non-negative unconstrained values, which makes the optimal setting very challenging; see Sermpinis et al., (2014). An alternative approach, the *ν*-SVR, can decrease the computational burden and simplify the parametrization.

The *ν*-SVR approach encompasses the ε parameter in the optimization process and controls it with a new parameter $v \in (0, 1)$. The optimization problem transforms to

$$\text{Minimize } C[v\varepsilon + \frac{1}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*)] + \frac{1}{2}\|w\|^2 \tag{A.3}$$

$$\text{subject to} \begin{Bmatrix} \xi_i \ge 0 \\ \xi_i^* \ge 0 \\ C \ge 0 \end{Bmatrix} \text{and} \begin{Bmatrix} y_i - w^T\varphi(x) - b \le +\varepsilon + \xi_i \\ w^T\varphi(x) + b - y_i \le +\varepsilon + \xi_i^* \end{Bmatrix}$$

The above quadratic optimization problem is transformed into a dual problem, and its solution is based on the introduction of two Lagrange multipliers α_i and $\alpha_i^*$and the mapping with kernel function $K(x_i, x)$:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + \text{b where } 0 \le \alpha_i, \alpha_i^* \le \frac{C}{n} \tag{A.4}$$

The application of the kernel function transforms the original input space into one with more dimensions, in which a linear decision border can be identified. Factor b is computed following Karush–Kuhn–Tucker conditions. A detailed mathematical explanation of the above solution can be found in Vapnik (1995). Support vectors (SVs) (x_i in Eq. (7)) lie outside the *ε*-tube[21], whereas non-SVs lie within the *ε*-tube. Increasing *ε* leads to less SV selection, whereas decreasing it results in ‘flatter’ estimates. The norm term $\|w\|^2$ characterizes the complexity (flatness) of the model. The term

[21] An SV is either a boundary vector ($(\alpha_i - \alpha_i^*) \in [-C/n, C/n]$, $\xi_i, = \xi_i^* = 0$), $\alpha_i, = \alpha_i^* = \frac{C}{n} and \; \xi_i, \; \xi_i^* > \; 0$).

$[v\varepsilon + \frac{1}{n}\sum_{i=1}^{n}(\xi_i + \xi_i^*)]$ is the training error, as specified by slack variables. In particular, in the 'v-trick', as presented by Scholkopf et al., (1999), increasing ε leads to a proportional increase of the first term (training error) in Eq. (7), whereas its second term decreases proportionally to the fraction of side the ε-tube. Hence, v can be considered the upper bound on the fraction of errors. Conversely, decreasing ε leads again to a proportional change of the first term, but the change in the second term is also proportional to the fraction of SVs. In other words, ε will shrink as long as the fraction of SVs is smaller than v, meaning that v is also the lower band in the fraction of SVs. Consequently, the introduction of the parameter C satisfies the need to trade model complexity for training error, and vice versa (Cherkassky and Ma, 2004). In general, both terms cannot be minimal or close to zero at the same time. The SVR algorithm estimates the w and b of the linear function of Eq. 4 with the predefined ε and C for the resulting regression function so as to achieve a good generalization ability. This result should not be too complex, while at the same time avoiding many training errors. If this balance is achieved, then the SVR offers a solution to the overfitting problem.

Appendix B. GA theoretical framework

GAs, introduced by Holland (1995), are search algorithms that are inspired by the principle of natural selection. They are useful and efficient if the search space is large and complicated or if there is no mathematical analysis of the problem available. A population of candidate solutions, called chromosomes, is optimized via a number of evolutionary cycles and genetic operations, such as crossovers or mutations[22]. Chromosomes consist of genes, which are the optimizing parameters. At each iteration (generation), a fitness function is used to evaluate each chromosome, measuring the quality of the corresponding solution, and the fittest chromosomes are selected to survive. This evolutionary process is continued until certain termination criteria are met. In general, GAs can address large search spaces and do not become trapped in local optimal solutions like other search algorithms.

[22] The specifications of the GA were based on the guidelines of Koza (1992).

The GA uses the *one-point crossov*er and the *mutation operator*. The *one-point crossover* creates two offspring from each pair of parents. The parents and a crossover point c_x are selected at random. The two offspring are made by concatenating the genes that precede c_x in the first parent with those that follow (and include) c_x in the second parent. The probability of selecting an individual as a parent for the crossover operator is called the *crossover probability*. The offspring produced by the crossover operator replace their parents in the population. Conversely, the mutation operator places random values in randomly selected genes with a certain probability, called the *mutation probability*. This operator is very important for avoiding local optima and ensuring the exploration of a larger surface of the search space. For the selection step of the GA, the *roulette wheel selection process* is used (Holland, 1995). In roulette wheel selection, chromosomes are selected according to their fitness. The better the chromosomes, the more chances they have of being selected. Usually, elitism is used to raise the evolutionary pressure on better solutions and to accelerate the evolution. Thus, we ensure that the best solution is copied to the new population without changes, so that the best solution found in a generation can survive at the end of that generation.

Appendix C. Cumulative return figures

Figs. C.1-C.3 present the cumulative returns of the best two models in terms of their profitability over time for the VIX futures, VXN futures and VXZ ETN.

[Fig. C.1.]

[Fig. C.2.]

[Fig. C.3.]

These figures show that all of the model strategies present relatively stable performances in terms of profitability, with no large drawdowns.

List of tables

Table 1. The VIX, VXN and VXD dataset-Neural Network's and GASVR algorithm training datasets.

	Name of period	Trading days	Start date	End date
Dataset 1	*Total dataset*	1830	05 August 2002	06 November 2009
	Training set	1538	05 August 2002	11 September 2008
	Out-of-sample dataset	292	12 September 2008	06 November 2009
Dataset 2	*Total dataset*	1830	03 January 2007	09 April 2014
	Training set	1538	03 January 2007	11 February 2013
	Out-of-sample dataset	292	12 February 2013	09 April 2014

Table 2. Descriptive statistics for the levels and logarithms of the implied volatility indices.

	VIX	VXN	VXD
Summary statistics (levels)			
Mean	20.627	24.284	19.082
Standard deviation	9.6471	10.132	8.8870
Skewness	2.1375	1.8177	2.0846
Kurtosis	9.2621	6.9029	8.7878
Jarque-Bera	0.0000	0.0000	0.0000
Summary statistics (logs)			
Mean	2.9442	3.1212	2.8671
Standard deviation	0.3863	0.3532	0.3838
Skewness	0.8226	0.8403	0.8715
Kurtosis	3.4344	3.2800	3.4034
Jarque-Bera	0.0000	0.0000	0.0000

The period examined is from August 5, 2002, to April 4, 2014. We report the sample mean, standard deviation, skewness and kurtosis, as well as the *p*-values of the Jarque–Bera test for normality.

Table 3. Unit root, stationarity and long memory test for the logarithms of VIX, VXN and VXD indices.

Tests	VIX	VXN	VXD
ADF	0.0000	0.0000	0.0000
PP	0.0000	0.0000	0.0000
KPSS	0.0690	0.0540	0.0650
V/S	5.1570	5.2570	5.3820

The *p*-values of the ADF and PP tests are reported. The table also shows the values of the KPSS test statistic for the stationarity property, the critical values of which are 0.119, 0.146 and 0.216 at the 10%, 5% and 1% significance levels, respectively. Finally, the values of the V/S test for long memory are reported, with the critical values being 1.36 and 1.63 at the 5% and 1% levels, respectively.

Table 4. Volatility Indices (VIX and VXN) futures contracts

Delivery month of the contract	Available trading days
01 April 2013	190
01 June 2013	188
01 August 2013	186
01 October 2013	167
01 December 2013	188
01 February 2014	185
01 April 2014	187

Table 5. Out-of-sample performances of model specifications for each of the implied volatility indices from September 12, 2008 to November 6, 2009.

12/09/2008-06/ 11/2009		VIX	VXN	VXD
RW	*MAE*	0.1791	0.1707	0.1874
	RMSE	0.2103	0.1974	0.2166
AR(1)	*MAE*	0.0517	0.0450	0.0524
	RMSE	0.0732	0.0620	0.0733
ARFIMA	*MAE*	0.0519	0.0456	0.0520
	RMSE	0.0730	0.0622	0.0725
ARFIMA-GASVR (res)	*MAE*	0.0524	0.0457	0.0518
	RMSE	0.0730	0.0633	0.0731
HAR	*MAE*	0.0470	0.0419	0.0472
	RMSE	0.0646	0.0557	0.0636
HAR-RNN	*MAE*	0.0471	0.0418	0.0473
	RMSE	0.0650	0.0565	0.0651
HAR-GASVR	*MAE*	0.0330	0.0418	0.0421
	RMSE	0.0452	0.0568	0.0579
HAR-GASVR (res)	*MAE*	0.0300	0.0392	0.0383
	RMSE	0.0430	0.0542	0.0521

We report the out-of-sample performances of the models under study for the period September, 12, 2008 to November, 6, 2009, based on the mean absolute error (MAE) and mean squared error (MSE) criteria computed for each model's out-of-sample forecasts. The smaller the value of each criterion the better the predictive ability of the model considered.

Table 6. Out-of-sample performances of model specifications for each of the implied volatility indices from February 2, 2013 to April 9, 2014.

12/02/2013-09/04/2014		VIX	VXN	VXD
RW	*MAE*	0.0809	0.0724	0.0732
	RMSE	0.1006	0.0850	0.0916
AR(1)	*MAE*	0.0490	0.0423	0.0451
	RMSE	0.0720	0.0580	0.0634
ARFIMA	*MAE*	0.0488	0.0420	0.0442
	RMSE	0.0716	0.0573	0.0623
ARFIMA-GASVR (res)	*MAE*	0.0480	0.0424	0.0459
	RMSE	0.0681	0.0580	0.0657
HAR	*MAE*	0.0489	0.0411	0.0425
	RMSE	0.0683	0.0545	0.0575
HAR-RNN	*MAE*	0.0490	0.0388	0.0395
	RMSE	0.0685	0.0543	0.0532
HAR-GASVR	*MAE*	0.0470	0.0358	0.0405
	RMSE	0.0610	0.0475	0.0548
HAR-GASVR (res)	*MAE*	0.0388	0.0317	0.0354
	RMSE	0.0522	0.0435	0.0489

We report the out-of-sample performances of the models under study for the period February, 2, 2013 to April, 9, 2014, based on the mean absolute error (MAE) and mean squared error (MSE) criteria computed for each model's out-of-sample forecasts. The smaller the value of each criterion the better the predictive ability of the model considered.

Table 7. Giacomini-White test for the mean squared error: the VIX index.

VIX	ARFIMA	ARFIMA-GASVR	HAR	HAR-RNN	HAR-GASVR
12/09/2008-06/ 11/2009					
ARFIMA-GASVR	0.205				
HAR	0.038**	0.048**			
HAR-RNN	0.033**	0.043**	0.225		
HAR-GASVR	0.001***	0.002***	0.000***	0.000***	
HAR-GASVR (res)	0.001***	0.000***	0.000***	0.000***	0.000***
12/02/2013-09/04/2014					
ARFIMA-GASVR	0.151				
HAR	0.186	0.133			
HAR-RNN	0.227	0.141	0.253		
HAR-GASVR	0.026**	0.004***	0.000***	0.001***	
HAR-GASVR (res)	0.000***	0.000***	0.000***	0.000***	0.000***

The out-of-sample periods covered run from September 12, 2008, to November 06, 2009, and from February 12, 2013, to April 9, 2014. The *p*-values of the GW statistic presented indicate agreement with the null hypothesis that the performance of the model in the column is equivalent to that of the model in the row in terms of mean squared errors.

* Denotes a rejection of the null hypothesis at the 10% level of significance.

** Denotes a rejection of the null hypothesis at the 5% level of significance.

*** Denotes a rejection of the null hypothesis at the 1% level of significance

Table 8. Giacomini-White test for the mean squared error: the VXN index.

VXN	ARFIMA	ARFIMA-GASVR	HAR	HAR-RNN	HAR-GASVR
12/09/2008-06/ 11/2009					
ARFIMA-GASVR	0.151				
HAR	0.030**	0.043**			
HAR-RNN	0.018**	0.031**	0.275		
HAR-GASVR	0.015**	0.028**	0.24	0.504	
HAR-GASVR (res)	0.001***	0.006***	0.024**	0.000***	0.000***
12/02/2013-09/04/2014					
ARFIMA-GASVR	0.137				
HAR	0.000***	0.001***			
HAR-RNN	0.040**	0.048**	0.627		
HAR-GASVR	0.000***	0.000***	0.000***	0.154	
HAR-GASVR (res)	0.000***	0.000***	0.000***	0.066*	0.000***

The out-of-sample periods covered run from September 12, 2008, to November 06, 2009, and from February 12, 2013, to April 9, 2014. The *p*-values of the GW statistic presented indicate agreement with the null hypothesis that the performance of the model in the column is equivalent to that of the model in the row in terms of mean squared errors.

* Denotes a rejection of the null hypothesis at the 10% level of significance.

** Denotes a rejection of the null hypothesis at the 5% level of significance.

*** Denotes a rejection of the null hypothesis at the 1% level of significance

Table 9. Giacomini-White test for the mean squared error: VXD index.

VXD	ARFIMA	ARFIMA-GASVR	HAR	HAR-RNN	HAR-GASVR
12/09/2008-06/ 11/2009					
ARFIMA-GASVR	0.255				
HAR	0.042**	0.052*			
HAR-RNN	0.011**	0.020**	0.128		
HAR-GASVR	0.002***	0.013**	0.000***	0.009***	
HAR-GASVR (res)	0.002***	0.003***	0.000***	0.000***	0.000***
12/02/2013-09/04/2014					
ARFIMA-GASVR	0.087*				
HAR	0.000***	0.014**			
HAR-RNN	0.000***	0.000***	0.000***		
HAR-GASVR	0.000***	0.000***	0.000***	0.139	
HAR-GASVR (res)	0.000***	0.000***	0.000***	0.000***	0.000***

The out-of-sample periods covered run from September 12, 2008, to November 06, 2009, and from February 12, 2013, to April 9, 2014. The *p*-values of the GW statistic presented indicate agreement with the null hypothesis that the performance of the model in the column is equivalent to that of the model in the row in terms of mean squared errors.

* Denotes a rejection of the null hypothesis at the 10% level of significance.

** Denotes a rejection of the null hypothesis at the 5% level of significance.

*** Denotes a rejection of the null hypothesis at the 1% level of significance

Table 10. SPA and MCS tests for the out-of-sample periods: the VIX, VXN and VXD indices.

	VIX				VXN				VXD			
	SPA		MSC		SPA		MCS		SPA		MSC	
12/09/2008-06/ 11/2009	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
RW	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AR(1)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ARFIMA	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000
ARFIMA-GASVR	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000
HAR	0.0000	0.0000	0.0000	0.0000	0.2287	0.0101	0.2923*	0.0573	0.0000	0.0000	0.0000	0.0000
HAR-RNN	0.0000	0.0000	0.0000	0.0000	0.1542	0.0292	0.2923*	0.0573	0.0000	0.0000	0.0002	0.0000
HAR-GASVR	0.2036	0.1870	0.4140*	0.0943	0.0981	0.0377	0.2306*	0.0573	0.0069	0.0065	0.0129	0.0107
HAR-GASVR (res)	0.7964	0.9987	1.0000*	1.0000*	0.9669	0.6016	1.0000*	1.0000*	0.5129	0.5092	1.0000*	1.0000*
12/02/2013-09/04/2014												
RW	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AR(1)	0.0000	0.0000	0.0013	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ARFIMA	0.0031	0.0010	0.0016	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ARFIMA-GASVR	0.0002	0.0000	0.0013	0.0003	0.0000	0.0000	0.0000	0.0000	0.0005	0.0000	0.0001	0.0000
HAR	0.0040	0.0010	0.0013	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
HAR-RNN	0.0000	0.0000	0.0013	0.0002	0.0812	0.0035	0.0793	0.0014	0.0123	0.0068	0.0237	0.0050
HAR-GASVR	0.0090	0.0010	0.0016	0.0002	0.0067	0.0001	0.0275	0.0009	0.0020	0.0002	0.0033	0.0008
HAR-GASVR (res)	0.7763	0.7670	1.0000*	1.0000*	0.9308	0.6303	1.0000*	1.0000*	0.5225	0.5130	1.0000*	1.0000*

The table reports the p-values of the SPA (Hansen, 2005) and MCS (Hansen et al., 2011) tests in terms of the MSE and MAE criteria. Low p-values indicate either that the respective benchmark model is inferior to at least one alternative (SPA) or that it is unlikely that the model will belong to the set of the 'best' models (MCS).

* Denotes that the model examined belongs to the set of 'best' models at the 95% confidence level.

Table 11. Trading performance of the VIX, VXN futures and the iPath S&P 500 VIX mid-term futures index ETN from February 12, 2013, to April 9, 2014.

	VIX		VXN
	Futures	ETN (VXZ)	Futures
ARFIMA			
Sharpe ratio	-0.046	-0.069	-0.037
95% CI	(-0.1)-0.01	(-0.12)-0.00	(-0.09)-0.02
Leland's Ap	-0.039	-0.016	-0.023
95% CI	(-0.09)-0.01	(-0.02)-0.00	(-0.06)-0.01
ARFIMA-GASVR (res)			
Sharpe ratio	-0.053	-0.017	0.006
95% CI	(-0.11)-0.0	(-0.07)-0.04	(-0.05)-0.12
Ap	-0.044	-0.004	0.004
95% CI	(-0.09)-0.00	(-0.01)-0.01	(-0.03)-0.04
HAR			
Sharpe ratio	0.088*	0.478*	0.084*
95% CI	0.02-0.14	0.41-0.54	0.02-0.014
Ap	0.088*	0.386*	0.060*
95% CI	0.03-0.14	0.37-0.40	0.02-0.10
HAR-RNN			
Sharpe ratio	-0.027	0.519*	0.087*
95% CI	(-0.08)-0.03	0.45-0.58	0.02-0.14
Ap	-0.024	0.398*	0.061*
95% CI	(-0.08)-0.03	0.38-0.41	0.04-0.10
HAR-GASVR			
Sharpe ratio	0.081*	0.096*	0.033
95% CI	0.02-0.14	0.03-0.15	(-0.02)-0.09
Ap	0.088*	0.294*	0.026
95% CI	0.03-0.14	0.27-0.30	(-0.02)-0.07
HAR-GASVR (res)			
Sharpe ratio	0.184*	0.721*	0.127*
95% CI	0.10-0.26	0.63-0.80	0.07-0.18
Ap	0.168*	0.451*	0.098*
95% CI	0.09-0.24	0.43-0.47	0.03-0.16

We report the out-of-sample annualized Sharpe ratio (*SR*) and the annual Leland's (1999) alpha (A_p) for the period February 12, 2013, to April 9, 2014, based on the examined models' forecasts. A simple trading rule is followed: The investor goes long (short) in the volatility futures and the ETN when the forecasted value of the implied volatility index is greater (smaller) than its current value. We also report the 95% confidence intervals of the *SRs* and A_p values for each forecasting model, based on bootstrap simulations, to assess the statistical significance of the returns.

* Denotes the rejection of the null hypothesis of a zero return at the 5% level of significance.

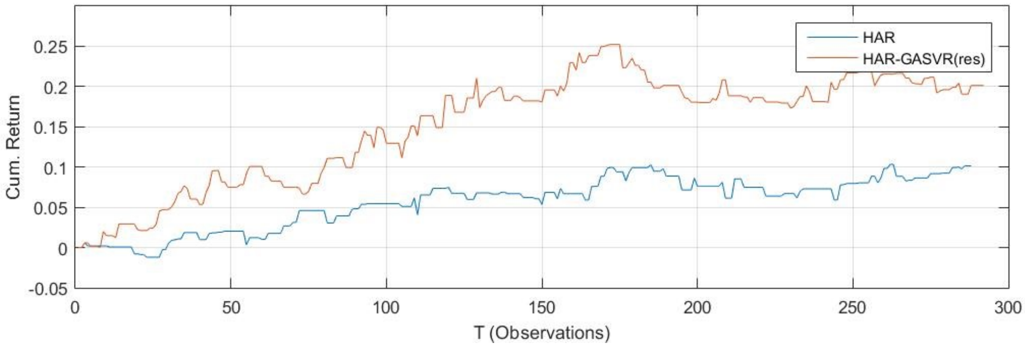

Fig. C.1. Cumulative returns of HAR and HAR-GASVR(res) in the out-of-sample for VIX futures.

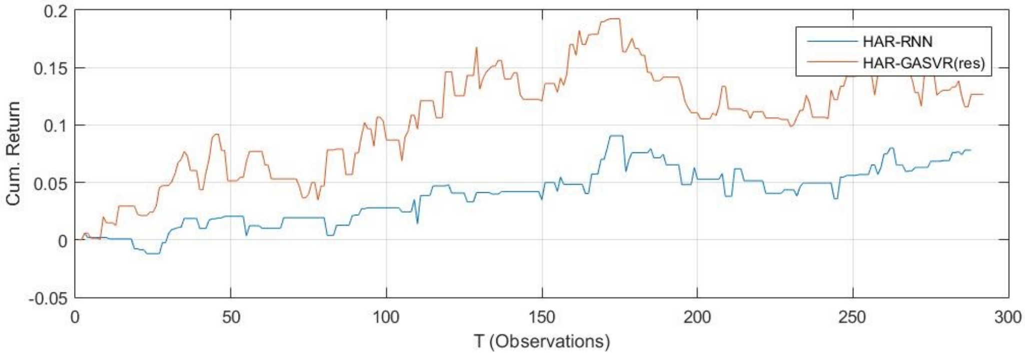

Fig. C.2. Cumulative returns of HAR-RNN and HAR-GASVR(res) in the out-of-sample for VXN futures.

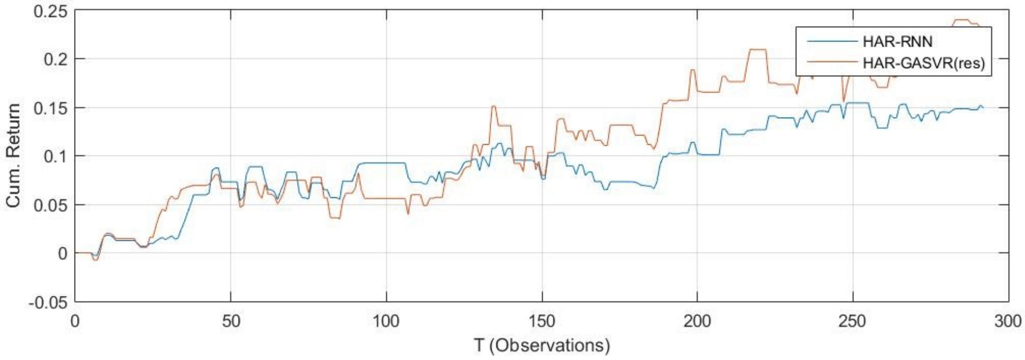

Fig. C.3. Cumulative returns of HAR and HAR-GASVR(res) in the out-of-sample for VXZ ETN.

CHAPTER 3
PERFORMANCE OF TECHNICAL TRADING RULES: EVIDENCE FROM THE CRUDE OIL MARKET

1. Introduction

Technical analysis (sometimes referred to as *chartism*) is believed to be one of the longest-established forms of investment analysis, being a set of graphical or mathematical techniques exploring future trading opportunities for financial assets just by analyzing the time-series history of their asset prices, volume data, and a summary of securities statistics. Brock et al. (1992) mention that technical trading rules "*beating the market*" is supposed to be as old as the U.S. stock market itself. Nowadays, investment funds, brokerage firms, and trading platforms from all over the world utilize numerous types of technical indicators and oscillators as prospective moneymaking tools.

On the other hand, despite the undisputable popularity of technical analysis among practitioners, academia has been rather skeptical for a long time now about its merits, and there is an ongoing debate as to whether the generated profits are just lucky. On the *effectiveness* of this form of analysis and its power to yield profits, Malkiel (1981) describes it pertinently as the "*anathema*" of the academic world, which usually loves to pick on it. This argument is derived from the Efficient Market Hypothesis (EMH), which expresses that security prices reveal all the available information to investors. However, since the 1960s, prominent academics and practitioners have claimed that predictable patterns do exist in returns (especially in certain periods of time), which can lead to *abnormal* profits.[23] In this regard, even Keynes (1936) outlines that most traders' decisions can be deemed a consequence of "*animal spirits*". This conclusion is closely connected with Lo's (2004) proposed Adaptive Market Hypothesis (ADH) that assumes that evolutionary market dynamics, such as cycles, trends, and market inefficiencies, can trigger occasionally arbitrage opportunities.

This fruitful debate has culminated in a large number of empirical studies employing technical trading rules in several markets and for different indices. Some have found results to support the notion that trading strategies are able to deliver superior returns, at

[23] Earlier studies of technical analysis and patterns in stock returns include Alexander (1961, 1964), Fama (1965, 1970), Fama and Blume (1966), Levy (1967), James (1968), Jensen and Benington (1970), and Sweeney (1980).

least in certain time periods (Neftci 1991; Brock et al. 1992; Neely et al. 1997; Conrad and Kaul 1998; Sullivan et al. 1999; Lo et al. 2000; Kavajecz and Odders-White 2004; Qi and Wu 2006; Hsu et al. 2010; Neely and Weller 2011; Shynkevich 2012; Taylor 2014). Despite this, other papers report that technical trading strategies are unable to predict future prices, especially when transaction costs are considered (Bessembinder and Chan 1998; Allen and Karjalainen 1999; Ready 2002; Marshall et al. 2008; Bajgrowicz and Scaillet 2012; Yamamoto 2012).

Inevitably, *data snooping* effects arise in most of the studies mentioned above, particularly when a large number of trading strategies are implemented and tested. Amongst the pioneers in studying the data snooping effects are Jensen and Benninghton (1970), defining it as "*selection bias*", as well as Lo and MacKinlay (1990) who summarize that more likely patterns can emerge when data are severely exploited. This is apparently true if one considers that, by exploring a sizeable universe of different trading rules, it is highly likely one will find a rule that works well, even by chance. Many efforts have been made to minimize the undesirable consequences of data snooping, which are illustrated in the studies of White (2000), Romano and Wolf (2005), Hansen (2005), Romano and Wolf (2007), Romano et al., (2008), Hsu et al., (2010), Bajgrowicz and Scaillet (2012), and Hsu et al., (2014).

In this paper, we evaluate the performance of the whole universe of 7846 technical trading rules (TTRs) proposed by Sullivan et al., (1999) on the crude oil market. This universe of rules is the most popular and common, and creates a connection with the previous literature in this field of research (Brajgowicz and Scaillet 2012; Marshall et al., 2008). In particular, we apply five families of strategies (i.e., *filter rules*, *moving averages*, *support and resistance*, *channel breakout* and *on-balance volume averages*) to the daily prices of West Texas Intermediate (WTI) light, sweet crude oil futures, as well as the United States oil (USO) fund, covering a period from April 2006 to January 2016.[24]

Crude oil futures offer the opportunity to trade one of the world's most liquid oil commodities on the New York Mercantile Exchange (NYMEX) for up to 108

[24] We chose this specific period in order to examine the TTRs' performance on the same trading days for crude oil futures and the USO, given that the inception date of the USO was in April 2006.

consecutive months.[25] In line with Marshall et al., (2008) and Wang and Yu (2004) we employ Datastream continuous price series of crude oil futures, which represent the price of the most actively traded contract. This guarantees that the underlying instrument should last longer than the observation period when analyzing the performance of the TTRs. Furthermore, futures markets are more attractive for pursuing active trading strategies than stock markets, since they involve much lower transaction costs (e.g. spreads and commissions), and short-selling is easily applied.

The USO is the largest and most liquid oil-related exchange traded fund (ETF) ($3.7 billion in assets), and is designed to track the daily price movements of WTI light, sweet crude oil.[26] It is exposed to crude oil prices by means of holding positions on front-month crude oil futures contracts. The USO is free of the substantial storage costs involved in other crude oil inventories, entailing low total costs of management, a feature that makes it very attractive to investors. As far as we are concerned, this is the first time that the effectiveness of TTRs will be tested specifically on the crude oil market. We believe that it is a rather interesting area of investigation, since crude oil prices have exhibited considerable fluctuations over the years in response to geopolitical and economic turmoil.[27] Today the oil industry is being shaped by one of its most dramatic price movements of recent times, having experienced almost a 70% fall since June 2014. These extreme fluctuations mark the crude oil market out as a *trending* market, potentially lucrative for applying the TTRs, since trend following is one of the key aspects of technical analysis. Furthermore, since previous empirical findings on the hedge fund industry and the Dow Jones Industrial Average (DJIA) index have shown that TTRs perform quite well when *strong* negative or positive returns occur in the market (Fung and Hsieh 1997; Bajgrowicz and Scaillet 2012), we have a strong motivation to explore them on the crude oil market as well.

[25] The final settlement date is the 4th U.S. business day prior to the 25th calendar day of the month preceding the contract month.

[26] The USO periodically "rolls over" its underlying futures contracts by selling those that are approaching expiration and buying those that expire farther into the future. The investment objective of the USO is publicized on its website (http://www.unitedstatesoilfund.com/).

[27] For instance, in 2008, crude oil reached its highest value, followed by a fall below $50 per barrel due to the Lehman Brothers crisis in 2009.

In this paper, the first contribution is to revisit the historical success of TTRs in the oscillated market of crude oil.[28] For this purpose, we divide and assess the rules' performance in four different subperiods, each one characterized mostly by having bearish or bullish trends. This allow us to access the power of multiple hypothesis testing methods in an environment in which momentum trading rules tend to work well by definition, and so measure the level of significance of rules yielding positive performance. As crude oil futures and ETFs have small expense ratios compared to other assets, there is a strong potential for trading rules to achieve gains. On the other hand, low transaction costs may help to increase market liquidity, leading to market efficiency (Hedge and McDermott 2004). Thus, the second contribution is to give an answer to the question of which one of the above two cases holds in the case of the crude oil market. With this aim, for the in-sample analysis, we compare the performance when applying TTRs to the USO and the crude oil futures to explore potential differences due to contango or backwardation effects as a result of rolling over crude oil futures in the case of the USO. An in-sample analysis including the impact of transaction costs is also reported. The transaction costs are embodied endogenously in the trading rule selection process, each time a buy or sell signal is generated. This helps investors foresee which TTRs' performance can outweigh transaction costs ex ante (Bajgrowicz and Scaillet 2012). The reason for this is that strategies' predictability is occasionally neutralized when TTRs are selected before the implementation of transaction costs because of frequent signals. Another contribution is that, instead of only using regular evaluation criteria, such as the mean return and the Sharpe ratio, we also employ the Calmar ratio criterion as an essential performance measure for technical traders, especially when momentum strategies are employed. Indeed, the Calmar ratio is an important indicator for the hedge fund industry in general since it displays the average annual return on an investment, per unit of maximum drawdown (Schuhmacher and Eling, 2011). In technical analysis its usage is particularly useful, especially when momentum strategies, which can suffer significant drawdown, are employed.

