Quantitative Finance
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rquf20

Quantitative spread trading on crude oil and refined products markets

Mark Cummins [a] & Andrea Bucca [b]

[a] DCU Business School, Dublin City University, Dublin 9, Ireland

[b] Glencore International Ltd, 50 Berkley Street, London, W1J 8HD, UK

Published online: 13 Sep 2012.

To cite this article: Mark Cummins & Andrea Bucca (2012) Quantitative spread trading on crude oil and refined products markets, Quantitative Finance, 12:12, 1857-1875, DOI: 10.1080/14697688.2012.715749

To link to this article: http://dx.doi.org/10.1080/14697688.2012.715749

Quantitative Finance, Vol. 12, No. 12, December 2012, 1857–1875

Routledge
Taylor & Francis Group

Quantitative spread trading on crude oil and refined products markets

MARK CUMMINS*† and ANDREA BUCCA‡

†DCU Business School, Dublin City University, Dublin 9, Ireland
‡Glencore International Ltd, 50 Berkley Street, London, W1J 8HD, UK

(*Received 27 August 2011; in final form 23 July 2012*)

Quantitative trading in oil-based markets is investigated over 2003–2010, with a focus on WTI, Brent, heating oil and gas oil. A total of 861 spreads are considered. A novel optimal statistical arbitrage trading model is applied, with generalised stepwise procedures controlling for data snooping bias. Aggregating upward and downward mean-reversion, profitable strategies are identified with Sharpe ratios greater than 2 in many instances. For the top categories, average daily returns range from 0.07 to 0.55%, with trade lengths of 9–55 days. A collapse in the number of profitable trading strategies is seen in 2008. Robustness to varying transactions costs is examined.

Keywords: Econometrics; Energy derivatives; Trading systems; Quantitative trading strategies

JEL Classification: C1, C5, C6, C12, C52, C63

1. Introduction

A significant increase in commodities trading has been observed over the past decade. This increased activity has emerged in a range of forms, including commodity index investment and the proliferation of commodity-based hedge funds. Mou (2010) reports that, by mid-2008, the value of long positions held by index investors reached \$256 bn, with the market ratio of index investment positions to total open interest at 44%. The collapse in commodity markets resulting from the credit crisis and global economic recession concerns led to a huge sell off of positions, with the total index investment value falling to \$112 bn and the market ratio falling to 39%, although both measures had recovered by the end of 2009 to \$211 bn and 52%, respectively. Gregeriou *et al.* (2009) report that assets under management by the Commodity Trading Advisor (CTA) industry reached \$208 bn by the end of 2008. This increased commodity trading activity has led to contentious debates within both academia and industry on price bubbles in commodity markets and the role of speculators in particular (see, for instance, Gilbert (2008)). Undisputed is the fact that speculative trading is a key feature of modern commodity markets as investors and funds seek to actively and often aggressively search out returns. In this context, this study makes a number of key contributions to the literature.

Quantitative trading opportunities in the crude oil and refined products markets are investigated over the period 2003–2010, with particular focus on WTI, Brent, heating oil and gas oil. A wide range of common- and cross-commodity spreads (including calendar, crack and locational spreads) are considered. Trading strategies are designed so as to exploit any predictability that exists in the unit volumetric spreads. Drawing on the current literature, the novel *optimal* statistical arbitrage trading model of Bertram (2010) is applied for the empirical analysis in this study.§ Few papers deal directly with the issue of optimal entry and exit trading signals for statistical arbitrage trading strategies. These include

*Corresponding author. Email: mark.cummins@dcu.ie

§A number of alternative studies in the area of statistical arbitrage trading have been conducted to date, including Burgess (1999, 2000), Trapletti *et al.* (2002), Vidyamurthy (2004), Whistler (2004), Elliott *et al.* (2005), Gatev *et al.* (2006), Do *et al.* (2006) and Avellaneda and Lee (2010). Andrade *et al.* (2005), Papadakis and Wysocki (2007) and Do and Faff (2009) provide independent verification of the trading rule proposed by Gatev *et al.* (2006). Aldridge (2009), Bowen *et al.* (2010) and Dunis *et al.* (2010) consider high-frequency applications. Other papers of interest include Shleifer and Vishny (1997), Hogan *et al.* (2004) and Lin *et al.* (2006).

Quantitative Finance
ISSN 1469–7688 print/ISSN 1469–7696 online © 2012 Taylor & Francis
http://www.tandfonline.com
http://dx.doi.org/10.1080/14697688.2012.715749

Vidyamurthy (2004), Elliott *et al.* (2005) and Do *et al.* (2006). The approach of Bertram (2010) is very much distinct from these papers. Specifically, modelling a given spread series as a mean-reverting Ornstein–Uhlenbeck process, analytic solutions exist for the optimal entry and exit levels determined through maximising the expected return *per unit time*, which is defined as the ratio of the deterministic return of the strategy to the expected trade cycle time. Statistical arbitrage trading strategies with defined entry and exit levels offer deterministic returns but uncertainty lies in the length of the trade cycles. Normalisation by the expected trade cycle time explicitly accounts for the alternative deterministic returns and stochastic trade cycle times associated with alternative strategies, allowing for consistent cross-comparison.

The full set of commodity data considered leads to 861 spreads in total. The objective of this study is to statistically test the performance of just over 2500 statistical arbitrage style strategies simultaneously, based on the constructed spreads. The hypothesis tests are designed to formally identify those trading strategies that, with statistical significance, outperform a given benchmark in terms of mean daily log-return. The benchmark is defined to have zero mean daily log-return, corresponding to taking no position in a given spread. This introduces the well-known issue of data snooping bias, whereby, under naive analysis, profitable trading strategies may be identified by pure chance alone. This links directly to the broader issue of multiple hypothesis testing in general statistical and econometric applications. Romano *et al.* (2009) provide a detailed exposition of the issues pertaining to multiple hypothesis testing, outlining the key literature in the area. Two recent generalised stepwise techniques proposed in the literature are used to control for data snooping bias. The stepdown procedure of Romano and Wolf (2007) and the balanced stepdown procedure of Romano and Wolf (2010) are applied, both serving as improvements over more conservative single-step approaches, such as the seminal reality check bootstrap test of White (2000) and the superior predictive ability test of Hansen (2005). The generalised procedures offer greater *power*, where power is loosely defined as the ability to reject false null hypotheses. The balanced stepdown procedure offers a further improvement in allowing for balance amongst the hypothesis tests in the sense that each is treated equally in terms of power. Applying these generalised stepwise procedures to control for data snooping bias within this quantitative trading study presents a significant contribution to the literature. Specifically, whereas the conservative single-step procedures allow for statistical inferences around the best-performing trading strategy only, the stepwise procedures recursively identify all of those trading strategies that are profitable with statistical significance.

The reality check bootstrap test of White (2000) has been applied in a number of quantitative trading studies to identify profitability while rigourously controlling for data snooping bias. Sullivan *et al.* (1999), Hsu and Kuan (2005), Qui and Wu (2006), Park and Irwin (2007), and Marshall *et al.* (2008) apply the reality check bootstrap test to evaluate the profitability of a wide range of technical trading rules commonly used in industry. Marshall *et al.* (2008) consider a data set of 15 commodities, including crude oil and heating oil, and conclude that, with the exception of oats, technical trading rules are not profitable. Hsu *et al.* (2009) apply a stepwise extension of the superior predictive ability test of Hansen (2005) to re-evaluate the profitability of technical trading rules. Our study is distinct from these papers in that statistical arbitrage techniques are employed to exploit predictability or mean-reversion in the spreads where it exists.

The study further contributes by allowing for practical comparison of the stepwise and the balanced stepwise procedures in the context of a trading application. The stepwise procedure of Romano and Wolf (2007) improves on single-step procedures with subsequent iterative steps that allow for additional profitable strategies to be identified. The balanced stepwise procedure of Romano and Wolf (2010) likewise improves on single-step procedures with subsequent iterative steps that allow for additional profitable strategies to be identified, but with the added benefit of balance such that trading strategies with large mean daily returns do not dominate those with lower mean daily returns. The balanced stepdown procedure is shown to identify many more profitable strategies than the non-balanced stepdown procedure. However, there exists broad consistency, in the sense that those trading strategies identified as being profitable by the stepdown procedure are also identified in general by the balanced stepdown procedure.

The remainder of the paper is organised as follows. Section 2 discusses the crude oil and refined products markets, describing in detail the data used for the empirical analysis. Section 3 presents the optimal statistical arbitrage trading model of Bertram (2010). Section 4 discusses the issue of data snooping bias and links this to the broader issue of multiple hypothesis testing. The details of the stepdown procedure of Romano and Wolf (2007) and the balanced stepdown procedure of Romano and Wolf (2010) are also given. The empirical analysis conducted in this study is outlined in section 5, defining the formal hypothesis tests and discussing transaction cost and liquidity issues. Section 6 presents the results of the empirical analysis and considers robustness to varying transaction costs. Section 7 concludes.

2. Crude oil and refined products markets

The importance of crude oil and refined products markets within the modern economy is well appreciated. Increased volatility and uncertainty around prices underlies the active trading of derivatives products to hedge and manage the wide variety of risk exposures of producers and consumers in the energy sector. The trading of futures spreads offer a range of flexible risk management solutions. Calendar spreads involve the trading of futures contracts on the same underlying commodity with different maturities. Such spreads allow producers and suppliers to trade maturity-based price differentials and hedge movements along the futures curve, in particular transitions between backwardation and contango. Crack spreads represent the price differential between a given crude oil and refined product and essentially represent the production margin for refineries. Locational spreads represent the price differential between the same or related commodities trading in different geographical regions. These spreads allow energy market participants to exploit locational arbitrage, subject to transport and associated costs in the case of physical deliveries.

Volatility in the crude oil and refined products markets has attracted much interest in speculative trading opportunities. The quantitative trading of futures spreads in these markets is the primary focus of this study, premised on the existence of predictability and mean-reversion in relative price movements. Under the seminal cost of carry theory (Kaldor 1939), a number of studies examine the term structure relationship between spot and futures prices from both a market efficiency perspective (Crowder and Hamed 1993, Kellard *et al.* 1999) and a price discovery perspective (Schwarz and Szakmary 1994, Silvapulle and Moosa 1999). Girma and Paulson (1999) and Gjolberg and Johnsen (1999) examine the relation between crude oil and a range of refined products. It is concluded that cointegrating relationships exist, implying predictability in the crack spreads. Kinnear (2002) and Alizadeh and Nomikos (2004) consider locational arbitrage opportunities in the crude oil markets. All of these studies provide motivation for exploring the existence of predictability and mean-reversion in spreads from a trading perspective. Rather than test for these features directly, the approach taken in this study is to apply mean-reversion-based trading strategies to a large data set of spreads and use advanced data snooping control procedures to identify statistically significant profitability.

For the empirical analysis to follow, a comprehensive data set of crude oil and refined product futures contracts are used, comprising WTI and Brent on the crude oil side and gas oil (GO) and heating oil (HO) on the refined products side. These commodities are chosen for this study on the basis of size, importance and liquidity. WTI and HO are both traded on the New York Mercantile Exchange (NYMEX), while Brent and GO are both traded on the Intercontinental Exchange (ICE). The data set covers the 11-year period from 3 January 2000 to 31 December 2010. Most notably, this period covers the record high crude oil prices recorded in 2008 and the subsequent collapse in the latter part of the same year resulting from the global economic crisis, in addition to the gradual recovery in crude oil prices over 2009–2010. All relevant conversions were done to ensure the time series are quoted consistently in dollars per barrel.

The data set includes futures curves for WTI, Brent and HO running from the prompt month up to month 12, with the GO futures curve running from the prompt month up to month 6. These choices are made to ensure sufficient liquidity from a trading perspective, with the construction of the individual time series explicitly taking into account the rolling of futures contracts. Transaction costs and contract liquidity are discussed in more detail later in section 5. For the quantitative spread trading analysis, the full range of common and cross-commodity spreads (including calendar, crack and locational spreads) are considered. With 42 different maturity contracts across the four commodity groups, a total of 861 individual spreads are available for analysis. Finally, non-synchronicity bias is avoided with all time series being observed at the same time of 5.15 p.m. EDT, coinciding with the close of the WTI crude oil market.

