Statistical Arbitrage and High-Frequency Data with an Application to Eurostoxx 50 Equities

By
Christian L. Dunis*
Gianluigi Giorgioni**
Jason Laws***
Jozef Rudy****
(Liverpool Business School, CIBEF****
Liverpool John Moores University)

March 2010

Abstract

The motivation for this paper is to apply a statistical arbitrage technique of pairs trading to high-frequency equity data and compare its profit potential to the standard sampling frequency of daily closing prices. We use a simple trading strategy to evaluate the profit potential of the data series and compare information ratios yielded by each of the different data sampling frequencies. The frequencies observed range from a 5-minute interval, to prices recorded at the close of each trading day.

The analysis of the data series reveals that the extent to which daily data are cointegrated provides a good indicator of the profitability of the pair in the high-frequency domain. For each series, the in-sample information ratio is a good indicator of the future profitability as well.

Conclusive observations show that arbitrage profitability is in fact present when applying a novel diversified pair trading strategy to high-frequency data. In particular, even once very conservative transaction costs are taken into account, the trading portfolio suggested achieves very attractive information ratios (e.g. above 3 for an average pair sampled at the high-frequency interval and above 1 for a daily sampling frequency).

Keywords

High-frequency data, statistical arbitrage, pairs trading, cointegration, time adaptive models

* ***Christian Dunis*** *is Professor of Banking and Finance at Liverpool Business School and Director of the Centre for International Banking, Economics and Finance (CIBEF) at Liverpool John Moores University (E-mail: c.dunis@ljmu.ac.uk)*

** ***Gianluigi Giorgioni*** *is a Senior Lecturer in Economics and Finance at Liverpool Business School and a member of CIBEF (E-mail: G.Giorgioni@ljmu.ac.uk)*

*** ***Jason Laws*** *is a Reader of Finance at Liverpool Business School and a member of CIBEF (E-mail: J.Laws@ljmu.ac.uk)*

**** ***Jozef Rudy*** *is an Associate Researcher with CIBEF (E-mail: J.Rudy@2009.ljmu.ac.uk) and currently working on his PhD thesis at Liverpool Business School.*

***** ***CIBEF*** *– Centre for International Banking, Economics and Finance, JMU, John Foster Building, 98 Mount Pleasant, Liverpool L3 5UZ.*

1. INTRODUCTION

In this article a basic pair trading (long-short) strategy is applied to the constituent shares of the Eurostoxx 50 index. A long-short strategy is applied to shares sampled at 6 different frequencies, namely 5-minute, 10-minute, 20-minute, 30-minute, 60-minute and daily sampling intervals. The high frequency data spans from 3rd July 2009 to 17th November 2009, our daily data spans from 3rd January 2000 to 17th November 2009.

We introduce a novel approach, which helps enhance the performance of the basic trading strategy. The approach consists in selecting the pairs for trading based on the best in-sample information ratios and the highest in-sample t-stat of the ADF test of the residuals of the cointegrating regression sampled a daily frequency. We form the portfolios of 5 best trading pairs and compare the performance with appropriate benchmarks.

Yet another improvement we introduce is the use of the high-frequency data. The advantage of using the high-frequency data is higher potentially achievable information ratio[1] compared to the use of daily closing prices, see Aldridge (2009) and thus higher attractivity for investors.

Market neutral strategies are generally known for attractive investment properties, such as low exposure to the equity markets and relatively low volatility, see Capocci (2006) but recently the profitability of these strategies has deteriorated, see Gatev *et al.* (2006). While Gatev *et al.* (2006) only go back to 2002, the Hedge Fund Equity Market Neutral Index (HFRXEMN Index in Bloomberg) which started one year later, i.e. 2003, does not show the supposed qualities for which market neutral strategies are attractive, i.e. steady growth and low volatility. The industry practice for market neutral hedge funds is to use a daily sampling frequency and standard cointegration techniques to find matching pairs, see Gatev *et al.* (2006, p. 10), who "use this approach because it best approximates the description of how traders themselves choose pairs." Thus, by modifying an already well-known strategy using intraday data we may obtain an "edge" over other traders and compare the results of simulated trading using intraday data on various sampling frequencies with daily data.

The rest of the paper is organized as follows. In section 2, we present the literature review, section 3 describes the data used and section 4 explains the methodology implemented. Section 5 presents the pair trading model, section 6 gives the out-of-sample performance results of the pair trading strategy taking transaction costs into account, while sections 7 presents our results in a diversified trading portfolio context. Section 8 concludes.