Finally, a comprehensive persistence analysis of TTRs is employed. For that analysis the *false discovery rate* (FDR) technique (Bajgrowicz and Scaillet 2012; Barras, Scaillet and Wermers 2010) is used to minimize data snooping effects. The performance of the

[28] Marshal et al. (2008) evaluate the performance of the Sullivan et al. (1999) universe of TTRs in 15 major commodities futures series, while considering naïve methods of accounting for data snooping effects. One of these series refers to crude oil futures covering the period 1984-2005.

FDR technique is compared with the equally powerful *k-familywise error rate* (*k*-FWER) technique of Romano and Wolf (2007) and Romano et al., (2008), rather than more conservative methods such as the *bootstrap reality check* (BRC) of White (2000) or its stepwise extension proposed by Romano and Wolf (2005). The rationale behind this is our desire to investigate whether a generalized version of the conservative FWER measure could demonstrate the same performance level as the powerful FDR. In particular, we are the first to assess the out-of-sample performance of a portfolio of genuine TTRs, while applying the FDR and *k*-FWER methods respectively, and also including transaction costs. The portfolio is constructed and rebalanced on a semi-annual basis, each time using data from the previous six months and evaluating its profitability in the next half of the year. We report that the powerful nature of the FDR approach to identifying genuine TTRs is also verified in the case of crude oil. Furthermore, we observe that the less conservative *k*-FWER than the strict FWER of Romano and Wolf (2005), can achieve similar results to the FDR in terms of trading performance, while also allowing a certain number of false selections to occur. Moreover, the FDR succeeds in selecting a larger amount of genuine TTRs than the *k*-FWER portfolio.

For the in-sample simulation period, the findings indicate that more than half of the TTRs exhibit great predictive power, especially in periods of substantial crude oil price movements. Additionally, the best TTRs are able to achieve high mean returns, as well as Sharpe and Calmar ratios, across the whole period considered. On the other hand, when it comes to persistence analysis, the TTRs selected by the data snooping methods show no persistent nature in the out-of-sample period. Although, the portfolios are able to generate positive performance in some periods, we observe that the superior returns are considerably small. However, there is only one period in which both portfolios achieve a Sharpe ratio slightly bigger than 1. This might be a justification of the ADH, which supports the notion that profits might occur over some horizons. Overall, we conclude that the best-performing TTRs are mostly accessible to investors observing the returns ex post, in an in-sample period, and therefore it is not easy for them to foresee truly out-of-sample profitable rules ex ante, without hindsight.

The remainder of the paper is structured as follows. In Section 2, a detailed description of the universe of TTRs proposed by Sullivan et al. (1999), as well the performance criteria, are provided. Section 3 describes the time series and the descriptive statistics of the data considered. Section 4 presents a synopsis of the existing methods accounting for data snooping. A detailed description of the FDR and *k*-FWER approaches, as well as the portfolios' characteristics, is provided in the same section. Section 5 provides evidence of the TTRs performance in the in-sample period, with and without consideration of transaction costs. In Section 6, the persistence analysis is presented, while accounting for transaction costs at the same time. Finally, Section 7 presents the concluding remarks.

2. Technical trading rules universe and performance measures

In Section 2.1, we review the universe of TTRs proposed by Sullivan et al. (1999). We briefly present the performance measurement tools of mean return, Sharpe ratio, and Calmar ratio in Section 2.2.

2.1. Technical trading rules universe

Technical analysis incorporates a large spectrum of approaches as a form of predictive modeling. Although these methods use mostly graphical rather than mathematical or statistical tools, they use time series of past prices, volumes, and other observables to define whether a buy (long), neutral (out of the market), or sell (short) strategy should be taken within the next time period. As stated earlier, we adopt the whole universe of 7846 TTRs for each subperiod for comparison purposes. The universe is separated into five categories of indicators, while different parameterizations can be employed for each rule.[29]

Filter rules: An investor *buys* if the price *increases* by a fixed percentage from a previous *low*, and he *sells* if the price *decreases* by a fixed percentage from a previous *high*. An alternative definition of subsequent highs (lows) can be defined as the highest (lowest) closing price observed over a prespecified number of previous days, excluding the current day. Thus, the filter rule allows the initiation of an investor's position only in response to major price trends. We also consider the impact of two extra filters. The first one allows a neutral position when the price increases (decreases) from a previous

[29] The interested reader can refer to the appendix of Sullivan et al., (1999) for a detailed description of each rule as well as the extra parameters used.

low (high) by a smaller percentage than the percentage needed to initiate a buy (sell) position. The second one assumes a position is held for a fixed number of periods.

Moving averages: Assumes a crossover between short-long moving averages to generate a trade. Usually, an investor buys (sells) when the short-moving average moves above (below) the long-moving average. These upside (downside) penetrations of a moving average help an investor to discover new trends and maintain his position as long as the crossover remains. Three extra filters are applied. The first one demands that the short-moving average penetrates the long-moving average by a fixed percentage, otherwise no position is initiated. The second one applies a delay filter, which requires a signal to remain valid for a prespecified number of days before an action is taken. The third one considers a holding period similar to the one employed in filter rules.

Support and resistance: A trader buys (sell) when the price rises above (below) the local maximum (minimum) over the previous *n* days. The intuition behind this rule is that usually investors think that sooner or later the movement of the equity's price will tend to stop and return to a certain level (sell at the peak and buy at the bottom). However, if the price breaks through a certain resistance (support) level, it is more likely to continue drifting upward (downward) until it finds a new resistance (support) level. Thus, a buy (sell) signal is triggered. An alternative definition of extreme highs/lows is also used, similar to the one employed in filter rules. In addition to that, fixed-percentage-band, delay and holding-period filters are imposed.

Channel breakouts: An investor buys (sells) when the price moves above (below) the channel. A channel occurs when the high over the previous, prespecified, days is within a fixed percentage of the low over the previous prespecified days. The graphical representation of a price channel is equal to a pair of parallel trend lines. As soon as one of these trend lines is "broken", a buy or sell signal is generated. A fixed percentage band is also exercised around the channel, as well as a holding period for each position triggered.

On-balance volume averages: These operate in a similar way to the moving-average rules (crossover between short/long on-balance volumes). However, the indicator here is the volume. The economic meaning is that the volume is greater on days when the price movement is an extreme fall (bearish) or an extreme rise (bullish). A technical trader adds (subtracts) the daily volume to (from) that of the previous day, when the current closing price has increased (decreased), in order to construct the new on-balance volume indicator. Then, a moving average is applied. Furthermore, the same filters as in the case of the moving average are used.

All of the trading rules described above are considered *momentum* or *trend-following* TTRs except for the support and resistance rules that can be deemed *contrarian* (mean-reverting) trading strategies.

2.2. Measuring performance

The performance of the TTRs is mainly assessed through the *mean return* and *Sharpe ratio* criteria. The mean return is the absolute criterion of each rule's returns, while the Sharpe ratio is a relative performance criterion since it represents the ratio of the average excess return to the total risk of the investment. Practically speaking, the TTRs earn the risk-free rate in periods when a neutral signal is triggered.[30] In our analysis, we are the first in the relevant literature to evaluate the performance of TTRs by also employing the Calmar ratio. The Calmar ratio[31] is an important indicator for investment banks as well as the hedge fund industry in general, since it displays the average annual return of an investment per unit of maximum drawdown. Furthermore, practitioners find it of great importance, especially when they are dealing with momentum strategies that can suffer a considerable drawdown. On the other hand, the Sharpe ratio is mostly applied for mean-reverting or contrarian strategies.[32]

Specifically, let $s_{j,t-1}$ denote the trading signal for each trading rule $j, 1 \leq j \leq l$ (where $l = 7846$) at the end of each prediction period $t - 1$ $(\tau \leq t \leq T)$, where $s_{j,t-1} = 1, 0,\ or - 1$ represents a long, neutral or short position taken at time t. In addition to that let r_t designate the return of the price series exercised, and r_t^f be the

[30] Actually, following the studies of Brock et al., (1992), Sullivan et al., (1999), and Bajgrowicz and Scaillet (2012) who implement the "double-or-out" trading strategy, a buy signal leads a trader to borrow money at the "risk-free" rate in order to double the investment in the commodity portfolio, a neutral signal leads to the trader simply holding the commodity, and when a sell signal occurs the trader liquidates and exits the market.

[31] Developed by Young (1991), the Calmar ratio stands for California Managed Account Reports. It is a performance measurement used to evaluate commodity trading advisors and hedge funds.

[32] We acknowledge Ernest P. Chan for pointing this out to us.

"risk-free" rate.[33] The mean return criterion $\overline{f}_{j,t}$ for the trading rule j at time t is defined by

$$\overline{f}_{j,t} = \frac{1}{N}\sum_{\tau=R}^{T} \ln\left(1 + s_{j,t-1} r_t\right),\ j = 1, \dots, l, \qquad (1)$$

where $N = T - \tau + 1$ is the number of days examined. We denote as τ the start date for each subperiod, and even for the first one, since lagged values up to 250 days are employed in the universe of rules. Then, the Sharpe ratio criterion expression SR_j for trading rule j at time t is defined by

$$SR_{j,t} = \frac{1}{N}\sum_{t=R}^{T} \frac{\ln(1+s_{j,t-1}r_t - r_t^f)}{\widehat{\sigma_j}},\ j = 1, \dots, l, \qquad (2)$$

where $\frac{1}{N}\sum_{t=R}^{T} \ln\left(1 + s_{j,t-1} r_t - r_t^f\right)$ and $\widehat{\sigma_j}$ are the mean excess return and the estimated standard deviation of the mean excess return respectively. Finally, the Calmar ratio criterion $Calmar_j$ is obtained as the annualized mean return of each rule j over its maximum drawdown (MDD_j):

$$Calmar_{j,t} = \frac{\overline{f}_{j,t}*252}{MDD_j},\ j = 1, \dots, l, \text{ where } MDD_j = \min[r_t - \max\{\textstyle\sum_{t=R}^{T} r_t\}]. \qquad (3)$$

3. Data description

In this section, the settlement prices and trading volumes for the USO and crude oil futures for four different subperiods are analyzed. Table 1 reports the intervals covered for each one of them, while Fig. 1 represents the time series dynamics for the two underlying instruments examined from April 2007 to January 2016.

[Fig. 1]

[Table 1]

Subperiod 1 is characterized by a sharp increase and subsequent fall in crude oil prices for both CL futures and USO due to the Lehman Brothers collapse and so it is characterized by mixed trends. Subperiod 2 reveals an upward trend for the CL futures series, but this is not quite observable for USO ones. However, we define this period as

[33] We use as a risk-free rate the daily effective federal funds rate, in accordance with all the previous literature.

a bullish due to the upward trend in crude oil spot prices. Subperiod 3 mainly dominated by mixed trends for both cases, while Subperiod 4 includes their recent extreme fall and for that reason is characterized as bearish in terms of trend.

The summary of descriptive statistics of daily buy-and-hold returns for the USO and crude oil futures for the four subperiods is reported in Table 2. We follow the existing literature and calculate the daily returns as the natural logarithm of price relatives. The distribution characteristics are described by the *mean*, *standard deviation*, *skewness* and *kurtosis* statistics as well as the first-order autocorrelation under the Ljung-Box (1978) Q statistics at the 5% significance level.

[Table 2]

The crude oil futures yield positive performance (with 10 basis points as the highest value) for the first three subperiods, with the only exception being the last one in which a highly negative mean return is reported, standing for the bearish market. The USO yields negative returns for all subperiods except the second. The standard deviation is also comparable across both USO and crude oil futures series. However, with regards to skewness, there is a split between negative and positive signs in the case of crude oil futures, while the USO exhibits mostly negative signs. Moreover, a considerable level of kurtosis is observed in both series for all subperiods. The Ljung-Box (1978) Q statistics test indicates that both crude oil futures and the USO have significant first-order autocorrelations in half of the subperiods.[34] Finally, for the last subperiod describing the extreme fall, the first-order autocorrelation is significant for both series.

4. Data snooping bias

Data snooping bias should always be adjusted for when examining the predictive ability of a large number of trading strategies (i.e., technical trading indicators). The issue emerges when the financial dataset is severely exploited by trading rules dependent to each other, such as in our case (i.e., weak dependence between same family rules). This may result in the identification of TTRs that generate profits purely out of luck, and for that reason multiple hypothesis testing is attempted to minimize data snooping bias. In addition, selecting one trading rule recognized as the best, without consideration of

[34] Most of the trading rules employed in this study are designed to capture momentum. Their effectiveness is mainly based on the existence of significant autocorrelation of returns series.

the entire universe of strategies that it is pooled from, when its statistical inference is tested can also lead to false discoveries. In our paper, we employ and compare the FDR and the k-FWER which are two of the most powerful data snooping methodologies in the relevant literature. In Section 4.1, we review the existing data snooping methods. The data snooping specifications employed in this study are outlined in Sections 4.2 and 4.3, respectively. The multiple hypothesis testing setup together with the construction of the portfolio of TTRs following the proposed methods are described in Sections 4.5 and 4.6.

4.1. Existing data snooping methods

The finance literature introduces several methods for mitigating data snooping bias. The majority of these focus on two main statistical approaches for testing multiple hypotheses: the FWER and the FDR. The difference between the two is mostly intuitive, rather than based on conscious reasoning. FWER is defined as the probability of making at least one false rejection (which is unacceptable), while FDR views "unacceptability" in terms of a proportion (Harvey and Liu 2014). For instance, a 10% false discovery rate denotes that more than 10 false discoveries in 100 tests would be unacceptable. Thus, the FWER is more conservative than the FDR, especially when the universe of rules is large.

In statistics, the most standard FWER method is the Bonferroni correction, in which individual null hypotheses (for each one of the total universe of rules) are rejected for each p-value less than a significance level of α/l in a single-step procedure. This structure is employed in the BRC of White (2000) and is carried out in such a way as to reassure that the significance level of the contemporaneous test of all l rules is less than α. In this way, the BRC evaluates whether the "best" performing strategy (drawn from l strategies) has significant predictive power with respect to the performance of the whole universe in a two-tailed-hypothesis framework. The null hypothesis tested is that the performance of the best trading rule is no better than the benchmark (e.g. "risk-free" rate):

$H_{0j}: \max_{j=1,..l} \varphi_j \leq 0$, where φ_j is the performance measure of the jth rule (4)

Even though the BRC is used by Sullivan et al. (1999), we believe that it is a rather conservative measure that lacks power since it focuses only to the best strategy. Now, by also applying the Bonferroni correction, Hansen (2005) presents his *superior predictive* *ability* (SPA) test, which minimizes the influence of poor and inconsistent strategies by using studentized instead of non-studentized test statistics[35]. However, it also focuses on the rule that appears best for the observed financial series. Furthermore, the call to identify more outperforming strategies and not relying only on the best strategy when undertaking investment decisions led to the Holm (1979) method. The Holm method works in a stepwise structure, with individual *p*-values ordered from smallest (most significant) to largest (least significant), and each one compared with a less strict significance level moving "down" the list. Following the Holm procedure, Romano and Wolf (2005) introduce their *stepwise multiple testing* (StepM) method as an improvement to the *single-step* BRC testing method of White (2000), while Hsu et al., (2010) develop a stepwise extension of the SPA test of Hansen (2005). Although stepwise approaches are powerful tools, their main drawback is that they do not select further rules once they have detected a rule whose performance is due to luck.

In practice, investors do not search only for the best rule, but invest money in all possible outperforming strategies. Romano and Wolf (2007) develop a generalized methodology, controlling for the stringent FWER criterion. Their goal is to reject at least a specific number of false hypotheses to maximize diversification. In a similar way, Hsu et al., (2014) apply this generalization to the stepwise method of Hsu et al., (2010) to minimize the data snooping effects on the performance of the Commodity Trading Advisory fund.

Moreover, the FDR tolerates a certain proportion of false rejections so as to construct a well-diversified portfolio of trading rules, while accounting for the data snooping effect. Thus, Bajgrowicz and Scaillet (2012) employ the modified FDR$^{+/-}$ version of Barras et al., (2010) in the context of identifying outperforming TTRs on the DJIA index. Their findings confirm the superiority of the FDR over the conservative FWER approach of Romano and Wolf (2005) in detecting and building a portfolio of genuine rules.

4.2. Multiple hypothesis testing framework

As data snooping techniques are actually multiple hypothesis testing procedures, in what follows we need first to define the test statistic. The Sharpe ratio criterion (as

[35] A studentized test statistic refers to a simple test statistic divided by the consistent estimator of its standard deviation. This helps one to compare objects in the same units of standard deviation.

defined in Section 2.2) is chosen as the test statistic when performing the multiple hypothesis testing using the k-FWER and FDR methods for data snooping.[36] We selected this ratio not only for comparison with previous studies, but also for its undoubtable popularity across traders. The test statistic for each rule j defines the setup under the null hypothesis $(H_{0j}\colon \varphi_j = 0)$ that rule j does *not* outperform the benchmark, where $\varphi_j = SR_j$ in this case. On the contrary, the alternative hypothesis assumes the presence of abnormal performance, positive or negative $(H_{Aj}\colon \varphi_j > 0 \; or \; \varphi_j < 0)$ in a two-tailed test. However, since we are mainly interested in identifying significantly outperforming rules, we define a technical trading rule j as significantly positive, if it displays abnormal performance (i.e., reject H_{0j}) and its performance metric is positive (i.e., $\varphi_j > 0$). The "risk-free" rate is used as a benchmark, describing an investor being out of the market.

4.3. The FDR$^{+/-}$ method

The *FDR$^{+/-}$* has its foundations in the FDR statistical criterion introduced by Benjamini and Hochberg (1995), which assumes that, by tolerating a small proportion of false discoveries amongst all rejections (e.g., significant TTRs), one obtains a more powerful multiple hypothesis testing tool than via the conservative FWER method.[37]

The *FDR$^{+/-}$* has some unique features that make it suitable for traders that are not just looking for the best rule, but also for a class of strategies with genuine predictive power that can help them diversify risk. In Bajgrowicz and Scaillet (2012), the FDR approach provides a sensible trade-off between significantly positive and false selections, making it less strict than the FWER method. Additionally, its comparative advantage is the ability to find the outperforming rules, even if the performance of the best rule in the sample is due to luck. In practice, it is not unusual for such a rule with no significant predictability to achieve the greatest performance in terms of profits. This feature is not

[36] We do not apply the Calmar ratio criterion since its formulation is based on at least a couple years of previous data, while in our persistence analysis we use a rebalancing period of six months (see Section 5).

[37] The initial FDR version of Benjamini and Hochberg (1995) adopted independence across multiple hypotheses. Later, studies by Benjamini and Yekuteli (2001), Storey (2002), and Storey et al. (2004) proved that the FDR holds under "weak dependence" conditions when the number of hypotheses is very large. Also, Bajgrowicz and Scaillet (2012) explain that the Sullivan et al. (1999) trading rules satisfy this feature, since the rules are dependent in small blocks (within the same family) and independent across different families.

available in the other methods, whose stepwise nature prevents them from detecting further outperforming rules once a "lucky" rule has been identified.

The FDR concentrates on estimating the expected value of the ratio of erroneous selections over the rules showing significant performance. Specifically, the *FDR*$^{+/-}$ is defined as the expected value of the proportion of false selections, *F*, among the significant rules, *R* (positive or negative). The latter are just the rules that perform either better or worse than the benchmark while at the same time their *p*-value rejects the null hypothesis of no abnormal performance under some threshold *γ*. Thus, the estimate is given by $\widehat{FDR}^{+/-} = \hat{F}^{+/-}/\hat{R}^{+/-}$, where $\hat{F}^{+/-}$ and $\hat{R}^{+/-}$ are the estimators of $F^{+/-}$and $R^{+/-}$, respectively. For instance, an *FDR*$^{+/-}$ 100% conveys that, among both the outperforming and underperforming trading strategies, no rule generates genuine performance on average and vice versa.

The estimation of *FDR*$^{+/-}$ is not very tedious, especially when the *p*-value of each rule's corresponding test statistic has already been computed. In order to acquire the individual *p*-values, we follow the resampling procedure of Sullivan et al., (1999). Using the stationary bootstrap method of Politis and Romano (1994) to resample the returns of each strategy, the corresponding test statistic for each bootstrap series of returns is calculated.[38] The *p*-value is obtained by comparing the original test statistic (φ_j) to the quantiles of each bootstrapped test statistic vector. The estimate of $\widehat{FDR}$ is given by

$$\widehat{FDR}(\gamma) = \hat{F}/\hat{R} = \frac{\widehat{\pi_0} l\gamma}{\#\{p_j \leq \gamma;\ j=1,\ldots,l\}}, \tag{5}$$

where l is the entire universe of TTRs, γ is the *p*-value cut-off and $\widehat{\pi_0} = \frac{\#\{p_j > \lambda;\ j=1,\ldots,l\}}{l(1-\lambda)}$ is an estimator of the proportion of rules that show no abnormal (either positive or negative) performance in the entire universe and for a two-sided framework. The estimation of $\widehat{\pi_0}$ requires us to define the tuning parameter *λ* by visually examining the histogram of all *p*-values.[39] Thus, in our study, *λ* is chosen by employing the same method.

[38] The block length used is equal to *q* = 0.1, and the number of bootstrap realizations is set to *B* = 1000, following previous studies.

[39] Bajgrowicz and Scaillet (2012) set the value of *λ* just by looking for the level above which the histogram of *p*-values becomes fairly flat, representing the region of null *p*-values. There is also an automated version of this process described by Storey (2002).

Following Barras et al. (2010), and after estimating $\widehat{\pi_0}$, we then focus on the right tail of the test statistic distribution (i.e. $\varphi_j > 0$), where the outperforming TTRs lie. Thus, we can compute a separate estimator for $\widehat{FDR}^+(\gamma)$.[40] This holds under the assumption that the false discoveries spread evenly between TTRs with positive and negative performance and with equal tail significance $\gamma/2$. Thus, the estimator is

$$\widehat{FDR^+}(\gamma) = \hat{F}^+/\hat{R}^+ = \frac{1/2\widehat{\pi_0}l\gamma}{\#\{p_j \leq \gamma, \varphi_j > 0;\ j = 1, \dots, l\}}, \tag{6}$$

Furthermore, the number of TTRs showing abnormal performance can be extrapolated as $\pi_A = 1 - \pi_0$. Now, defining the positive, π_A^+, and negative, π_A^-, proportions of rules in the population, we acquire $\pi_A^+ = \frac{T(\gamma)^+ + A(\gamma)^+}{l}$ and $\pi_A^- = \frac{T(\gamma)^- + A(\gamma)^-}{l}$, where $T(\gamma)^+$ and $T(\gamma)^-$ symbolize the number of strategies with positive and negative returns, respectively, and *p*-values less than *γ*. On the other hand, $A(\gamma)^+$and $A(\gamma)^-$ indicate the size of alternative models showing positive and negative performance without rejecting the null hypothesis (*p*-value greater than *γ*), respectively.

To conclude, $\hat{T}(\gamma)^+$ (likewise $T(\gamma)^-$) is defined as the estimator of the significantly positive rules minus the estimator of false selections:

$$\hat{T}(\gamma)^+ = \widehat{R^+(\gamma)} - \widehat{F^+(\gamma)} = \#\{p_j \leq \gamma, \varphi_j > 0;\ j = 1, \dots, l\} - \tfrac{1}{2}\widehat{\pi_0}l\gamma. \tag{7}$$

However, the most crucial part of identifying the genuine TTRs is the method of controlling a predetermined level of *FDR*$^+$ (i.e. 10%) or, in other words, finding the right *p*-value cutoff *γ* above which lie the rules with no statistically significant performance. We achieve this by following Storey et al., (2004), while using point estimates of the FDR. In particular, the *p*-values of the TTRs with positive performance are placed in ascending order. Then, starting with the smallest one, while adding the next *p*-value corresponding to the second rule, the *FDR*$^+$ is recomputed. This procedure is repeated until the desired *FDR*$^+$ is attained.

4.4. The k-FWER method

The second method employed for data snooping bias in our study is the *k*-FWER approach developed by Romano and Wolf (2007) and Romano et al., (2008). The rationale for implementing it in our case is its flexibility in detecting a great number of genuine trading strategies once the strict FWER criterion is eased, making it more

[40] The *FDR*$^-$ part can be calculated in a similar way.

suitable for investors who want to identify as many outperforming strategies as possible. However, we also want to examine whether a generalized version of the conservative FWER measure, allowing some false rejections, would achieve the same results as the powerful FDR^{+}. Contrary to its predecessor, the k-FWER criterion is defined as

$$k - FWER_p = P\{Reject\ at\ least\ k\ of\ the\ H_{0j}\}, \tag{8}$$

which is the probability of rejecting at least k true null hypotheses. In the multiple hypothesis testing setup, under a statistical significance level of a, the k-FWER is controlled if

$$k - FWER_p \leq a. \tag{9}$$

The *k-FWER* framework has an analogous structure to that of the *StepM*-BRC technique of Romano and Wolf (2005). However, it allows for at least a small number of false selections to be retained. Moreover, a resampling mechanism also needs to be used. Thus, for comparison purposes, we also employ the stationary bootstrap of Politis and Romano (1994), while using the same procedure to calculate the bootstrapped test statistics and thus the critical values for the BRC, *StepM*-BRC, and FDR tests. Each bootstrap test statistic vector also needs to be centered on its original value.

After the computation of the empirical bootstrapped distribution and the critical values, since the setup of the *k-FWER* approach is similar to that of the *StepM*-BRC test, only a few other steps need to be modified. The more general k-FWER approach needs to satisfy the criterion that at least k hypotheses will be rejected, instead of just one. Specifically, the TTR test statistics, φ_j, of the strategies showing positive performance are relabelled in descending order, with the first referring to the largest. During the first stage, individual decisions are executed for each rule, the null hypothesis H_{0j} being rejected if the test statistic, φ_j, is greater than the critical value

$$\widehat{c_1} = c_{\{1,\dots,l\}}(1 - a, k, \widehat{P_T}) \text{ for } 1 \leq j \leq l, \tag{10}$$ [41]

where $\widehat{c_1}$ is the estimated smallest $(1 - a)$ quantile of the re-centered sampling distribution of the k^{th} largest rule under the bootstrapped probability measure $\widehat{P_T}$. Then, denote by R_1 the number of statistically significant rules (hypotheses rejected) during

[41] The critical value $\widehat{c_1}$ asymptotically controls the k-FWER criterion. According to the theory $c_1 = c_K(1 - a, k, P)$. However, the set K and the probability mechanism P are unknown. Therefore, K is replaced by the set of all rules $\{1, \dots, l\}$ and the probability measure $\widehat{P_T}$ of the bootstrapped distribution is used instead of P.

the first stage. If $R_1 < k$ the procedure is terminated. This happens because it is feasible that all significant rules are true rejections. On the other hand, if $R_1 > k$ there is a strong possibility that some false rejections will have been included in the total number of rejections. Therefore, we need to move on to the second stage, excluding the test statistics of the rejected strategies. The remaining ones are tested in a new hypothesis testing setup. This time, each of the test statistics, φ_j, is compared with the critical value

$$\widehat{c_2} = \max\{c_K(1 - a, k, \widehat{P_T})\} \text{ for } R_1 + 1 \leq j \leq l \quad (11)$$

while individual decisions are also made, $\widehat{c_2}$ depicts the maximum quantile of the set of quantiles, including the rejected $k - 1$ hypotheses from the first step, as well as all the hypotheses that have not been rejected yet. The intuition is that we are not certain which of the rejected hypotheses might be true so both rejected and non-rejected hypotheses, together with the largest quantile, must be considered. Finally, if no further hypotheses are rejected in the second step, the procedure terminates. Otherwise, the stepwise setup is maintained, and new decisions are carried out involving new $\widehat{c_m}$ maximum critical values, until no other rejections occur.

The steps described above reflect the one-sided framework, which is meaningful when searching for genuine TTRs among the entire set of rules, which display positive performance and fall within the right tail of the distribution.

4.5. Portfolio construction

We construct the portfolios of rules by selecting them in accordance with the *FDR*$^+$ and *k*-FWER. In particular, we set the $\widehat{FDR^+}$ and *k*-FWER equal to 10%, as a good trade-off between truly outperforming TTRs and wrongly chosen ones (Bajgrowicz and Scaillet 2012). Despite the fact that *k* is an integer in the case of *k*-FWER, we adjust it to a number that is equal to 10% of the rules showing positive performance for each interval examined. Thus, we acquire 10%-FDR$^+$ and 10%-FWER portfolios, which means that which means that 90% of the total number of the portfolio's rules, significantly outperform the benchmark. The signals of the chosen rules are pooled with equal weight, similarly to a forecast averaging technique. We do not attribute more weight to more effective rules since this would result in reducing the FDR$^+$ and *k*-FWER portfolios below the desired level. We finally treat the neutral signals as totally liquidating our positions and do not invest a proportion of wealth, corresponding to them, at the “risk-free” rate. This assumption helps us to measure the true performance of the FDR portfolios.

5. In-sample performance

5.1. In-sample performance with no transaction costs

Table 3 reports the number of TTRs displaying a positive performance[42] under the *mean return*, *Sharpe* ratio, and *Calmar* ratio criteria, for crude oil futures and USO, Subperiods 1-4.

[Table 3]

Concerning Subperiods 1 (18 April 2007 – 29 May 2009) and 4 (1 August 2013 – 1 January 2016), that contain strong trends or an unstable environment, it seems that a significant proportion of the TTRs considered are able to achieve a positive performance for both crude oil futures and USO. This outcome could almost have been anticipated since the majority of the strategies are momentum or trend-following rules capturing extreme movements. On the other hand, the number of outperforming rules is reduced for the mean return and Sharpe ratio for Subperiods 2 (1 June 2009 – 31 May 2011) and 3 (1 April 2011 – 31 July 2013), which show a more balanced evolution of prices. Regarding the Calmar ratio, the number of outperforming rules is even less (at most 25%), as a TTR needs to achieve a Calmar ratio above 1 to be considered a good strategy.[43] It is also worth mentioning that, during Subperiod 3, the profitable TTRs on crude oil futures are almost half of those on the USO for the mean return, the Sharpe and the Calmar ratio. Moreover, the number of outperforming rules identified, based on the Sharpe ratio, is consistently less than or (almost) equal to those identified by the mean return criterion, for crude oil futures and USO. This feature stems from the subtraction of the risk-free rate, which results in the elimination of returns of a very small magnitude.