3. Optimal statistical arbitrage trading model

This section provides a detailed mathematical exposition of the novel optimal statistical arbitrage trading model of Bertram (2010). The issue of optimal statistical arbitrage trading is approached by first assuming that the spread between two asset log-price series, denoted s_t, is given by the following zero-mean OU process:†

$$\mathrm{d}s_t = -\alpha s_t \mathrm{d}t + \sigma \mathrm{d}W_t, \tag{1}$$

with $\alpha, \sigma > 0$ and W_t denoting a Wiener process. Defining the entry and exit levels of the trading strategy by a and m, respectively, a complete trade cycle is the time taken for the spread process to transition from a to m and then return back to a. Formally, the trade cycle time is defined as follows:

$$\mathcal{T} \equiv \mathcal{T}_{a\to m} + \mathcal{T}_{m\to a},$$

where $\mathcal{T}_{a\to m}$ is the time to transition from a to m and $\mathcal{T}_{m\to a}$ is the time to transition from m to a, and the independence of the two times follows from the Markovian property of the OU process. Given relative transaction costs c, the total log-return from one trade cycle of the statistical arbitrage trading strategy is given by $r(a, m, c) \equiv m - a - c$, which is deterministic but for which the associated trading cycle time is stochastic. In this context, Bertram (2010) proposes the following expected return per unit time and variance of return per unit time measures:

$$\xi(a, m, c) \equiv \frac{r(a, m, c)}{E(\mathcal{T})},$$

$$\varsigma(a, m, c) \equiv \frac{r^2(a, m, c)V(\mathcal{T})}{E^3(\mathcal{T})},$$

where $E(\mathcal{T}) = E(\mathcal{T}_{a\to m}) + E(\mathcal{T}_{m\to a})$ is the expected trade cycle time and $V(\mathcal{T}) = V(\mathcal{T}_{a\to m}) + V(\mathcal{T}_{m\to a})$ is the variance of the trade cycle time. Following a transformation of the OU process to a dimensionless system, and drawing on the first-passage time theory of Thomas (1975), Sato (1977) and Ricciardi and Sato (1988), Bertram (2010)

†The zero-mean assumption does not present any issue in practice. The optimal entry and exit levels obtained can be easily translated to account for a non-zero mean in empirical data.

derives the following analytic expressions for $E(\mathcal{T})$, $V(\mathcal{T})$, $\xi(a, m, c)$ and $\varsigma(a, m, c)$:

$$
\begin{aligned}
E(\mathcal{T}) &= \frac{\pi}{\alpha}\left(\mathrm{Erfi}\left(\frac{m\sqrt{\alpha}}{\sigma}\right) - \mathrm{Erfi}\left(\frac{a\sqrt{\alpha}}{\sigma}\right)\right),\\
V(\mathcal{T}) &= \frac{w_1\left(\frac{m\sqrt{2\alpha}}{\sigma}\right) - w_1\left(\frac{a\sqrt{2\alpha}}{\sigma}\right) - w_2\left(\frac{m\sqrt{2\alpha}}{\sigma}\right) + w_2\left(\frac{a\sqrt{2\alpha}}{\sigma}\right)}{\alpha^2},\\
\xi(a,m,c) &= \frac{\alpha(m-a-c)}{\pi\left(\mathrm{Erfi}\left(\frac{m\sqrt{\alpha}}{\sigma}\right) - \mathrm{Erfi}\left(\frac{a\sqrt{\alpha}}{\sigma}\right)\right)},\\
\varsigma(a,m,c) &= \alpha(m-a-c)^2\\
&\quad\times \frac{w_1\left(\frac{m\sqrt{2\alpha}}{\sigma}\right) - w_1\left(\frac{a\sqrt{2\alpha}}{\sigma}\right) - w_2\left(\frac{m\sqrt{2\alpha}}{\sigma}\right) + w_2\left(\frac{a\sqrt{2\alpha}}{\sigma}\right)}{\pi^3\left(\mathrm{Erfi}\left(\frac{m\sqrt{\alpha}}{\sigma}\right) - \mathrm{Erfi}\left(\frac{a\sqrt{\alpha}}{\sigma}\right)\right)^3},
\end{aligned}
$$

where $\mathrm{Erfi}(z)$ is the imaginary error function,

$$
\begin{aligned}
w_1(z) &\equiv \left(\frac{1}{2}\sum_{g=1}^{\infty}\Gamma\left(\frac{g}{2}\right)(\sqrt{2}z)^g/g!\right)^2\\
&\quad - \left(\frac{1}{2}\sum_{g=1}^{\infty}(-1)^g\Gamma\left(\frac{g}{2}\right)(\sqrt{2}z)^g/g!\right)^2,\\
w_2(z) &\equiv \sum_{g=1}^{\infty}\Gamma\left(\frac{(2g-1)}{2}\right)\Psi\left(\frac{(2g-1)}{2}\right)(\sqrt{2}z)^{(2g-1)}/(2g-1)!
\end{aligned}
$$

and $\Psi(z) \equiv \psi(z) - \psi(1)$, with $\Gamma(z)$ and $\psi(z)$ the gamma and digamma functions, respectively.

With these analytic results in place, it is shown that the optimal entry and exit levels a^* and m^* may be derived by either maximising the expected return per unit time $\xi(a, m, c)$ or the associated per unit time Sharpe ratio. In the former case, it is established that $m^* = -a^*$, with $a^* < 0$ being the root of the equation

$$
\exp\left(\frac{\alpha a^2}{\sigma^2}\right)(2a+c) - \sigma\sqrt{\frac{\pi}{\alpha}}\mathrm{Erfi}\left(\frac{a\sqrt{\alpha}}{\sigma}\right) = 0.
$$

In the latter case, the risk-free rate of interest r_f is introduced and the associated per unit time Sharpe ratio is given by

$$
\begin{aligned}
S(a,m,c,r_\mathrm{f}) &\equiv \frac{\xi(a,m,c) - (r_\mathrm{f}/E(\mathcal{T}))}{\sigma}\\
&= (m-a-c-r_\mathrm{f})\sqrt{\frac{E(\mathcal{T})}{(m-a-c)^2 V(\mathcal{T})}}\\
&= (m-a-c-r_\mathrm{f})\\
&\quad\times\sqrt{\frac{\frac{\pi}{\alpha}\left(\mathrm{Erfi}\left(\frac{m\sqrt{\alpha}}{\sigma}\right) - \mathrm{Erfi}\left(\frac{a\sqrt{\alpha}}{\sigma}\right)\right)}{(m-a-c)^2\frac{w_1\left(\frac{m\sqrt{2\alpha}}{\sigma}\right) - w_1\left(\frac{a\sqrt{2\alpha}}{\sigma}\right) - w_2\left(\frac{m\sqrt{2\alpha}}{\sigma}\right) + w_2\left(\frac{a\sqrt{2\alpha}}{\sigma}\right)}{\alpha^2}}}.
\end{aligned}
$$

It is established similarly that $m = -a$, $a < 0$, and the optimal entry level a^* follows from maximising the Sharpe ratio expression with this substitution for m.

3.1. Remark on model specification

The OU process described in equation (1) for the log-price spread series is simplistic in its specification, although it has been used extensively in the finance literature, particularly for interest rate modelling (Vasicek 1977). The OU process assumes a Gaussian distribution for the spread series, with constant volatility, constant long-run mean level, constant speed of mean reversion and symmetry in adjustments from above and from below the long-run mean level. Given the non-Gaussian nature of financial data, the OU process is inherently mispecified. However, the purpose of this study is to dynamically exploit predictability in the spread between the log-prices of related oil and oil-based products. The key advantage of using the OU process to describe this predictability is that it allows for analytic trade execution signals to be determined, which in the broader context provides significant computational efficiencies that are of particular benefit for high-frequency statistical arbitrage trading. As noted by Bertram (2010), the simplicity of the OU process further offers ease in investigating how the important system variables relate to each other, in particular time. Model specification issues for the Bertram (2010) trading model are discussed by Cummins (2011), where it is concluded that the closer a spread series is to normal then the less the model mispecification error will be in general.

4. Multiple hypothesis testing: Data snooping bias

The objective of the study is to formally and statistically test the performance of the optimal statistical arbitrage trading model of Bertram (2010) described in the previous section for the quantitative trading of spreads in the crude oil and refined products markets. This will inevitably involve the testing of a large number of implementations of the trading model simultaneously. This introduces the well-established issue of data snooping bias, whereby, under naive analysis, profitable trading strategies may be identified by pure chance alone. This links directly to the broader issue of multiple hypothesis testing in statistical and econometric applications. The remainder of the section expands on this, and introduces two key generalised techniques that will be used to control for the problem of data snooping bias.

The issue with multiple hypothesis testing is that the probability of false discoveries, i.e. the rejection of true null hypotheses by chance alone, is often significant. There are a number of approaches described in the literature to deal with this multiple comparisons problem and control for the familywise error rate and variants. Romano *et al.* (2009) provide an excellent summary of the issues and the literature. The familywise error rate (FWER) is defined as the probability that at least one or more false discoveries occur. Consistent with the notation of Romano *et al.* (2009), the following definition is made:

$$
\begin{aligned}
FWER_\theta = P_\theta\{&\text{reject at least one null hypothesis } H_{0,s}\\
&: s \in \mathcal{I}(\theta)\},
\end{aligned}
$$

where $H_{0,s}$, $s = 1, \ldots, S$, is a set of null hypotheses, and $\mathcal{I}(\theta)$ is the set of true null hypotheses. Controlling the FWER involves setting a significance level $\tilde{\alpha}$ and requiring that $FWER_\theta \leq \tilde{\alpha}$. Controlling for the FWER in this way is particularly conservative given that it does not allow even for one false discovery and so is criticised for lacking *power*, where power is loosely defined as the ability to reject false null hypotheses, i.e. identify true discoveries (Romano *et al.* 2009). The greater S, the more difficult it is to make true discoveries.

To deal with this weakness, generalised FWER approaches have been proposed in the literature. The generalised FWER seeks to control for k (where $k \geq 1$) or more false discoveries and, in so doing, allows for greater power in multiple hypothesis testing. The generalised k-FWER is defined as follows:

$$k\text{-}FWER_\theta = P_\theta\{\text{reject at least k null hypotheses } H_{0,s} \\ : s \in \mathcal{I}(\theta)\}.$$

Towards building a framework to identify profitable trading strategies, with statistical significance, on the set of calendar and crack spreads in this study, the following one-sided hypothesis tests are considered:

$$H_{0,s} : \theta_s \leq 0 \quad \text{vs.} \quad H_{1,s} : \theta_s > 0.$$

The objective is to control for the multiple comparisons in this scenario through the generalised familywise error rate, which offers greater power while also implicitly accounting for the dependence structure that exists between the tests. This section continues as follows. Section 4.1 presents a single-step procedure as described by Romano *et al.* (2009). Section 4.2 presents the stepwise procedure of Romano and Wolf (2007), which serves as an improvement on the single-step approach by allowing for subsequent iterative steps to identify additional hypothesis rejections. Finally, section 4.3 presents the balanced stepwise procedure of Romano and Wolf (2010), which is again a marked improvement that allows for balance amongst the hypothesis tests in the sense that each is treated equally in terms of power, i.e. in the identification of true discoveries.

4.1. *Single-step procedure*

Assume a set of test statistics $T_{n,s} = \hat{\theta}_{n,s}$ associated with the hypothesis tests, where n is introduced to denote the sample size of the data used for estimation. Letting $A \equiv \{1, \ldots, S\}$, the single-step procedure proceeds by rejecting all hypotheses where $T_{n,s} \geq c_{n,A}(1 - \tilde{\alpha}, k)$, and where $c_{n,A}(1 - \tilde{\alpha}, k)$ represents the $(1 - \tilde{\alpha})$-quantile of the distribution of k-max$(\hat{\theta}_{n,s} - \theta_s)$ under P_θ. With P_θ unknown, the critical value $c_{n,A}(1 - \tilde{\alpha}, k)$ is also unknown. However, an estimate critical value may be determined using appropriate bootstrapping techniques. That is, the critical value $\hat{c}_{n,A}(1 - \tilde{\alpha}, k)$ is estimated as the $(1 - \tilde{\alpha})$-quantile of the distribution of k-max$(\hat{\theta}^*_{n,s} - \hat{\theta}_{n,s})$ for $\hat{P}_\theta$ an unrestricted estimate of P_θ. See Romano *et al.* (2009) for further technical details.

4.2. *Stepdown procedure*

The stepwise procedure of Romano and Wolf (2007) improves on the single-step procedure described in the previous section by allowing for subsequent iterative steps to identify additional hypothesis rejections. The stepdown procedure is constructed such that, at each stage, information on the rejected hypotheses to date is used in re-testing for significance on the remaining hypotheses. Romano and Wolf (2007) describe the following steps to the algorithm.