2. LITERATURE REVIEW

a. Market neutral strategies

Pair trading is a well-known technique, having been developed in 1980 by a team of scientists lead by a Wall Street quant Nunzio Tartaglia, see Gatev *et al.* (2006). The

[1] Information ratio is calculated as the ratio of annualized return to annualized standard deviation.

strategy is widely documented in current literature including Enders and Granger (1998), Vidyamurthy (2004), Dunis and Ho (2005), Lin *et al.* (2006) and Khandani and Lo (2007).

The general description of the technique is that a pair of shares is formed, where the investor is long one and short another share. The rationale is that there is a long-term equilibrium (spread) between the share prices, and thus the shares fluctuate around that equilibrium level (the spread has a constant mean). The investor evaluates the current position of the spread based on its historical fluctuations and when the current spread deviates from its historical mean by a pre-determined significant amount (measured in standard deviations), the spread is subsequently altered and the legs are adjusted accordingly. The investor bets on the reversion of the current spread to its historical mean by shorting/going long an appropriate amount of each share in a pair. The appropriate amount of each share is expressed by the variable beta, which tells the investor the number of the shares X he has to short/go long, for each 1 share Y. There are various ways of calculating beta. Either it can be fixed, or it can be time-varying. In this paper to make beta time-varying, we will use rolling ordinary least squares (OLS) regression, double exponential smoothing prediction (DESP) model and the Kalman filter. More about these methods can be found in the section about time adaptive models and methodology.

b. Market neutral strategies and high frequency data

From an extensive review of literature there appears to be only one relevant study regarding high frequency market neutral trading systems, see Nath (2003). Nath (2003) looks at market neutral strategies in US fixed-income market, nevertheless literature pertaining to high frequency market neural trading systems is extremely limited.

c. Cointegration

Cointegration is a quantitative technique based on finding long-term relations between asset prices introduced in a seminal paper by Engle and Granger (1987). Thus, cointegration might help identify potentially related pairs of assets. However, in this paper we will consider all the possible pairs from the same industry, not only cointegrated ones. The reason is that in this way we will be able to measure whether the cointegrated pairs in the in-sample period perform better in the out-of-sample period than the non-cointegrated ones.

Another approach was developed by Johansen (1988), which can be applied to more than two assets at the same time. The result is a set of cointegrating vectors that can be found in the system. The spread between the assets is not the one with the lowest variance, as was the OLS case, but the most stable one in the long-term, see Alexander (2001). According to Alexander (2001, p. 361) the Engle and Granger (1987) methodology is preferred in financial applications due to its simplicity and lower variance, important point to consider from the risk management perspective. As in this paper we only deal with pairs of shares, we also prefer the simpler Engle and Granger (1987) methodology.

There are also many applications of cointegration in the world of investing, for instance index replication, which exploits long-term qualities of cointegration

requiring only occasional portfolio rebalancing [see e.g., Dunis and Ho (2005), Alexander and Dimitriu (2002)]. Then, there are market neutral arbitrage strategies based on cointegration, where one enters the trade when the relationship is away from long-term mean, and exits it when it has returned to the long-term mean again. Burgess (2003), Lin *et al.* (2006) or the work of Galenko *et al.* (2007), who term their work high-frequency trading, just use daily closing prices among 4 world indexes, instead of real intraday continuous or intraday minute data.

d. Time adaptive models

Dunis and Shannon (2005) use time adaptive betas with the Kalman filter methodology (see Hamilton (1994) or Harvey (1981) for a detailed description of the Kalman filter implementation). The Kalman filter is a popular technique when time varying parameters in the model need to be estimated (see Choudhry and Wu (2009), Giraldo Gomez (2005), Brooks *et al.* (1998) and Burgess (1999)). These papers support the Kalman filter method as a superior technique for adaptive parameters. It is a forward looking methodology, as it tries to predict the future position of the parameters as opposed to using a rolling OLS regression, see Bentz (2003).

Alternatively DESP models can be used for adaptive parameter estimation. According to LaViola (2003a) and LaViola (2003b) DESP models offer comparable prediction performance to the Kalman filter, with the advantage that they run 135 times faster.

e. Hedge funds

The pair trading technique is used primarily by hedge funds and there is a whole distinct group bearing the name “market neutral funds”, see Khandani and Lo (2007) for the definition or Capocci (2006) for a closer examination of their properties. Hedge funds employ dynamic trading strategies, see Fung and Hsieh (1997). Those strategies are dramatically different from the ones employed by mutual funds and this enables them to offer investors more attractive investment properties (expressed by e.g. information ratio), see Liang (1999).