For the same sample periods, Table 4 shows the in-sample performance of the best rule under the mean return, Sharpe ratio and Calmar ratio criteria respectively, free of

[42] Positive performance means a mean return or Sharpe ratio above zero, or a Calmar ratio above one.

[43] A Calmar ratio value of 1-2 is assumed a good strategy, a value between 2-5 very good, and a value greater than 5 recognized as excellent (Young 1991).

transaction costs, for crude oil futures (Panel A) and USO (Panel B). The corresponding *p*-value of the BRC test for the best rule and for each performance criterion is also displayed in parenthesis. The buy-and-hold strategy for the crude oil futures and the USO is also displayed in the columns on the right-hand side of the table.

[Table 4]

For both Panels A and B, the best rule results seem very encouraging compared with the buy-and-hold strategies, for crude oil futures and USO. The best rule's performance indicates that profitable trading strategies exist in all subperiods, according to all criteria. Specifically, the Sharpe and Calmar ratios are high enough that the best rules can be characterized as very good trading opportunities in the majority of the years covered. However, we should mention that the evidence provided above has no economic value, since transaction costs are not considered, and it is just a trivial experiment concerning predictability. Moreover, the corresponding BRC *p*-values are quite high in half the cases, indicating that the performance of some of the best rules is not significant. Also, the information reported in Table 4 relies only on findings that are discovered ex post, and there is no guarantee that a trader will have selected the potentially best rule in advance only by looking at its long-term historical behavior. In practice, investors rebalance their positions more frequently to capture any changes in the economic and financial milieu. Despite the above, Tables 3 and 4 still reveal the existence of technical indicators that are able to capture patterns in the daily prices for both crude oil futures and the USO.

5.2. In-sample performance including transaction costs

Since predictive power is not always synonymous with profitability, an investor should always check carefully whether the returns gained from trading strategies are sufficient to cancel out the transaction costs. Indeed, trading rules pooled before transaction costs are more likely to generate frequent signals, thus increasing the probability of their performance benefits being eliminated once the transaction costs are included.

The majority of the previous studies examine the performance of TTRs through a breakeven analysis, wherein the effect of transaction costs is computed ex post, once outperforming rules have been identified. However, this undoubtedly makes it more complicated for a trader to foresee profitable rules that will offset transaction costs a priori. Contrary to that, we again follow Bajgrowicz and Scaillet (2012), handling transaction costs "*endogenously*" and not "*exogenously*" to the selection process. In particular, we subtract the transaction costs every time a buy or sell signal is triggered. Following the study of Locke and Venkatesh (1997), who estimate that futures markets' one-way transaction costs range from 0.04 to 3.3 basis points, we consider the second, larger amount for the crude oil futures. Furthermore, we assume that an investor funds their position with 100% equity rather than using a margin, since we measure daily returns as the log of the difference in price relatives (Bessembinder, 1992; Miffre and Rallis, 2007; Marshall et al., 2008). For the case of the USO, we incorporate one-way transaction costs of 5 basis points on each trade. This level of transaction costs is justified based on the literature, as well as information from floor traders (Hsu et al. 2010), for the trading of ETFs.

Tables 5 and 6 display the number of outperforming rules as well as the in-sample performance for the crude oil futures and the USO when one-way transaction costs of 3.3 and 5 basis points are considered, respectively, ex ante. Comparing the results with Table 3, in Table 5 we can see that the number of corresponding outperforming rules has decreased considerably for all evaluation methods, especially when the Calmar ratio is employed.

[Table 5]

[Table 6]

In Table 6, the in-sample performance with transaction costs provides a similar picture. Thus, as expected, the values for all evaluation criteria are reduced. However, in Table 6, we also observe that the best trading rules are still able to achieve better performance than the buy-and-hold strategy, for all criteria and across all subperiods. Interestingly, compared with the values in Table 4, the Sharpe ratios for the best rules are lower. However, the trading strategies remain very promising, as the Sharpe ratios are still above 1.5. Similarly, in Table 6, the Calmar ratios are high enough that we can conclude that the generated returns are sufficient to outweigh the transaction costs. When it comes to the statistical significance of the best rules' performance, Table 6 demonstrates that none of the *p*-values from the BRC test are significant. However, this does not mean that TTRs with genuinely good performance do not exist among the most profitable rules.

Tables 7 and 8 demonstrate the impact of transaction costs on the historically best TTRs, selected with respect to the mean return, Sharpe ratio and Calmar ratio criteria, while accounting for zero and non-zero one-way transaction costs, for the crude oil futures and the USO, respectively, in each subperiod.

[Table 7]

[Table 8]

Generally speaking, the best trading strategy nominated remains within the same family of rules, before and after one-way transaction costs are considered, in most cases, and for both the crude oil futures and the USO. In addition to this, Table 7 demonstrates that, in the case of crude oil futures, the small magnitude of transaction costs does not have a strong impact on the chosen best rule, and this applies to all criteria and across all subperiods. However, Table 8 portrays a contradictory picture, especially under the mean return and Sharpe ratio criteria, for the case of the USO. TTRs selected without consideration of the transaction costs produce more frequent trading signals than those for which the transaction costs have been taken into account endogenously. For instance, under the Sharpe ratio measure and Subperiod 1, a 25-50 day on-balance volume rule is more likely to trigger frequent signals than a channel breakout rule with a 20-100 window of days, 0.075 channel width and a five-day holding period, which suffers more constraints. The best rules after the inclusion of transaction costs are *not* usually among the best ones before their inclusion. One explanation might be that trading the USO entails larger transaction costs than trading crude oil futures.

Moreover, the successful rules do not trade on longer-term price movements once transaction costs are incorporated, in either case. While there are cases (under the Sharpe ratio and in Subperiod 2 for the USO) where the best rule, in-sample, uses a larger window of 200 days of data when transaction costs are included, compared to a window of 10 days used by the best rule under zero transaction costs, this is not true in most of the cases.

Another interesting finding emerges when employing the Calmar ratio criterion. The best rules derived before and after the inclusion of transaction costs are closely related, perhaps due to the maximum drawdown factor employed. Searching for the best strategy, while minimizing the maximum drawdown of its returns, increases an investor’s probability of ending up with a rule that generates less frequent signals, even before the inclusion of transaction costs to avoid larger drawdowns. This might be the reason for TTRs selected under the Calmar ratio approach being almost the same before and after consideration of transaction costs in each sample period.

Finally, the most important evidence gleaned from observing both tables together, is that the rules selected as the best ones based on technical analysis for the crude oil futures belong to a different family from those selected when trading the USO, under all criteria and across all subperiods considered. This may be a potential justification for the different dynamics that characterize the crude oil futures compared to the USO, with contango or backwardation outcomes having a significant effect on the calculation and redemption procedures for ETFs. On the other hand, the above findings may just reveal the considerable effects of data snooping bias, in that the best rule's performance may be achieved merely through luck in most cases, leading to different rules being identified as the best for the two assets and for the different subperiods.

6. Persistence analysis

The economic evaluation of TTRs' performance in the crude oil market is covered in this section. One of the fundamental questions that technical traders must answer when evaluating TTRs' predictive power is whether the rules selected as superior *ex ante* during backtesting are also able to generate abnormal returns once the transaction costs are considered, for an out-of-sample period. We shed light on whether some of the outperforming rules would have been able to produce profits in practice due to the high volatility in the crude oil market, using only past price data.

A persistence (out-of-sample) analysis of TTRs' performance in the crude oil market is applied here for the very first time. With this aim, we build portfolios of outperforming rules, and re-evaluate the portfolios' performance on a semi-annual basis. In the first six months, when the total universe of rules' performance is tested, we construct equally weighted portfolios while accounting for data snooping bias using the FDR^{+} and *k*-FWER methods as described in Sections 4.2 - 4.5. Specifically, every six months, two portfolios are constructed employing price data from the previous six months. Then, the out-of-sample performance of the chosen rules is evaluated over the following half of the year. As mentioned earlier, the "risk-free" rate is considered as the benchmark. In particular, we assess the performance of rules in-sample (IS), before constructing portfolios of the “genuine” (statistically significant) in-sample rules and measuring their performance out-of-sample (OOS). This structure matches how investors in practice set up their own strategies based only on a priori information.

Table 9 displays the results of the persistence analysis under the annualized Sharpe ratio and 3.3 and 5 basis points of transaction costs for the crude oil futures and the USO respectively, in accordance with the 10%-FDR$^+$ and 10%-FWER rules selection criteria, as well as the best rule’s performance, observed across the different in-sample and out-of-sample periods. The table also includes each portfolio’s median size as well as its percentage amount over the total universe of TTRs in brackets.

[Table 9]

The results clearly indicate that there is *no* persistence in the trading rules’ performance, as both selection criteria verify. However, in periods when both portfolios are able to generate positive performance out-of-sample, the Sharpe ratio levels are considerably smaller than those of the in-sample performance. For instance, no investor will choose a portfolio whose Sharpe ratio level is below one. However, the only case of a Sharpe ratio exceeding one is in Subperiod 2 for the trading of crude oil futures contracts, for both portfolios. The ADH, which describes how profit opportunities might be available during some periods, but then disappear in later ones, might shed some light on the above results. Overall, the picture is opposed to the evidence found regarding the in-sample performance in Section 5, which implies that the best-performing rules are accessible *only* to investors observing the returns ex post. Additionally, comparing the 10%-FWER and 10%-FDR$^+$ portfolios with the best rule’s performance, we notice that, in most cases, both portfolios achieve better performance out-of-sample than just employing the best rule, verifying the benefits of employing the proposed data snooping methods as portfolio construction techniques. To summarize, interestingly, the *no hot hands* phenomenon that is confirmed in Bajgrowicz and Scaillet (2012) for the DJIA also appears in the crude oil market.

As presented in Table 9, one of the most important findings is the overall performance of the 10%-FWER and 10%-FDR$^+$ specifications for data snooping as portfolio compilers. We could say that the two approaches seem able to achieve almost equal performance for the trading of crude oil futures, except in Subperiod 1 where the 10%-FWER portfolio outperforms its counterpart. On the other hand, the FDR$^+$ portfolio seems to generate slightly better performance in the case of the USO. The generalization of the FWER measure (to allow for some false selections) improves its performance, which is demonstrated by the achievement of more or less similar Sharpe ratios to the FDR$^+$ portfolio. The *k*-FWER circumvents the lucky rules that do not produce significantly good performance, while keeping only the genuine ones. Furthermore, it does not suffer from preventing the selection of further rules once it has identified a lucky one, as its predecessor did. When it comes to median portfolio size, the findings indicate that the median size of the 10%-FWER portfolios for both the crude oil futures and the USO, is less than that of the FDR$^+$ portfolios at all times, with this difference considerable during specific subperiods.

Finally, Table 10 presents both the 10%-FWER and 10%-FDR$^+$ portfolios' average decomposition according to each of the five families of TTRs for the crude oil futures and the USO respectively. In particular, we report the number of rules selected from each family divided by the total number of rules included in each portfolio, as a percentage[44].

[Table 10]

We observe that the 10%-FWER and 10%-FDR$^+$ portfolios show different selection preferences among the five families of rules. Furthermore, the preferences seem to differ depending on whether the crude oil futures or the USO are being traded. For instance, the 10%-FWER portfolio seems to mostly choose TTRs from the on-balance volume and channel breakout families, followed by the support and resistance rules, while the 10%-FDR$^+$ portfolio selects TTRs mostly from the support and resistance and channel breakout families, with the on-balance volume rules coming next in order of preference. The filter rules family displays the smallest percentages for both portfolios, crude oil futures and USO, as well as across all subperiods. In general, the support and resistance, channel breakouts and on-balance volume rule families are the most significant in capturing the patterns of the crude oil market.

[44] We should mention that the number of rules chosen varies substantially from one six-month period to the next. Sometimes, the portfolio consists of almost exclusively new rules, even after the first rebalancing.

7. Conclusion

Evidence of the historical success of technical trading rules is revisited, this time in the trending crude oil market. Although, numerous efforts have been made in the field of evaluating the performance of technical trading rules (forex, stock, large/small cap markets, etc.), this is the first time it has been done for crude oil. Findings from previous studies are divided on whether technical analysis can achieve genuine abnormal performance. The motivation of this study was to examine whether technical trading indicators and oscillators could benefit from the severe fluctuations characterizing the crude oil market lately. The majority of these rules are designed to capture such patterns, being momentum and trend following strategies. We focus on the crude oil futures and the United States Oil fund (USO) as the largest and most liquid crude oil exchange traded fund (ETFs), developed to track the daily price movements of West Texas Intermediate ("WTI") light, sweet crude oil.

First, we reassess the predictive power of Sullivan et al.'s (1999) universe of trading rules in the overall crude oil market, to verify that patterns exist. Evidence of the rules' performance on crude oil futures as well as the USO in an in-sample simulation demonstrates that, during periods of dramatic crude oil price movements, more than half of the rules show great predictive power. However, the corresponding *p*-values of the best rules of the *bootstrap reality check* (BRC) test of White (2000) are not statistically significant most of the time. Popular performance measures used by fund managers and traders, such as the Sharpe and Calmar ratios, are employed to measure the profitability of rules. All of them show the best rule for each period to be a very good trading opportunity.

Second, we endogenously incorporate transaction costs when evaluating trading rules' performance in the case of crude oil futures contracts and the USO. Strategies employed by Brock et al., (1992) and Sullivan et al., (1999) can trigger very frequent signals, which might lead to the elimination of superior returns when transaction costs are considered. However, technical trading rules are still able to achieve profits, although their performance is decreased, because of the relatively small transaction costs applied when trading commodities futures and ETFs.

Third, we employ two of the most powerful techniques for accounting for data snooping, in order to identify significantly profitable trading strategies. The *false discovery rate* (FDR) approach, as described by Bajgrowicz and Scaillet (2012) when evaluating technical trading rules on the DJIA index, is used to control false discoveries. The *k-familywise error rate* (*k*-FWER) methodology developed by Romano and Wolf (2007) is also applied to check whether a generalized version of the conservative FWER criterion allowing for some false rejections performs equally well to the powerful FDR method. Both specifications are able to select more rules and better diversify against model uncertainty than the previous BRC and Romano and Wolf approaches that are prevented from searching for more rules once a “lucky” one has been detected.

Finally, a persistence analysis is carried out for the purpose of economically evaluating the rules’ performance. The question that needs to be answered here is whether investors can foresee which rules will generate future returns – that will outweigh transaction costs – without prior knowledge. We respond to this argument by creating portfolios with the FDR and *k*-FWER approaches, using only past data in an in-sample period, and evaluating their performance out-of-sample. The findings show that there is no persistent nature to the rules’ performance, contrary to the outstanding in-sample results, although tiny profits can be achieved in some periods. The results seem to be in favor of the Adaptive Market Hypothesis (ADH) of Lo (2004). Moreover, the FDR$^+$ and *k*-FWER approaches show almost equal performance.

In terms of further research, some manipulation-proof versions of performance metrics can be used instead of regular ones. So far, we focus our experiments on the performance of multiple hypothesis testing frameworks on measures such as the mean return and Sharpe ratio. However, these measures are subject to manipulation and generate spectacular results to an uninformed investor employing dynamic strategies. For that reason, manipulation-proof measurement methods being consistent with standard financial market equilibrium conditions such as those of Ingersoll, Spiegel, Goetzmann and Welch (2008) and the Morningstar Risk-Adjusted Rating introduced in July 2002, should be further investigated in our applications.

List of tables

Table 1. Sample periods for crude oil futures and USO

Sample period	Dates	Trading days	Market trend
Subperiod 1:	18 April 2007- 29 May 2009	534	Mixed
Subperiod 2:	1 June 2009- 31 March 2011	464	Bullish
Subperiod 3:	1 April 2011- 31 July 2013	586	Mixed
Subperiod 4:	1 August 2013- 1 January 2016	618	Bearish

Table 2. Descriptive statistics of daily returns on crude oil futures and USO

Instrument	Subperiod	Mean (%)	St.dev. (%)	Skewness	Kurtosis	First AC
Crude oil futures	Subperiod 1	0.01	3.03	0.44	11.4	0.01
	Subperiod 2	0.10	1.59	-0.01	4.55	0.18**
	Subperiod 3	0.01	1.42	-0.35	4.88	0.08
	Subperiod 4	-0.19	1.77	0.66	5.97	0.12**
USO	Subperiod 1	-0.06	2.91	-0.19	4.28	-0.11**
	Subperiod 2	0.04	1.92	-0.07	3.23	0.00
	Subperiod 3	-0.02	1.75	-0.57	6.24	-0.06
	Subperiod 4	-0.22	1.99	0.00	5.36	-0.10**

This table reports the descriptive statistics of daily returns on the crude oil futures and the USO. Means and standard deviations are reported in percentage points (%). "First AC" stands for first-order autocorrelation. Asterisks (**) denote significant first-order autocorrelation for returns at the 5% level according to the Ljung-Box (1978) Q statistics.

Table 3. Numbers of outperforming rules for in-sample performance with no transaction costs

Instrument	Sample period	Outperforming rules		
		Mean return	Sharpe ratio	Calmar ratio
Crude oil futures	Subperiod 1	4049	3979	574
		[51%]	[50%]	[7%]
	Subperiod 2	2226	2222	406
		[28%]	[28%]	[5%]
	Subperiod 3	840	838	82
		[10%]	[10%]	[1%]
	Subperiod 4	5148	5147	485
		[65%]	[65%]	[6%]
USO	Subperiod 1	4114	4096	2025
		[52%]	[52%]	[25%]
	Subperiod 2	1774	1774	341
		[22%]	[22%]	[4%]
	Subperiod 3	1920	1919	166
		[24%]	[24%]	[2%]
	Subperiod 4	5000	5000	1445
		[63%]	[63%]	[18%]

This table presents the outperforming rules in levels, and as percentages over the total universe in brackets, identified according to the positive daily mean return, annualized Sharpe ratio, and Calmar ratio criteria respectively, for the crude oil futures and the USO, across the different subperiods.

Table 4. In-sample performance with no transcription costs

Sample period	Best rule			Buy-and-hold strategy		
Panel A *Crude oil futures*	Mean return (%)	Sharpe ratio	Calmar ratio	Mean return (%)	Sharpe ratio	Calmar ratio
Subperiod 1	0.34	1.82	8.21	0.01	0.30	0.25
	(0.09)*	(0.05)**	(0.03)**			
Subperiod 2	0.14	2.30	7.70	0.10	0.95	1.45
	(0.99)	(0.39)	(0.83)			
Subperiod 3	0.12	1.71	8.83	0.01	0.09	0.07
	(0.97)	(0.80)	(0.43)			
Subperiod 4	0.26	2.10	9.80	-0.19	-1.3	-0.62
	(0.08)*	(0.06)*	(0.10)*			
Panel B *USO*						
Subperiod 1	0.38	2.05	6.95	-0.06	-0.13	-0.04
	(0.09)*	(0.07)*	(0.11)			
Subperiod 2	0.17	2.07	8.92	0.04	0.37	0.54
	(0.96)	(0.71)	(0.69)			
Subperiod 3	0.13	1.78	7.31	-0.02	-0.09	-0.05
	(0.98)	(0.74)	(0.23)			
Subperiod 4	0.23	2.19	9.35	-0.22	-1.69	-0.69
	(0.08)*	(0.07)*	(0.06)*			

This table reports the performance results of the best rule, and its corresponding BRC *p*-value in-sample in parenthesis, as well as the buy-and-hold strategy for the crude oil futures and the USO respectively, under the daily mean return, annualized Sharpe ratio, and Calmar ratio criteria, and across the different subperiods. * denotes a rejection of the null hypothesis at the 10% level of significance, ** denotes rejection at the 5% level.

Table 5. Numbers of outperforming rules for in-sample performance with transaction costs

Instrument/ Transaction costs	Sample period	Outperforming rules		
Crude oil futures *(3.3 bps)*		Mean return	Sharpe ratio	Calmar ratio
	Subperiod 1	3623 [46%]	3558 [45%]	320 [4%]
	Subperiod 2	1555 [19%]	1559 [19%]	274 [3%]
	Subperiod 3	525 [6%]	525 [6%]	69 [0.8%]
	Subperiod 4	4569 [58%]	4565 [58%]	250 [3%]
USO *(5 bps)*				
	Subperiod 1	3777 [48%]	3739 [47%]	1341 [17%]
	Subperiod 2	1066 [13%]	1065 [13%]	168 [2%]
	Subperiod 3	931 [11%]	929 [11%]	79 [1%]
	Subperiod 4	4346 [55%]	4346 [55%]	363 [4%]

This table presents the outperforming rules in levels, and in percentages over the total universe in brackets, identified according to the positive daily mean return, annualized Sharpe ratio, and Calmar ratio criteria, for the crude oil futures and the USO, assuming 3.3 and 5 basis points (bps) one-way transaction costs respectively, across the different subperiods.

Table 6. In-sample performance with transaction costs

Sample period	Best rule			Buy-and-hold strategy		
Panel A *Crude oil futures*	Mean return (%)	Sharpe ratio	Calmar ratio	Mean return (%)	Sharpe ratio	Calmar ratio
Subperiod 1	0.30	1.73	7.76	0.01	0.30	0.25
	(0.25)	(0.15)	(0.29)			
Subperiod 2	0.11	2.21	7.13	0.10	0.95	1.45
	(1.00)	(0.49)	(0.82)			
Subperiod 3	0.08	1.69	7.80	0.01	0.09	0.07
	(0.99)	(0.81)	(0.42)			
Subperiod 4	0.23	1.89	8.53	-0.19	-1.3	-0.62
	(0.31)	(0.27)	(0.35)			
Panel B *USO*						
Subperiod 1	0.33	1.87	6.00	-0.06	-0.13	-0.04
	(0.44)	(0.68)	(0.39)			
Subperiod 2	0.13	1.89	6.05	0.04	0.37	0.54
	(1.00)	(0.88)	(0.74)			
Subperiod 3	0.08	1.60	5.11	-0.02	-0.09	-0.05
	(1.00)	(0.91)	(0.53)			
Subperiod 4	0.18	2.09	8.67	-0.22	-1.69	-0.69
	(0.33)	(0.22)	(0.28)			

This table reports the performance results of the best rule, and its corresponding BRC p-value in-sample in parenthesis, including one-way transaction costs for the crude oil futures and the USO, as well as the buy-and-hold strategies for the crude oil futures and the USO respectively, under the daily mean return, annualized Sharpe ratio, and Calmar ratio criteria, across the different subperiods.

Table 7. Best in-sample technical trading rules for crude oil futures

Sample period	Costs	Best rule
Mean return		
Subperiod 1	Zero	Sup.&Res. (10-day alt. extrema, 0.05 band)
	3 bps	Sup.&Res. (20-day alt. extrema, 0.05 band)
Subperiod 2	Zero	Filter (0.01 pos. initiation, 5-day hold. per.)
	3bps	Filter (0.01 pos. initiation, 5-day hold. per.)
Subperiod 3	Zero	Moving Avg. (25-30 day, 25-day hold. per.)
	3 bps	Moving Avg. (25-30 day, 25-day hold. per.)
Subperiod 4	Zero	On-Bal.-Vol. (40-50 day, 10-day hold. per.)
	3 bps	On-Bal.-Vol. (40-50 day, 10-day hold. per.)
Sharpe ratio		
Subperiod 1	Zero	Chan.Br. (20-day, 0.075 width, 10-day hold. per.)
	3 bps	Chan.Br. (20-day, 0.075 width, 0.01 band, 10-day hold. per.)
Subperiod 2	Zero	Sup.&Res. (10-day, 0.1 band)
	3 bps	Sup.&Res. (20-day, 0.03 band)
Subperiod 3	Zero	Chan.Br. (15-day, 0.05 width, 0.03 band 10-day hold. per.)
	3 bps	Chan.Br. (15-day, 0.05 width, 0.03 band 10-day hold. per.)
Subperiod 4	Zero	Chan. Br. (10-day, 0.05 width, 0.01 band 25-day hold. per.)
	3 bps	Chan. Br. (10-day, 0.03 width, 0.001 band 25-day hold. per.)
Calmar ratio		
Subperiod 1	Zero	Sup.&Res. (4-day al. extrema, 4-day delay, 5-day hold. per.)
	3 bps	Sup.&Res. (4-day alt. extrema 4-day delay, 25-day hold. per.)
Subperiod 2	Zero	Sup.&Res. (200-day, 0.03 band, 5-day hold. per.)
	3 bps	Sup.&Res. (200-day, 0.03 band, 5-day hold. per.)
Subperiod 3	Zero	Chan.Br. (100-day, 0.1 width, 0.001 band, 50-day hold. per.)
	3 bps	Chan.Br. (100-day, 0.15 width, 0.01 band, 5-day hold. per.)
Subperiod 4	Zero	Sup.&Res. (150-day, 2-day delay, 25-day hold. per.)
	3 bps	Sup.&Res. (200-day, 2-day delay, 50-day hold. per.)

This table presents the historically best-performing trading rule, chosen under the daily mean, annualized Sharpe ratio, and Calmar ratio criteria, for the crude oil futures, in each sample period and for either zero or 3.3 basis points (bps) one-way transaction costs.

Table 8. Best in-sample technical trading rules for USO

Sample period	Costs	Best rule
Mean return		
Subperiod 1	Zero	Moving Avg. (5-25 day, 50-day hold. per.)
	5 bps	Moving Avg. (1-250 day, 50-day hold. per.)
Subperiod 2	Zero	On-Bal.-Vol. (10-50 day, 0.001 band)
	5 bps	On-Bal.-Vol. (5-50 day, 0.003 band)
Subperiod 3	Zero	On-Bal.-Vol. (15-200 day, 0.005 band)
	5 bps	On-Bal.-Vol. (50-125 day, 0.005 band)
Subperiod 4	Zero	Moving Avg. (1-25 day, 10-day hold. per.)
	5 bps	Moving Avg. (1-25 day, 10-day hold. per.)
Sharpe ratio		
Subperiod 1	Zero	On-Bal.-Vol. (25-50 day)
	5 bps	Chan.Br. (20-100 day, 0.075 width, 5-day hold. per.)
Subperiod 2	Zero	Sup.&Res. (10-day, 0.01 band)
	5 bps	Sup.&Res. (200-day, 0.04 band, 10-day hold. per.)
Subperiod 3	Zero	On-Bal.-Vol. (75-100 day, 0.04 band)
	5 bps	Chan.Br. (25-day, 0.03 width, 0.05 band 10-day hold. per.)
Subperiod 4	Zero	Chan.Br. (10-day, 0.05 width, 0.1 band 25-day hold. per.)
	5 bps	Chan.Br. (10-day, 0.05 width, 0.1 band 25-day hold. per.)
Calmar ratio		
Subperiod 1	Zero	Chan.Br. (15-day, 0.075 width, 5-day hold. per.)
	5 bps	Chan.Br. (50-day, 0.1 width, 0.005 band 5-day hold. per.)
Subperiod 2	Zero	Sup.&Res. (5-day, 3-day delay, 50-day hold. per.)
	5 bps	Sup.&Res. (5-day, 3-day delay, 50-day hold. per.)
Subperiod 3	Zero	Chan.Br. (25-day, 0.03 width, 0.05 band 10-day hold. per.)
	5 bps	Sup.&Res. (20-day, 4-day delay, 5-day hold. per.)
Subperiod 4	Zero	Chan.Br. (10-day, 0.01 width, 5-day hold. per.)
	5 bps	Chan.Br. (10-day, 0.01 width, 5-day hold. per.)

This table presents the historically best-performing trading rule, chosen under the daily mean, annualized Sharpe ratio, and Calmar ratio criteria, for the USO, in each sample period and for either zero or 5 basis points (bps) one-way transaction costs.

Table 9. Performance persistence analysis under the Sharpe ratio criterion with transaction costs and risk-free rate as benchmark.

Sample period	10%-FWER portfolio			10%-FDR$^+$ portfolio			Best rule	
Crude oil futures	*Median size*	IS	OOS	*Median size*	IS	OOS	IS	OOS
Subperiod 1	47 [0.6%]	3.37	0.19	50 [0.6%]	2.71	-0.32	3.64	-0.64
Subperiod 2	26 [0.3%]	4.32	1.06	37 [0.4%]	4.44	1.01	3.46	0.68
Subperiod 3	24 [0.3%]	3.27	-1.13	40 [0.4%]	3.44	-0.62	3.05	-0.72
Subperiod 4	25 [0.3%]	3.20	-0.24	112 [1.4%]	3.98	-0.48	3.72	-1.20
USO	*Median size*	IS	OOS	*Median size*	IS	OOS	IS	OOS
Subperiod 1	123 [1.5%]	2.82	0.28	170 [2.1%]	2.71	0.37	3.64	-0.64
Subperiod 2	28 [0.3%]	4.19	-0.15	73 [0.9%]	3.98	-0.29	3.46	0.68
Subperiod 3	42 [0.5%]	3.30	-1.27	70 [0.9%]	3.00	-0.42	3.05	-0.72
Subperiod 4	167 [2.1%]	3.37	-0.26	300 [3.8%]	3.66	0.23	3.72	-1.20

This table indicates the in-sample (IS) and out-of-sample (OOS) annualized Sharpe ratio of trading rules chosen according to the 10%-FWER portfolio of Romano and Wolf (2007), and the 10%-FDR$^+$ portfolio of Bajrowicz and Scaillet (2012), with a semi-annual rebalancing and a risk-free rate as benchmark, as well as the best rule in-sample. The table also displays the portfolios' median sizes in levels, and in percentages of the total rules universe in brackets, across different subperiods.