- **Step 1:** Let A_1 denote the full set of hypothesis indices, i.e. $A_1 \equiv \{1, \ldots, S\}$. If the maximum test statistic observed, i.e. max$(T_{n,s})$, is less than or equal to the estimated critical value $\hat{c}_{n,A_1}(1 - \tilde{\alpha}, k)$, then fail to reject all null hypotheses and stop the algorithm. Otherwise, proceed to reject all null hypotheses $H_{0,s}$ for which the associated test statistics exceed the critical value level, i.e. where $T_{n,s} > \hat{c}_{n,A_1}(1 - \tilde{\alpha}, k)$.
- **Step 2:** Let R_2 denote the set of indices for the hypotheses rejected in Step 1 and let A_2 denote the indices for those hypotheses not rejected. If the number of elements in R_2 is less than k, i.e. $|R_2| < k$, then stop the algorithm, as the probability of k or more false discoveries is zero in this case. Otherwise, the appropriate critical value to be applied at this stage is calculated as follows:

$$\hat{d}_{n,A_2}(1 - \tilde{\alpha}, k) = \max_{I \subseteq R_2,\, |I| = k-1} \{\hat{c}_{n,K}(1 - \tilde{\alpha}, k) : K \equiv A_2 \cup I\}.$$

Hence, additional hypotheses from A_2 are rejected if $T_{n,s} > \hat{d}_{n,A_2}(1 - \tilde{\alpha}, k)$, $s \in A_2$. If no further rejections are made, then stop the algorithm.

⋮

- **Step j:** Let R_j denote the set of indices for the hypotheses rejected up to Step $(j - 1)$ and let A_j denote the indices for those hypotheses not rejected. The critical value to be applied at this stage is calculated as follows:

$$\hat{d}_{n,A_j}(1 - \tilde{\alpha}, k) = \max_{I \subseteq R_j,\, |I| = k-1} \{\hat{c}_{n,K}(1 - \tilde{\alpha}, k) : K \equiv A_j \cup I\}.$$

Hence, additional hypotheses from A_j are rejected if $T_{n,s} > \hat{d}_{n,A_j}(1 - \tilde{\alpha}, k)$, $s \in A_j$. If no further rejections are made, then stop the algorithm.

⋮

From the description of the above algorithm, at each stage j in the stepwise procedure, the hypotheses that are not rejected thus far are re-tested over a smaller population of hypothesis tests than previously. The size of this smaller population is given by $(|A_j| + k - 1)$, which includes all the hypotheses within A_j, in addition to $(k - 1)$ hypotheses drawn from those already rejected, i.e. drawn from R_j. Given that control of the generalised k-FWER is the premise of the procedure, it is expected that there are at most $(k-1)$ false discoveries amongst the set of hypotheses rejected R_j. However, it is not known which of the rejected hypotheses may represent false discoveries. Hence, it is necessary to circulate through all combinations of R_j, of size $(k-1)$, in order to obtain the maximum critical value $\hat{d}_{n,A_j}(1-\tilde{\alpha},k)$ against which to test the hypotheses within A_j. See Romano and Wolf (2007) for further technical details.

4.2.1. Operative method. In requiring to circulate through all subsets of R_j, of size $(k-1)$, in order to obtain the maximum critical value to apply at each stage of the stepdown procedure, the algorithm can become highly, if not excessively, computationally burdensome. Depending on $|R_j|$ and the value of k, the number of combinations ${}^{|R_j|}C_{k-1}$ can become very large. Romano and Wolf (2007) therefore suggest an operative method that reduces this computational burden, while at the same time maintaining much of the attractive properties of the algorithm.†

For this, first consider the hypothesis tests rejected up to step $(j-1)$ and place these in descending order of test statistic, i.e.

$$T_{n,r_1} \geq T_{n,r_2} \geq \cdots \geq T_{n,r_{|R_j|}},$$

where $\{r_1, r_2, \ldots, r_{|R_j|}\}$ is the appropriate permutation of associated hypothesis test indices that gives this ordering. Now consider a user-defined maximum number of combinations, $N_{\max}$, at each step of the algorithm. Then choose an integer value such that ${}^{M}C_{k-1} \leq N_{\max}$ and replace the critical value calculation at each step j of the algorithm with the following:

$$\hat{d}_{n,A_j}(1-\tilde{\alpha},k) = \max_{I\subseteq\{r_{\max(1,|R_j|-M+1)},\ldots,r_{|R_j|}\},\,|I|=k-1} \{\hat{c}_{n,K}(1-\tilde{\alpha},k) : K \equiv A_j \cup I\}.$$

What this serves to do is to replace circulating through all the hypothesis tests rejected to date with that of circulating through only the M least-significant hypothesis tests rejected. Of course, in the case where $M \geq |R_j|$, then this amounts to circulating through all the hypotheses rejected. Although this approach is premised on the assumption that the (up to $k-1$) false discoveries lie within the least-significant hypotheses rejected so far, it does offer significant computational efficiencies for the algorithm. It is this operative method that is used for the empirical analysis in subsequent sections.

4.3. Balanced stepdown procedure

Whereas the stepwise procedure of the previous section is an improvement on the single-step procedure of section 4.1, it does not offer by construction balance in the sense that each hypothesis test is treated equally in terms of power. The balanced stepwise procedure of Romano and Wolf (2010) addresses this issue.

Introducing some notation, let $H_{n,s}(\cdot,\, P_\theta)$ denote the distribution function of $(\hat{\theta}_{n,s}-\theta_s)$ and let $c_{n,s}(\tilde{\gamma})$ denote the $\tilde{\gamma}$-quantile of this distribution. The confidence interval

$$\{\theta_s : \hat{\theta}_{n,s} - \theta_s \leq c_{n,s}(\tilde{\gamma})\}$$

then has coverage probability $\tilde{\gamma}$. Balance is the property that the marginal confidence intervals for a population of S simultaneous hypothesis tests have the same probability coverage. Within the context of controlling the generalised k-FWER, the overall objective is to ensure that the simultaneous confidence interval covers all parameters θ_s, $s=1,\ldots,S$, except for at most $(k-1)$ of them, for a given limiting probability $(1-\tilde{\alpha})$, while at the same time ensuring balance (at least asymptotically). So, what is sought is that

$$\begin{aligned}
&P_\theta\{\hat{\theta}_{n,s}-\theta_s \leq c_{n,s}(\tilde{\gamma}) \text{ for all but at most } (k-1)\\
&\quad\text{of the hypotheses}\}\\
&\quad\equiv P_\theta\{H_{n,s}(\hat{\theta}_{n,s}-\theta_s, P_\theta) \leq \tilde{\gamma} \text{ for all but at most } (k-1)\\
&\quad\text{of the hypotheses}\}\\
&\quad\equiv P_\theta\{k\text{-}\max(H_{n,s}(\hat{\theta}_{n,s}-\theta_s, P_\theta)) \leq \tilde{\gamma}\} = 1-\tilde{\alpha}.
\end{aligned}$$

Letting $L_{n,\{1,\ldots,S\}}(k, P_\theta)$ denote the distribution of $k\text{-}\max(H_{n,s}(\hat{\theta}_{n,s}-\theta_s, P_\theta))$, the appropriate choice of the coverage probability $\tilde{\gamma}$ is then $L^{-1}_{n,\{1,\ldots,S\}}(1-\tilde{\alpha},k,P_\theta)$.

As before, given that P_θ is unknown, it is necessary to use appropriate bootstrapping techniques to generate an estimate of the coverage probability $L^{-1}_{n,\{1,\ldots,S\}}(1-\tilde{\alpha},k,P_\theta)$, under $\hat{P}_\theta$. Therefore, from this development it is possible to define the simultaneous confidence interval

$$\{\theta_s : \hat{\theta}_{n,s}-\theta_s \leq H^{-1}_{n,s}(L^{-1}_{n,\{1,\ldots,S\}}(1-\tilde{\alpha},k,\hat{P}_\theta),\hat{P}_\theta)\}.$$

The right-hand side of the above inequality will form the basis of the critical value definitions used within the stepdown procedure. See Romano and Wolf (2010) for further technical details. Note that although the above development was made assuming the full set of hypothesis tests, it equally applies to any subset $K \subseteq \{1,\ldots,S\}$. Hence, the balanced stepwise algorithm may now be described as follows.

- **Step 1:** Let A_1 denote the full set of hypothesis indices, i.e. $A_1 \equiv \{1,\ldots,S\}$. If, for each hypothesis test, the associated test statistic $T_{n,s}$ is less than or equal to the corresponding critical value estimate $\hat{c}_{n,A_1,s}(1-\tilde{\alpha},k) \equiv H^{-1}_{n,s}(L^{-1}_{n,A_1}(1-\tilde{\alpha}, k,\hat{P}_\theta),\hat{P}_\theta)$, then fail to reject all null hypotheses and stop the algorithm. Otherwise, proceed to reject all null hypotheses $H_{0,s}$ for which the associated test statistics exceeds the critical value level, i.e. where $T_{n,s} > \hat{c}_{n,A_1,s}(1-\tilde{\alpha},k)$.

†The generic algorithm offers a number of attractive features. Firstly, the generic algorithm is conservative in its rejection of hypotheses. Secondly, the generic algorithm also allows for finite sample control of the k-FWER under P_θ. And thirdly, the bootstrap construction is such that the generic algorithm provides asymptotic control in the case of contiguous alternatives. Romano and Wolf (2007) provide a more detailed discussion.

- **Step 2:** Let R_2 denote the set of indices for the hypotheses rejected in Step 1 and let A_2 denote the indices for those hypotheses not rejected. If the number of elements in R_2 is less than k, i.e. $|R_2|<k$, then stop the algorithm, as the probability of k or more false discoveries is zero in this case. Otherwise, the appropriate critical value to be applied for each hypothesis test s at this stage is calculated as follows:

$$\hat{d}_{n,A_2,s}(1-\tilde{\alpha},k) = \max_{I\subseteq R_2,\,|I|=k-1} \{\hat{c}_{n,K,s}(1-\tilde{\alpha},k) : K \equiv A_2 \cup I\}.$$

  Hence, additional hypotheses from A_2 are rejected if $T_{n,s} > \hat{d}_{n,A_2,s}(1-\tilde{\alpha},k)$, $s \in A_2$. If no further rejections are made, then stop the algorithm.

  ⋮

- **Step j:** Let R_j denote the set of indices for the hypotheses rejected up to Step $(j-1)$ and let A_j denote the indices for those hypotheses not rejected. The appropriate critical value to be applied for each hypothesis test s at this stage is calculated as follows:

$$\hat{d}_{n,A_j,s}(1-\tilde{\alpha},k) = \max_{I\subseteq R_j,\,|I|=k-1} \{\hat{c}_{n,K,s}(1-\tilde{\alpha},k) : K \equiv A_j \cup I\}.$$

  Hence, additional hypotheses from A_j are rejected if $T_{n,s} > \hat{d}_{n,A_j,s}(1-\tilde{\alpha},k)$, $s \in A_j$. If no further rejections are made, then stop the algorithm.

  ⋮

Similar to the stepwise algorithm of the previous section, at each stage j in the stepwise procedure, the hypotheses that are not rejected thus far are re-tested over a smaller population of hypothesis tests than previously. The size of this smaller population is given by $(|A_j|+k-1)$, which includes all the hypotheses within A_j, in addition to $(k-1)$ hypotheses drawn from those hypotheses already rejected, i.e. drawn from R_j. Given that control of the generalised k-FWER is the premise of the procedure, it is expected that there are at most $(k-1)$ false discoveries amongst the set of hypotheses rejected R_j. However, it is not known which of the rejected hypotheses may represent false discoveries. Hence, it is necessary to circulate through all combinations of R_j, of size $(k-1)$, in order to obtain the appropriate critical values. Where the algorithm departs significantly from the previous section is that a maximum critical value $\hat{d}_{n,A_j,s}(1-\tilde{\alpha},k)$ must be determined for each hypothesis test s. This adds an additional layer of computational burden on the algorithm.

4.3.1. Operative method.

Similar to the stepdown procedure of section 4.2, the need to circulate through all subsets of R_j, of size $(k-1)$, in order to obtain, in this case, a set of maximum critical values to apply at each stage of the stepdown procedure, means the algorithm can become excessively computationally burdensome. Romano and Wolf (2010) therefore suggest an operative method that reduces this computational burden in the spirit of that proposed by the authors for the stepdown procedure (Romano and Wolf 2007).