3. THE EUROSTOXX 50 INDEX AND RELATED FINANCIAL DATA

We use 50 stocks that formed the Eurostoxx 50 index as of 17th November 2009, see Appendix f for the names of shares we used. The data downloaded from Bloomberg includes 6 frequencies: 5-minute, 10-minute, 20- minute, 30-minute and 60-minute data (high-frequency data) and daily prices. We call all the data related with the minute dataset high-frequency for brevity purposes.

Our database of minute data spans from 3rd July 2009 to 17th November 2009, both dates included[2]. We download the data from Bloomberg, which only stores the last 100 business days worth of intraday data. We downloaded the data on 17th

[2] The high-frequency database includes prices of transactions for the shares that take place closest in time to the second 60 of particular minute-interval (e.g. transaction recorded just before the end of any 5-minute interval, or whichever selected interval in case of other high-frequencies), but not having taken place after second 60, so that if one transaction took place at e.g. 9:34:58 and the subsequent one at 9:35:01, the former transaction would be recorded as of 9:35.

November 2009 and that is why our intraday data span from 3rd July 2009. Intraday stock prices are not adjusted automatically by Bloomberg for dividend payments and stock splits and we had to adjust them ourselves.[3] Our database only includes the prices at which the shares were transacted, we do not dispose of bid and ask prices. Therefore some of our recorded prices are bids and some of them asks depending on which transaction was executed in each particular case. As for the number of data points we have at our disposal, we have as many as 8.000 data points when data are sampled at 5-minute interval for the last 5 months. For lower frequencies, the amount of data falls linearly with decreasing frequency. For example, in the case of 10-minute data we have around 4.000 data points whereas we only have 2.000 data points for 20-minute data.

The database that includes daily closing prices spans from 3rd January 2000 to 17th November 2009, including the dates mentioned. The data are adjusted for dividend payments and stock splits[4]. Some shares do not date back as far as 3rd January 2000, and as a consequence the pairs that they formed contain lower amount of data points.[5]

In Table 1 below we show the start and the end of the in- and out-of-sample periods for all the frequencies. For high-frequency data the in- and out-of-sample periods have the same lengths. For daily data, the in-sample period is much longer than the out-of-sample period. The start of the out-of-sample period is not aligned between daily and high-frequency data. If the out-of-sample period for daily data started at the same date as is the case for high-frequency data, it would not contain enough data points for the out-of-sample testing (had it started on 10th September, it would have contained only as little as 50 observations and this is why we start the out-of-sample period for daily data at the beginning of 2009, yielding 229 data points).

	In-sample		No. points	Out-of-sample		No. points
5-minute data	03 July 2009	09 September 2009	4032	10 September 2009	17 November 2009	4032
10-minute data	03 July 2009	09 September 2009	2016	10 September 2009	17 November 2009	2016
20-minute data	03 July 2009	09 September 2009	1008	10 September 2009	17 November 2009	1008
30-minute data	03 July 2009	09 September 2009	672	10 September 2009	17 November 2009	672
60-minute data	03 July 2009	09 September 2009	336	10 September 2009	17 November 2009	336
Daily data	03 January 2000	31 December 2008	2348	01 January 2009	17 November 2009	229

Table 1. Specification of the in- and out-of-sample periods and number of data points contained in each

We used the Bloomberg sector classification with the "industry_sector" ticker. We divide the shares in our database into 10 industrial sectors: Basic Materials, Communications, Consumer Cyclical, Consumer Non-cyclical, Diversified, Energy, Financial, Industrial, Technology and Utilities. Also note that there is only one share in the category "diversified" and "technology" in Appendix e, which prevents both these shares from forming pairs.

[3] Daily data are adjusted automatically by Bloomberg. Concerning intraday data, first we obtain the ratio of daily closing price (adjusted by Bloomberg) to the last intraday price for that day (representing the unadjusted closing price). Then we multiply all intraday data during that particular day by the calculated ratio. We repeat the procedure for all days and shares for which we have intraday data.

[4] Daily data are automatically adjusted by Bloomberg.

[5] In particular, four shares do not date back from 3rd January 2000 (Anheuser-Busch starts from 30th November 2000, Credit Agricole S.A. starts from 13th December 2001, Deutsche Boerse AG starts from 5th February 2001 and GDF Suez starts from 7th July 2005).