Table 10. Portfolio decomposition into families of rules

Sample period	F	MA	SR	CB	OBV
Crude oil futures					
Subperiod 1	2% (1%)	18% (28%)	5% (22%)	30% (8%)	45% (39%)
Subperiod 2	1% (2%)	0% (3%)	56% (64%)	40% (31%)	3% (0%)
Subperiod 3	6% (0%)	8% (3%)	2% (66%)	26% (25%)	58% (6%)
Subperiod 4	4% (1%)	16% (16%)	15% (49%)	43% (16%)	21% (18%)
USO					
Subperiod 1	13% (1%)	34% (43%)	10% (24%)	21% (14%)	22% (17%)
Subperiod 2	2% (2%)	1% (3%)	64% (81%)	9% (11%)	22% (2%)
Subperiod 3	11% (0%)	17% (2%)	5% (95%)	25% (0%)	42% (2%)
Subperiod 4	1% (1%)	15% (16%)	11% (26%)	29% (31%)	42% (25%)

This table reports the average percentage of rules belonging to each family, in each portfolio constructed using the 10%-FWER (10%-FDR) methods, for the crude oil futures and the USO respectively, across each subperiod. F: filter rules, MA: moving averages, SR: support and resistance rules, CB: channel breakouts, and OBV: on-balance volume averages.

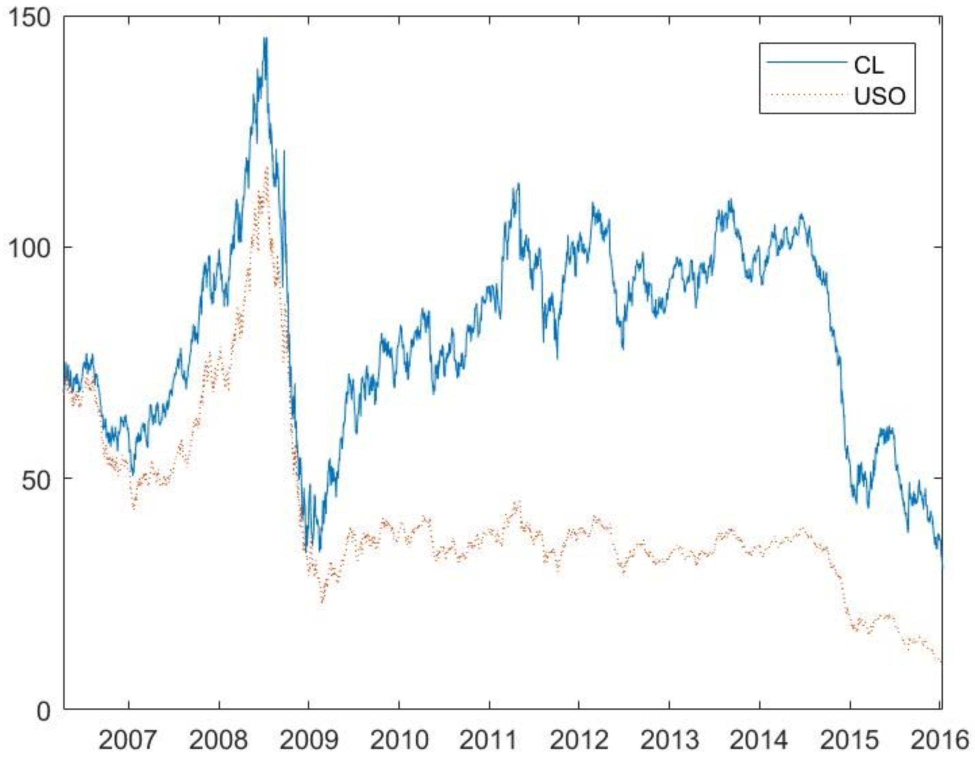

Fig.1. The time series dynamics of crude oil futures (CL) and United States oil fund (USO)

CHAPTER 4
PAIRS TRADING, TECHNICAL ANALYSIS & DATA SNOOPING: MEAN REVERSION VS MOMENTUM

1. Introduction

This study revisits the excess profitability of 'pairs trading', as one of the most popular short-term speculation strategies, through technical analysis by analysing time series history and employing self-financing trading rules, while controlling for data snooping effects. Being a representative of the general class of proprietary 'statistical arbitrage' techniques, severely exploited by hedge funds and investment banks, 'pairs trading' has a long history, back to mid-1980s, when the famous quant Nunzio Tartaglia tried to discover arbitrage opportunities in the equities markets by using sophisticated statistical methods through automated trading systems. His high-tech systems characterized by the successful recognition of pairs of securities whose prices tended to move together historically. The concept of 'pairs-trading' is quite simple then, go long the first component while going short the second when their spread widens and if the history repeats itself the prices will converge yielding profits. Thus, cointegration between the long and short components' prices, contrarian principles and past prices dynamics are major mechanisms for a statistical arbitrage strategy expected to work. Those mechanisms and their potential profitability on pairs based on trading rules have been addressed by several studies in the past. Hogan et al., (2004) find that value and momentum trading strategies can constitute arbitrage opportunities in the equities market, while Gatev et al., (2006) verify the profitability of simple divergence-convergence trading rules in terms of standard deviation due to temporary mispricing of securities.

We investigate the evolution of excess profitability of pairs trading on daily data including spreads of commodities, equity indices and foreign exchange rates over the period January 1990 through December 2016, while employing a large universe of disciplined, consistent technical trading rules. We find an annualized mean excess return and Sharpe ratio of 17.6% and 1.20 respectively, for the top pair portfolio generated by the best technical trading rule over that period.

Technical analysis (sometimes referred to as chartism) is believed to be one of the longest-established forms of investment analysis, being a set of graphical or mathematical techniques exploring future trading opportunities for financial assets just by analysing the asset prices' time-series history. Although there is an ongoing debate on the effectiveness of this form of analysis rooted in the lack of economic theory, numerous studies have revisited technical analysis' predictability across several markets (e.g., Brock et al., 1992; Neely et al., 1997; Allen and Karjalainen 1999; Sullivan et al., 1999; Lo et al., 2000; Kavajecz and Odders-White 2004; Marshall et al., 2008; Neely and Weller 2011; Bajgrowicz and Scaillet 2012; Hsu et al., 2016). More than half of them plead for technical trading significant profitability, while the rest are placed against.

Nevertheless, we introduce for the first time a comprehensive and up-to-date analysis of technical analysis' performance on pairs trading as being developed in practice by statistical arbitrageurs. In particular, we examine the predictability and excess profitability of over 18,000 momentum and mean-reverting technical trading rules on 'famous' pairs being frequently advertised by trading websites or launched by financial market companies. Consistent with the findings of Gatev et al., (2006) we show that simple mean reversion is not the only factor responsible for generating significant profits but also momentum, while we examine a group of other 'stylized facts' derived from the literature on technical analysis in all the three markets examined (see Menkhoff and Taylor, 2007; Marshall et al., 2008; Brajgowicz and Scaillet, 2012), such as that technical analysis' profitability has shrunk over time due to informational efficiency, and that transaction costs do not necessarily neutralize its excess returns. Moreover, in cases where the predictability and excess profitability of technical analysis on pairs trading is significant, we try also to justify the potential reasons for the outperformance.

We also consider a multiple hypothesis testing framework accounting for data snooping effects arising when recruiting such a big dataset whose number of variables (i.e., trading rules) is larger than the number of observations. Multiple hypothesis frameworks focusing on the estimation of adjusted p-values and developed to limit such occurrences are more than necessary nowadays. Classical statistical inference of single hypothesis is highly likely to trigger false discoveries due to the enormous amount of information constantly utilized by investors (Harvey, 2016). We use the false discovery rate control (FDR) of Barras et al., (2010) as one of the most powerful and suitable data snooping tool for investment decisions, although there is an ongoing effort in constructing such frameworks over the years by a series of methodological studies (see White, 2000; Hansen, 2005; Romano and Wolf, 2007; Hsu et al., 2010). The FDR successfully identifies the significantly profitable trading rules among those achieving positive performance, while allowing for a small number of false rejections to guarantee diversification benefits against risk.

In addition, we explore the robustness of our results by conducting a break-even analysis of transaction costs as well as subperiod analysis to assess the time-varying predictability of technical trading rules by separating our dataset into five subperiods based on major historical events. Break-even transaction costs exceed even conservative historical estimates of actual transaction costs and so do not necessarily eliminate the chance of generating significant profits. Subperiod analysis rejects the existence of monotonic downward trend in the selection of outperforming technical trading rules on more than half of the pairs examined over the years. Thus, a natural question arising is whether our results imply violation of equilibrium asset pricing and of course of the efficient market hypothesis. Our findings are more in favour of the Adaptive Market Hypothesis (Lo, 2004), in which evolutionary market dynamics create arbitrage opportunities periodically, leading to the selection of a greater number of significant rules even in recent years.

Finally, we perform a true out-of-sample analysis, which uses each subperiod’s last year of daily data as the out-of-sample period, providing an extra evaluation of the time-varying excess profitability of technical analysis on pairs trading. We compare our results with statistical arbitrageurs’ common strategies, which set the ‘lookback’ period of contrarian rules equal to the spread’s half-time of mean-reversion. Although the major trend of the out-of-sample performance shows a decay in excess profitability, portfolios of significant rules on commodity pairs can still achieve a very healthy Sharpe ratio 1.83 in recent years. Moreover, we create combined portfolios consisting of commodity, equity and foreign exchange spreads to explore the out-of-sample performance and diversification benefits of a pairs trader exposed in every single market separately as well as in all markets together. Our results represent a continuance of the findings of Gatev et al., (2006) and so, they uncover something about the performance of relative-price arbitrage activities in practice. In addition to this, they can help financial economists understand the risk and return characteristics of one of hedge fund's actively trading strategy and provide empirical evidence on how market efficiency is maintained in practice.

2. Data and descriptive statistics

2.1. Data and pairs formation

Our data consists of 'famous' spreads either actively traded by statistical arbitrage investors or being frequently advertised by trading websites or launched by financial market companies such as the CME and ICE groups [45]. We consider daily data on pairs employed by using the spreads between the closing prices of commodities, equity indices and foreign exchange rates. In total, we examine 15 pairs including four commodity pairs (Brent-WTI crude oil, platinum-gold, platinum-palladium and corn-ethanol), six equity indices pairs, three European and three U.S., (FTSE 100-CAC 40, Euro Stoxx 50-DAX, FTSE100-FTSE250, DJIA-Russell 1000, S&P 500-Russell 2000, Russell 1000-Russell 2000) and five foreign exchange rate pairs (CHF-EUR, CAD-AUD, EUR-JPY, AUD-ZAR and CAD-ZAR). We denote the foreign exchange rates employed as the rates between the U.S. dollar and the foreign currency (i.e., U.S. dollars per unit of foreign currency), while for the formulation of the equity spreads we use directly the equity indices instead of any corresponding ETFs, following the previous literature. Furthermore, for our commodity spreads we employ the continuous price series of each commodity's future contracts, which represent the daily closing price of the most actively traded contract. This ensures that the underlying instrument should last longer than the observation period when analysing the performance of technical trading. The sample dataset for all the pairs covers the period from January 1, 1990 to December 12, 2016, except from the corn-ethanol, whose sample period starts from March 30, 2006, due to data availability. We also consider daily data on short-term interest rates for every currency, since they constitute a major concern for currency traders, due to their effect on the overall return generated from a trading strategy. We

[45] By using the term 'famous' pairs, we refer to pairs, which present a sufficient liquidity in terms of the number of quotations, as well as they are actively traded by investors. We have communicated with fund managers and algorithmic traders who suggested to us the majority of pairs, described above, for our application.

used the Datastream Thomson Reuters database to acquire all the closing prices explained above.

To evaluate the technical analysis in pairs trading, we need first to define every single spread relative to its corresponding instruments. In the meantime, it is very essential to tackle the issue of non-simultaneous pricing plaguing such simulations, by ensuring that the closing times of each leg occur at the same time. Indeed, all the examined commodities contracts have the same trading hours as they are listed in the CBOT (agriculture), NYMEX (energy and precious metals) and COMEX (precious metals) exchanges respectively, which constitute the CME group derivatives marketplace. For example, the closing times of agricultural futures are 13:20 p.m. and 7:45 a.m., while for the energy and precious metals ones is 17:00 p.m. eastern time. We should mention that the rolling forward procedure from a commodity futures contract which is near to maturity to a new month contract in the future it is also very important here. We use the same rollover procedure for both the underlying commodities futures, which is the one offered by the Datastream software to guarantee that no issues of non-synchronous trading exist. This procedure rollovers the old futures contract to the first day of the new month's most actively traded contract. As for the European equity spreads, each one consists of equity indices issued by the same or different stock exchanges (i.e., London and Frankfurt Stock Exchange and Euronext Paris), which have the same actual closing times after taking into account the hour time difference. The same holds for each of the U.S. indices considered, which are mainly issued by NYSE. In addition to this, since we obtain our foreign exchange rates data from the WM/Reuters FX benchmarks, all of them represent the closing spot rates fixed daily at 16:00 p.m. London time.

There are various ways for a market participant to construct a spread depending on what his principal goal is in terms of investment. In our study, we pair any two assets by just subtracting the closing price of the one underlying leg from the other since our aim is mainly to capture their dominant trends using technical analysis. This means that we allocate an equal proportion of our wealth to each side. Thus, the formation of a pair P_t, in which we go long a risky asset P_1 and short another risky asset P_2 at time t, is $S_t = P_{1,t} - P_{2,t}$. However, in the case of commodity spreads, we must take into consideration the fact that both commodity futures contracts are traded at different units before we start calculating the spreads and therefore we need to adjust for that[46]. In the cases of equity and exchange rate spreads it doesn't make a great difference the order we place each leg during the subtraction. Despite this, we decide to employ the rule that small cap usually lead the large cap segments, since the value of the former tend to change more regularly compared to the latter indices showing a more stable pattern (see Simons, 2010). Thus, for all the equity spreads investigated we go long the large cap, while we go short the small cap indices. For the foreign exchange rate spreads, we don't follow a specific rule and just calculate the difference between the spot prices of the underlying exchange rates in any order.

To estimate the daily gross return (without interest rates for the currency pairs) and therefore the investment performance from pairs trading, we employ the formula

$$r_t = \ln\left(\frac{P_{1,t}}{P_{1,t-1}}\right) - \ln\left(\frac{P_{2,t}}{P_{2,t-1}}\right) \tag{1}$$

where r_t is the daily gross return from buying the pair and holding it for one day, while $P_{1,t}$ and $P_{2,t}$ are the spot prices of the first and second components respectively, on day t, while $P_{1,t-1}$ and $P_{2,t-1}$ are the spot prices of the two components on day $t-1$.

2.2. Statistical behaviour and descriptive statistics

Table 1 reports the descriptive statistics of the daily returns on all spreads formed along with the statistical behaviour between the time series of each pair's underlying legs. Regarding the statistical behaviour, we employ the fractionally cointegrated vector autoregressive model (FCVAR) of Johansen and Nielsen (2012) to test for cointegration

[46] For the Brent-WTI crude oil spread or the also called "crack" spread we use the 1:1 ratio by definition, since both are traded in U.S. $/barrel units. For the platinum-gold as well as for the platinum-palladium spread we use the 2:1 ratio, since the gold and palladium futures contract unit is 100 troy ounces, while for the platinum contracts is 50 troy ounces. The price quotation for the three of them is U.S. $/troy ounce. CME group has very recently announced the launch of first-ever platinum-gold and platinum-palladium spread futures. However, due to data unavailability we need to construct these spreads on our own, as indicated above. As for the corn-ethanol spread, we need to consider that the corn futures are priced in U.S. cents/bushel, while the ethanol futures are priced in U.S. dollars/gallon. At the time of writing 1 bushel of corn generates approximately 2.8 gallons of ethanol. Therefore, we define a spread traded at U.S. cents/bushel at time t as $S_t = (2.8 * E_t * 100) - C_t$.

ranking[47]. As we have already mentioned, cointegration is synonymous to the mean-reversion of a pair's spread and therefore important to perform statistical arbitrage techniques, since it justifies mean-reversion in the long run equilibrium. For that reason, we present the *p*-values of the likelihood ratio (*LR)* statistic, also called as 'trace statistic', testing for cointegrating combinations between the underlying time series, under the null hypothesis of zero cointegration rank[48]. We also report the *p*-values testing for first-order autocorrelation under the Ljung-Box Q statistics of the daily returns of all spreads.

[Table 1]

Among the four commodity pairs only the Brent-WTI crude oil yields positive performance, on average (0.3 basis points per day or 0.78% annually). The rest of the commodity pairs display a negative performance, with the corn-ethanol generating the most negative returns (-9.2 basis points per day or -23.8 annually). Among the equity spreads three out of six yield positive mean returns, with the FTSE100-FTSE250 generating the biggest ones (1.2 basis points per day or 3.04% annually), while S&P 500-Russell 2000 pair yields the minimum performance (-0.8 per day or -2.16% annually). In the case of currency spreads the CAD-ZAR performs the most on average (2 basis points per day or 5.04% annually) and the CHF-EUR underperforms the most (-0.7 basis points per day or -1.76% annually).

In terms of standard deviation of daily gross returns, commodity spreads are in general more volatile than equity and exchange rate spreads respectively. The most volatile commodity spread is the corn-ethanol (0.19%), while the least volatile is the platinum-gold (0.12%). The most and least volatile spreads among the equity ones are the FTSE100-CAC 40 (0.76%) and the DJIA-Russell 1000 (0.33%) respectively. The currency pairs display similar levels, with the AUD-ZAR and CAD-ZAR being associated with the highest standard deviations (0.89%), and the EUR-CHF with the lowest (0.41%).

[47] This method is an enhanced version of the CVAR model of Johansen (1995) accommodating for both fractional integration and cointegration. For instance, the FCVAR includes two additional fractional parameters *d* and *b* denoting the fractional integration order, and the degree of fractional cointegration respectively. For more information see Nielsen and Johansen (2012).

[48] We also have also computed the results from the Johansen (1995) cointegration test, which are available upon request.

The Ljung-Box test for residual autocorrelation per daily gross returns indicates persistence in almost the half of the cases at least at 10% significance level. We translate this into the existence of trends for the majority of spreads considered. Indeed, only the return series of platinum-gold doesn't display a significant first order autocorrelation among all commodity pairs. In the case of equity pairs three out of five (Eurostoxx 50-DAX, DJIA-Russell 1000 and Russell 1000-Russell 2000) show residual autocorrelation at least at 5% significance level, while among the six currency pairs three appear significantly first order autocorrelated, of which two (EUR-CHF, CAD-AUD) present this property at 1% level of significance. This evidence also justifies the exercise of trend-following technical trading rules and the comparison of their performance with this of contrarian trading rules.

Finally, the computed *p*-values of the trace statistic testing for cointegrating relationships between the two legs of a pair, reveal the rejection of the null hypothesis for zero rank cointegration for the vast majority of all the spreads considered. However, this result was somehow expected, since all these spreads are actively used for statistical arbitrage trading by many trading institutions. For example, all commodity spreads reveal at least one cointegrating relation between their corresponding commodities at least at 5% statistical significance level. The same holds for the equity pairs, in which three out of six spreads (Euro Stoxx 50-DAX, FTSE100-FTSE250 and Russell 1000-Russell 2000) present cointegration at 5% nominal level, while the rest of them validate at least one rank of cointegration at 1% statistical significance level. Contrary to the above, the currency spreads show the weakest cointegration relations between their corresponding foreign currencies, compared to commodity and equity pairs. Among all currency spreads two, namely the CAD/USD-AUD/USD and the AUD/USD-ZAR/USD, do not reject the null hypothesis of zero cointegration rank, while the rest exchange rate pairs validate cointegration at least at 10% significance level, from which the EUR-CHF and the EUR-JPY show cointegration at 5% nominal level. However, we will not discard the two non-cointegrating pairs from our technical trading rules' exercise since one of the main goals of this study is to explore whether technical analysis yields a significant performance on 'popular' spreads heavily traded by investors or advertised by trading platforms.[49] Another reason for trading those two non-cointegrating pairs, is the

[49] The above examined currency pairs were suggested from hedge fund managers and statistical arbitrageurs performing pairs trading after making contact with them.

empirical links between the real exchange rates of Canada, Australia and South Africa and the real prices of commodities that they export (see Chen and Rogoff, 2003). The primary reason is that commodities form a significant component of those countries' exports and so of their economies, while the considerable connection between their exchange rates and commodity prices potentially explain a co-movement between their real exchange rates.

3. Technical trading rules universe

Technical analysis in its qualitative and most known form incorporates mainly graphical tools, such as chart analysis of past prices dynamics in order to capture specific patterns in the data, which investors use for future predictions. However, in our study, we focus on its strictly quantitative form, which uses quantitative modelling of time series data to generate forecasts and so trading signals. This type of modelling exploits the excess profitability of technical trading rules constructed in an algorithmic framework, while employing time series of past prices, volumes, and other observables to define whether a buy (long), neutral (out of the market), or sell (short) signal should be taken in the next trade.

We consider seven families of technical trading rules based on past price data of the computed pairs as they are widely used by commodities, equities and forex traders[50]. Those classes of rules are categorized in momentum/trend-following rules and contrarian/mea-reverting rules usually employed by pair traders and try to identify 'overbought/oversold levels'. The momentum/trend-following rules include: *filter rules*, whose main characteristic is to follow strong trends by taking long (short) positions accordingly; *moving averages*, attempting to ride trends and taking positions when crossovers between the pair and a moving average of a given length or between two moving averages of different lengths occur signifying a break in the trend; *support and resistance rules,* which try to identify breaches of a pair's price through local maximums (minimums) triggering further price movements towards the same direction and leading to long (short) signals; *channel breakouts* similar to having time-varying

[50] We do not apply rules utilizing volumes transactions, such as the on-balance volume averages (see Sullivan et al., 1999) since it is not easy to accurately observe the exact volume of a pair of two assets, which do not actively traded into market.

support and resistance levels which form a channel of fixed percentage, leading to a signal when pair's price penetrates the channel from above or below. The contrarian/mean-reverting rules include: *relative strength indicators (RSI)*, which belongs to the general family of 'overbought/oversold' indicators and attempt to capture a correction towards the opposite direction of a pair's price extreme movement; *Bollinger band reversals*, attempting to identify overbought and oversold market levels, defined as a particular distance of price from its moving average of a given length in terms of standard deviation; *Commodity channel index rules (CCI)*, similar to a combination of RSIs and Bollinger band reversals, they try to quantify the connection among a pair's price, its corresponding moving average and standard deviation, however a specific inverse factor is used to scale the index.

Following previous studies (Sullivan et al., 1999; Hsu et al., 2016) we consider numerous variations of the above technical trading rules as well as a spectrum of different plausible parameterizations of each variation. These possible trading rules consist a large universe summing up to a total of 18,412 apparent technical trading rules including, 1932 filter rules, 7920 moving averages, 2310 support and resistance rules, 2250 channel breakouts, 730 relative strength indicators, 2160 Bollinger bands and 1110 commodity channel index rules. We present the exact details of each class of trading rules, their variations as well as the various parameterizations examined in Appendix A.

4. Excess returns, transaction costs and performance metrics

Before we assess the performance of technical trading rules, we must first compute the daily excess return from buying and holding each spread, (i.e., buying the first underlying asset and selling the second simultaneously) for each prediction period. For commodity and equity spreads the calculation of their daily excess return is the daily gross return, as defined before, net of the risk-free rate r_f

$$r_t = [\ln\left(\frac{P_{1,t}}{P_{1,t-1}}\right) - \ln\left(\frac{P_{2,t}}{P_{2,t-1}}\right)] - \ln(1 + r_f) \qquad (2)$$

We use as the risk-free rate the daily effective federal funds rate, in accordance with the previous literature. In order to calculate the daily excess return for currency spreads we follow Hsu et al., (2016) and take into account the short-term interest rates of each currency, in which case the excess return is defined as

$$r_t = [\ln\left(\frac{P_{1,t}}{P_{1,t-1}}\right) - \ln\left(\frac{(1+i_{t-1})}{(1+i^*_{1,t-1})}\right)] - [\ln\left(\frac{P_{2,t}}{P_{2,t-1}}\right) - \ln\left(\frac{(1+i_{t-1})}{(1+i^*_{2,t-1})}\right)] \tag{3}$$

where i_{t-1} and $i^*_{1,t-1}$ designate the daily interest rates on U.S. dollar deposits and the first foreign currency deposits, respectively, while $i^*_{2,t-1}$ denote the daily interest rates on the second foreign currency deposits on day $t-1$. Thus, the excess return consists of the spread between the appreciations of each of the two foreign currencies against the domestic currency (U.S. dollar) over the holding period, $\ln\left(\frac{P_{1,t}}{P_{1,t-1}}\right) - \ln\left(\frac{P_{2,t}}{P_{2,t-1}}\right)$, less the difference of the interest rate carries related with borrowing one unit of domestic currency and lending one unit of foreign currency overnight, $\ln[\frac{(1+i_{t-1})}{(1+i^*_{1,t-1})}] - \ln[\frac{(1+i_{t-1})}{(1+i^*_{2,t-1})}]$. We transform the annualized risk-free and short-term interest rates downloaded from Datastream, into daily rates for our application. We achieve this by using the formula $i_t = \ln(1+i^a_t)/360$.

Now let $s_{j,t-1}$ denote the trading signal generated from a trading rule $j, 1 \leq j \leq l$ (where $l = 18{,}412$) at the end of each prediction period $t-1$ $(\tau \leq t \leq T)$ depending on the information given, where $s_{j,t-1} = 1, 0,\ or -1$ represents a long, neutral or short position taken at time t. We mainly use three performance metrics throughout this study, the *mean excess return* and *Sharpe ratio* criteria, as well as the compounded annual growth rate of an investment in our out-of-sample market portfolio simulation (see subsection *5.3*). The mean excess return is the absolute criterion of each rule's returns, while the Sharpe ratio is a relative performance criterion since it represents the ratio of the average excess return to the total risk of the investment estimated standard deviation of excess returns. Practically speaking, the technical trading rules earn the risk-free rate in periods when a neutral signal is triggered. The mean excess return criterion $\overline{f}_{j,t}$ for the trading rule j is given by

$$\overline{f}_{j,t} = \frac{1}{N}\sum_{t=R}^{T} s_{j,t-1} r_t\,,\ j = 1, \ldots, l, \tag{4}$$

where $N = T - \tau + 1$ is the number of days examined. We denote as τ the start date for each subperiod, since some of the trading rules employ lagged values up to one year (252 days), which we need to take into account. Then, the Sharpe ratio criterion expression SR_j for trading rule j at time t is defined by

$$SR_{j,t} = \frac{\overline{f}_{j,t}}{\widehat{\sigma_{j,t}}},\ j = 1, \dots, l, \quad (5)$$

where $\overline{f}_{j,t}$ and $\widehat{\sigma_{j,t}}$ are the mean excess return and the estimated standard deviation of the mean excess return respectively. Except from measuring units of mean excess returns per unit of risk of an investment, the Sharpe ratio is strictly connected with the actual *t*-statistic of the empirical distribution of a strategy's returns, which places this metric appropriate for our multiple hypothesis testing framework (see Harvey and Liu, 2015)[51]. Compound annual growth rate (CAGR) is slightly different from the annualized mean excess return by assuming that we do not withdraw profits or add extra cash for offsetting losses and thus, we hold cash into our account each time period, while we maintain the same leverage throughout the process (see Chan, 2017). In such a way, we achieve a compounded growth rate over time for every asset, in which the gains and losses are rolled over, similar to investing $1, which grows daily at the rate of daily mean excess return.

$$CAGR_{j,t} = (1 + \textstyle\sum_{t=R}^{T} s_{j,t-1} r_t)^{252/(T-R)} - 1,\ j = 1, \dots, l \quad (6)$$

So far, we haven't assumed the impact of transaction costs on the technical trading rules' performance over the examined spreads. In practice, these may be quite significant especially for statistical arbitrage traders, which take long and short positions on two assets simultaneously. On the other hand, estimating the effect of transaction costs ex post, once outperforming rules have been identified, makes it more complicated for an investor to foresee which outperforming rules will offset transaction costs a priori. A potential predictability of a selected strategy before the implementation of transaction costs, can be easily neutralized when those are adjusted through the selection process, sometimes due to the impact of frequent signals (Timmerman and Granger, 2004). Thus, we handle transaction costs "*endogenously*" to the selection process. In particular, we subtract the transaction costs every time a buy or sell signal is triggered based on the prediction of the corresponding spread. This comes down to taking into consideration the one-way transaction costs of each component separately.

[51] The t-statistic of a given sample of historical returns $(r_1, r_2, \dots r_t)$, testing the null hypothesis that the average excess return is zero, is usually defined as $t = \frac{\widehat{\mu}}{\widehat{\sigma}/\sqrt{T}}$, while the corresponding Sharpe ratio is given by the formula $SR = \frac{\widehat{\mu}}{\widehat{\sigma}}$.

Following the study of Locke and Venkatesh (1997), we consider 3.3 basis points one-way transaction costs for a position taken on each of the commodity futures employed to construct the commodity spreads. Furthermore, we assume that an investor funds their position with 100% equity rather than using a margin, since we measure daily returns as the log of the difference in price relatives (Miffre and Rallis, 2007; Marshall et al., 2008). After speaking with several brokerage firms, we assume 5 basis points one-way transaction costs for the corresponding European equity indices and 2 basis points for the U.S. ones, which are the costs charged for institutional investors. Trading exchange rates doesn’t incorporate any fixed amount of brokerage costs such in equities or commodities markets. However, the only transaction costs investors facing, arise from the bid-ask spread in spot exchange rates and interest rates. Following Neely and Weller (2013) and Hsu et al., (2016), we calculate the one-way transaction costs of each currency from their corresponding bid-ask spread in forward exchange rates on any particular day. Specifically, we use the one third of the quoted one-month forward rate bid-ask spread in each currency, since several studies have realized that posted bid-ask spreads are usually larger than the rates that the effective ones are actually traded (see Lyons, 2001; Neely and Weller 2003). This results to average one-way transaction costs of 2.9 basis points for developed countries and 17.4 basis points for emerging countries (i.e., South Africa).

5. The issue of data snooping bias

5.1. Definition and existing data snooping methods

The investigation of significant excess profitability of such a sizeable universe of technical trading rules involves the control of data snooping bias. The also called data mining issue has nowadays become even more urgent because of the severe usage of large datasets by investors and researchers, leading to promising results sometimes even by chance. We deem such data replication costly since it is quite easy to incorrectly discover a profitable trading rule. Classical statistical inference focusing on single hypothesis testing for each rule, without paying attention to the performance of the rest strategies, can lead to false rejections or the so called, Type I error due to extensive specification search. Multiple hypothesis frameworks developed to limit such occurrences are more than necessary nowadays. Recently, Harvey (2017) raises this issue as the p-hacking phenomenon (i.e. frequent falsely significant p-values) and explains that new, adjusted p-values reflecting genuine significance for an investment strategy should be defined.