It is first necessary to be able to order the hypothesis tests rejected up to step $(j-1)$ in terms of significance. To this end, it is noted that marginal p-values can be obtained as follows:

$$\hat{p}_{n,s} \equiv 1 - H_{n,s}(\hat{\theta}_{n,s}, \hat{P}_\theta).$$

This gives the following ascending order for the significance of the hypothesis tests:

$$\hat{p}_{n,r_1} \le \hat{p}_{n,r_2} \le \cdots \le \hat{p}_{n,r_{|R_j|}},$$

where $\{r_1, r_2, \ldots, r_{|R_j|}\}$ is the appropriate permutation of associated hypothesis test indices that gives this ordering. As before, a maximum number of combinations, $N_{\max}$, at each step of the algorithm is defined. Then an integer value M is chosen such that ${}^{M}C_{k-1} \le N_{\max}$, leading to the calculation of the critical values as follows:

$$\hat{d}_{n,A_j,s}(1-\tilde{\alpha},k) = \max_{I\subseteq\{r_{\max(1,|R_j|-M+1)},\ldots,r_{|R_j|}\},\,|I|=k-1} \{\hat{c}_{n,K,s}(1-\tilde{\alpha},k) : K \equiv A_j \cup I\}.$$

The rationale for this approach is exactly the same as that described in section 4.2.1, with the same effect being the introduction of significant computational efficiencies to the algorithm. It is this operative method that is used for the empirical analysis in subsequent sections.

5. Outline of empirical analysis

With the objective being to formally test the performance of the optimal statistical arbitrage trading model for the quantitative trading of spreads in the crude oil and refined products markets, data snooping bias presents a real issue that needs to be controlled. Furthermore, the performance tests are correlated by nature of the pairings and the markets considered, and so any testing procedure needs to be sufficiently flexible to account for this dependence structure. The stepdown and balanced stepdown procedures described in the previous section offer a more generalised and flexible approach to controlling data snooping bias than the commonly used reality check bootstrap test of White (2000) or the superior predictive ability test of Hansen (2005). As identified by Romano *et al.* (2009), both tests lack power in the sense that they control only the familywise error rate (i.e. $k=1$), and are further restrictive by being only single-step procedures. Additionally, the reality check bootstrap test of White (2000) does not consider studentised test statistics and so by construction suffers from lack of balance. The stepdown and balanced stepdown procedures, in contrast, control the generalised familywise error rate using stepwise procedures, with the latter also offering balance by construction. Given that the test statistics used in this study are not studentised, the balanced stepdown procedure is considered superior to the stepdown procedure as it is asymptotically equivalent to studentisation. Hence, as each test is treated equally, from a practitioner perspective the balanced stepdown procedure shows no bias between the trading strategies considered. This novel application of the stepdown and balanced stepdown procedures to the problem of quantitative trading within the energy markets represents a key contribution of this study to the literature.

In implementing the stepdown and balanced stepdown procedures, it is first necessary to identify the formal hypothesis tests that are to be be applied. The approach taken in this study is to formally identify those spread trading strategies that, with statistical significance, outperform a given benchmark in terms of mean daily log-return. The benchmark is defined as in Hsu and Kuan (2005) as equivalent to taking no position in the spread, and so the hypothesis tests look to identify departures from the zero mean daily log-return. Letting ζ_s denote the daily log-return of trading strategy s and $\theta_s \equiv E(\zeta_s)$, then the hypothesis tests may be formalised as follows:

$$H_{0,s} : \theta_s \leq 0 \quad \text{vs.} \quad H_{1,s} : \theta_s > 0,$$

for a full set of hypothesis tests $\{1, \ldots, S\}$, where S is set equal the number of tests considered. The appropriate estimate of θ_s is the mean daily log-return observed on trading strategy s over a given historical period with n daily observations. Letting $\zeta_{t,s}$, $t = 1, \ldots, n$, denote the daily log-return of trading strategy s at time t, the estimate $\hat{\theta}_{n,s} = \sum_{t=1}^{n} \zeta_{t,s}/n$. Given that θ_s is unknown, implementing the stepdown and balanced stepdown procedures requires use of appropriate bootstrapping techniques. This involves replacing the true specification $(\hat{\theta}_{n,s} - \theta_s)$ with the estimates $(\hat{\theta}^*_{n,s}(b) - \hat{\theta}_{n,s})$, where $\hat{\theta}^*_{n,s}(b)$ are bootstrap estimates of θ_s and $b = 1, \ldots, B$ are the indices for the bootstrap samples.

With the hypothesis tests formalised, the next stage is to implement the optimal statistical arbitrage trading model on each spread within the data set. For this it is necessary to fit the following general OU process to each spread time series:

$$\mathrm{d}s_t = \alpha(\mu - s_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \qquad (2)$$

in order to obtain parameter estimates $\hat{\alpha}$, $\hat{\mu}$ and $\hat{\sigma}$ for generating the entry and exit trading signals as described in section 3. Adjusting the signals from the trading model for the non-zero long-run mean level $\hat{\mu}$ is straightforward, giving the effective entry and exit levels $\hat{a} \equiv a + \hat{\mu}$ and $\hat{m} \equiv m + \hat{\mu}$, respectively, where by definition of a and m, $\hat{a} < \hat{m}$. To optimise trading performance and exploit the mean-reversion fully (where it exists in the data), trades will be executed to exploit the transition from $\hat{a} \rightarrow \hat{m}$ and the transition back from $\hat{m} \rightarrow \hat{a}$. Therefore, $\hat{a}$ and $\hat{m}$ will herein be referred to merely as trading signals, alternating interpretation between entry and exit signals. The trading strategies are constructed to go long (short) the spread when the trading signal $\hat{a}$ ($\hat{m}$) is breached and then close out the position once $\hat{m}$ ($\hat{a}$) is breached. In this context, it is important to emphasise that the results reported in the next section reflect the aggregation of these *two* trading approaches.

A number of alternative in-sample and out-of-sample periods are considered for estimation and evaluation of the spread trading strategies. Specifically, eight separate one-year out-of-sample periods are considered for evaluation of trading strategy performance, namely, each year over the period 2003–2010. These choices allow for a wider testing of the optimal statistical arbitrage trading model and an examination of trading profitability over time. Three separate in-sample periods are then considered for the estimation of the trading models and the generation of trading signals to be applied out-of-sample. One-year, two-year and three-year periods are considered for estimation, where, for convenience of construction, 252 trade days are assumed in each year. So, for example, the two-year in-sample period includes the 2×252 trade dates prior to the start of the out-of-sample period. Given the daily frequency of the data these choices are seen as reasonable to capture consistent mean-reversion effects, while at the same time examining the impact of alternative estimation periods. Hence, for each out-of-sample period, a total of 2583 (i.e. 3×861) individual trading strategies are tested simultaneously.

To complete the set up of the empirical analysis, it is finally necessary to discuss the choice of generalising parameter k and the probability parameter $\tilde{\alpha}$ to be used within the stepwise and balanced stepwise procedures. To ensure tight control of the number of false discoveries, while at the same time offering power to the tests, k is chosen to ensure that no more than 1% of the tests considered represent false discoveries. Hence, based on a population of $S = 2583$ tests, k is set equal to $\lceil S \times 1\% \rceil = 26$. The significance level $\tilde{\alpha}$ chosen is 5%, such that the implementation of the stepwise and balanced stepwise procedures amount to ensuring that

$$P_\theta\{k = 26 \text{ or more false discoveries}\} \leq \tilde{\alpha} = 5\%.$$

5.1. *Remark on transaction costs*

Transaction costs are considered and incorporated into the analysis via the transaction costs parameter c detailed in section 3. Given the unavailability to the authors of detailed bid–ask spread information on the underlying futures contracts over the full sample period, the authors draw on the literature to estimate transaction costs. Laws *et al.* (2008) apply filter techniques for the trading of spreads comprised from WTI, Brent and HO using data over the period 1995–2005. The relative transactions costs reported by the authors for these three commodities are presented in table 1. These transaction costs represent the average of four intraday spreads. Without information on bid–ask spreads in the GO market, the relative transaction costs of the HO market are assumed to be a reasonable proxy. Hence, for each spread transaction, the total relative transactions costs are given by the sum of the costs for the individual legs.

The transaction costs are assumed to apply across the full futures curve, which eases the analysis but is clearly restrictive in the sense that transaction costs are likely to increase with increasing maturity. Additionally, given that

Table 1. Estimated relative transactions costs.

WTI	0.0940%
Brent	0.0289%
Heating oil	0.2482%
Gas oil	0.2482%

the sample period of Laws *et al.* (2008) does not coincide with this study, the transactions costs assumed may not be representative. In light of these limitations, the robustness of the empirical results to varying the transaction costs parameter *c* is examined at the close of the next section. In particular, a range of higher transaction cost levels are assumed and the analysis repeated in order to determine the impact on average daily returns and the number of profitable trades identified.

5.2. *Remark on liquidity*

An additional issue of practitioner relevance for this study is the level of liquidity across the various futures curves. The ability to execute large volumes of trades based on quantitative trading signals is a key consideration for industry. With only volume data available to the authors, volume-based measures are used here to provide insights into average liquidity across the sample data. Two common measures are calculated for each commodity and maturity futures contract, broken down by each of the 2003–2010 trading years. These measures are the *five-day turnover rate* and the *five-day Hui-Heubel liquidity ratio*. A more comprehensive study of liquidity would be extremely beneficial but, due to data constraints, is beyond the scope of this paper. The issue of liquidity is not considered a source of bias for the stylised trading analysis of this paper given the unit volumetric assumption made in the spread trading design.

Sarr and Lybek (2002) provide a detailed overview of liquidity measures, including the two volume-based measures considered here. The turnover rate captures the number of times the outstanding volume changes hands over a specified period. The five-day turnover rate (TR) is used here to give a measure of liquidity over five consecutive trade dates and is defined formally as follows:

$$TR = V/(S \times \bar{P}),$$

where S is the outstanding volume, $\bar{P}$ is the average closing price over the five days and V is the monetary volume value, given by

$$V = \sum_i P_i \times Q_i,$$

where P_i and Q_i are the closing prices and trade volume on date i. The greater the turnover rate, then the greater the depth to the market.

The Hui-Heibel liquidity ratio, on the other hand, seeks to incorporate the impact of trading on price. It considers the percentage price differential between the highest and lowest prices achieved over a specified period relative to the monetary volume value. The five-day Hui-Heubel ratio is used and is defined formally as follows:

$$L_{\text{hh}} \equiv [(P_{\max} - P_{\min})/P_{\min}]/TR,$$

where $P_{\max}$ and $P_{\min}$ are the maximum and minimum closing prices recorded over the five trade days and TR is the five-day turnover rate already defined. The lower the value of the Hui-Heubel liquidity ratio, then the greater the level of liquidity.

Tables A1–A4 present, for each commodity and each trading year, the average values of the calculated turnover and Hui-Heubel measures. For the most part, all four commodities show declining liquidity across the futures curve, with high liquidity over the short-dated contracts in particular. Despite the differences in the size of the various markets, there are relatively comparable levels of liquidity across the futures curves based on both measures. The turnover rates also appear to suggest increasing levels of liquidity over the eight-year period, with the Hui-Heubel liquidity ratio results showing broad agreement with this observation.

6. Empirical results

Performing the empirical analysis as set out in the previous section leads to a number of interesting observations. Tables 2 and 3 present trading performance results based on applying the stepdown and balanced stepdown procedures respectively to control for data snooping bias. For each trading year, average daily returns, Sharpe ratios and trade lengths are reported for three specific categories, namely the top 10, top 20 and all trading strategies identified by the procedures as being profitable with statistical significance. Also reported for each category are the average estimates for the structural parameters of the underlying OU process. The final column in each table gives the total number of profitable trading strategies identified in each year.

It is again important to emphasise that the results reported reflect the aggregation of *two* trading approaches, one taking long positions in the spreads to exploit upward movements between the trading signals and the second taking short positions to exploit downward movements between the trading signals. Many profitable trading strategies are identified, which in many instances report Sharpe ratios that exceed 2 and in some instances are even close to 4. For the top 10 and top 20 categories, average daily returns fall within the approximate range of 0.07–0.35% for most years, with 2009 showing exceptional results in the range of 0.5–0.55% driven by the locational HO–GO spreads. The associated trade lengths lie in the approximate range of 9–55 trade dates, with the shortest horizons being reported for 2009. The years 2003 and 2009 prove to be particularly successful relative to other years with Sharpe ratios close to or in excess of 3. For the years 2008 and 2010, the stepdown procedure actually fails to identify

Table 2. Average empirical results: Stepdown procedure.