Many efforts have been made so far to minimize the above undesirable consequences and we broadly divide them into two different categories: the first one controls the family-wise error rate (FWER), while the second one controls the false discovery rate (FDR). Their difference is mainly intuitive; the FDR calculates the false rejections in terms of proportion and it is defined as the proportion of false discoveries among the total number of rejections. On the other hand, the FWER estimates the probability of making at least one false rejection, which is more conservative by definition, especially when the number of hypotheses is large.

A great number of the existing studies exercise their proposed data snooping methods mainly on technical trading rules performance (see among others Sullivan et al., 1999; Romano and Wolf, 2005; Hansen, 2005; Hsu et al., 2010; Bajgrowicz and Scaillet, 2012). Large universes of technical trading rules provide a breeding ground to test the power of multiple hypothesis methods, since it is quite likely to discover a rule working well, even by chance, especially within the same classes of rules.

White (2000) introduces the so called 'bootstrap reality check' (BRC), which focuses on the statistical significance of the 'best' performing strategy, drawn from l number of strategies, while contemporaneously tests whether the significance of all strategies is less than the nominal significance level α. Thus, the null hypothesis of the BRC test in a joint hypothesis framework is

$$H_{0j}: \max_{j=1,..l} \varphi_j \leq 0, \tag{7}$$

where φ_j is the performance measure of the jth trading rule, no better than the benchmark. An important innovation of this procedure is the estimation for the first time of the empirical distribution of the reality check statistic through bootstrapping. Sullivan et al., (1999) also use the BRC test, to select significant technical trading rules, however we believe that it is a rather conservative approach since it concentrates only on the 'best' performing rule, and not to all rules showing positive performance.

Focusing also to the *maximal*, trading rule, Hansen (2005) proposes the *superior predictive ability* (SPA) test, to correct drawbacks of the BRC test such as the influence of poor and inconsistent strategies, especially when the l is quite large, which leads to low power. In this way, the SPA test employs studentized test statistics, while sets less weights to the test statistics of rules showing poor performance[52]. Both White's (2000) BRC test and the Hansen's (SPA) test use the Bonferroni correction method, a special case of the FWER, for multiple hypothesis testing, in which the individual null hypotheses are rejected for each p-value less than a significance level of α/l in a single-step procedure.

Investors on the other hand are keen on discovering all the statistically significant trading rules showing positive performance instead of investing their wealth in the maximal one only, involving model risk. Incorporating this assumption, Romano and Wolf (2005) suggest their *stepwise multiple testing* (*StepM*) method as an improvement to the *single-step* BRC testing method of White (2000). They actually embrace the Holm (1979) procedure following a stepwise structure, in which individual p-values are placed in an acceding order, after bootstrapping the empirical distribution of each rule, similar to White (2000). During the first step, they compare each p-value with a nominal significance level, leading to selecting the significant rules. In the second step, they replicate the same mechanism after excluding the statistically significant rules of the first step. This helps to identify more significant trading rules likely appeared insignificant during the first step due to correlation and dependence among the trading rules. They repeat the whole procedure until no significant trading rule is identified. In a similar manner, Hsu et al., (2010) develop a stepwise extension of the SPA test of Hansen (2005) in order to minimize the influence of poor performers on the power of the test.

Although the above stepwise approaches are quite powerful as multiple hypothesis testing tools, their main drawback is that they do not select further rules once they have detected a rule whose performance is due to luck. This may help in effectively controlling the Type I error on the one hand, but it may increase the probability of a Type II error (i.e. missing true discoveries) on the other. To achieve a good balance between those two types of errors, Romano and Wolf (2007) developed a generalized methodology, controlling for the stringent FWER criterion, while allowing for more than one false rejections to happen. Their goal is to reject at least a specific number of

[52] A studentized test statistic refers to a simple test statistic divided by the consistent estimator of its standard deviation. This helps one to compare objects in the same units of standard deviation.

false hypotheses to have a good tradeoff between Type I and II errors, as well as selecting more significant rules.

While in the case of FWER numerous developments have been made in the literature to achieve the above tradeoff, the FDR tolerates by definition a certain proportion of false rejections, so as to identify every statistically significant outperforming rule, while having a good balance in controlling Type I and Type II errors (see also Abramovich et al., 2006). This feature makes FDR a more powerful multiple hypothesis testing tool than the other conservative FWER methods. In financial literature, Barras et al., (2010) propose a modified $FDR^{+/-}$ version based on Storey's (2004) FDR approach, to discover significant alphas in mutual fund performance. This modified version allows for the first time the separate quantification of the proportion of false discoveries among trading rules performing better or worse than the benchmark. Bajgrowicz and Scaillet (2012) employ the $FDR^{+/-}$ approach in the context of identifying outperforming technical trading rules on the DJIA index, while accounting for data snooping. Their findings confirm the power of the specific FDR approach over the conservative FWER method in detecting more significant technical trading rules, while it provides a reasonable balance between true positives and erroneous discoveries in diversifying against model risk. Another comparative advantage of the $FDR^{+/-}$ approach against the FWER methods, is the ability to find the outperforming rules, even if the performance of the best rule in the sample is due to luck. In practice, it is quite regular for the best rule, in terms of the highest excess returns, to possess no statistically significant profitability. For the above reasons, we adopt this FDR test as the most suitable multiple hypothesis testing specification in the context of controlling the data snooping effects arising from the application of our large technical trading rules' universe in pairs trading. In the next subsections, we briefly describe the $FDR^{+/-}$ multiple hypothesis testing setup as well as its usage as a portfolio construction tool.

5.2. Multiple hypothesis testing framework and the FDR methodology

Elaborating on the FDR approach as a multiple hypothesis testing procedure, in what follows we need first to define the null hypothesis, H_0, according to a test statistic, φ. We use the mean return and the Sharpe ratio criteria as the test statistics for conducting our multiple hypothesis testing. The test statistic for each rule j defines the setup under the null hypothesis (i.e., $H_{0j}\colon \varphi_j = 0$), in which the rule j does *not* outperform the benchmark[53]. On the contrary, the alternative hypothesis assumes the presence of abnormal performance, positive or negative (i.e., H_{Aj}: $\varphi_j > 0$ or $\varphi_j < 0$). The *FDR*$^{+/-}$ method requires p-values, p_j for $1 \leq j \leq l$, from a two-tailed test, since the main parameter we need to estimate is the proportion of rules with no abnormal performance, π_0, in the total universe, satisfying the H_{0j}. However, since we are mainly interested in identifying significantly outperforming rules, we define a technical trading rule j as significantly positive, if it displays abnormal performance (i.e., reject H_{0j}) and its performance metric is positive (i.e., $\varphi_j > 0$).

The initial FDR version of Benjamini and Hochberg (1995) adopts independence across multiple hypotheses, while later, studies by Benjamini and Yekuteli (2001), Storey (2002), and Storey et al., (2004) prove that the FDR holds also under "weak dependence" conditions when the number of hypotheses is very large. Also, Bajgrowicz and Scaillet (2012) explain that technical trading being built likewise to our universe satisfy this feature, since the rules are dependent in small blocks, within the same family (e.g. filter rules), but essentially independent across different families. Thus, the more "local" the dependence, the more likely is to meet the weak dependence principle.

The FDR concentrates on estimating the expected value of the ratio of erroneous selections over the rules showing significant performance. Specifically, define the *FDR*$^{+/-}$ as the expected value of the proportion of false selections, $F^{+/-}$, among the significant rules, $R^{+/-}$ (positive or negative). The latter are just the rules that perform either better or worse than the benchmark while at the same time their p-values reject the H_{0j} under some nominal significance level α. Thus, the estimate of *FDR*$^{+/-}$ is given by $\widehat{FDR}^{+/-} = \frac{\widehat{F^{+/-}}}{R^{+/-}}$, where $\hat{F}^{+/-}$ and $\hat{R}^{+/-}$ are the estimators of $F^{+/-}$and $R^{+/-}$, respectively. For example, an *FDR*$^{+/-}$ 100% conveys that, among both the outperforming and underperforming trading strategies, no rule generates genuine performance on average and vice versa.

After acquiring all the individual p-values, using resampling procedures relevant to the previous literature, such as the stationary bootstrap of Politis and Romano (1994), the specific formula for the estimate of $\widehat{FDR^{+/-}}$ under a threshold γ is as follows

[53] Since we use the excess return and Sharpe ratio as the performance metrics, our benchmark is by definition the "risk-free" rate, describing an investor being out of the market.

$$\widehat{FDR^{+/-}}(\gamma) = \widehat{F^{+/-}}/\widehat{R^{+/-}} = \frac{\widehat{\pi_0} l\gamma}{\#\{p_j \leq \gamma;\ j=1,\dots,l\}}, \tag{8}$$

where l is the entire universe of technical trading rules, γ is the p-value cut-off and $\widehat{\pi_0} = \frac{\#\{p_j > \lambda;\ j=1,\dots,l\}}{l(1-\lambda)}$ is an estimator of the proportion of rules that show no abnormal performance. The estimation of $\widehat{\pi_0}$ requires us to define the tuning parameter λ by visually examining the histogram of all p-values.

Focusing now on the significantly positive technical trading rules, we can compute a separate estimator for the FDR^+. This holds under the assumption that the false discoveries spread evenly between technical trading rules with positive and negative performance and with equal tail significance $\gamma/2$, due to symmetry assumptions. Thus, the estimator is

$$\widehat{FDR^{+}}(\gamma) = \hat{F}^{+}/\hat{R}^{+} = \frac{1/2\widehat{\pi_0} l\gamma}{\#\{p_j \leq \gamma, \varphi_j > 0;\ j=1,\dots,l\}}, \tag{9}$$

where $\hat{R}^+$is now the number of trading rules chosen as significantly positive, while among them $\hat{F}^+$denotes those rules which have been selected falsely. Similarly, we can compute a separate estimator of the FDR^- among the rules generating negative returns. However, this is out of the scopus of this paper.

Finally, we can also extrapolate the proportion of trading rules displaying nonzero performance as $\pi_A = 1 - \pi_0$. This may be useful for an investor who wants to divide π_A into the proportions of positive, π_A^+, and negative, π_A^-, rules in the population. Nevertheless, the critical part of the FDR method is to identify the right p-value cutoff γ by controlling the FDR^+ at a predetermined level (i.e. 10%) in order to isolate the genuinely outperforming rules from the total population. We describe the precise steps of achieving this, the estimation of λ and so of $\widehat{\pi_0}$, as well as the computation of π_A^+ and π_A^- in the Appendix B.

5.3. Portfolio construction

We construct the portfolios of rules by selecting them in accordance with the FDR^+. In particular, we set the $\widehat{FDR^+}$ equal to 10%, as a good trade-off between truly outperforming technical trading rules and wrongly chosen ones (Bajgrowicz and Scaillet 2012). Thus, we built a 10%-FDR$^+$ portfolio of trading rules for each pair, which means that 90% of the total number of the portfolio's rules, significantly outperform the benchmark. We pool the signals of the chosen rules with equal weight, similarly to a forecast averaging technique. We do not attribute more weight to more effective rules since this would result in decreasing the FDR$^+$ portfolios below the desired level, similar to selecting fewer strategies. We finally treat the neutral signals as totally liquidating our positions and do not invest a proportion of wealth, corresponding to them, at the "risk-free" rate. This assumption helps us to measure the true performance of the FDR portfolios.

6. The full-sample performance of technical trading rules

We try now to measure the empirical predictability of technical analysis on pairs trading, based on our full sample of 25-year period. In doing so, we need first to examine the statistical significance of such predictability by using the FDR approach based on the performance metrics of every technical trading rule described in the previous sections. Table 2 provides evidence not only for the highest performance metric and its corresponding *p*-value generated by the best among all rules, but also highlights the power of the FDR approach by reporting the number of predictive rules yielding statistically significant positive performance.

[Table 2]

Firstly, we present the findings based on the mean excess returns in panel A, which conveys that the ability of technical trading rules to forecast the examined pairs is limited in general in sample. Only three out of fifteen pairs are predictable at 1% significance level (i.e., the number of pairs with one asterisk) based on the mean excess returns' performance, while two pairs are also predictable at 5% and 10% nominal level respectively.

Of these, the three pairs belong to the commodity market, which leads to three out of four commodity pairs (i.e., 66%) being predictable, for instance the Brent-WTI, the platinum-gold and the corn-ethanol. The Brent-WTI seems to be the most predictable pair not only for the commodity market but also for the rest of the markets examined, based on the mean excess return criterion. The FDR method identifies 442 significantly predictive rules for this commodity pair, while in terms of the economic magnitude of this predictability, the best-performing rule yields an outstanding mean excess return of

17.6% per annum at the 1% level of statistical significance. In the corn-ethanol case, the second best most predictive pair, there exist 12 significantly predictive rules, from which the best-performing rule also generates a very promising 8.57% annualized mean excess return, while the platinum-gold is predictable by one rule only yielding a 2.03% annualized mean excess return.

On the other hand, the evidence for the technical analysis' forecasting ability, based on mean excess returns, drops for the constructed pairs in equity and foreign exchange markets. Only one out of the six equity pairs (i.e., 16%) is predictable at the 1% significance level, while the predictability comes mainly from the best-performing technical trading rule, generating a significant annual return of 1.70% (S&P 500-Russell 2000). We get an analogous picture for technical rules' predictiveness on foreign exchange pairs. In particular, there is also one out of five exchange rate pairs (i.e., 20%), which is predictable this time at the 10% level. In addition to this, only the best rule is able of providing an excess profitability of 1.41% (JPY/USD-EUR/USD) per annum.

Panel B now reports the predictability and excess profitability of technical analysis based on the Sharpe ratio performance metric. Specifically, we allow for risk adjusted returns in terms of volatility and explore whether technical trading rules can yield significantly positive annualized Sharpe ratios on the investigated pairs. Opposite to the results associated with the mean excess returns, technical analysis predictability appears significantly strong in pairs trading when we adjust for risk. All commodity pairs are significantly profitable under the Sharpe ratio criterion mainly at the 1% significance level, except from the corn-ethanol whose profitability is significant at the 5% level. Again, the most predictable pair is the Brent-WTI with 563 trading rules producing positive performance, compared to twelve for the gold-platinum, ten for the platinum-palladium and six for the corn-ethanol. The annualized Sharpe ratios for the statistically significant, best-performing rules for commodity pairs range from 0.45 (corn-ethanol) to a very healthy 1.20 (Brent-WTI).

In the case of the six equity pairs, five are significantly profitable after employing the technical trading rules at least at the 10% level, of which one is profitable at the 5% level and two at the 1% level. Although, the maximum Sharpe ratio of 0.54% is achieved by the best-performing among the nine outperforming rules of S&P 500-Russell 2000, the FTSE 100-CAC 40 is the most predictable pair under this criterion with 21 significantly outperforming technical trading rules and a maximum Sharpe ratio of 0.31. The rest top Sharpe ratios for each pair range from 0.31 to 0.45, while all of them reveal more than one predictive trading rule.

Among the five exchange rate pairs, four generate genuinely positive Sharpe ratios using technical analysis, of which two are statistically significant at the 10% level and two at the 1% level. Similar to the technical analysis performance on equity pairs, the top generated Sharpe ratio in not synonymous with the strongest predictability of an exchange rate pair. For example, the CAD/USD-ZAR/USD generates the top Sharpe ratio of 0.50% with twelve outperforming trading rules, while the JPY/USD-EUR/USD is the most predictable pair with sixteen outperforming rules but the Sharpe ratio of the best one is only 0.34%.

In overall terms, Table 2 reveals a considerable difference in selecting a trading rule depending on the significant mean excess return in contrast with the Sharpe ratio. Especially for equity and exchange rate pairs, we can hardly find any predictive rules based on the mean excess return criterion even with the powerful FDR method. However, when we adjust for risk the picture is totally different and we can identify numerous predictive rules amongst the total universe for almost every single pair. There are of course cases (i.e., FTSE 100-FTSE 250 and AUD/USD-ZAR/USD), in which we cannot select any significantly predictive trading rules with either the mean excess return or the Sharpe ratio metric. Moreover, there is only one case, such as the corn-ethanol pair, in which we observe the number of predictive rules decreasing after taking risk into consideration.

Another interesting investigation, and one of the main objectives of this study, is to pinpoint which technical trading rules are the best-performing for each pair according to our both criteria and especially the families of rules which concentrate the lion's share on the number significantly predictive rules. Table 2 provides us with the relevant information about the best-performing trading rules for commodity, equity and exchange rate pairs. In particular, the majority of best-performing rules on commodity pairs belong to two mean-reverting families, for instance the relative strength indicators and Bollinger bands. Using the mean excess return as our criterion, the three cases that generate significant returns reveal three different rules as best-performing, namely a Bollinger band, a relative strength indicator and a support and resistance rule. Considering on the other hand the Sharpe ratio criterion all the best-performing rules belong either to Bollinger band rules or relative strength indicators across all commodities spreads.

For equity spreads, the highest performing rule, which produces the significant mean excess return for one case only, belong to the class of Bollinger bands, while a support and resistance rule assuming a holding period is the best-performing rule in the single case, in which a significant mean excess return is generated among the currency spreads. When we use the Sharpe ratio metric among all equity pairs, three cases indicate commodity channel indices as highest performing rules, while one is a channel breakout rule and one is a Bollinger band rule. Among foreign exchange rate pairs, there is a diversity on the rules generating significant Sharpe ratios. For example, we observe a moving average rule, a support and resistance rule, a channel breakout, a commodity channel index and a filter rule as best-performing.

In overall, there is a penchant for mean-reverting rules to be best-performing and especially for those assuming a holding period *c* across all spreads examined. This was somehow anticipated since we trade pairs which cointegrate and mean-reversion is imminent in the long-run.

Table 3 provide us now with more in-depth information about the families, in which the total number of significantly predictive rules of each pair belong in percentage terms. Using a powerful multiple hypothesis framework, such as the FDR, allow us to identify the families of all rules performing significantly best on each spread, giving a clearer picture for the patterns existing in each spread alone. We mostly focus on the families of rules producing significant Sharpe ratios under which we are able to identify a biggest number of predictive rules compared to those based on the mean excess returns. We however display the percentages of the latter metric in parenthesis.

[Table 3]

For commodity spreads, in the three out of four cases generating significant mean excess returns, the majority of predictive rules belong to mean-reverting rules except from the corn-ethanol case, in which the highest percentage of predictive rules is attributed to the support and resistance family. Moreover, when we employ the Sharpe ratio criterion we observe two cases (i.e., platinum-gold and platinum-palladium), in which the highest percentages of predictive rules concentrate on trend-following rules, for example the channel breakouts and support and resistance rules, while for the rest of the cases, contrarian rules and especially the Bollinger bands remain dominant in terms of significance.

Although there is only one case generating a significant mean excess return among equity pairs, this doesn't also happen for the significant rules selected under the Sharpe ratio metrics, which needs further investigation. In four out of six spreads, contrarian technical trading rules seem to dominate in terms of predictability, substantially those belonging to the commodity channel indices. However, there are two occasions, namely the FTSE100-CAC 40 and Euro Stoxx 50-DAX, in which the predictability of technical trading rules is mostly attributed to momentum rules, even though their best-performing rules exist in the Bollinger bands family. For instance, channel breakouts appear dominant in the first case, while channel breakouts, moving averages and support and resistance rules concentrate the highest percentages in the second.

The above findings are even more profound for currency spreads. For the majority of the cases (four out of five), momentum rules have the lion's share among the significant rules, with moving averages and support and resistance concentrating the highest percentages, while filter rules and channel breakouts having the rest. There are also cases (i.e. CHF-EUR and EUR-JPY), in which no mean-reverting rules exists among the significant ones.

7. Robustness checks

7.1. Break-even transaction costs

So far, we have reported the results of technical trading rules performance by handling transaction costs (brokerage fees and forward rate bid-ask spreads) as endogenous to the selection process. Moving one step forward and following previous studies, such as those of Bessembinder and Chan (1998), Neely and Weller (2013) and Bajgrowicz and Scaillet (2012), we try to identify the break-even transaction costs of the most significant technical trading rule for each corresponding pair. Typically, one-way break-even transaction costs represent the level of transaction costs that neutralize the profits generated by a trading rule and so minimize the excess profitability exactly to zero. Thus, we compute the break-even transaction costs by increasing them up to the level that the most predictive trading rule is not able to generate positive performance under the FDR framework. This approach tackles the exogeneity problem of transaction costs, described by Bajgrowicz and Scaillet (2012), by deriving ex ante break-even transaction costs computed endogenously. The overall procedure helps us to identify level of robustness of the results generated in the previous section by comparing the difference of the break-even transaction costs with the actual transaction costs we consider. The greater the difference the more robust the performance of a technical trading rule is deemed.

Table 4 displays the break-even transaction costs as well as the actual one-way transaction costs employed in our study (in basis points). We also report the number of trades for the most significant trading rules. The first column corresponds to the actual transaction costs used for each pair, i.e. brokerage fees for commodity and equity pairs as well as the mean of estimated forward rate bid-ask spread for currency pairs. The next four columns relate to the one-way break-even transaction costs, as described above, and the number of trades triggered for the best significant technical trading rule selected under the mean excess return and the Sharpe ratio criterion respectively, over the full sample period.

[Table 4]

In terms of trading activity, the significant mean excess return-selected rules generate a considerably larger number of trades compared to those of the significant rules chosen under the Sharpe ratio metric. For instance, in the case of mean excess return criterion for commodity pairs the number of trades triggered span from 291 (platinum-gold) to 2432 (Brent-WTI), while for the rules selected under the Sharpe ratio metric it spans from 3 (corn-ethanol) to 434 (Brent-WTI). This phenomenon becomes more intense when looking at the equity or the exchange rate pairs. The nature of Sharpe ratio to capture the average excess return per unit of total risk is probably the reason of choosing rules triggering less trades as best significant, aiming to minimize the total risk of the investment. Moreover, the trading rules of each three markets examined by both the mean excess return and Sharpe ratio criterion split into three different groups. Commodity pairs' significant rules tend to trade on a higher frequency, while equity and currency pairs' rules trade on medium and low frequency respectively, on average. However, given that we build our analysis on daily data covering a 25-year period (January 1990 to December 2016), which corresponds to a total of 7045 trading days (except from the corn-ethanol pair), the overall trading activity denotes that the genuine trading rules are quite prudent in overall. This picture stems from treating the transaction costs endogenously to the selection process and so rules generating less signals are chosen. In other words, a considerable amount of transaction costs can offset the performance of trading rules triggering more frequent signals.

Break-even transaction costs on the other hand, surpass by far the actual transaction costs in most cases, and especially when we consider the Sharpe-ratio-selected rules, since their number of trades is quite small compared to the full sample trading days. For instance, the break-even transaction costs of the best-performing rules selected by mean excess returns for commodity pairs range from 24 (corn-ethanol) to 26 (platinum-gold), while for the corresponding Sharpe ratio-selected ones range from 25 (gold-platinum) to 552 (corn-ethanol) basis points. It is also worth to mention that, the commodity pairs' rules achieve the highest break-even transaction costs on average, with second and third best those of foreign exchange and equity pairs' rules respectively. Despite this there are quite few cases (e.g. JPY/USD-EUR/USD), in which the break-even transaction costs of significant mean excess return-selected rules are slightly above the actual transaction costs or less than ten basis points. In general, we are aware that it is difficult to measure the transaction costs precisely since they have declined over time. Nevertheless, the break-even transaction costs displayed in Table 4 exceed even conservative (high) historical estimates of actual transaction costs on average, for example Allen and Karjalainen (1999) and Ready (2002) use transaction costs ranging from 10 to 25 basis points to trade U.S. stock indices. Thus, it is noteworthy that technical predictability can be transformed to excess profitability given a fair level of transaction cost in pairs trading at least in a "backtesting" framework of genuinely selected rules under the FDR method. This robustness check also highlights that transaction costs do not necessarily eradicate the possibility of yielding significant profits on pairs trading using technical analysis.

7.2. Subperiod analysis

In this subsection we focus on shorter periods of time in order to assess the time-varying predictability of technical trading rules on our corresponding pairs. Previous empirical studies on the performance of technical analysis reveal a considerable decay in the excess profitability and the evolution of predictability over the recent years as a sign of informational efficiency improvement across investors. For example, Sullivan et al. (1999) and Bajgrowicz and Scaillet (2012) provide relevant evidence from equities markets, Menkhoff and Taylor (2007) and Neely et al., (2009) express lower profitability of technical analysis overtime in foreign exchange market, while Marshall et al., (2008) come to similar conclusions in commodities market. Moreover, Gatev et al., (2006) evidence lower profits in stock market pairs trading during more recent periods as a potential outcome of increased hedge fund activity. Our 25-year historical dataset allow us to revisit the evolution of predictability of technical trading rules, this time on pairs traded in different markets.

We separate the whole sample into five subperiods: 1991-1996, 1997-2001, 2002-2007, 2008-2011 and 2012-2016. (Although, our sample starts from 1990, this year is not included in our first subsample since we require data back to one year to generate some of the trading rules). Despite the fact that the above subperiods may not have the same size, they are closely related to major historical events for all markets considered, for example the Maastricht Treaty in 1992, the East Asian currency crisis in 1997, the "dotcom" bubble in 1999-2000 and the upcoming 2002 credit crunch, the appearance of euro in 2002 and the 2003-2007 energy crisis, the global financial crisis of 2008 and finally the recent crude oil downturn in 2014.

Table 5 presents the numbers of significant rules in terms of predictability and after controlling for data snooping bias. We carry out the FDR procedure under the Sharpe ratio criterion in every subperiod and for every pair separately. For the rest of the paper we will mostly concentrate on the Shape ratio as our *t*-statistic for the multiple hypothesis testing. The reason behind this, is the Sharpe ratio's strong linkage and equivalence with the true *t*-statistic testing for the null hypothesis, compared to the excess return (see Harvey and Liu, 2015; for details). This equivalence explains the use of Sharpe ratio as a more suitable investment attractiveness metric.

[Table 5]

In the first panel of the table, we notice that technical analysis is able to predict the commodities spreads, almost across all subperiods. However, the number of significant rules differs from period to period. Except from the Brent-WTI crude oil case, which seems to be consistent with the findings of Marshall et al., (2008), and so highlights the decay on rules predictability in more recent decades, on the other hand this is not the case for the rest of commodities spreads. The overall picture indicates that technical predictability seems also strong during the more recent subperiods and for specific spreads. For instance, the platinum-gold and corn-ethanol pairs are more predictable with technical analysis as we move towards the last year of our dataset, compared to the earlier years. There is also the case of platinum-palladium, in which we observe a more stable performance of technical trading rules over all the examined subperiods. The picture seems even more diverse for the equities pairs in the second panel. For all the pairs considered, there are times of considerable predictability, when the number of significant rules is larger, and times of weak performance, without revealing a specific pattern. Only in the FTSE100-FTSE250 spread we can detect an evolutionary pattern in technical analysis predictability as we move towards the last two subperiods. Thus, previous evidence as those described by Brock et al., (1992), Sullivan et al., (1999) and Bajgrowicz and Scaillet (2012) are not consistent with those of equities on pairs trading. Finally, we remark a very similar picture in exchange rates case, however in pairs such as the CAD/USD-ZAR/USD and the AUD/USD-ZAR/USD the number or predictable rules is quite small, sometimes zero, over the majority of subperiods. The reason behind the low performance of technical analysis in these specific exchange rate pairs may be the large amount of transaction costs used, especially those for ZAR/USD currency. In overall, the commodities pairs appear to be more predictable using the technical trading rules than both the equities and exchange rate pairs across all the subperiods. The second best predictable pairs are those of the equities market.

The most important finding of Table 5 lies on the fact that the predictability of technical analysis on pairs trading has different characteristics than on single assets. We show that mainly there is no uniformly monotonic downward trend in the performance of technical trading rules on pairs. As we mention above, this is opposite to the findings of other relevant studies focusing on single commodities, equity indices and foreign exchange markets. The evidence above generally supports the Adaptive Markets Hypothesis (Lo, 2004) instead of the Efficient Market Hypothesis, according to which evolutionary market dynamics create arbitrage opportunities periodically, when more significant rules are generated. Nevertheless, since model predictability is not always synonymous with excess profitability, we investigate the validity of the Adaptive Market Hypothesis in our out-of-sample analysis section in a realistic simulation. Finally, the subperiod analysis exercise also emphasize the powerfulness and flexibility of FDR method in selecting significant rules even in short periods of time compared to previous approaches.

Fig. 1 is a scatterplot of each spread's predictive rules against their evolution across every subperiod. Specifically, it provides information on the decomposition of each of the FDR portfolios of predictive rules into their corresponding families across all spreads through time.

[Figure 1]

We notice a specific pattern in the selection of significant trading rules, which remains the same almost across every subperiod. Interestingly, this pattern involves a large concentration of momentum rules, such as support and resistance, channel breakouts and filter rules, except from mean-reverting ones. In fact, the latter ones consist a smaller percentage of the total number of predictive rules, despite the fact that the best-performing rules of each spread belong in these families most of the times. The Fig.1 also validates the decline of the total number of significantly predictive rules as we reach the last subperiod.

For instance, support and resistance, channel breakouts and filter rules appear to dominate in the excess profitability of pairs during the first subperiod. In the two following periods, the overall picture seems the same, however as we approach towards their end (i.e., 2001 and 2007) the Bollinger bands and commodity channel indices gain more ground in the predictability of the examined pairs, a feature which also appears during the last subperiod (i.e., 2012-2016). On the other hand, we can hardly find any contrarian technical trading rules in the fourth subperiod (2008-2011). Another important finding provided by the figure above, is the very sporadic appearance of moving averages among the significant outperforming rules across all spreads and subperiods.

8. Out-of-sample analysis

8.1. Out-of-sample performance for each pair

So far, we have concentrated our analysis only on in sample (IS) performance which is mostly based on backtesting procedures, which emphasize on the predictive ability of technical trading rules. Empirical studies (see, Brajgowicz and Scaillet, 2012; Harvey et al., 2016; Hsu et al., 2016, among others) employ also an out-of-sample (OOS) analysis to further address the issue of data snooping, as well as to economically evaluate the performance of a portfolio of rules, selected ex ante, likewise institutional investors would have done in practice in an upcoming period. However, OOS forecast has several constraints despite its universal acceptance. In fact, there is no genuine OOS analysis using historical data, since we already know what really happened in the economy (see Harvey and Liu, 2015) and even if this exists, Mclean and Pontiff (2016) demonstrate that OOS return predictability declines dramatically during the post-publication period of a study. Nevertheless, it remains a robust approach to assess the performance persistence, and so controlling for data snooping, of technical trading rules chosen under a multiple hypothesis testing method, such as the FDR, in the OOS period. Despite that are aware it is not always consistent with genuine profitability.