Year		Avg. daily ret.	Avg. Sharpe ratio	Avg. trade length (days)	Avg. α	Avg. μ	Avg. σ	# Profitable strategies
2003	Top 10	0.00361	2.94	15.09	59.367	−0.094	0.259	
	Top 20	0.00345	2.92	15.81	53.443	−0.103	0.252	
	All	0.00345	2.92	15.81	53.443	−0.103	0.252	15
2004	Top 10	0.00299	2.44	18.74	50.529	−0.098	0.295	
	Top 20	0.00275	2.30	18.88	52.578	−0.096	0.279	
	All	0.00265	2.24	19.58	54.128	−0.099	0.275	28
2005	Top 10	0.00259	2.33	19.08	26.167	−0.169	0.275	
	Top 20	0.00253	2.22	19.81	23.461	−0.172	0.280	
	All	0.00253	2.22	19.81	23.461	−0.172	0.280	14
2006	Top 10	0.00241	2.41	19.36	34.984	−0.078	0.287	
	Top 20	0.00220	2.19	21.28	34.765	−0.092	0.289	
	All	0.00195	1.98	24.22	33.899	−0.102	0.279	35
2007	Top 10	0.00293	2.93	12.62	83.288	−0.127	0.233	
	Top 20	0.00272	2.75	13.33	87.855	−0.120	0.226	
	All	0.00233	2.32	16.22	79.593	−0.121	0.227	54
2008	Top 10	–	–	–	–	–	–	–
	Top 20	–	–	–	–	–	–	–
–	All	–	–	–	–	–	–	–
2009	Top 10	0.00547	3.95	9.13	92.993	0.010	0.260	–
–	Top 20	0.00499	3.66	10.05	86.505	0.007	0.256	–
–	All	0.00456	3.32	11.14	95.267	0.008	0.258	31
2010	Top 10	–	–	–	–	–	–	–
–	Top 20	–	–	–	–	–	–	–
–	All	–	–	–	–	–	–	–

Table 3. Average empirical results: Balanced procedure.

Year		Avg. daily ret.	Avg. Sharpe ratio	Avg. trade length (days)	Avg. α	Avg. μ	Avg. σ	# Profitable strategies
2003	Top 10	0.00354	2.96	14.28	65.130	−0.089	0.263	
	Top 20	0.00314	2.87	14.84	66.133	−0.105	0.253	
	All	0.00164	2.16	27.24	28.997	−0.078	0.183	142
2004	Top 10	0.00298	2.47	17.58	53.573	−0.096	0.281	
	Top 20	0.00272	2.34	20.09	47.738	−0.103	0.279	
	All	0.00174	1.84	28.21	33.896	−0.068	0.197	52
2005	Top 10	0.00258	2.33	19.46	25.491	−0.167	0.273	
	Top 20	0.00246	2.22	20.71	23.742	−0.160	0.276	
	All	0.00202	2.05	23.71	23.444	−0.170	0.251	71
2006	Top 10	0.00241	2.41	19.36	34.984	−0.078	0.287	
	Top 20	0.00220	2.19	21.28	34.765	−0.092	0.289	
	All	0.00164	2.03	24.93	36.322	−0.093	0.237	46
2007	Top 10	0.00293	2.93	12.62	83.288	−0.127	0.233	
	Top 20	0.00272	2.75	13.33	87.855	−0.120	0.226	
	All	0.00199	2.18	23.83	66.403	−0.120	0.219	82
2008	Top 10	0.00244	1.86	18.26	99.863	0.001	0.232	
	Top 20	0.00244	1.86	18.26	99.863	0.001	0.232	
	All	0.00244	1.86	18.26	99.863	0.001	0.232	8
2009	Top 10	0.00547	3.95	9.13	92.993	0.010	0.260	
	Top 20	0.00499	3.66	10.05	86.505	0.007	0.256	
	All	0.00264	2.43	24.38	46.336	0.011	0.174	63
2010	Top 10	0.00092	1.91	43.88	55.796	−0.014	0.262	
	Top 20	0.00066	1.82	55.43	30.522	−0.004	0.184	
	All	0.00060	1.73	56.54	26.737	−0.005	0.164	23

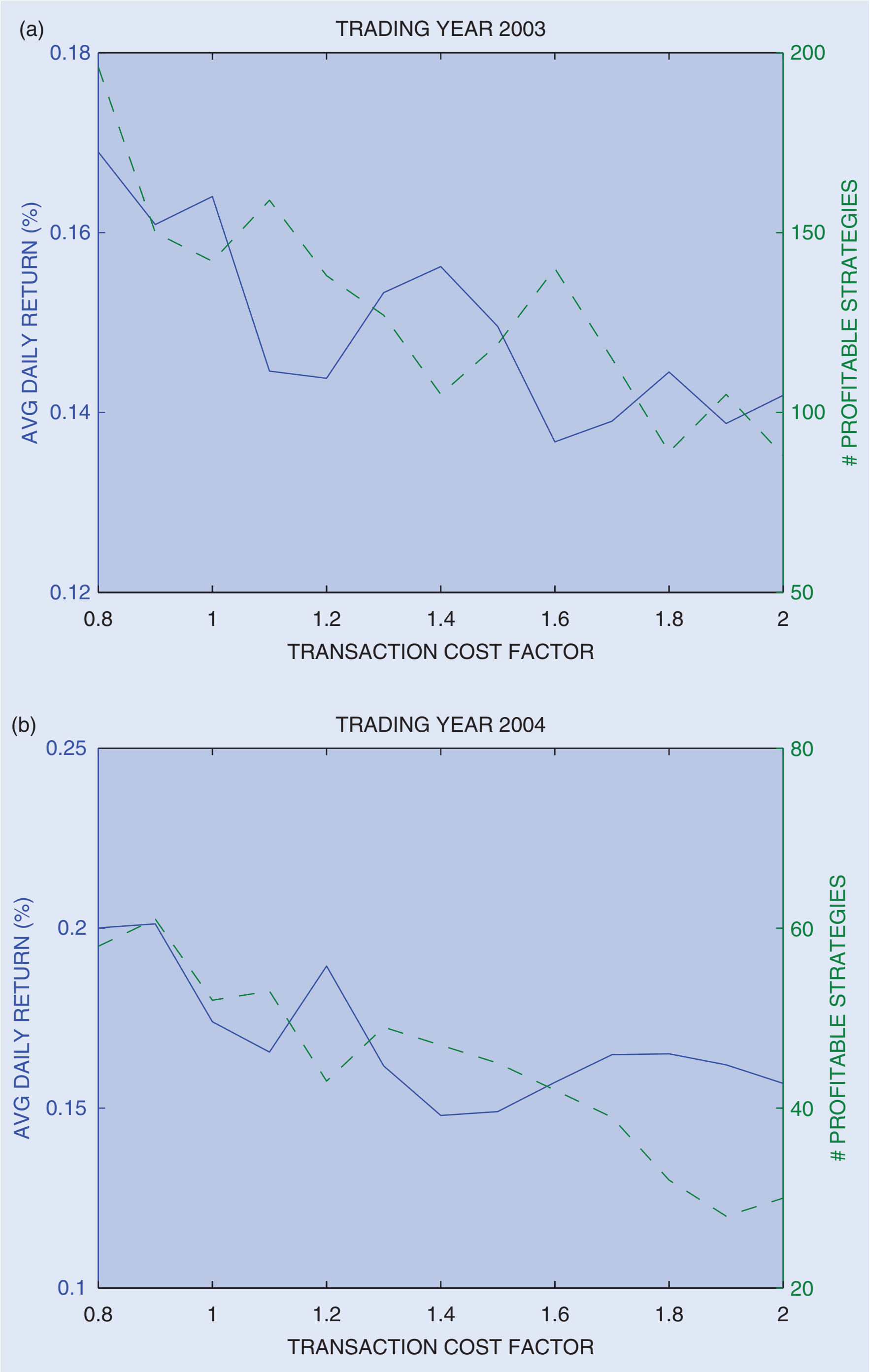

Figure 1. Transaction cost robustness: 2003 and 2004.

any profitable strategies, with the balanced stepdown procedure only extracting eight and 23 profitable strategies, respectively. The numbers are much lower than in other years. Average daily returns in excess of 0.2% are reported for 2008 but the associated Sharpe ratios are seen to fall below 2 on average. The discussion to follow will explore further these observations for 2008 in the context of the differences between the stepdown and balanced stepdown procedures. Similar poor performance is observed for the year 2010, where average daily returns can be seen to be at the lower range of approximately 0.07%, with Sharpe ratios again below 2 on average.

Substantial differences are observed between the stepdown and balanced stepdown procedures in terms of the total number of profitable trading strategies identified, despite showing broad consistency within the top 10 and top 20 categories. The balanced stepdown procedure manages to identify between 1.5 and 9.5 times the number of profitable strategies and succeeds in identifying strategies in 2008 and 2010 where the stepdown procedure fails. This observation reflects the manner in which the balanced stepdown procedure treats each hypothesis test equally in terms of power, whereas the stepdown procedure is biased towards those trading strategies with larger

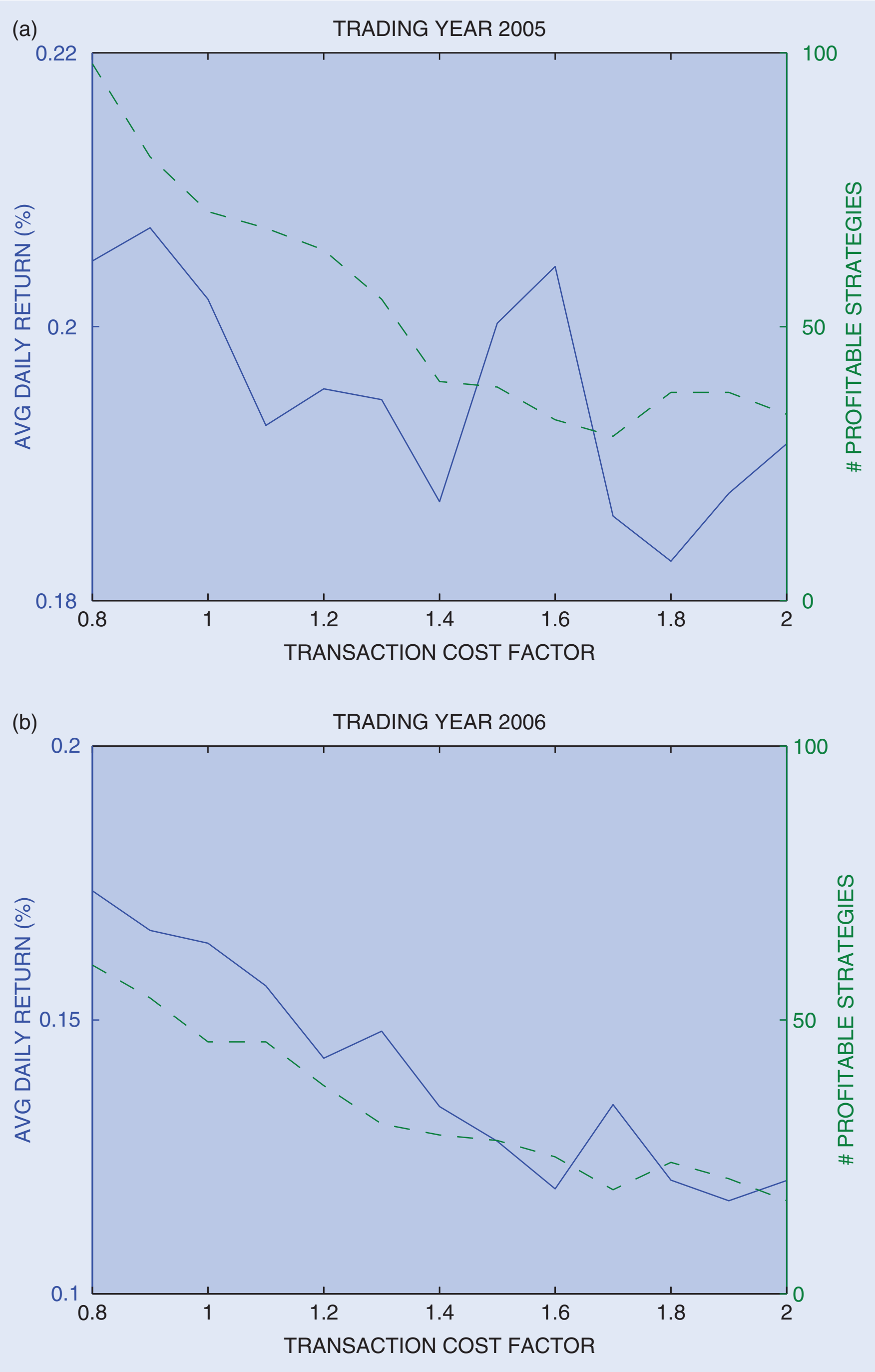

Figure 2. Transaction cost robustness: 2005 and 2006.

average daily returns. The dramatic reduction in the number of profitable strategies in 2008 relative to 2007 and the years previous reflects the collapse in crude oil and other commodity prices resulting from the credit crisis shock and concerns over global commodity demand. The effect of the collapse in commodity prices manifests as a structural shift in the range of the spreads over 2008 relative to 2007, 2006 and 2005. Therefore, large divergences from the long-run mean levels estimated in-sample are observed for the majority of spreads. These divergences lead to significant losses for the majority of trading strategies out-of-sample in 2008.