Another important issue with the OOS estimation raised by Harvey and Liu (2015) is the data splitting in in-sample and out-of-sample intervals. This estimation procedure usually comes down to a tradeoff between type I (false discoveries) and type II errors (missed discoveries), which is closely related to the testing power of these periods. In particular, the shorter the in-sample dataset the greater the chance of missing true discoveries (type II error) and vice versa. For instance, a 90-10 split of the data leads to an increase of type I error, while similarly a 50-50 split leads to an increase of II error. Although, multiple hypothesis testing frameworks, such as the FDR help to resolve the above issue, we adopt a 70-30 split for the IS and OOS intervals respectively to achieve a good balance between those two type of errors.

We therefore accommodate an OOS investment performance by exercising the specific split in our dataset, while the FDR procedure is employed during the IS period for each pair in order to select the technical trading rules for evaluation in OOS. Specifically, we construct equally weighted portfolios (allocate \$1 evenly) of significant rules, while accounting for data snooping bias under the FDR method using 70% percent of each subperiod's historical data considered in subsection *7.2*. The last 30% of the remaining data is used for the OOS estimation. This approach, provide us with almost the last year of every subperiod as out-of-sample horizon, while the previous years (no more than four years) constitute the in-sample period. Although we understand that this is still a stringent out-of-sample evaluation, it better matches to what traders do in practice, instead of using a single long-term in-sample horizon, dated back to early 90s, when totally different dynamics existed. The above structure will help as exploit the different dynamics of each subperiod and invest in an out-of-sample horizon accordingly.

Table 6 demonstrates the in-sample performance of the equally-weighted FDR portfolios of significant rules as well as their number, based on the Sharpe ratio criterion and for each corresponding pair.

[Table 6]

The in-sample evidence in Table 6 validates again the ability of the FDR method in selecting a sufficient number of predictive rules across all subperiods and for each pair, except from a number of few cases. The overall picture is similar to the one provided during the subperiod analysis in sub-section *7.2*, in which we generally find no specific trend, upward or downward, in the evolution of the number of predictive rules throughout the years.

What is more important in this table is in-sample performance measured under the Sharpe ratio metric. Consider first the findings for the commodity pairs. The Sharpe ratios are far above 1 in most cases, except from the platinum-palladium pair during the last subperiod. There is also no clear trend or shrinkage in the Sharpe ratios performance, validating the evolutionary dynamics of the Adaptive Market Hypothesis. In support of this evidence, there is also the case of the Brent-crude oil pair, in which the Sharpe ratio criterion is not able to identify any significant rule during the 2008-2011 period. For the equity pairs, the evidence shows a similar performance. The Sharpe ratios range from 0.67 to 2.56, while this metric selects at least one predictive rule under the FDR method for each pair across all subperiods. Finally, for exchange rate pairs the Sharpe ratios range from 0.64 to 3.60, achieving the highest one, JPY/USD-EUR/USD pair, in the first subperiod considered. Nevertheless, the average number of genuine rules selected under the Sharpe ratio measure in every subperiod is the smallest compared to the ones of commodity and equity pairs, while there are two cases, CAD/USD-ZAR/USD and CAD/USD-AUD/USD, in which the Sharpe ratio criterion does not select any predictive rules during the first and last periods respectively.

Moving now to the key evidence for investors, Table 7 reports the out-of-sample performance of the equally-weighted FDR portfolios of significant rules, as well as the best significant one's, based on the Sharpe ratio criterion and for each corresponding pair.

[Table 7]

Concentrate first on the findings for commodity pairs. There is at least one post-sample period for each pair, in which either the FDR portfolio of significant technical trading rules or the best significant rule yield a positive Sharpe ratio. Those ratios span from 0.22 to 1.83 for the FDR portfolio and from 0.56 to 1.79 for the best rule. The Brent-crude oil pair seems the most promising one, yielding constantly positive Sharpe ratios above 1, consistent with both the FDR portfolio and the best rule. Notwithstanding this performance, it appears only during the first two post-sample periods, 1997 and 2001, and then decays to negative or zero Sharpe ratios over the recent periods. On the other hand, we observe commodity pairs, such as the platinum-gold, whose predictive rules achieve positive Sharpe ratios as we move towards the more recent out-of-sample years, reaching the highest Sharpe ratio of 1.83 during 2011 and then falling back to Sharpe ratios of 0.60 on average during 2016. It is also worth mentioning that the FDR portfolio of significant rules generates on average better performance than the best significant rule across all commodity pairs and post-sample periods, even when negative Sharpe ratios discovered. This result highlights the diversification benefits of the FDR method in constructing portfolios of significant rules, while minimizing the downside risk.

Considering the results for equity pairs the overall picture is analogous to commodity pairs, showing no trend and excess profitability in specific time periods. Despite that, the technical trading rules yield a poorer performance with positive Sharpe ratios ranging from 0.22 to 1.48 for FDR portfolios, while Sharpe ratios exceeding 1 only in two cases, the FTSE 100-CAC 40 pair in 1996 and the S&P 500-Russell 2000 in 2007. The poor performance is also justified by numerous negative Sharpe ratios spanning from -0.06 to –2.23 for specific out-of-sample horizons, as well as by pairs, such as the DJIA-Russell 1000, demonstrating negative Sharpe ratios across all the five post-sample years. Interestingly, the highest negative Sharpe ratios concentrate on the earliest post-sample periods. Moreover, it is unclear whether the FDR portfolio of significant rules achieves a better performance compared to the best significant rule across all equity pairs.

In terms of foreign exchange rate pairs' out-of-sample performance, the results seem more encouraging since the significant rules yield positive Sharpe ratios for the majority of pairs, at least during the first three out-of-sample periods. Specifically, cases such as the JPY/USD-EUR/USD pair show a decay in the performance through the years, while others like the EUR/USD-CHF/USD pair reports considerable Sharpe ratios cyclically. There is also the special case of CAD/USD-ZAR/USD, whose Sharpe ratios are being cancelled out by the large transaction costs most of the times. Moreover, the out-of-sample positive Sharpe ratios range from 0.15 to 1.32 for the FDR portfolio of significant rules and from 0.52 to 1.46 for the best significant rule for each pair. However, considering also the negative performance of the FDR portfolio and the best significant rule it is once again uncertain whether one of these two approaches depicts better performance.

In general, the out-of-sample performance of technical trading rules on commodity pairs is on average higher than the almost equal performance of trading rules on equity and foreign exchange pairs. In addition to this, the FDR framework provides an efficient portfolio construction tool mostly in the case of commodity pairs, in which the diversification benefits appear to be strong. Nevertheless, we can also grasp FDR method's benefits on equity and exchange rate pairs when the best significant rule is out of the market, generating zero Sharpe ratios, the FDR portfolio produces even a small amount of profits in many cases. Finally, the Adaptive Market Hypothesis seems more appropriate in explaining the overall out-of-sample performance of technical trading rules on the corresponding pairs in compliance with Table 7. Pair traders can exploit and arbitrage away returns in specific periods of time, as this is also the key feature of statistical arbitrage. On the other hand, these returns tend to diminish in the periods following especially when more traders deploy their strategies alleviating the existing arbitrage opportunities and leading to negative returns. This happens only until the market and business conditions change over time in an evolutionary rate creating new profit opportunities in future periods of time.

8.2. Out-of-sample performance under mean reversion

As John Maynard Keynes mentions "in the long run we are all dead". So, pair traders often try to identify what is the expected holding period for a mean-reversal trade to take advantage of it. For instance, they may not want to trade pairs with long holding period as a way of minimizing their exposure on the market, i.e. large drawdowns. Half-life of mean-reversion usually defines this holding period. However, it is an intrinsic property of the price time series, rather than a trading strategy.

In order to model a mean-reverted process we adopt the Ornstein-Uhlenbeck formula accounting for mean-reversion. In such a way, we can compute the expected time (half-life) of mean-reversion using daily prices via this formula. For instance, let $z(t)$ be the mean-reverting spread, then

$$dz(t) = -\theta(z(t) - \mu)dt + dw \tag{10}$$

where dz is the change of the spread value during dt, dw is some Gaussian noise and μ is the spread's mean. This process defines a stochastic differential equation familiar to a derivatives trader. Thus, the expected value of $z(t)$ after integrating the process and using Ito's lemma is $E(z(t)) = \mu + \exp(-\theta t)(z_0 - \mu)$. The expected value of $z(t)$ now follows an exponential decay to μ at the rate θ, while the half-life of this decay is equal to $\frac{\ln(2)}{\theta}$, if $\theta > 0$[54]. This actually corresponds to the time required for the path above to progress half way toward its long-term expectation, and so the time needed before we can cash in any profit (see Meucci, 2009). Moreover, a higher mean-reversion θ, indicates a stronger cointegration between the time series and vice versa. Now, discretizing the Ornstein-Uhlenbech formula, we can obtain both μ and θ by using a linear OLS regression on historical spreads:

$$z(t+1) - z(t) = -\theta(z(t) - \mu) \tag{11}$$

The new formulation leads to a vector autoregressive model of order one, denoted as VAR(1), which also includes cointegrated dynamics as we have already described in the application of the fractional cointegrated vector autoregressive analysis of Johansen and Nielsen (2012) in *Section 2*.

Technical traders employing contrarian trading rules usually set their lookback period (number of lags of daily values) equal to the half-life as an optimal mean-reverting strategy. Following this approach, we proceed to an out-of-sample analysis measuring the performance of our contrarian families of rules only (RSIs, Bollinger bands and CCIs) for robustness purposes. In particular, we consider the same in-sample

[54] In case where $\theta < 0$ the trajectory follows an exponential explosion instead of mean reversion.

and out-of-sample data sets for every subperiod as in the previous subsection. We then calculate the half-life of each pair using the in-sample data given. In the next step, we exercise the families of mean-reverting rules by setting as a lookback period the half-time of mean-reversion of each corresponding pair. This comes down to a new universe of 400 contrarian technical trading rules. During this backtesting procedure we also conduct the FDR test for multiple hypothesis testing to identify the significant rules for each pair and across all subperiods. Finally, we measure the excess profitability of the significant rules in the following out-of-sample periods by creating equally weighted portfolios of genuine rules for each pair in a similar way as in subsection *8.1*.

We report first the in-sample performance of this approach in the table below. In particular, Table 8 displays the number of significantly predictive contrarian rules created by taking into account the half-life of mean-reversion of each pair, as well as the performance of equally-weighted FDR portfolios of those rules, under the Sharpe ratio criterion and for every subperiod in-sample.

[Table 8]

All parts the three parts, upper, middle and lower, in Table 8 display a weak predictability of the specific statistical arbitrage traders' approach for the commodity, equity and foreign exchange pairs respectively. The FDR method hardly selects a significant number of predictive rules, while for a considerable number of pairs, for all the three markets considered, it does not even select a single rule. For instance, the technical predictability of mean-reverting rules under the half-time criterion for commodity pairs, concentrates mostly in the case of Brent-crude oil spread. In this case, the FDR test chooses at least a single trading rule except from the fourth subperiod, while for the other pairs the predictability is close to zero. In addition, the in-sample Sharpe ratios of predictive rules range from 0.68 to 1.96, which are substantially below the relevant ones in the previous analysis using the whole universe of technical trading rules.

The in-sample performance of "half-time" contrarian rules on equity spreads reveals a similar picture. The number of predictive rules remains in single digits for every pair and across all subperiods, while their corresponding Sharpe ratios span from 0.47 to 1.27. In the case of currency pairs, the picture seems slightly better with positive Sharpe ratios ranging from 0.47 to 1.98, while the FDR method manages to select more predictive rules for more pairs in almost every period. Those findings are far from encouraging for institutional traders in case they use the rule of thumb of a 50% haircut of in-sample Sharpe ratios in predicting the out-of-sample ones. Moreover, we observe that technical predictability is very weak during the second and fifth subperiods for both equity and exchange rate pairs, as we can only find one predictable pair for each market.

We report the out-of-sample findings of the half-time of mean-reversion strategy in Table 9. Again, the table demonstrates the equally-weighted FDR portfolios' performance of significant rules as well as the best significant one, based on the Sharpe ratio criterion and for each corresponding pair, for comparison purposes. We also report the median of half-time of mean-reversion of each pair computed in the in-sample period as an extra column.

[Table 9]

For commodity pairs, we obtain the only promising results by trading the Brent-crude oil spread, whose returns also decay to zero after reaching the third post-sample year. For all other spreads the trading simulation reveals almost zero or negative returns over the rest out-of-sample periods. The positive Sharpe ratios range only from 0.15 to 1.35, while the FDR portfolio of significant rules shows almost equal performance with the best significant rule. There are also cases, such as the platinum-gold pair, in which the significant rules produce no signals and stay out-of-the market across all post-sample periods. The medians of half-time of mean reversion span from 10 to 201.

Regarding the equity spreads, the overall message we get is the same as above with medians of half-time ranging however at higher levels (i.e. 167-416). Pairs trading via technical analysis can generate positive Sharpe ratios, even small ones, only in half of the cases and for specific out-of-sample periods. Trading the S&P 500-Russell 2000 and the Russell 1000-Russell 2000 pairs yields the most encouraging Sharpe ratios, namely 1.31 and 0.89, for the 1996 and 2011 post-sample years, respectively. For all the other pairs, out-of-sample technical trading generates zero or negative returns. Again, the FDR portfolio's performance is equivalent with that of best rule since for most of the cases the FDR portfolio consists only with the best strategy.

The lower part of Table 9 indicates a shrinkage in any out-of-sample performance over the years on pairs yielding positive Sharpe ratios (EUR/USD-CHF/USD, AUD/USD-ZAR/USD), whose levels range from 0.92 to 1.14. Only in the case of

CAD/USD-AUD/USD spread the FDR portfolio and the best significant rule produce a positive Sharpe ratio of 0.73 in 2007, while technical profitability for the rest of the pairs shows zero or negative performance. Moreover, the average median of half-time for exchange rate pairs is higher than the relevant ones of commodity and equity pairs.

We anticipated the above performance for a couple of reasons. First of all, in-sample cointegration between assets it is not always synonymous with out-of-sample results. According also to Atilio Meucci (2010), eigen series, which are relative to smallest eigen values and they are responsible for making quicker profits, are regularly least robust out-of-sample, while the most mean-reverting series leading to lesser potential returns, are usually neutralized by the transaction costs.

8.3. Out-of-sample performance for portfolios of pairs

The final out-of-sample simulation involves the performance examination of integrated market portfolios of technical trading rules across all post-sample periods, as a pairs trader would have done in practice. Usually, pairs' traders expose themselves not only at a single market, but they are constantly searching for arbitrage opportunities across several markets and assets. We construct four portfolios (one for commodity pairs, one for equity pairs, one for foreign exchange pairs and a global one), using the all technical trading rules and so the pairs selected as significant in terms of Sharpe ratio for every in-sample period data. We apply this portfolio management procedure not only to all significant FDR rules, but also to the best significant rule for each pair, similar to Hsu et al., (2016), as we have presented in subsection *8.1* for comparison purposes. Furthermore, we also compare their performance with identical integrated portfolios constructed under the half-time, mean-reverting significant rules described this time in subsection *8.2*. We assume no optimization for the integrated portfolios by assigning equal weight of our total wealth to every single rule identified as significant across all pairs, similar to constructing the FDR portfolios of significant rules in previous subsections. For instance, in the first post-sample period we invest $1 of total wealth and distribute it evenly across all pairs in a particular portfolio (e.g. commodities portfolio). This comes down to a further equal allocation of a specific pair's wealth proportion across the signals of its corresponding significant FDR rules. In case where only one best significant rule considered, we allocate it the whole wealth proportion.

Therefore, the above portfolio construction approach is equivalent to simply averaging the excess return of its constituent assets.

We present the out-of-sample performance of the four market portfolios across every out-of-sample subperiod and for every alternative portfolio composition employed, in Table 10.

[Table 10]

According to the first panel, using all significant rules of each pair to build the market representing portfolios, commodities achieve the best performance across all out-of-sample years, yielding mostly Sharpe ratios of up to 1.73 and CAGR of 6.44% during 2001, while it shows a positive performance in three out of five post-sample periods. On the other hand, the equity market portfolio does much worse yielding positive Sharpe ratios of only 0.36 and 0.25 during 2001 and 2007, with their corresponding CAGRs are considerable low, i.e. 0.24 % and 0.01%, while for the rest of post-sample periods achieves negative returns. The foreign exchange market portfolio shows a slightly better performance, but only in specific periods, with Sharpe ratios being close to 1 (i.e. 0.98 and 0.83) in 2001 and 2011, even though their CAGRs are quite low to constitute attractive investments. In terms of global portfolio's performance now, the diversification benefits we obtain are quite observable. The global portfolio actually retains the positive performance attributed to our best commodities' portfolio during the same years, with Sharpe ratios remaining high and ranging from 0.83 to 1.82 but this time the CAGRs have fallen to levels ranging from 0.81 to 2.30%.

Employing now only the best significant rules of each pair to construct our four portfolios, the overall picture of the second panel doesn't seem very encouraging in terms of performance. Although, all market portfolios are able to yield a positive performance from time to time, this is considerably low to act in support of excess profitability. Only during the first out-of-sample period the commodities' portfolio achieves an outstanding Sharpe ratio of 1.56 and a CAGR of 8.37%, while the foreign exchange market portfolio produces a Sharpe ratio of 1.17 and a CAGR of 0.43%. The global portfolio also depicts this performance yielding a Sharpe ratio of 1.43 and a CAGR of 2.57% in 1996, but this performance decays to zero or turns to negative returns for the rest of the periods. Comparing the second panel with the first panel results, we once again validate the power of FDR method in selecting a sufficient number of significant rules, while diversifying against model uncertainty and achieving a better overall performance over time, instead of only using the best significant rule.

Considering the analogous market portfolios constructed using FDR-significant rules under the half-time of mean-reversion criterion, we find that commodities' portfolio yields a positive performance at least during the first three periods. Especially in 1996 and 2001 the corresponding Sharpe ratios (0.86 and 1.05) and CAGRs (i.e. 4.32% and 7.80%) indicate some good arbitrage opportunities, which then decay over the next periods. In the case of equities and foreign exchange currencies, the out-of-sample performance is again not very encouraging for most of the post-sample years. Although, we can find some periods of positive Sharpe ratios and CAGRs, those are not statistically significant different from zero in most cases. The highest Sharpe ratio and CAGR (i.e. 1.31 and 0.18%) is achieved by the equities portfolio in 1996. Moreover, the global portfolio is mainly driven by the commodities' portfolio showing a similar performance with Sharpe ratios ranging from 0.32 to 1.05 and CAGRs from 0.03% to 7.80%.

In general, the profitability of pairs trading using technical analysis has shrunk over the years as justified from the all three panels of the table, especially for equities and foreign exchange rates markets. Consistent with the previous literature (Gatev et al., 2006; Marshal et al., 2008; Neely et al 2009; Bajgrowicz and Scaillet, 2012), most the above findings emphasize the decomposition of technical analysis' excess profitability over time, presumably due to the increased hedge fund and trading activity. Only in the case of commodities market, there might be some windows of excess profitability in more recent periods but its level is still questionable. Notwithstanding this profitability sets commodities' market more lucrative for a pairs' trader as revealed by the post-sample performance of our three different commodity portfolios over time, and compared to the relevant performance for rest of markets reported.

9. Conclusion

We investigate a hedge fund trading strategy based on the assumption of cointegrated prices in an informational efficient market, widely known as pairs trading, while we employ technical analysis to predict the prices movements of formatted spreads. The long-debated issue of whether and why technical analysis is still profitable is explored for the first time under statistical arbitrage conditions.

We conduct a large-scale research of the predictability and excess profitability of technical trading rules across a large set of 'famous' commodity, equity and currency pairs, being actively traded by statistical arbitrageurs, in long sample periods. Our analysis involves a quite large number of technical trading rules split in generic momentum and contrarian classes. We also adopt recently developed multiple hypothesis testing methods, totally adequate in such applications as they allow us to create statistical inferences generating new, adjusted thresholds preserving against data mining issues.

Our findings reveal that technical trading has predictive power for most of the spreads considered, especially in terms of yielding significant Sharpe ratios. Commodity pairs are in general more predictable with technical analysis compare to equity and currency spreads. Moreover, using technical analysis to trade our suitable formed pairs exhibits significant returns, which are robust to even conservative one-way transaction costs. In addition to this, we contradict previous explanations for the pairs trading profits generated only by rules based on contrarian principles. A realistic out-of-sample analysis made across five different subperiods reveals that although excess profitability of technical analysis has shrunk over time, some commodity pairs display an encouraging performance during recent periods.

One possible reason for the declined profitability of pairs trading in recent years may be the increased hedge fund activity squeezing out potential returns. However, abnormal returns achieved using technical trading rules on certain spreads and periods may be a compensation to arbitrageurs for a temporarily no-fully-rational behaviour. Those findings seem to favour the Adaptive Market Hypothesis (Lo, 2004), which assumes that investors severely exploit arbitrage opportunities and therefore make them diminish as soon as the overall market learns and milks the profitable trading strategies until new opportunities arise again in evolutionary cycles.

In terms of further research, the manipulation-proof versions of performance metrics mentioned in chapter 3 could also employed in technical analysis performance on spread trading. Some possible candidates could be the measurement method developed by Ingersoll et al, (2008) and the Morningstar Risk-Adjusted Rating in 2002. Moreover, information derived by the fractionally cointegrating relationship of the examined pairs could be used further in the technical trading rules exercise. For example, we could calculate and take into account the optimal hedge ratio while forming a stationary spread portfolio based on the eigenvectors' information or by just using the OLS coefficient between the components of a pair.

Appendix A. Details of technical trading rules parameters

In this section, we describe in precise detail the universe of our total technical trading rules, following the previous studies of Sullivan et al., (1999) and Hsu et al., (2016).

A.1. Filter rules

The filter rule allows the initiation of a pairs' trader position only in response to major price trends. Therefore, an investor *buys* a pair if its price *increases* by a fixed percentage from a previous *low*, and he *sells* if the price decreases by a fixed percentage from a previous *high*. We assume three different filter rule variations as described below, while we set the previous low (high) as the most recent pair value between the daily closing prices of two assets that is less (greater) than the *n* previous daily pair value, for a given value of *n.*

F1*: If the daily pair value between the daily closing spot prices of two assets increases (decreases) by at least x percent from its previous low (high) and remains so for d days, then go long (short) the pair.*

The second variant allows also for neutral positions in lieu of always being either long or short according to the basic filter rule.

F2: If the daily pair value increases by at least x percent from its previous low and remains so for d(x) days, then go long the pair until its daily value decreases at least y percent from its subsequent high and remains so for d(y) days at which time liquidate the long position. If the daily pair value decreases by at least x percent from its previous high and remains so for d(x) days, then go short the pair until its daily value increases at least y percent from its subsequent low and remains so for d(y) days at which time liquidate the short position.

The third variation assumes a position is held for a fixed number of periods ignoring all other signals.

F3: *If the daily pair value increases (decreases) by at least x percent from its previous low (high) and remains so for d days, then go long (short) the pair for c days and then neutralize the position.*

$n = 1, 2, 5, 10, 15, 20$ [6 values]

$x = 0.05, 0.1, 0.5, 1.0, 5.0, 10.0, 20.0$ in % [7 values]

$y = 0.05, 0.1, 0.5, 1.0, 5.0, 10.0, 20.0$ in % [7 values]. Noting that $y < x$, there are 21 $x - y$ combinations.

$d = 0, 1, 2, 5$ [4 values]. Noting that $d(y) < d(x)$, there are 6 $d(x) - d(y)$, combinations.

$c = 1, 5, 10, 15, 20, 25$ [6 values]

The total number of rules is: $(x * d * n) + ((x - y) * (d(x) - d(y)\ combinations) * n) + (x * d * c * n) = 168 + 756 + 1008 = 1932.$

A.2. Moving averages

One of the most popular and actively traded class of technical trading rules across several markets, which assumes a crossover between the pair value and a moving average of a given length or between short-long moving averages of a different length in order to generate a trade. These upside (downside) penetrations of a moving average help an investor to discover the emergence of new trends and maintain his position as long as the crossover remains. We provide four different variations as follows:

MA1: If the daily pair value moves at least x percent above (below) the moving average, i.e. MA(n), and remains so for d days then go long (short) the pair until its closing value moves at least x percent below (above) MA(n) and remains so for d days, ay which time go short (long) the pair.

Similar to the *filter rules* a variant of a *moving average rule* assuming we hold a position for a fixed number of periods, ignoring all other signals would be as described below:

MA2: If the daily pair value moves at least x percent above (below) the MA(n), and remains so for d days then go long (short) the pair for c days and then neutralize the position.

A short/long double moving average as well as a similar one including a fixed number holding period are described in the next two paragraphs as:

MA3: If the short moving average, i.e. MA(m), moves at least x percent above (below) the long moving average, i.e. MA(n), and remains so for d days then go long (short) the pair until MA(m) moves at least x percent below (above) MA(n) and remains so for d days, at which time go short (long) the pair.

MA4: *If the MA(m) moves at least x percent above (below) the MA(n) and remains so for d days then go long (short) the pair for c days and then neutralize the position.*

$n = 2, 5, 10, 15, 20, 25, 50, 100, 150, 200, 250$ [11 values]

$m = 2, 5, 10, 15, 20, 25, 50, 100, 150, 200$ [10 values]. Noting that $m < n$, there are 55

$m - n$ combinations.

$x = 0, 0.05, 0.1, 0.5, 1.0, 5.0$ in % [6 values].

$d = 0, 2, 3, 4, 5$ [5 values].

$c = 5, 10, 25$ [3 values]

The total number of rules is: $(n * x * d) + (n * x * d * c) + \big((m - n\ combinations) * x * d\big) + ((m - n\ combinations) * x * d * c) = 330 + 990 + 1650 + 4950 = 7920.$

A.3. Support and resistance

Likewise filter trading rules, support and resistance rules try to discover major price movements beyond certain levels, which are difficult to been breached, rather than a most recent high or low. The intuition behind this rule is that usually investors think that sooner or later the movement of the pair's price will tend to stop and return to a certain level. However, if the price breaks through a certain resistance or support level by a certain amount, it is more likely to continue moving to the same direction until it finds new such levels. In this way, a long or short signal is generated in terms of rule's construction. Again we pre-define the support and resistance levels as the minimum and maximum closing value of a pair over the previous *n* closing values respectively.

SR1: *If the daily pair value rises above (below) by at least x percent from the local maximum (minimum) over the n previous pair values and remains so for d days, then go long (short) the pair.*

In addition to the above, we impose a holding-period filter:

SR2: *If the daily pair value rises above (below) by at least x percent from the local maximum (minimum) over the n previous pair values and remains so for d days, then go long (short) the pair for c days and then neutralize the position.*

$n = 2, 5, 10, 15, 20, 25, 50, 100, 150, 200, 250$ [11 values].

$x = 0.05, 0.1, 0.5, 1.0, 2.5, 5.0, 10.0$ in % [7 values].

$d = 0, 1, 2, 3, 4, 5$ [6 values].

$c = 1, 5, 10, 25$ [4 values].

The total number of rules is: $(n * x * d) + (n * x * d * c) = 462 + 1848 = 2310.$

A.4. Channel breakouts

Similar to having a time-varying support and resistance rule, a trading channel occurs when the highest value of a pair over the previous prespecified days is within a fixed percentage (*b%*) of the lowest value over the previous prespecified days. The graphical representation of a price channel is equal to a pair of parallel trend lines drifting together within a certain width. As soon as one of these trend lines is "broken", a buy or sell signal is generated. Thus, an investor goes long (short) when the price moves above (below) the channel. The above time-varying support and resistance levels represent the lower and upper bounds of the channel, while their difference from the high and low respectively, over the previous prespecified days, doesn't exceed the *b%.*

CB1: If a b% trading channel occurs and if the daily pair value rises above (below) by at least x percent from the upper (lower) bound over the n previous days and remains so for d days, then go long (short) the pair.

We consider also a holding period for each position triggered:

CB2: If a b% trading channel occurs and if the daily pair value rises above (below) by at least x percent from the upper (lower) bound over the n previous days and remains so for d days, then go long (short) the pair for c days and then neutralize the position.

$n = 5, 10, 15, 20, 25, 50, 100, 150, 200, 250$ [10 values].

$b = 0.1, 0.5, 1.0, 5.0, 10.0$ in % [5 values].

$x = 0.05, 0.1, 0.5, 1.0, 5.0$ in% [5 values]. Noting that $x < b$, there are 15 $(x - b)$ combinations.

$d = 0, 1, 2$ [3 values].

$c = 1, 5, 10, 25$ [4 values].

The total number of rules is: $(n * (x - b)\ combinations * d) + (n * (x - b)\ combinations * d * c) = 450 + 1800 = 2250.$

A.5. Relative Strength Indicator rules

Introduced by Levy (1967), relative strength rules (RSI) belong to the general family of 'overbought/oversold' indicators, or commonly called as oscillators, from which we also pool the rest of our proposed 'reversal' trading rules (i.e. Bollinger bands, CCIs). As already mentioned, RSIs attempt to reveal upcoming price corrections towards the opposite direction of extreme upward or downward movements, in which a short or long signal is executed accordingly. The generic formula of RSI is

$$RSI_t(n) = 100 - \frac{100}{1+\frac{U_t(n)}{D_t(n)}} = 100\left[\frac{U_t(n)}{U_t(n)+D_t(n)}\right] \qquad \text{(A.1)}$$

where $U_t(n)$ and $D_t(n)$ represent the cumulated upward and downward trend, calculated as the sum of first differences between monotonically increasing or decreasing closing prices in absolute terms over the previous *n* days. In other words, the $U_t(n)$ denotes the total gains of a potential upward movement, while the $D_t(n)$ denotes the total losses of a potential downward movement over the previous *n* days. Thus, normalized to the scale of 100, the RSI estimates the dominance of an upward relative to the dominance of a downward trend. In its simplest version, an RSI of a value 70 characterize a specific pair as overbought, while a value of 30 rates the pair as overbought[55]. Except from this naïve RSI variant we also consider two more modifications as have been introduced by Hsu et al, (2016):

RSI1: If $RSI_t(n)$ rises above 70 then go short the pair. Alternatively, if $RSI_t(n)$ falls below 30 then go short the pair.

RSI2: If $RSI_t(n)$ rises above $50 + k$ for at least d days and then subsequently falls below $50 + k$, go short the pair. Alternatively, if $RSI_t(n)$ falls below $50 - k$ for at least d days and then subsequently rises above $50 - k$, go long the pair.

RSI3: If $RSI_t(n)$ rises above $50 + k$ for at least d days and then subsequently falls below $50 + k$,go short the pair for c days and then neutralize the position. Alternatively, if $RSI_t(n)$ falls below $50 - k$ for at least d days and then subsequently rises above $50 - k$, go long the pair for c days and then neutralize the position.

$n = 5, 10, 15, 20, 25, 50, 100, 150, 200, 250$ [10 values].

$k = 10, 15, 20, 25$ [4 values].

$d = 1, 2, 5$ [3 values].

$c = 1, 5, 10, 20, 25$ [5 values].

The total number of rules is: $n + (n * k * d) + (n * k * d * c) = 10 + 120 + 600 = 730$.

[55] This is the most naïve RSI rule and it can be found in several trading websites and platforms.

A.6. Bollinger Bands

Developed by the famous technical trader John Bollinger in the 1980s, Bollinger bands are volatility indicators trying to take advantage of unjustifiably high or low prices and their imminent corrections. To achieve this, they consider upper and lower bands of a pair's price in terms of standard deviations from a moving average over the previous prespecified days. Considering a pair's moving average as well as its moving standard deviation over the *n* previous days we have

$$MA_t(n) = \frac{1}{n}\sum_{i=1}^{n} P_{t-1+i}\,,\ \sigma_t(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{P_{t-1+i}-P_{t-1}}{P_{t-1}})^2} \tag{A.2}$$

We define the upper/lower bands of a given width z as $MA_t(n) \pm z * \sigma_t(n)$, which are the specific moving average plus/minus z times the specific moving standard deviation.

Almost always the price of a pair trades between the two bands except from cases in which extreme conditions occur. Thus, any breakout above or below the bands is a major event. Many investors believe the closer the prices move to the upper band, the more overbought the market and vice versa. This can lead to a pullback of prices captured by such an 'reversal' trading rule. We consider rules based on prices leaving the bands in order to trigger a position, and possibly then crossing of the moving average to neutralize the position.

BB1: If the daily closing spot price (spread) moves above the upper band of a given width z and remains so for d days go short the pair until it moves back to the moving average, at which time neutralize the short position. If the daily pair value moves below the lower band of a given width z and remains so for d days, go long the spread until it moves back to the moving average, at which time neutralize the position.

BB2: If the daily pair value moves above the upper band of a given width z and remains so for d days, go short the pair for c days and then neutralize the position. If the daily pair value moves below the lower band of a given width z and remains so for d days, go long the pair for c days and then neutralize the position.

$n = 5, 10, 15, 20, 25, 50, 100, 150, 200, 250$ [10 values].

$z = 0.5, 1, 1.5, 2, 2.5, 3$ [6 values].

$d = 0, 1, 2, 3, 4\ 5$ [6 values].

$c = 1, 5, 10, 20, 25$ [5 values].

The total number of rules is: $(n * z * d) + (n * z * d * c) = 360 + 1800 = 2160.$

A.7. Commodity Channel Index rule

Introduced by Donald Lambert in 1980, the Commodity Channel Index (CCI) also belongs to the family of oscillators attempting to capture cyclical trends and so to determine 'overbought/oversold' levels. The CCI was initially developed for discovering such levels in the commodities market, but its prominent applicability soon attracted technical traders to also use it in equities and currencies markets. Likewise Bollinger bands, the CCI not only uses extreme upper and lower bands to trigger long/short signals, but also takes into account the volatility of a pair. The CCI is defined as

$$CCI_t(n) = \frac{P_t - MA_t(n)}{0.015 * \sigma_t(n)} \qquad \text{(A.3)}$$

where P_t is the price of a pair at a specific time *t*, while $MA_t(n)$ and $\sigma_t(n)$ denote the pair's moving average and standard deviation over the previous *n* days, calculated as in the case of Bollinger bands. Thus, the CCI measures the current price level relative to an average price level over a specific period of time, while it is fairly high when prices are far above the moving average and vice versa. The constant 0.015 just ensures that the majority of CCI values will lie in between -100 and +100, which represent the upper and lower bounds of this trading rule.

As a 'reversal' indicator, CCI searches over overbought (i.e. above +100) or oversold conditions (i.e. below -100) foretelling a mean reversion. Similarly, bullish and bearish divergences can be used to detect early momentum shifts and anticipate trend reversals. We employ two simple 'reversal' variants of CCI as well as a CCI discovering Bullish/Bearish divergence breakouts.

CCI1: If the $CCI_t(n)$ remains above $(+100 + k)$ for at least d days and the subsequently moves below $+100$, go short the pair. If $CCI_t(n)$ remains below $(-100 - k)$ for at least d days and then subsequently moves above -100 , go long the pair.

Assuming a holding period *c* we have:

CCI2: If the $CCI_t(n)$ remains above $(+100 + k)$ for at least d days and the subsequently moves below $+100$, go short the pair for c days and then neutralize the position. If $CCI_t(n)$ remains below $(-100 - k)$ for at least d days and then *subsequently moves above* -100 *, go long the pair for c days and then neutralize the position.*

We finally consider a special case of a CCI and divergence breakout. Divergences can foresee a potential trend reversal point as they usually reflect a change in momentum. We examine two types of divergence, bullish and bearish. A bullish divergence appears when the pair performs a lower low (i.e. support break) and the CCI shapes a higher low, over the previous n days, indicating a less downside momentum. A bearish divergence appears when the pair performs a higher high (i.e. resistance break) and the CCI forms a lower high over the previous *n* days, indicating a less upside momentum. In other words, looking for a breach in support and resistance levels of a pairs' price, while in the meantime searching for a direction change of the CCI. In particular, we have:

CC3: *If the daily pair value moves below by at least x percent from the local minimum over the n previous pair values and CCI remains below*$(-100-k)$*, while its local minimum over the moves above its previous value n previous days, then go long the pair. If the daily pair value rises above by at least x percent from the local maximum over the n previous pair values, and the CCI remains above* $(+100+k)$*, while its local maximum moves below its previous value, over the n previous days then go short the pair.*

$n = 5, 10, 15, 20, 25, 50, 100, 150, 200, 250$ [10 values].

$k = 0, 50, 100$ [3 values].

$d = 1, 2, 3, 4\ 5$ [5 values].

$c = 1, 5, 10, 20, 25$ [5 values].

$x = 0.05, 0.1, 0.5, 1.0, 2.5, 5.0, 10.0$ in% [7 values].

The total number of rules is: $(n * k * d) + (n * k * d * c) + (n * x * k) = 150 + 750 + 210 = 1110.$

Appendix B. FDR implementation

We employ the Sullivan et al., (1999) approach as well as the stationary bootstrap of Politis and Romano (1994) to obtain the individual p-values for each trading rule. We also follow the "point estimates" procedure of Storey et al., (2004) for controlling the *FDR* under weak dependence conditions. The algorithm for implementation of the $FDR^{+/-}$ method then is as follows:

1. Calculate the return matrix V**,** in which each column V_{jT}represents the excess return daily series yielded by each of l technical trading rules for the specific time period of each days T examined.

2. Compute the vector of the performance measures $\Phi = (\varphi_1, \varphi_2, \dots, \varphi_j)$ of all trading rules, based on in each V_{jT} of the return matrix V**.**

3. Use the stationary bootstrap method of Politis and Romano (1994) to resample the returns V and create $b = 1, \dots, B$ bootstrap realizations V_{jb} for each trading rules' return series, while employing an average block size ${}^{1}/_{q}$.

4. For each bootstrap realization b calculate the vector of performance metrics $\Phi_b = (\varphi_{1b}, \varphi_{2b}, \dots, \varphi_{jb})$ based on every V_{jb}, where $j = 1, \dots, l$ and $b = 1, \dots, B$.

5. Obtain the p-values of each trading rule, p_j, by comparing the absolute value of each performance metric $T^{1/2}|\varphi_j|$ with the absolute value of its corresponding quantiles of $T^{1/2}|\varphi_{jb} - \varphi_j|$, for $b = 1, \dots, B$[56].

6. Plot the histogram of the total p-values in order to set the λ parameter equal to the level above which the p-values become fairly flat, representing the region of null p-values.

7. Compute the estimator of the proportion of rules with no abnormal performance as $\widehat{\pi_0} = \frac{\#\{p_j > \lambda;\ j=1,\dots,l\}}{l(1-\lambda)}$

[56] The initial multiple hypothesis setup is based on absolute values since we conduct the hypothesis testing in a two-tailed framework in order to estimate the proportion of rules with no abnormal performance $\widehat{\pi_0}$ based on the total p-values.

8. Focus on the right tail and short the p-values of trading rules showing positive performance in an ascending order.

9. Compute the FDR^{+} of the rule having the smallest p-value as $\widehat{FDR}^{+}(\gamma) = \frac{1/2\widehat{\pi_0}l\gamma}{\#\{p_j \leq \gamma, \varphi_j > 0;\ j=1,\ldots,l\}}$, then add the next rule corresponding to second smallest p-value and
re-compute the FDR^{+}.

10. Repeat the above process until reaching the desired FDR^{+} target. The trading rules added up to this level represent the significant ones, while the cutoff, γ, is the corresponding p-value of the last trading rule added.

During our empirical simulations, we set the stationary bootstrap parameters as $B = 1000$ and the average block length equal to 0.1 (i.e. $q = 10$).

List of tables

Table 1. Descriptive statistics and statistical behavior on spreads' daily spot prices and returns

Spreads	Mean(%)	Std. dev.	1st autoc.	Fract. Coint.
Commodities				
Brent-WTI crude oil	0.0031	0.0139	0.00***	0.023**
Platinum-Gold	-0.0005	0.0121	0.93	0.004***
Platinum-Palladium	-0.0292	0.0172	0.07*	0.049**
Corn-Ethanol	-0.0921	0.0191	0.00***	0.008***
Equities				
FTSE100-CAC 40	-0.0019	0.0076	0.28	0.000***
Euro Stoxx 50-DAX	0.0116	0.0058	0.00***	0.011**
FTSE100-FTSE250	0.0121	0.0067	0.11	0.028**
DJIA-Russell 1000	-0.0003	0.0033	0.00***	0.000***
S&P500-Russell 2000	-0.0086	0.0066	0.11	0.001***
Russell 1000-Russell 2000	0.0076	0.0062	0.04**	0.032**
Exchange rates				
EUR-CHF	-0.0076	0.0041	0.00***	0.031**
CAD-AUD	0.0002	0.0064	0.00***	0.531
EUR-JPY	-0.0003	0.0064	0.06*	0.025**
AUD-ZAR	0.0089	0.0089	0.18	0.938
CAD-ZAR	0.0205	0.0089	0.14	0.088*

We present descriptive statistics of daily returns on holding spreads based on commodities, equities and foreign currencies against the U.S. dollar, as well as the *p*-values for cointegration ranking between the spot prices of the underlying legs of each spread, based on the FCVAR analysis of Johansen and Nielsen (2012). * denotes a rejection of the null hypothesis at the 10% level, ** denotes rejection at the 5% level and *** denotes rejection at 1% level of significance.

Table 2. The predictive ability and excess profitability of technical trading rules

Spreads	A. Mean excess return			B. Sharpe ratio		
	# predictive rules	Highest return (%) (p-values)	Best rule	# predictive rules	Highest ratio (p-values)	Best rule
Commodities						
Brent-WTI crude oil	442	17.6 (0.00)***	BB1	563	1.20 (0.00)***	BB2
Platinum-Gold	1	2.03 (0.03) **	RSI2	12	0.61 (0.00)***	RSI2
Platinum-Palladium	0	3.70 (0.20)	F2	10	0.56 (0.00)***	BB2
Corn-Ethanol	12	8.57 (0.00)***	SR2	6	0.45 (0.03)**	RSI2
Equities						
FTSE100-CAC 40	0	1.56 (0.11)	BB1	21	0.31 (0.07)*	CCI3
Euro Stoxx 50-DAX	0	0.33 (0.23)	BB2	9	0.45 (0.00)***	CB2
FTSE100-FTSE250	0	0.66 (0.17)	SR2	0	0.25 (0.38)	CCI3
DJIA-Russell 1000	0	0.41 (0.26)	MA1	3	0.31 (0.09)*	CCI3
S&P500-Russell 2000	1	1.89 (0.00)***	BB2	9	0.54(0.00)***	BB2
Russell 1000-Russell 2000	0	1.23 (0.21)	BB2	10	0.39 (0.04)**	CCI2
Exchange rates						
EUR-CHF	0	0.84 (0.18)	SR2	9	0.34 (0.06)*	MA1
CAD-AUD	0	0.65 (0.24)	BB2	4	0.49 (0.00)***	SR2
EUR-JPY	1	1.41 (0.09) *	SR2	16	0.34 (0.08)*	CB1
AUD-ZAR	0	0.55 (0.28)	BB2	0	0.24 (0.70)	CCI1
CAD-ZAR	0	0.56 (0.30)	BB2	12	0.50 (0.00)***	F3

We impose transaction costs in returns and examine the performance of total 18,412 technical rules over the full sample period. We implement the FDR test to select technical rules providing significantly positive performance. We considered mean excess return and Sharpe ratio as two performance metrics. "#predictive rules" denotes the number of technical rules that provide significantly positive mean excess returns and Sharpe ratios, while controlling the FDR at 10% of false rejections. "Highest return/ratio" denotes the best rule's mean excess return and Sharpe ratios with p-values in parenthesis, while the specific best rules are reported in the "Best rule" section. All mean excess returns and Sharpe ratios are annualized, "***" denotes statistical significance at the 1% level.

Table 3. Portfolio decomposition into families of rules

Spreads	RSI	Filter	MA	SR	Ch.Br.	BB	CCI
Commodities							
Brent-WTI crude oil	1.42 (0.90)	- (-)	- (-)	- (-)	- (-)	92.3 (97.0)	6.21 (2.03)
Platinum-Gold	8.30 (100)	- (-)	- (-)	41.6 (0.00)	- (-)	25.0 (0.00)	25.0 (0.00)
Platinum-Palladium	- (-)	- (-)	- (-)	- (-)	80.0 (0.00)	10.0 (0.00)	10.0 (0.00)
Corn-Ethanol	16.6 (8.30)	- (-)	- (-)	0.00 (66.6)	0.00 (16.6)	0.00 (8.33)	83.3 (0.00)
Equities							
FTSE100-CAC 40	4.76 (0.00)	- (-)	- (-)	- (-)	57.1 (0.00)	- (-)	38.1 (0.00)
Euro Stoxx 50-DAX	- (-)	- (-)	11.1 (0.00)	11.1 (0.00)	66.6 (0.00)	11.1 (100)	- (-)
FTSE100-FTSE250	- (-)	- (-)	- (-)	- (-)	- (-)	- (-)	100 (0.00)
DJIA-Russell 1000	33.3 (0.00)	- (-)	- (-)	- (-)	- (-)	33.3 (0.00)	33.3 (0.00)
S&P500-Russell 2000	11.1 (0.00)	- (-)	- (-)	- (-)	22.2 (0.00)	33.3 (100)	33.3 (0.00)
Russell 1000-Russell 2000	- (-)	- (-)	- (-)	- (-)	- (-)	9.09 (0.00)	90.9 (0.00)
Exchange rates							
EUR-CHF	- (-)	- (-)	11.1 (0.00)	88.8 (0.00)	- (-)	- (-)	- (-)
CAD-AUD	25.0 (0.00)	- (-)	0.00 (0.00)	75.0 (0.00)	- (-)	- (-)	- (-)
EUR-JPY	- (-)	- (-)	6.25 (0.00)	62.5 (100)	31.25 (0.00)	- (-)	- (-)
AUD-ZAR	- (-)	- (-)	- (-)	- (-)	- (-)	- (-)	100 (0.00)
CAD-ZAR	8.33 (0.00)	50.0 (0.00)	8.33 (0.00)	25.0 (0.00)	- (-)	8.33 (0.00)	- (-)

This table reports the average percentage of rules belonging to each family of rules examined, according to the 10%-FDR portfolio, for each pair and for the full sample. We consider the Sharpe ratio as performance metric, while the corresponding results using the mean excess return are reported in parenthesis. Our seven families of rules include: RSI: relative strength indicators, Filter: filter rules, MA: moving averages, SR: support and resistance rules, Ch.Br: channel breakouts, BB: Bolinger bands and CCI: commodity channel indices.

Table 4. Break-even transaction costs for predictive technical rules

Spreads		A. Mean excess return		B. Sharpe ratio	
	Cost (bps)	Break-even cost (bps)	# trades	Break-even cost (bps)	# trades
Commodities					
Brent-WTI crude oil	6.6	26	2432	77	434
Platinum-Gold	6.6	25	291	25	291
Platinum-Palladium	6.6	-	-	69	75
Corn-Ethanol	6.6	24	487	552	3
Equities					
FTSE100-CAC 40	10	-	-	121	3
Euro Stoxx 50-DAX	10	-	-	29	8
FTSE100-FTSE250	10	-	-	-	-
DJIA-Russell 1000	4.0	-	-	34	4
S&P500-Russell 2000	4.0	13	526	27	50
Russell 1000-Russell 2000	4.0	-	-	42	5
Exchange rates					
EUR-CHF	6.1	-	-	69	5
CAD-AUD	8.0	-	-	28	6
EUR-JPY	3.4	7.7	725	14	3
AUD-ZAR	22	-	-	-	-
CAD-ZAR	21	-	-	145	8

We report highest one-way break-even transaction costs (in basis points) that will reduce the performance metrics of the most predictive rules to zero. Mean excess return and Sharpe ratio are considered as performance metrics. “#trades” denotes the number of trades triggered by each trading rule over the sample period. “-“ denotes that for given the pair and performance metric, it does not exist any significantly profitable trading rule.

Table 5. The number of technical rules with significantly positive Sharpe ratios in five subsample periods

Subsample	1991-1996	1997-2001	2002-2007	2008-2011	2012-2016
Commodities					
Brent-WTI crude oil	793	86	104	14	11
Platinum-Gold	6	9	42	51	15
Platinum-Palladium	19	11	10	16	9
Corn-Ethanol				9	32
Equities					
FTSE100-CAC 40	12	22	12	7	18
Euro Stoxx 50-DAX	8	6	1	12	8
FTSE100-FTSE250	4	4	5	14	14
DJIA-Russell 1000	2	6	7	61	6
S&P500-Russell 2000	12	11	13	5	21
Russell 1000-Russell 2000	0	25	6	11	7
Exchange rates					
EUR-CHF	22	3	2	15	1
CAD-AUD	6	1	8	28	0
EUR-JPY	26	3	17	7	18
AUD-ZAR	4	2	7	8	0
CAD-ZAR	0	2	2	11	4

This table presents the number of technical rules (out of a total of 18,142) that provide significantly positive Sharpe ratios (based on the FDR test) over five subsample periods: 1991-1996, 1997-2001, 2002-2007, 2008-2011, 2012-2016. We design the subsample periods based on historical events including, the Maastricht Treaty in 1992, the East Asian currency crisis in 1997, the "dotcom" bubble in 1999-2000 and the upcoming 2002 credit crunch, the appearance of euro in 2002 and the 2003-2007 energy crisis, the global financial crisis of 2008 and finally the recent crude oil downturn in 2014. Historical transaction costs are imposed in returns for the test profitability.

Table 6. The predictability and profitability of technical trading rules in the in-sample subsample periods

	1991-1995		1997-2000		2002-2006		2008-2010		2012-2015	
	# predictive rules	Sharpe ratio	# predictive rules	Sharpe ratio	# predictive rules	Sharpe ratio	# predictive rules	Sharpe ratio	# predictive rules	Sharpe ratio
Commodities										
Brent-WTI crude oil	2214	2.11	1204	1.79	2151	2.09	0	1.47	6	1.59
Platinum-Gold	35	3.02	26	2.57	7	1.39	23	1.13	10	2.15
Platinum-Palladium	23	3.09	19	1.98	19	2.14	14	2.23	2	0.60
Corn-Ethanol							8	2.61	22	1.15
Equities										
FTSE100-CAC 40	15	1.71	48	1.18	12	1.43	7	1.72	10	1.56
Euro Stoxx 50-DAX	45	1.31	6	1.53	10	1.62	26	1.46	6	1.47
FTSE100-FTSE250	60	1.61	1	0.74	5	1.56	47	1.44	14	1.17
DJIA-Russell 1000	8	1.14	24	1.80	36	1.19	8	2.03	12	1.03
S&P500-Russell 2000	39	2.47	11	2.41	19	2.56	7	2.09	31	1.08
Russell 1000-Russell 2000	24	1.09	43	1.78	2	0.67	8	1.40	115	0.99
Exchange rates										
EUR-CHF	32	2.10	7	1.10	4	1.14	20	1.61	11	1.07
CAD-AUD	28	2.61	2	1.11	7	1.67	4	1.49	0	1.08
EUR-JPY	52	3.60	19	1.74	13	1.91	11	2.17	5	1.43
AUD-ZAR	1	1.22	3	1.15	10	0.88	7	1.21	2	0.64
CAD-ZAR	0	1.52	2	0.71	6	1.77	8	1.24	5	1.45

We report the in-sample performance of FDR portfolios of significant rules for each pair as those computed by using 70% of the available data for each subperiod. We implement the FDR test to select the technical rules with significantly positive performance and to construct equally weighted portfolios of the significant rules for each pair. We considered the Sharpe ratio as performance metric and 10% of false rejections. We also impose real historical transaction costs in the returns of total 18,412 technical rules. “#predictive rules” denote the number of technical rules generating significantly positive Sharpe ratios under the FDR test. “Sharpe ratio” indicates the annualized Sharpe ratio of each FDR portfolio of significant rules employed on each pair.

Table 7. Out-of-sample annualized Sharpe ratio of the FDR portfolio and the best significant in-sample rules

	1996		2001		2007		2011		2016	
	FDR port.	Best rule	FDR port.	Best rule	FDR port.	Best rule	FDR port.	Best rule	FDR port.	Best rule
Commodities										
Brent-WTI crude oil	1.14	1.79	1.72	1.06	-0.01	-1.21	0.00	0.00	0.00	0.00
Platinum-Gold	-0.59	1.23	-0.69	-1.19	-0.25	-1.24	1.83	-0.81	0.63	0.56
Platinum-Palladium	0.77	-1.08	0.65	-0.77	-0.49	1.38	0.24	0.61	-0.22	-0.26
Corn-Ethanol							0.22	-0.01	0.48	-0.02
Equities										
FTSE100-CAC 40	1.28	0.00	-0.91	-1.26	-0.81	0.00	-1.47	-1.18	-1.21	-1.11
Euro Stoxx 50-DAX	0.08	-1.34	-0.94	-0.34	0.00	0.00	-0.44	0.29	-0.06	0.47
FTSE100-FTSE250	-2.23	0.00	0.83	0.83	-0.15	0.00	0.09	0.89	-0.93	0.34
DJIA-Russell 1000	-1.11	-1.06	-0.94	-1.16	-0.95	0.00	-0.10	-0.72	-1.01	-1.02
S&P500-Russell 2000	-1.04	-0.41	0.22	0.86	1.48	0.00	0.28	0.78	-0.83	-0.82
Russell 1000-Russell 2000	-1.48	0.73	-0.16	0.28	0.00	0.00	-0.20	0.00	-0.49	-0.57
Exchange rates										
EUR-CHF	1.32	0.00	-0.07	-1.09	0.73	0.00	1.00	-0.86	0.15	-0.81
CAD-AUD	-0.64	1.46	0.82	0.82	0.81	0.52	-0.26	1.17	0.00	0.00
EUR-JPY	1.01	1.07	0.85	1.15	0.47	0.59	-0.92	-1.85	0.70	0.00
AUD-ZAR	-1.03	-1.03	0.88	0.88	-0.43	-0.02	0.00	0.00	-0.98	-0.98
CAD-ZAR	0.00	0.00	0.00	0.00	-1.88	-0.54	0.00	0.00	-0.43	0.00

We report out-of-sample annualized Sharpe ratio for the last year of each subperiod based on the FDR portfolios of significant rules and the best-predictive rule for each pair as those computed in the in-sample testing of each subperiod covering 70% of the dataset. The best rules are defined as technical rules providing the highest Sharpe ratio among all trading rules in the in-sample and under the 10%-FDR test. We impose historical transaction costs in computation.

Table 8. The predictability and profitability of technical trading rules in the in-sample subsample periods under the HMR approach

	1991-1995		1997-2000		2002-2006		2008-2010		2012-2015	
	# predictive rules	Sharpe ratio	# predictive rules	Sharpe ratio	# predictive rules	Sharpe ratio	# predictive rules	Sharpe ratio	# predictive rules	Sharpe ratio
Commodities										
Brent-WTI crude oil	7	1.94	2	1.94	31	1.96	0	1.09	1	0.68
Platinum-Gold	0	0.49	0	0.56	0	0.53	0	1.09	0	0.79
Platinum-Palladium	0	0.52	1	0.85	0	0.79	0	1.04	0	0.68
Corn-Ethanol							2	0.75	0	0.73
Equities										
FTSE100-CAC 40	1	0.49	0	0.76	2	0.63	0	0.75	0	0.52
Euro Stoxx 50-DAX	1	0.65	0	0.54	0	0.64	0	0.71	0	0.78
FTSE100-FTSE250	0	0.55	0	0.76	0	0.51	5	0.88	0	0.83
DJIA-Russell 1000	1	0.68	0	0.87	1	0.74	0	1.12	0	0.91
S&P500-Russell 2000	4	0.97	0	0.81	1	0.47	0	0.69	0	0.25
Russell 1000-Russell 2000	0	0.78	3	1.27	1	0.48	4	0.82	2	1.27
Exchange rates										
EUR-CHF	2	0.91	0	0.52	0	0.24	0	0.77	4	0.65
CAD-AUD	3	1.98	0	0.55	1	0.47	3	1.02	0	0.52
EUR-JPY	0	0.76	0	0.59	1	1.19	2	1.73	0	1.33
AUD-ZAR	3	1.14	1	0.91	0	0.72	1	1.26	0	0.45
CAD-ZAR	1	0.94	0	0.55	0	0.64	0	0.58	0	0.60

We report the in-sample performance of FDR portfolios of significant rules for each pair under the half-time of mean reversion approach as those computed by using 70% of the available data for each subperiod. We implement the FDR test to select the technical rules with significantly positive performance and to construct equally weighted portfolios of the significant rules for each pair. We considered the Sharpe ratio as performance metric and 10% of false rejections. We also impose real historical transaction costs in the returns of 400 contrarian technical rules. "#predictive rules" denote the number of technical rules generating significantly positive Sharpe ratios under the FDR test. "Sharpe ratio" indicates the annualized Sharpe ratio of each FDR portfolio of significant rules employed on each pair.

Table 9. Out-of-sample annualized Sharpe ratio of the FDR portfolio and the best significant in-sample rules under the half-time mean of reversion approach

		1996		2001		2007		2011		2016	
	Median HMR	FDR port.	Best rule	FDR port.	Best rule	FDR port.	Best rule	FDR port.	Best rule	FDR port.	Best rule
Commodities											
Brent-WTI crude oil	10	0.36	-0.47	1.05	1.35	0.11	0.07	0.00	0.00	0.00	0.00
Platinum-Gold	201	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Platinum-Palladium	160	0.00	0.00	0.15	0.15	0.00	0.00	0.00	0.00	0.00	0.00
Corn-Ethanol	49.5							-0.89	-0.89	0.00	0.00
Equities										0.00	0.00
FTSE100-CAC 40	275	0.28	0.28	0.00	0.00	-0.73	-0.73	0.00	0.00	0.00	0.00
Euro Stoxx 50-DAX	167	0.51	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FTSE100-FTSE250	416	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.43	0.00	0.00
DJIA-Russell 1000	310	0.00	0.00	0.00	0.00	-0.21	-0.21	0.00	0.00	0.00	0.00
S&P500-Russell 2000	234	1.32	1.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Russell 1000-Russell 2000	225	0.00	0.00	0.00	0.00	-0.88	-0.88	0.89	0.89	0.00	0.00
Exchange rates											
EUR-CHF	532	0.92	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.29
CAD-AUD	140	-0.93	-1.21	0.00	0.00	0.73	0.73	0.00	0.00	0.00	0.00
EUR-JPY	454	0.00	0.00	0.00	0.00	-0.36	-0.36	-1.80	-1.80	0.00	0.00
AUD-ZAR	197	1.14	1.14	0.00	0.00	0.00	0.00	-0.65	-0.65	0.00	0.00
CAD-ZAR	84	-0.02	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

We report out-of-sample annualized Sharpe ratio for the last year of each subperiod based on the FDR portfolios of significant rules and the best-predictive rule for each, under the half-time of mean reversion approach as those computed in the in-sample testing of each subperiod covering 70% of the dataset. The best rules are defined as technical rules providing the highest Sharpe ratio among all trading rules in the in-sample and under the 10%-FDR test. We impose historical transaction costs in computation.