For further insights, appendix B presents the top 10 tradable spreads identified each year and gives the associated average daily return, Sharpe ratio and trade length results. The simplicity of the OU process means parameter estimation may be performed using OLS linear regression. In addition to parameter estimates, *F*-test *p*-values are provided to give an indication of goodness-of-fit of the model. Crack spreads and HO–GO locational spreads can be seen to dominate the composition of the top 10 strategies. The crack spreads include the common-exchange WTI–HO and Brent–GO cracks in addition to the cross-exchange WTI–GO crack. For the most part, the spreads involve either common, adjacent or at least

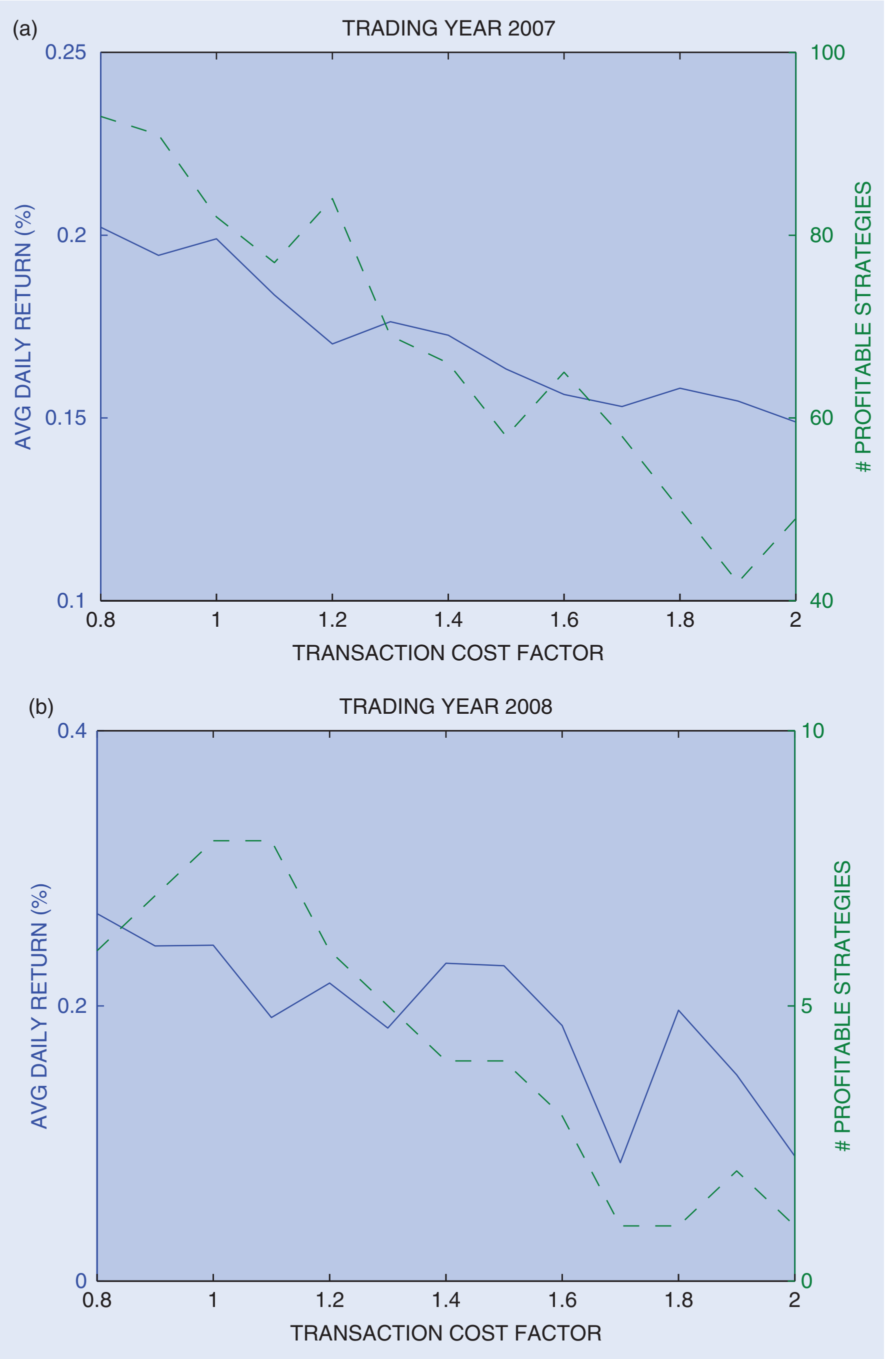

Figure 3. Transaction cost robustness: 2007 and 2008.

near-by dated contracts. Interestingly, not a single calendar spread makes up the top 10 strategies for any of the years, while locational Brent–WTI spreads only appear in the mix for 2010 (with the exception of one observation in 2003). In terms of the in-sample estimation period used to generate the trading signals, the top 10 profitable strategies do not show any particular pattern. The one-, two- and three-year in-sample estimation periods are all represented within the top 10 strategy groups. Across the eight years, the one-year in-sample period does marginally come out on top, representing approximately 40% of the profitable strategies, with more or less an even split between the two- and three-year estimation periods.

6.1. Robustness to transaction costs

As outlined in section 5, the levels of transaction costs assumed for the analysis thus far have been taken from Laws *et al.* (2008), with the HO transaction costs assumed to apply for the GO market as well. Given that the sample period used by these authors does not coincide with this study, the transactions costs assumed may not be representative. Therefore, this section now examines the robustness of trading strategy performance to varying transaction cost levels. The transaction parameter *c* used in the trading model is allowed to range between lower and higher levels than those outlined in table 1, with the

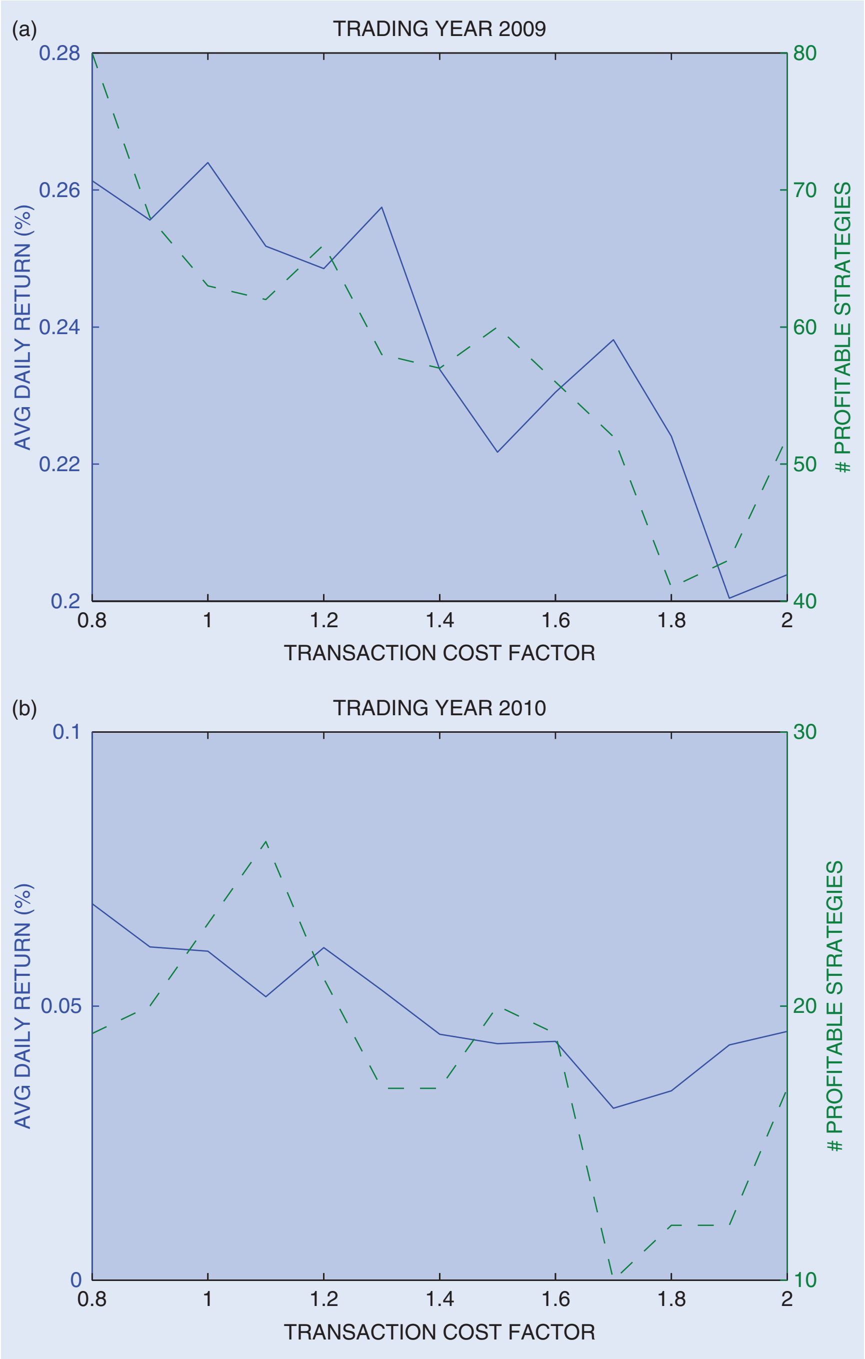

Figure 4. Transaction cost robustness: 2009 and 2010.

objective being to determine the impact on average daily returns and the number of profitable trades identified. For ease of analysis and exposition, a set of transaction cost factors is considered. Each factor is applied to the transaction costs in table 1 and the empirical analysis of the previous section is then repeated. The set of transaction cost factors assumed is $\{0.8, 0.9, \ldots, 1.9, 2\}$, allowing transaction costs to vary between 80% and 200% of those assumed so far.

Figures 1–4 present plots of the average daily returns and the number of profitable trading strategies for each out-of-sample trading year considered. Broadly speaking, it can be seen that the returns and the number of strategies decline with increasing transaction cost levels. However, these declines are not monotonic, reflecting that the calculation of the trading signals is a function of the transaction cost parameter c as outlined in section 3. So as the transaction cost parameter c varies, then so too does the positioning of the signals. Between the lowest and highest transaction cost factors (i.e. 0.8 and 2, respectively), average daily returns in annualised terms decline by approximately 10–15% for 2004, 2006, 2007 and 2009 and by 4–7% for 2003, 2005 and 2010. In the case of trading year 2008, it has been seen already that the number of profitable strategies collapses relative to other years, with only eight trading strategies identified as profitable from the full data set. For the highest transaction cost factor this reduces to a single strategy. The decline in average daily returns in annualised terms between the lowest and highest transaction cost factors is a significant 44%.

7. Conclusion

This study examines the quantitative trading of spreads in the crude oil (WTI and Brent) and refined products (heating oil and gas oil) markets, making a number of contributions to the literature. Firstly, the novel statistical arbitrage trading model of Bertram (2010) is applied to a wide range of spreads (including calendar, crack and locational spreads), representing a comprehensive empirical analysis of this model. The model leads to profitable spread trading and it is shown that performance is quite robust to varying transaction costs. Secondly, generalised stepwise procedures are used to control for data snooping bias within the quantitative trading application. The stepdown procedure of Romano and Wolf (2007) and the balanced stepdown procedure of Romano and Wolf (2010) are applied, both serving as improvements over more conservative single-step approaches, such as the reality check bootstrap test of White (2000) and the superior predictive ability test of Hansen (2005). The generalised procedures offer greater power to reject false null hypotheses, with the balanced stepdown procedure offering equal treatment in the identification of profitable strategies. Profitable trading strategies are identified, with results reflecting the aggregation of taking long and short positions in the spreads. For the top 10 and top 20 categories, average daily returns fall within the approximate range of 0.07–0.55%, with trade lengths of 9–55 days and Sharpe ratios of between 2 and 4 in many cases. Thirdly, the study allows for practical comparison of the stepwise and the balanced stepwise procedures in the context of a trading application. The balanced stepdown procedure is unbiased in its approach and is shown to identify many more profitable trading strategies compared with the non-balanced stepdown procedure. For instance, a collapse in the number of profitable trading strategies is seen in 2008, reflecting the impact of the credit crisis and the distortion of spreads relative to previous years. Whereas the stepdown procedure fails to identify any profitable strategies, the balanced procedure is successful in doing so.

References

Aldridge, I., *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*, 2009 (Wiley: Hoboken, NJ).

Alizadeh, A.H. and Nomikos, N.K., Cost of carry, causality and arbitrage between oil futures and tanker freight markets. *Transport. Res. Part E*, 2004, **40**, 297–316.

Andrade, S., di Pietro, V. and Seasholes, M., Understanding the profitability of pairs trading. Working Paper, UC Berkley, 2005.

Avellaneda, M. and Lee, J.H., Statistical arbitrage in the US equities market. *Quant. Finance*, 2010, **10**, 761–782.

Bertram, W.K., Analytic solutions for optimal statistical arbitrage trading. *Physica A*, 2010, **389**, 2234–2243.

Bowen, D., Hutchinson, M.C. and O'Sullivan, N., High-frequency equity pairs trading: Transaction costs, speed of execution, and patterns in returns. *J. Trading*, 2010, **5**, 1–38.

Burgess, A.N., A computational methodology for modelling the dynamics of statistical arbitrage. PhD Thesis, London Business School, 1999.

Burgess, A.N., Statistical arbitrage models on the FTSE 100. In *Computational Finance*, edited by Y. Abu-Mustafa, B. LeBaron, A.W. Lo and A.S. Weigend, 2000 (MIT Press: Cambridge, MA).

Crowder, W. and Hamed, A., A cointegration test for oil futures market efficiency. *J. Fut. Mkts*, 1993, **13**, 933–941.

Cummins, M., Optimal statistical arbitrage: A model specification analysis on ISEQ equity data. *Irish Account. Rev.*, 2011, **17**, 21–40.

Do, B. and Faff, R., Does naïve pairs trading still work? Working Paper, Monash University, 2009.

Do, B., Faff, R. and Hamza, K., A new approach to modeling and estimation for pairs trading. Working Paper, Monash University, 2006.

Dunis, C.L., Giorgini, G., Laws, J. and Rudy, J., Statistical arbitrage and high-frequency data with an application to Eurostoxx 50 equities. Working Paper, Liverpool Business School, 2010.

Elliott, R.J., Van der Hoek, J. and Malcolm, W.P., Pairs trading. *Quant. Finance*, 2005, **5**, 271–276.

Gatev, E., Goetzmann, W.N. and Rouwenhorst, K.G., Pairs trading: Performance of a relative-value arbitrage rule. *Rev. Financ. Stud.*, 2006, **19**, 797–827.

Gilbert, C.L., Commodity speculation and commodity investment. Working Paper, Universita Degli Studi di Trento, 2008.

Girma, P.B. and Paulson, A.S., Risk arbitrage opportunities in petroleum futures spreads. *J. Fut. Mkts*, 1999, **19**, 931–955.

Gjolberg, O. and Johnsen, T., Risk management in the oil industry: Can information on long-run equilibrium prices be utilized? *Energy Econ.*, 1999, **21**, 517–527.

Gregeriou, G.N., Huber, G. and Kooli, M., Performance and persistence of commodity trading advisors: Further evidence. *J. Fut. Mkts*, 2009, **30**, 725–752.

Hansen, P.R., A test for superior predictive ability. *J. Bus. Econ. Statist.*, 2005, **23**, 365–380.

Hogan, S., Jarrow, R., Teo, R. and Warachka, M., Testing marking efficiency using statistical arbitrage with applications to momentum and value strategies. *J. Financ. Econ.*, 2004, **73**, 525–565.

Hsu, P.H. and Kuan, C.M., Re-examining the profitability of technical analysis with White's reality check. Working Paper, 2005.

Hsu, P.H., Hsu, Y.C. and Kuan, C.M., Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. Working Paper, 2009.

Kaldor, N., Speculation and economic stability. *Rev. Econ. Stud.*, 1939, **7**, 1–27.

Kanamura, T., Rachev, S.T. and Fabozzi, F.J., A profit model for spread trading with an application to energy futures. *J. Trading*, 2010, **5**, 48–62.

Kavussanos, M.G. and Alizadeh, A., The expectations hypothesis of the term structure and risk premia in dry bulk shipping freight markets. *J. Transp. Econ. Policy*, 2002, **36**, 267–304.

Kellard, N., Newbold, P., Rayner, A. and Ennew, C., The relative efficiency of commodity futures markets. *J. Fut. Mkts*, 1999, **19**, 413–432.

Kinnear, K., The Brent WTI arbitrage: Linking the world's key crudes. Working Paper, 2002.

Laws, J., Dunis, C.L. and Evans, B., Trading and filtering futures spread portfolios: Futher applications of threshold and correlation lters. Working Paper, 2008.

Lin, Y., McRae, M. and Gulati, C., Loss protection in pairs trading through minimum profit bounds: A cointegration approach. *J. Appl. Math. Decis. Sci.*, 2006, **6**, 1–14.

Marshall, B.R., Cahan, R.H. and Cahan, J.M., Can commodity futures be protably traded with quantitative market timing strategies? Working Paper, 2008.

Mou, Y., Limits to arbitrage and commodity index investment: Front-running the Goldman roll. Working Paper, 2010.

Papadakis, G. and Wysocki, P., Pairs trading and accounting information. Working Paper, 2007.

Park, C.H. and Irwin, S.H., What do we know about the profitability of technical analysis? *J. Econ. Surv.*, 2007, **21**, 786–826.

Qui, M. and Wu, Y., Technical trading-rule profitability, data snooping, and reality check: Evidence from the foreign exchange market. *J. Money, Credit, Bank.*, 2006, **38**, 2135–2158.

Ricciardi, L.M. and Sato, S., First-passage-time density and moments of the Ornstein-Uhlenbeck process. *J. Appl. Probab.*, 1988, **25**, 43–57.

Romano, J.P., Shaikh, A.M. and Wolf, M., Hypothesis testing in econometrics. Working Paper, 2009.

Romano, J.P. and Wolf, M., Control of generalized error rates in multiple testing. *Ann. Statist.*, 2007, **35**, 1378–1408.

Romano, J.P. and Wolf, M., Balanced control of generalized error rates. *Ann. Statist.*, 2010, **38**, 598–633.

Sarr, A. and Lybek, T., Measuring liquidity in financial markets. Working Paper, 2002.

Sato, S., Evaluation of the first-passage time probability to a square root boundary for the Wiener process. *J. Appl. Probab.*, 1977, **14**, 850–856.

Schwarz, T.V. and Szakmary, A.C., Price discovery in petroleum markets: Arbitrage cointegration and the time interval of analysis. *J. Fut. Mkts*, 1994, **14**, 147–167.

Shleifer, A. and Vishny, R., The limits of arbitrage. *J. Finance*, 1997, **52**, 35–55.

Silvapulle, P. and Moosa, I., The relationship between spot and futures prices: Evidence from the crude oil market. *J. Fut. Mkts*, 1999, **19**, 175–193.

Sullivan, R., Timmermann, A. and White, H., Data-snooping, technical trading rule performance, and the bootstrap. *J. Finance*, 1999, **54**, 1647–1691.

Thomas, M.U., Some mean first-passage time approximations for the Ornstein-Uhlenbeck process. *J. Appl. Probab.*, 1975, **12**, 600–604.

Trapletti, A., Geyer, A. and Leisch, F., Forecasting exchange rates using cointegration models and intraday data. *J. Forecast.*, 2002, **21**, 151–166.

Vasicek, O.A., An equilibrium characterization of the term structure. *J. Forecast.*, 1977, **5**, 177–188.

Vidyamurthy, G., *Pairs Trading: Quantitative Methods and Analysis*, 2004 (Wiley: Hoboken, NJ).

Whistler, M., *Trading Pairs: Capturing Profits and Hedging Risk with Statistical Arbitrage Strategies*, 2004 (Wiley: Hoboken, NJ).

White, H., A reality check for data snooping. *Econometrica*, 2000, **68**, 1097–1126.

Appendix A. Liquidity measures

Table A1. Liquidity measures: WTI.

Year	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
Average 5-day turnover rate												
2003	4.05	3.01	1.78	1.08	0.69	0.54	0.54	0.44	0.38	0.33	0.27	0.25
2004	4.15	2.85	1.64	1.02	0.73	0.60	0.60	0.45	0.39	0.34	0.31	0.29
2005	4.15	2.55	1.85	1.25	0.84	0.67	0.67	0.49	0.37	0.32	0.32	0.24
2006	4.51	2.41	1.71	1.06	0.76	0.60	0.60	0.53	0.48	0.44	0.43	0.30
2007	7.39	3.17	2.20	1.44	0.95	0.74	0.74	0.44	0.37	0.28	0.26	0.21
2008	9.09	3.59	2.48	1.80	1.33	0.94	0.94	0.50	0.41	0.32	0.31	0.34
2009	10.04	3.60	2.45	1.82	1.50	1.24	1.24	0.82	0.73	0.68	0.55	0.53
2010	11.86	3.87	2.63	2.11	1.64	1.39	1.39	1.02	0.83	0.73	0.67	0.61
Average 5-day Hui-Heubel ratio												
2003	0.018	0.017	0.026	0.040	0.058	0.072	0.072	0.092	0.116	0.120	0.143	0.191
2004	0.019	0.019	0.031	0.048	0.067	0.092	0.092	0.121	0.147	0.188	0.226	0.244
2005	0.018	0.019	0.025	0.036	0.053	0.066	0.066	0.110	0.170	0.233	0.182	0.512
2006	0.015	0.018	0.023	0.035	0.050	0.065	0.065	0.076	0.093	0.102	0.123	0.191
2007	0.011	0.014	0.019	0.028	0.041	0.055	0.055	0.095	0.127	0.162	0.240	0.267
2008	0.016	0.024	0.031	0.040	0.057	0.077	0.077	0.168	0.250	0.289	0.314	0.474
2009	0.015	0.022	0.029	0.037	0.044	0.049	0.049	0.076	0.089	0.094	0.117	0.145
2010	0.006	0.011	0.016	0.020	0.025	0.030	0.030	0.042	0.048	0.054	0.061	0.073

Table A2. Liquidity measures: Brent.

Year	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
Average 5-day turnover rate												
2003	3.01	1.86	1.69	1.11	0.77	0.53	0.40	0.41	0.41	0.56	1.06	2.05
2004	2.85	2.07	1.75	1.01	0.65	0.53	0.36	0.43	0.38	0.31	0.40	2.01
2005	4.48	2.10	2.11	1.45	0.94	0.64	0.47	0.44	0.39	0.33	0.39	0.69
2006	5.81	2.09	2.12	1.44	0.96	0.73	0.54	0.47	0.48	0.42	0.43	0.52
2007	5.12	2.35	2.41	1.63	1.07	0.81	0.63	0.49	0.37	0.37	0.36	0.33
2008	7.24	2.64	2.55	2.08	1.53	1.11	0.90	0.71	0.59	0.45	0.34	0.31
2009	5.82	2.59	2.38	1.87	1.49	1.25	1.04	0.92	0.86	0.71	0.58	0.47
2010	6.19	2.69	2.26	1.94	1.71	1.50	1.29	1.11	1.02	0.89	0.73	0.67
Average 5-day Hui-Heubel ratio												
2003	0.185	0.093	0.030	0.041	0.062	0.089	0.123	0.163	0.125	0.172	0.257	0.156
2004	0.028	0.026	0.029	0.049	0.078	0.108	0.194	0.355	0.321	0.527	0.415	0.220
2005	0.013	0.022	0.021	0.029	0.048	0.078	0.129	0.262	0.340	0.903	0.355	0.651
2006	0.010	0.020	0.019	0.027	0.040	0.059	0.099	0.155	0.199	0.177	0.629	1.621
2007	0.009	0.017	0.017	0.023	0.037	0.049	0.069	0.102	0.154	0.173	0.845	0.623
2008	0.013	0.036	0.034	0.052	0.061	0.082	0.089	0.124	0.165	0.275	0.363	1.926
2009	0.016	0.026	0.028	0.033	0.042	0.048	0.058	0.069	0.073	0.092	0.108	0.155
2010	0.009	0.015	0.017	0.020	0.023	0.027	0.030	0.037	0.041	0.047	0.061	0.067

Table A3. Liquidity measures: Heating oil.