Table 10. Out-of-sample performance of technical trading portfolios based on the annualized Sharpe ratio

	1996		2001		2007		2011		2016	
FDR method	Sharpe ratio	CAGR%	Sharpe ratio	CAGR%	Sharpe ratio	CAGR%	Sharpe ratio	CAGR%	Sharpe ratio	CAGR%
Commodities portfolio	1.16	4.36	1.73	6.44	-0.11	-0.21	1.02	1.11	-0.19	-0.51
Equities portfolio	-1.57	-0.45	0.36	0.23	0.25	0.01	-1.23	-0.24	-1.01	-0.44
FOREX portfolio	-0.67	-0.14	0.98	0.26	-0.89	-0.16	0.83	0.66	-0.94	-0.35
Global portfolio	0.83	1.04	1.82	2.30	-0.20	-0.11	1.13	0.81	-0.50	-0.42
Best rule										
Commodities portfolio	1.56	8.37	-0.11	-0.82	-0.55	-1.33	0.25	0.60	-0.08	-0.35
Equities portfolio	-0.61	-0.63	0.54	0.72	0.00	0.00	-0.59	-0.39	-0.98	-0.45
FOREX portfolio	1.17	0.43	0.72	0.37	0.49	0.48	-1.10	-2.40	-0.98	-1.44
Global portfolio	1.43	2.57	0.06	0.13	-0.32	-0.42	-0.83	-1.08	-0.59	-0.73
HMR FDR port.										
Commodities portfolio	0.86	4.32	1.05	7.80	0.11	0.43	-0.89	-0.34	0.00	0.00
Equities portfolio	1.31	0.18	0.00	0.00	-1.16	-0.41	0.29	0.18	0.00	0.00
FOREX portfolio	-0.14	-0.04	0.00	0.00	-0.31	-0.45	-1.37	-1.56	0.32	0.03
Global portfolio	0.57	1.78	1.05	7.80	-0.04	-0.09	-1.31	-0.56	0.32	0.03

We report out-of-sample annualized Sharpe ratio for the last year of each subperiod and for portfolios composed of the corresponding FDR portfolios of significant rules, the best-predictive rule and the FDR portfolios constructed under the half-time of mean reversion approach for each pair as those computed in the in-sample testing of each subperiod covering 70% of the dataset. The best rules are defined as technical rules providing the highest Sharpe ratio among all trading rules in the in-sample and under the 10%-FDR test. We impose historical transaction costs in computation.

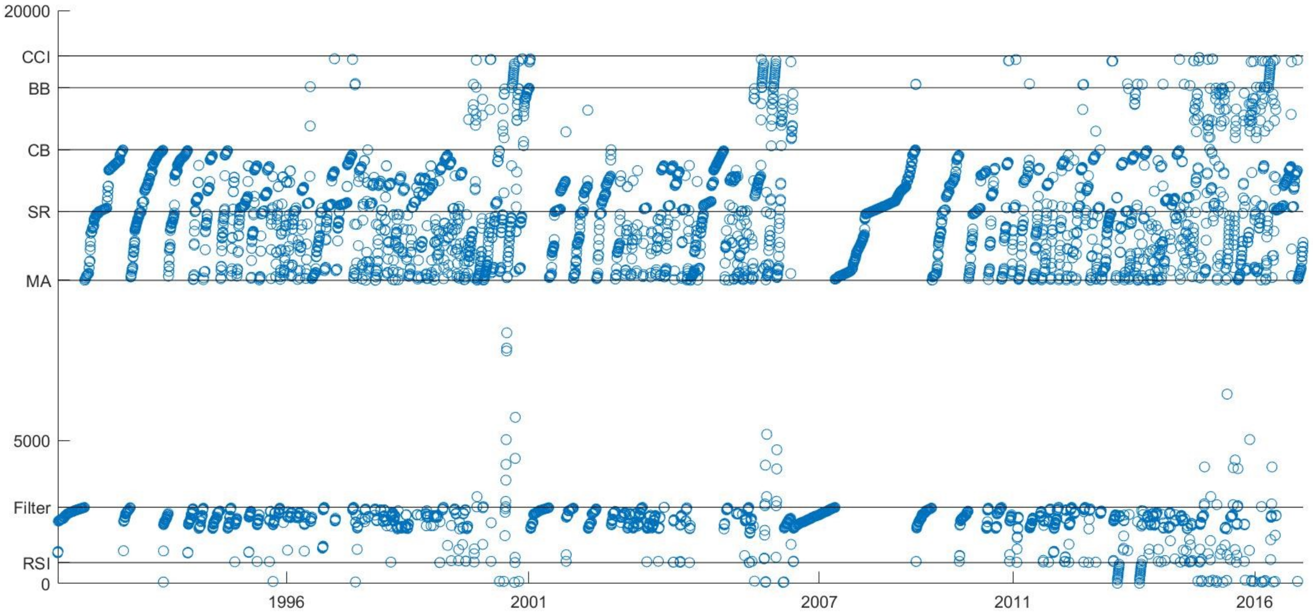

Fig.1. FDR portfolios decomposition for each pair and across all subperiods. The horizonal lines split the different families of trading rules, which add up in 18,412 trading rules in total. We display the categories of technical trading rules in the following order: RSIs, filter rules, moving averages, support and resistance rules, channel breakouts, Bollinger bands and CCIs.

CHAPTER 5
CONCLUSION

This thesis studies the predictability and excess profitability of financial markets employing up-to-date quantitative techniques derived from the fields of machine learning and data science, towards an effort to discover new scientific theories by analysing big and complex data patterns and correlations. The first objective is to review the financial market predictability by revisiting methods and trading strategies commonly exercised by trading desks to define up to what level the markets are predictable, while we try to reconcile our findings with existing market efficiency theories at the same time. The second and main aim is to produce new insights in the empirical dynamics of asset pricing. We think that we achieve this in two ways, firstly by investigating new technologies, such as machine learning and artificial intelligence, highly appraised by major market players (J.P. Morgan, Goldman Sachs, CITADEL, McKinsey) for dominating the functioning of financial marketing over the next years. Secondly by revisiting statistical inference in big data sets consisting of numerous significant trading strategies. Recent studies highlight the necessity of new, modified statistical inference approaches when various significant variables occur (see Harvey, 2017; Cochrane, 2011; among others). We apply adjusted multiple hypothesis testing frameworks, adjusting for data snooping issues, while establishing a good balance between Type I and Type II errors, in order to select significantly predictive trading rules. We present three essays trying to meet the above aims and motivations in compliance with the regulations of this thesis.

In the first essay, we investigate the existence of nonlinearities in the evolution of the implied volatility by providing evidence on the daily settlement of three U.S. market volatility indices, namely the VIX, VXN and VXD. We develop two semiparametric methodologies as a blend of the HAR specification of Corsi (2009) and one of the most promising heuristic techniques, a hybrid genetic algorithm–support vector regression (GASVR) model. We choose the HAR process due to its long-range dependence and persistent nature in modelling implied and realized volatilities. The first semiparametric approach involves an extra optimization term in the HAR model, in which the GASVR algorithm tries to optimize the three volatility components of the HAR specification. A residual analysis is also executed in a second specification, expressing potential asymmetric effects, which may be prevalent among the residuals. At this stage, we applied a heuristic regression between the residuals of HAR and its lagged values to test for further persistence.

The empirical findings indicate that the HAR-GASVR(res) approach produces more accurate predictions than those of its competitors by a significant margin. The HAR-GASVR model achieves the second-best performance. We perform robustness checks on the results by applying the SPA test (Hansen, 2005), the MCS procedure (Hansen et al., 2011) and the Giacomini and White (2006) test. The forecasting superiority of hybrid models confirms that the VIX, VXN and VXD indices exhibit nonlinear characteristics, while the also significant predictability of the HAR processes is justified by their persistent nature.

Finally, we employ the produced forecasts to exercise simple trading strategies on VIX and VXN futures contracts, as well as an S&P 500 VIX midterm futures index ETN for the economic evaluation of the methodologies proposed. The generated returns reveal that the HAR processes optimized using the GASVR algorithm, are capable to some extent of yielding statistically significant profits in normal conditions, not only in the case of trading the futures contracts, but also when we trade the ETN, in which case we achieve much higher gains due to their lower investor fee rates.

In the second essay, we reassess the evidence of the historical success of technical analysis by exercising the universe of technical trading rules of Sullivan et al., (1999), in the trending crude oil market. In particular, we try to investigate whether technical trading indicators capturing trends and momentum could benefit from the severe fluctuations characterizing the crude oil market lately. We focus our study on the crude oil futures and the United States Oil fund (USO), developed to track the daily price movements of West Texas Intermediate ("WTI") light, sweet crude oil.

Evidence of the rules’ performance on crude oil futures as well as the USO in an in-sample simulation demonstrates that during periods of dramatic crude oil price movements, more than half of the rules demonstrate a considerable predictability. However, the corresponding *p*-values of the best rules of the *bootstrap reality check* (BRC) test of White (2000) are not statistically significant most of the time, even though their corresponding performance metrics distinguish them as outstanding trading opportunities. We also endogenously incorporate transaction costs in both the cases of crude oil futures contracts and the USO, since trading rules' frequent signals, might neutralize superior returns when transaction costs are considered. However, the universe of technical trading rules still retains its profitability, although their performance shows a decay.

In terms of statistical inference, we employ two of the most powerful techniques accounting for data snooping to identify significantly outperforming trading strategies. The *false discovery rate* (FDR) approach of Barras et al., (2010) and the *k-familywise error rate* (*k*-FWER) methodology developed by Romano and Wolf (2007) both controlling for false rejections. We conclude that both specifications perform equally well, while they select a sufficient amount of rules to better diversify against model uncertainty than their predecessors.

We construct portfolios of significant rules using the FDR and *k*-FWER methods, by utilizing only past data in an in-sample period, while we assess their performance out-of-sample in a persistence analysis. The results reveal no persistent nature to the rules' performance, contrary to the very healthy in-sample results. However, and consistent with the Adaptive Market Hypothesis, tiny profits can be achieved in specific periods.

Finally, in the third essay, we revisit pairs trading by utilizing technical trading rules to predict the prices movements of formatted spreads based on cointegration assumptions. We construct 18,412 technical trading rules and we examine their predictability and excess profitability across a large set of 'famous' commodity, equity and currency pairs, being actively traded by statistical arbitrageurs, in long sample periods. Our technical trading rules are separated in momentum and contrarian classes. We also adopt the FDR method as a multiple hypothesis testing tool, which allows us to perform accurate statistical inferences in large data sets preserving against data mining issues.

Empirical evidence indicates significant predictability of technical analysis for most of the spreads considered, especially in terms of Sharpe ratio metrics, with commodity spreads being in general more predictable compared to the equity and currency ones. Furthermore, the generated returns are robust to even conservative one-way transaction costs in a break-even analysis. We also argue previous assumptions, which mention that pairs trading returns are explained by contrarian principles only. A five subperiod out-of-sample analysis reports the diminish of technical analysis' excess profitability over time maybe due to increased hedge fund activity. However, specific commodity spreads retain an encouraging performance even during recent periods.

A temporarily no-fully-rational investor's behaviour or the Adaptive Market Hypothesis may provide a possible explanation of abnormal returns achieved using technical trading rules on certain spreads and periods. The above theories assume that investors can benefit from arbitrage opportunities arising through time, and therefore make them diminish as soon as the overall market learns and milks the profitable trading strategies until new opportunities arise again in evolutionary cycles.

BIBLIOGRAPHY

Abramovich, F., Benjamini, Y., Donoho, D. L. & Johnstone, I. M., 2006. Special invited lecture: adapting to unknown sparsity by controlling the false discovery rate. The Annals of Statistics, 584-653.

Ahn, J., J., Kim, D., H., Oh, K., J., & Kim, T. Y., 2012. Applying option Greeks to directional forecasting of implied volatility in the options market: An intelligent approach. Expert Systems with Applications, 39(10), 9315–9322.

Ahoniemi, K., 2006. Modeling and forecasting implied volatility: An econometric analysis of the VIX index. Working paper, Helsinki School of Economics.

Aiolfi, M., & Timmermann, A., 2006. Persistence in forecasting performance and conditional combination strategies. Journal of Econometrics, 135 31-35.

Alexander, C., & Korovilas, D., 2013. Volatility Exchange-Traded Notes: Curse or Cure? The Journal of Alternative Investments, vol. 16, no.2, 52-70.

Alexander, S., 1961. Price movements in speculative markets: Trends or random walks: In-dustrial Management Review 2, 7-26.

Alexander, S., 1964. Price movements in speculative markets: Trends or random walks, no 2. The Random Character of Stock Market Prices (MIT Press, Cambridge, Mass.).

Allen, F., & Karjalainen, R., 1999. Using genetic algorithms to find technical trading rules. Journal of Financial Economics, 51(2), 245–271.

Andrews, D., W., K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.

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

Bandi, F., M., & Perron, B., 2006. Long memory and the relation between implied and realized volatility. Journal of Financial Econometrics 4, 636–670.

Barras, L., Scaillet, O., Wermers, R., & Stulz, R., 2015. False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas. Journal of Finance, 65(1), 179–216.

Benjamini, Y., & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 57(1), 289–300.

Benjamini, Y., & Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29(4), 1165–1188. doi:10.1214/aos/1013699998

Bessembinder, H. & Chan, K., 1998. Market efficiency and the returns to technical analysis. Financial management, 5-17.

Blair, B., J., Poon, S., H., & Taylor, S., J., 2001. Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns. Journal of Econometrics, 105 (1), 5–26.

Blair, B., J., Poon, S., H., & Taylor, S., J., 2001. Modelling S&P 100 volatility: The information content of stock returns. Journal of Banking & Finance, 25 (9), 1665–1679.

Britten-Jones, M., & Neuberger, A., 2000. Option prices, implied price processes, and stochastic volatility. Journal of Finance 55, 839-866.

Brock, W., Lakonishok, J., & LeBaron, B., 1992. Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. Journal of Finance, 47(5), 1731-1764.

Brooks, C., & Oozeer, M., C., 2002. Modeling the implied volatility of options on long gilt futures. Journal of Business Finance and Accounting 29, 111–137.

Buckland, S., T., Burnham, K., P., & Augustin, N., H., 1997. Model selection: an integral part of inference, Biometrics 53 (2), 603–618.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., Henke, N., & Trench, M., 2016. How artificial intelligence can deliver real value to companies. McKinsey Global Institute Report.

Busch, T., Christensen, B. J., & Nielsen, M., Ø., 2011. The role of implied volatility in forecasting future realized volatility and jumps in foreign exchange, stock, and bond markets. Journal of Econometrics, 160(1), 48–57.

Cao L., J, Chua K., S, & Guan L., K., 2003. C-ascending support vector machines for financial time series forecasting. In 2003 IEEE International Conference on Computational Intelligence for Financial Engineering. IEEE: New York; 317–323.

Chan, E. P., 2017. Machine trading. Willey trading series.

Chen, Y. C. & Rogoff, K., 2003. Commodity currencies. Journal of international Economics. 60(1), 133-160.

Cherkassky V., & Ma Y., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks 17:113–126.

Clements, A., C., & Fuller, J., 2012. Forecasting Increases in the VIX: A Time-Varying Long Volatility Hedge for Equities. NCER Working Paper Series.

Cochrane, J., H., 2011. Presidential address: Discount rates. The Journal of Finance, 66 (4), 1047-1108.

Conrad, J., & Kaul, G., 1998. An anatomy of trading strategies. Review of Financial Studies, 11(3), 489–519.

Corsi, F., 2009. A simple approximate long memory model of realized volatility. Journal of Financial Econometrics 7, 174–196.

Dotsis, G., Psychoyios, D., & Skiadopoulos, G., 2007. An empirical comparison of continuous-time models of implied volatility indices. Journal of Banking and Finance, 31,3584–3603.

Duan K., Keerthi S., & Poo A., N., 2003. Evaluation of simple performance measures for tuning SVM hyperparameters. Neurocomputing 51:41–59.

Dumas, B., Fleming, J., & Whaley, R., E., 1998. Implied volatility functions: Empirical tests. Journal of Finance 53, 2016–2059.

Dunis, C., Likothanassis, S., Karathanasopoulos, A., Sermpnis, G. & Theofilatos, K., 2013. A hybrid genetic algorithm–support vector machine approach in the task of forecasting and trading. Journal of Asset Management.

Engle, R., & Figlewski S., 2015. Modelling the Dynamics of Correlations among Implied Volatilities. Review of Finance, 19: pp. 991-1018

Fama E., & Marshall, B., 1966. Filter Rules and Stock-Market Trading. Journal of Business 39(1), 226–241.

Fama, E., 1965. The Behavior of Stock Market Prices. Journal of Business, 38, 34–105.

Fama, E., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. Journal of Finance, 25(2), 383–417.

Fernandes, M., Medeiros, M., C., & Scharth, M., 2014. Modeling and predicting the CBOE market volatility index. Journal of Banking & Finance, 40, 1–10.

Fleming, J., Ostdiek, B., & Whaley, R., E., 1995. Predicting stock market volatility: a new measure. Journal of Futures Markets 15, 265–302.

Frisman, R., 2001. Estimating the Value of Political Connections. The American Economic Review 91, 4, 1095-1102.

Froot, K., Schaferstein, D., & Stein, J., 1992. Herd on the street: Informational inefficiencies in a market with short-term speculation, Journal of Finance, 47, 4, 1461–1484.

Gatev, E., Goetzmann, W.N., & Rouwenhorst, K.G., 2006. Pairs Trading: Performance of a Relative-Value Arbitrage Rule. The Review of Financial Studies, 19 (3) 797–827.

Gerbert, P., Justus, J., & Hecker, M., 2017. Competing in the age of artificial intelligence. Henderson Institute, Boston Consulting Group.

Giacomini, R., & White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.

Giraitis, L., Kokoszka, P., Leipus, R., & Teyssière, G., 2003. Rescaled variance and related tests for long memory in volatility and levels. Journal of Econometrics 112, 265–294.

Gonçalves, S., & Guidolin, M., 2006. Predictable dynamics in the S&P 500 index options implied volatility surface. Journal of Business 79, 1591–1635.

Gonzalez Miranda, F., & Burgess, N., 1997. Modelling market volatilities: the neural network perspective. The European Journal of Finance, 3 (2), 137–157.

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

Hansen, P. R., Lunde, A., Nason, J. M., 2011. The Model Confidence Set. Econometrica, 79(2), 453–497.

Harvey, C. R. and Liu, Y., 2015. Backtesting. The Journal of Portfolio Management, 42(1),
13-28.

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

Harvey, C. R., 2017. The Scientific Outlook in Financial Economics. Duke I&E Research Paper No. 2017-05. dx.doi.org/10.2139/ssrn.2893930

Harvey, C. R., Liu, Y., & Zhu, H., 2016. …and the cross-section of expected returns. Review of Financial Studies, 29, 5-68.

Harvey, C., R., & Whaley, R., E., 1992. Market volatility prediction and the efficiency of the S&P 100 index option market. Journal of Financial Economics 31, 43–73.

Hedge, S. P., & McDermott, J. B., 2004. The market liquidity of DIAMONDS, Q's, and their underlying stocks. Journal of Banking and Finance, 28(5), 1043–1067.

Hogan, S., Jarrow, R., Teo, M., & Warachka, M., 2004. Testing market efficiency using statistical arbitrage with applications to momentum and value strategies. Journal of Financial Economics, 73 (3), 525-565.

Holland J., 1995. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. MIT Press: Cambridge, MA.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2), 65-70.

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

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

Hsu, Y., Kuan, C., & Yen, M., 2014. A Generalized Stepwise Procedure with Improved Power for Multiple Inequalities Testing. Journal of Financial Econometrics, 12(4), 730–755.

Ince, H., & Trafalis, T., B., 2006. Kernel methods for short-term portfolio management. Expert Systems with Applications 30, 535–542.

Ince, H., & Trafalis, T., B., 2008. Short term forecasting with support vector machines and application to stock price prediction. International Journal of General Systems 37, 677–687.

James, F., 1968. Monthly Moving Averages—An Effective Investment Tool. Journal of Financial and Quantitative Analysis, 3(3), 315–326.

Jegadeesh, N. and Titman, S., 1995. Overreaction, delayed reaction, and contrarian profits. Review of Financial Studies, 8(4), 973-993.

Jensen, M., & Benington, G., 1970. Random Walks and Technical Theories: Some Addition-al Evidence. Journal of Finance, 25(2), 469-482.

Jiang, G., & Tian, Y., 2005. Model-free implied volatility and its information content. Review of Financial Studies 18, 1305–1342.

Johansen, S., 1995. Likelihood-based inference in cointegrated vector autoregressive models. Oxford University Press.

Johansen, S., & Nielsen, M. Ø., 2012. Likelihood inference for a fractionally cointegrated vector autoregressive model. Econometrica, 80(6), 2667-2732.

Kavajecz, K., & Odders-White, E. R., 2004. Technical analysis and liquidity provision. Re-view of Financial Studies, 17(4), 1043–1071.

Keynes, J. M., 1936. The General Theory of Employment, Interest, and Money. Harcourt, Brace & Co., New York.

Kolanovic, M., & Krishnamachari, R., T., 2017. Big data and AI strategies: Machine learning & alternative data approach to investing. Quantitative Investing and Derivatives Strategy, J. P. Morgan.

Konstantinidi, E., & Skiadopoulos, G., 2011. Are VIX futures prices predictable? An empirical investigation. International Journal of Forecasting 27, 543–560.

Konstantinidi, E., Skiadopoulos, G., & Tzagkaraki, E., 2008. Can the evolution of implied volatility be forecasted? Evidence from European and US implied volatility indices. Journal of Banking & Finance, 32 (11), 2401–2411.

Koopman, S., J., Jungbacker, B., & Hol, E., 2005. Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. Journal of Empirical Finance 12, 445–475.

Koza, J., R., 1992. Genetic programming: on the programming of computers by means of natural selection (Vol. 1). MIT press.

LeBaron, B., 2000, The stability of moving average technical trading rules on the Dow Jones Index. Derivatives Use, Trading and Regulation, 5 (4), 324-338.

Lee, D., & Schmidt, P., 1996. On the power of the KPSS test of stationarity against fractionally-integrated alternatives. Journal of Econometrics 73, 285-302.

Leland, H., E., 1999. Beyond mean-variance: Performance measurement in a nonsymmetrical world. Financial Analysts Journal 55, 27–35.

Levy, R., 1967. Relative Strength as a Criterion for Investment Selection. Journal of Finance, 22(4), 595–610.

Ljung, G.M. & Box, G.E.P., 1978. On a measure of lack of fit in time series models. Biometrika, 65(2), 297-303.

Lo, A. W., & MacKinley, 1990. Data snooping biases in tests of financial asset pricing models. Review of Financial Studies, 3(3), 431–467.

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

Lo, A. W., Mamaysky, H., & Wang, J., 2000. Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. Journal of Finance 55(4), 1705-1765.

Locke, P., R., & Venkatesh, P., C., 1997. Futures market transaction costs. Journal of Futures Markets, 17(2), 229-245.

Lyons, R. K., 2001. The Microstructure Approach to Exchange Rates. MIT Press, Cambridge, MA.

Malkiel, B., 1981. A Random Walk Down Wall Street. Norton New York, 2ed.

Malliaris, M., & Salchenberger, L., 1996. Using neural networks to forecast the S&P 100 implied volatility. Neurocomputing, 10 (2), 183–195.

Marshall, B. R., Cahan, R. H., & Cahan, J. M., 2008. Can commodity futures be profitably traded with quantitative market timing strategies? Journal of Banking and Finance, 32(9), 1810-1819.

Marshall, B. R., Cahan, R. H., & Cahan, J. M., 2008. Does intraday technical analysis in the U.S. equity market have value? Journal of Empirical Finance, 15(2), 199–210.

McAleer, M., & Medeiros, M,. C., 2008. A multiple regime smooth transition Heterogeneous Autoregressive model for long memory and asymmetries. Journal of Econometrics, 147(1), 104–119.

McLean, R. D., & Pontiff, J., 2016. Does academic research destroy stock return predictability? The Journal of Finance, 71, 5–32.

Menkhoff, L., Taylor, M.P, 2007. The Obstinate Passion of Foreign Exchange Professionals: Technical Analysis. Journal of Economic Literature, 45, 936-972. doi:10.1257/jel.45.4.936

Meucci, A., 2009. Review of statistical arbitrage, cointegration, and multivariate Ornstein-Uhlenbeck. Working paper, Symmys.

Miffre, J., & Rallis, G., 2007. Momentum strategies in commodity futures markets. Journal of Banking & Finance, 31(6), 1863-1886.

Müller, U., Dacorogna, M., D., R., Olsen, R., Pictet, O., & Ward, J., 1993. Fractals and intrinsic time: a challenge to econometricians. In: Proceedings of the XXXIX International AEA Conference on Real Time Econometrics.

Müller, U., Dacorogna, M., D., R., Olsen, R., Pictet, O., & Weizsacker, J., 1997. Volatilities of different time resolutions: analysing the dynamics of market components. Journal of Empirical Finance 4, 213–239.

Neely, C. J., & Weller, P. A., 2003. Intraday technical trading in the foreign exchange market. Journal of International Money & Finance, 22, 223–237.

Neely, C. J., & Weller, P., 2011. Technical Analysis in the Foreign Exchange Market. Working paper 2011-001B, Federal Reserve Bank of St. Louis.

Neely, C. J., & Weller, P., 2013. Lessons from the evolution of foreign exchange trading strategies. Journal of Banking and Finance, 37, 3783-3798.

Neely, C. J., Weller, P. A. & Ulrich, M., 2009. The adaptive markets hypothesis: evidence from the foreign exchange market. Journal of Financial & Quantitative Analysis, 44, 467–488.

Neely, C., Weller, P., & Dittmar, R., 1997. Is technical analysis in the foreign exchange market profitable? A genetic programming approach. Journal of Financial and Quantitative Analysis 32(4), 405-426.

Neftci, S. N., 1991. Naive Trading Rules in Financial Markets and Wiener-Kolmogorov Pre-diction Theory: A Study of "Technical Analysis." Journal of Business, 64(4), 549-571.

Pai, P., F., Lin, C., S., Hong, W., C., & Chen, C., T., 2006. A Hybrid Support Vector Machine Regression for Exchange Rate Prediction. International Journal of Information and Management Sciences, 17 (2), 19- 32.

Politis, D., & Romano, J., 1994. The stationary bootstrap. Journal of the American Statistical Association 89, 1303–1313.

Psaradellis, I., & Sermpinis, G., 2016. Modelling and trading the U.S. implied volatility indices: Evidence from the VIX, VXN and VXD indices. International Journal of Forecasting 32 (2016) 1268–1283.

Psychoyios, D., & Skiadopoulos, G., 2006. Volatility Options: Hedging Effectiveness, Pricing and Model Error. Journal of Futures Markets, vol. 26, no. 1, 1-31.

Qi, M., & Wu, Y., 2006. Technical trading-rule profitability, data snooping, and reality check: Evidence from the foreign exchange market. Journal of Money, Credit and Banking, 38(8), 2135–2158.

Ready, M. J., 2002. Profits from technical trading rules. Financial Management 31, 43-61.

Refenes, A., N., & Holt, W., T., 2001. Forecasting Volatility with Neural Regression: A Contribution to Model Adequacy. IEEE Transactions on Neural Networks, 12, no. 4: 850–864.

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

Romano, J. P., & Wolf, M., 2007. Control of generalized error rates in multiple testing. Annals of Statistics, 35(4), 1378–1408.

Romano, J. P., Shaikh, A. M., & Wolf, M., 2008. Formalized Data Snooping Based on Generalized Error Rates. Econometric Theory, 24(2), 404–447.

Scholkopf B., & Smola A., 2002. Learning with Kernels. MIT Press: Cambridge, MA.

Scholkopf B., Bartlett P., Smola A., & Williamson R., 1999. Shrinking the tube: a new support vector regression algorithm. In Advances in Neural Information Processing Systems 11, Kearns MJ (ed.). MIT Press: Cambridge MA; 330–336.

Schuhmacher, F. & Eling, M., 2011. Sufficient conditions for expected utility to imply draw-down-based performance rankings. Journal of Banking & Finance, 35(9), 2311-2318.

Sepp, A., 2016. Volatility modelling and trading. Global Derivatives Workshop, Global Derivatives Trading & Risk Management 2016.

Sermpinis G., Laws J., Karathanasopoulos A., & Dunis C., L., 2012. Forecasting and trading the EUR/USD exchange rate with gene expression and psi sigma neural networks. Expert Systems with Applications 39: 8865–8877.

Sermpinis, G., Stasinakis, C., Theofilatos, K., & Karathanasopoulos, A., 2014. Inflation and Unemployment Forecasting with Genetic Support Vector Regression. Journal of Forecasting, 33 (6), 471-487.

Shapiro A., F., 2000. A Hitchhiker's guide to the techniques of adaptive nonlinear models. Insurance: Mathematics and Economics 26:119–132.

Shu, J., & Zhang, J., E., 2011. Causality in the VIX Futures Market. The Journal of Futures Markets, vol. 32, no. 1, 24-46.

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

Siedlecki W., & Sklansky J., 1989. A note on genetic algorithms for large-scale feature selection. Pattern Recognition Letters 10:335–347.

Simons, H., 2010. Russell index futures spreads: Trading the large cap market Segment in U.S. equities against the small cap segment. ICE.

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

Storey, J., Taylor, J. & Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society, Series B 66(1), 187–205.

Sullivan, R., Timmermann, A., & White, H., 1999. Data-Snooping, Technical Trading Rule Performance, and the Bootstrap. Journal of Finance, 54(5), 1647–1691.

Sun Z., Bebis G., & Miller R., 2004. Object detection using feature subset selection. Pattern Recognition 37: 2165–2176.

Suykens, J., A., K., Brabanter, J., D., Lukas, L., & Vandewalle, L., 2002. Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing 48, 85–105.

Sweeney, R., 1988. Some New Filter Rule Tests: Methods and Results. Journal of Financial and Quantitative Analysis, 23(3), 285–300.

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

Tenti P., 1996. Forecasting foreign exchange rates using recurrent neural networks. Applied Artificial Intelligence 10: 567–582.

Timmermann, A., & Granger, C. W. J., 2004. Efficient market theory and forecasting. International Journal of Forecasting, 20, 15–27.

Trafalis T., B, & Ince H., 2000. Support vector machine for regression and applications to financial forecasting. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN 2000, Vol. 1. IEEE Press: New York; 348–353.

Vapnik V., N., 1995. The Nature of Statistical Learning Theory. Springer: Berlin.

Wang, C., & Yu, M., 2004. Trading activity and price reversals in futures markets. Journal of Banking and Finance, 28(6), 1337-1361.

Whaley, R., E., 2000. The investor fear gauge. Journal of Portfolio Management 26, 12-17.

Whaley, R., E., 2009. Understanding the VIX. Journal of Portfolio Management 35 (3), 98-105.

White, H., 2000. A Reality Check for Data Snooping. Econometrica, 68(5), 1097–1126.

Yamamoto, R., 2012. Intraday technical analysis of individual stocks on the Tokyo Stock Exchange. Journal of Banking and Finance, 36(11), 3033–3047.

Yuang, F., C., 2012. Parameters Optimization Using Genetic Algorithms in Support Vector Regression for Sales Volume Forecasting. Applied Mathematics, 3 (1), pp. 1480 - 1486.