Year	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
Average 5-day turnover rate												
2003	4.13	2.29	1.24	0.77	0.60	0.48	0.39	0.39	0.35	0.33	0.29	0.36
2004	3.84	1.98	1.12	0.73	0.60	0.46	0.38	0.35	0.35	0.37	0.29	0.30
2005	3.88	1.89	1.16	0.78	0.57	0.49	0.44	0.39	0.36	0.36	0.32	0.38
2006	4.05	1.83	1.19	0.79	0.56	0.48	0.53	0.40	0.38	0.39	0.34	0.31
2007	4.27	2.02	1.43	1.02	0.79	0.65	0.58	0.48	0.45	0.40	0.45	0.45
2008	5.09	2.39	1.73	1.32	1.06	0.87	0.82	0.75	0.66	0.49	0.41	0.41
2009	4.90	2.25	1.33	1.01	0.85	0.75	0.73	0.67	0.64	0.58	0.56	0.50
2010	6.87	2.33	1.53	1.18	1.01	0.77	0.68	0.64	0.57	0.46	0.39	0.43
Average 5-day Hui-Heubel ratio												
2003	0.020	0.026	0.041	0.064	0.084	0.110	0.122	0.134	0.184	0.317	0.517	0.466
2004	0.021	0.030	0.050	0.072	0.090	0.118	0.139	0.159	0.186	0.227	0.253	0.377
2005	0.021	0.030	0.045	0.066	0.088	0.122	0.166	0.156	0.203	0.191	0.234	0.378
2006	0.017	0.025	0.037	0.053	0.073	0.087	0.079	0.116	0.135	0.229	0.388	0.332
2007	0.013	0.020	0.027	0.036	0.047	0.059	0.064	0.094	0.089	0.121	0.143	0.228
2008	0.018	0.029	0.038	0.051	0.062	0.076	0.086	0.089	0.108	0.147	0.213	0.253
2009	0.019	0.029	0.046	0.061	0.070	0.078	0.081	0.090	0.093	0.109	0.127	0.119
2010	0.012	0.018	0.025	0.033	0.040	0.052	0.059	0.072	0.077	0.094	0.115	0.113

Table A4. Liquidity measures: Gas oil.

Year	M1	M2	M3	M4	M5	M6
Average 5-day turnover rate						
2003	5.32	1.43	1.79	0.61	0.41	0.38
2004	5.72	0.95	0.94	0.57	0.38	0.31
2005	5.49	1.54	1.32	0.72	0.45	0.38
2006	17.31	1.90	1.57	0.97	0.55	0.42
2007	11.61	2.08	1.63	0.99	0.60	0.45
2008	15.73	2.69	2.19	1.35	0.94	0.74
2009	8.29	2.54	1.76	1.11	0.78	0.64
2010	6.22	2.92	2.12	1.44	1.03	0.80
Average 5-day Hui-Heubel ratio						
2003	0.047	0.043	0.045	0.102	0.149	0.384
2004	0.149	0.169	0.090	0.116	0.171	0.330
2005	0.025	0.034	0.041	0.073	0.155	0.228
2006	0.013	0.022	0.027	0.040	0.073	0.108
2007	0.012	0.017	0.022	0.036	0.061	0.084
2008	0.017	0.033	0.037	0.064	0.088	0.115
2009	0.021	0.026	0.037	0.056	0.078	0.099
2010	0.014	0.013	0.019	0.028	0.038	0.050

Appendix B. Top 10 profitable trading strategies

Table B1. Top 10 profitable trading strategies: 2003 and 2004.

Series	Avg. daily ret. (%)	Sharpe ratio	Avg. trade length (days)	α	μ	σ	*F*-Test *p*-value
2003							
GO M01–HO M02 (1Y)	0.444	3.06	9.67	142.90	−0.04	0.24	0.00
WTI M04–GO M04 (1Y)	0.404	4.12	9.21	105.55	−0.09	0.24	0.00
GO M01–HO M01 (1Y)	0.373	2.31	10.44	106.44	−0.03	0.26	0.00
GO M01–HO M02 (3Y)	0.362	2.46	17.40	51.36	−0.03	0.31	0.00
Brent M11–WTI M01 (1Y)	0.355	3.32	18.64	3.76	−0.20	0.20	0.02
Brent M01–GO M03 (3Y)	0.327	3.04	18.50	17.00	−0.13	0.31	0.00
Brent M01–GO M03 (2Y)	0.327	3.03	16.19	23.87	−0.13	0.28	0.00
WTI M05–GO M04 (1Y)	0.320	3.22	10.88	76.15	−0.10	0.24	0.00
WTI M04–GO M03 (2Y)	0.319	3.01	14.50	30.24	−0.10	0.28	0.00
GO M01–HO M02 (2Y)	0.306	2.07	17.40	94.04	−0.03	0.28	0.00
2004							
GO M04–HO M04 (3Y)	0.392	3.22	12.48	89.72	−0.04	0.26	0.00
GO M05–HO M05 (3Y)	0.328	2.76	12.38	78.74	−0.04	0.26	0.00
GO M04–HO M05 (3Y)	0.310	2.57	12.48	64.79	−0.03	0.27	0.00
GO M05–HO M05 (2Y)	0.292	2.44	13.68	73.71	−0.04	0.25	0.00
GO M04–HO M04 (2Y)	0.283	2.27	17.47	82.53	−0.04	0.26	0.00
Brent M12–GO M01 (1Y)	0.281	2.33	26.10	10.89	−0.28	0.38	0.02
WTI M04–GO M06 (3Y)	0.281	2.37	17.47	28.99	−0.08	0.26	0.00
Brent M11–GO M01 (1Y)	0.274	2.28	26.10	11.28	−0.27	0.37	0.02
GO M05–HO M04 (3Y)	0.273	2.20	20.15	53.24	−0.04	0.27	0.00
WTI M05–GO M06 (3Y)	0.270	2.31	17.47	41.83	−0.10	0.25	0.00

Table B2. Top 10 profitable trading strategies: 2005 and 2006.

Series	Avg. daily ret. (%)	Sharpe ratio	Avg. trade length (days)	α	μ	σ	*F*-Test *p*-value
2005							
Brent M12–GO M06 (3Y)	0.288	2.99	13.68	33.48	−0.21	0.24	0.00
Brent M01–GO M01 (1Y)	0.280	2.19	17.13	20.35	−0.20	0.30	0.00
Brent M11–GO M06 (3Y)	0.267	2.76	15.29	36.58	−0.21	0.24	0.00
WTI M08–GO M02 (1Y)	0.265	2.23	19.92	22.79	−0.17	0.30	0.00
Brent M03–GO M01 (2Y)	0.255	2.04	19.69	19.45	−0.20	0.31	0.00
Brent M10–GO M06 (2Y)	0.249	2.57	15.29	41.76	−0.20	0.23	0.00
WTI M06–GO M02 (3Y)	0.246	2.10	21.25	17.17	−0.14	0.29	0.00
WTI M04–HO M01 (1Y)	0.244	2.66	28.89	6.68	−0.15	0.18	0.05
GO M01–HO M01 (3Y)	0.243	1.77	19.92	42.34	−0.03	0.34	0.00
WTI M08–GO M02 (3Y)	0.242	2.03	23.55	14.30	−0.16	0.29	0.00
2006							
GO M02–HO M03 (2Y)	0.300	3.07	14.44	52.70	−0.02	0.29	0.00
GO M01–HO M02 (2Y)	0.265	2.46	18.57	39.03	−0.02	0.33	0.00
GO M02–HO M03 (3Y)	0.263	2.67	16.25	45.65	−0.03	0.30	0.00
GO M04–HO M04 (1Y)	0.238	2.51	18.57	43.18	−0.01	0.26	0.00
GO M03–HO M03 (1Y)	0.235	2.41	18.57	35.95	−0.01	0.27	0.00
WTI M02–GO M05 (1Y)	0.228	2.32	20.00	16.83	−0.19	0.26	0.00
WTI M03–GO M01 (2Y)	0.228	2.13	23.64	22.05	−0.14	0.32	0.00
WTI M02–GO M01 (2Y)	0.222	2.04	23.64	18.38	−0.14	0.33	0.00
GO M05–HO M05 (1Y)	0.220	2.37	16.25	57.80	−0.01	0.25	0.00
WTI M01–GO M05 (1Y)	0.217	2.09	23.64	18.25	−0.20	0.28	0.00

Table B3. Top 10 profitable trading strategies: 2007 and 2008.

Series	Avg. daily ret. (%)	Sharpe ratio	Avg. trade length (days)	α	μ	σ	*F*-Test *p*-value
2007							
WTI M02–GO M01 (2Y)	0.319	2.78	16.31	34.48	−0.16	0.29	0.00
Brent M02–GO M03 (1Y)	0.309	3.16	9.00	101.02	−0.17	0.21	0.00
Brent M05–GO M03 (1Y)	0.295	3.23	10.88	94.83	−0.15	0.21	0.00
Brent M03–GO M03 (1Y)	0.295	3.10	9.00	108.19	−0.16	0.20	0.00
Brent M05–GO M04 (1Y)	0.292	3.22	11.35	96.84	−0.16	0.20	0.00
Brent M04–GO M04 (1Y)	0.288	3.11	11.35	106.38	−0.17	0.20	0.00
GO M01–HO M01 (2Y)	0.287	2.66	15.35	56.23	0.00	0.29	0.00
GO M02–HO M02 (1Y)	0.283	2.88	12.43	122.85	0.00	0.22	0.00
Brent M02–GO M01 (1Y)	0.282	2.62	11.86	79.97	−0.15	0.23	0.00
WTI M03–GO M01 (2Y)	0.278	2.50	18.64	32.09	−0.15	0.28	0.00
2008							
GO M02–HO M01 (2Y)	0.283	2.26	15.35	91.53	0.01	0.23	0.00
GO M03–HO M02 (2Y)	0.261	2.07	15.35	113.21	0.00	0.22	0.00
GO M03–HO M04 (1Y)	0.248	1.99	16.38	115.36	−0.01	0.21	0.00
GO M02–HO M01 (3Y)	0.247	1.96	20.08	67.23	0.01	0.26	0.00
GO M02–HO M01 (1Y)	0.239	1.89	17.40	176.82	0.00	0.20	0.00
GO M01–HO M01 (2Y)	0.229	1.49	18.29	116.17	0.00	0.24	0.00
GO M01–HO M01 (3Y)	0.223	1.45	21.42	70.32	0.00	0.28	0.00
GO M03–HO M05 (1Y)	0.219	1.75	21.83	48.26	−0.01	0.23	0.00

Table B4. Top 10 profitable trading strategies: 2009 and 2010.

Series	Avg. daily ret. (%)	Sharpe ratio	Avg. trade length (days)	α	μ	σ	*F*-Test *p*-value
2009							
GO M01–HO M02 (2Y)	0.66	4.77	7.46	90.81	0.01	0.26	0.00
GO M02–HO M03 (2Y)	0.59	4.32	9.00	91.26	0.01	0.25	0.00
GO M01–HO M02 (3Y)	0.56	3.96	9.00	93.50	0.01	0.25	0.00
GO M02–HO M02 (1Y)	0.56	3.99	9.67	102.62	0.02	0.29	0.00
GO M01–HO M01 (3Y)	0.55	3.85	8.42	83.40	0.01	0.25	0.00
GO M01–HO M03 (3Y)	0.54	3.85	9.67	64.28	0.00	0.25	0.00
GO M03–HO M03 (1Y)	0.52	3.86	9.67	117.69	0.02	0.28	0.00
GO M02–HO M03 (3Y)	0.51	3.65	10.44	86.65	0.00	0.24	0.00
GO M01–HO M02 (1Y)	0.49	3.44	9.00	121.83	0.02	0.28	0.00
GO M04–HO M05 (3Y)	0.49	3.79	9.00	77.89	0.00	0.24	0.00
2010							
WTI M01–GO M05 (1Y)	0.114	1.43	65.25	14.94	−0.16	0.43	0.00
GO M05–HO M06 (1Y)	0.104	1.58	23.64	164.86	−0.01	0.27	0.00
GO M04–HO M05 (1Y)	0.103	1.55	23.64	160.37	−0.01	0.28	0.00
Brent M09–WTI M01 (3Y)	0.102	2.31	46.60	4.12	0.06	0.27	0.00
Brent M10–WTI M01 (3Y)	0.101	2.25	45.40	3.83	0.06	0.28	0.00
Brent M08–WTI M01 (3Y)	0.099	2.29	46.60	4.49	0.05	0.27	0.00
WTI M02–HO M10 (1Y)	0.083	2.31	52.20	7.75	−0.18	0.21	0.00
GO M05–HO M07 (1Y)	0.081	1.22	31.13	184.38	−0.02	0.26	0.00
Brent M05–WTI M02 (2Y)	0.070	2.13	52.20	7.23	0.03	0.17	0.01
Brent M06–WTI M02 (2Y)	0.067	1.99	52.20	5.99	0.04	0.17	0.00