For our pair trading methodology, we select all the possible pairs from the same industry. This is not a problem with daily data, as we have daily closing prices for the same days for all the shares in the sample. In contrast, at times an issue of liquidity with high-frequency data occurs where, for a certain pair, one share has a price related to a particular minute whilst no price is recorded for the other due to no transaction having taken place in that minute. In such an event, spare prices were dropped out so that we were left with two price time series with the same number of data points in each, where the corresponding prices were taken at approximately the same moment (same minute). However, such a situation presents itself only rarely, as these 50 shares are the most liquid European shares listed.

4. METHODOLOGY

In this part we describe in detail the techniques which we use in simulated trading. First we describe the Engle and Granger (1987) cointegration approach. Then, in order to make the beta parameter adaptive, we describe techniques which we used, namely rolling OLS, the DESP model and the Kalman filter.

However, as using the Kalman filter proves to be a superior technique for the beta calculation as will be shown later, only the Kalman filter is used for the calculation of the spread to obtain the final results presented in the paper.

a. Cointegration model

First, we form the corresponding pairs of shares from the same industry. Once these are formed, we evaluate whether the pairs are cointegrated in the in-sample period. We investigate in the empirical part whether the fact that pairs are cointegrated or not helps improve the profitability of the pairs selected. Thus, we do not disqualify any pairs at first and also take into account the ones that are not cointegrated.

The 2-step approach proposed by Engle and Granger (1987) is used for the estimation of the long-run equilibrium relationship where first the OLS regression shown below is performed.

$$Y_t = \beta X_t + \varepsilon_t \tag{1}$$

In the second step the residuals of the OLS regression are tested for stationarity using the Augmented Dickey-Fuller unit root test (hereinafter ADF) at 95% confidence level, see Said and Dickey (1984).

b. Rolling OLS

To calculate the spread, first we need to calculate the rolling beta using rolling OLS. Beta at time t is calculated from n previous points.

$$Y_t = \beta_t X_t + \varepsilon_t \tag{2}$$

However, the rolling OLS approach is the least favoured by the literature due to “ghost effect”, “lagging effect” and “drop-out effect”, see Bentz (2003).

We optimized the length of the OLS rolling window using genetic optimization.[6] For more details on the genetic optimization, see Goldberg (1989) and Conn *et al.* (1991). The objective of the genetic optimization was to maximize the average in-sample information ratio for 6[7] randomly chosen pairs[8] at a 20-minute sampling frequency. The optimized parameter was the length of the rolling window for the OLS regression in the in-sample period. Thus, the genetic algorithm was searching for the optimum length of the rolling window in the in-sample period with the objective to maximize the in-sample information ratio. The best values found for the in-sample period were subsequently used in the out-of-sample period as well. The same 6 pairs at the same sampling frequency with the same objectives were optimized also in case of the DESP model and Kalman filter.

The average OLS rolling window length for the 6 pairs found using genetic algorithm was 200 points, which was then used for all the remaining pairs and frequencies in the out-of-sample period.

c. Double exponential-smoothing prediction model

Double exponential smoothing-based prediction (DESP) models are defined by two series of simple exponential smoothing equations.

First, we calculate the original β_t series, where $\beta_t = \frac{Y_t}{X_t}$ at each time step. Once we have β_t series, we smooth it using the DESP model. DESP model is defined by the following 2 equations.

$$S_t = \alpha\beta_t + (1-\alpha)S_{t-1} \tag{3}$$

$$T_t = \alpha S_t + (1-\alpha)T_{t-1} \tag{4}$$

where β_t is an original series at time t, S_t is a single exponentially smoothed series, T_t a double exponentially smoothed series and α the smoothing parameter. At each point t in time, the prediction of the value of β_t in time period t+1 is given by:

$$\tilde{\beta}_{t+1} = a_t + kb_t \tag{5}$$

$$a_t = 2S_t - T_t \tag{6}$$

$$b_t = \frac{\alpha}{1-\alpha}(S_t - T_t) \tag{7}$$

[6] The optimization was performed in MATLAB. The genetic algorithm was run with default options. The optimization started with 100 generations and both, mutation and crossover, were allowed.
[7] We only optimized the parameters for 6 pairs due to the length of the genetic optimization process.
[8] MATLAB function rand was used to generate 6 random numbers from 1 to 176 (as rand only generates numbers from 0 to 1, the result of rand was multiplied by 176 and rounded to the nearest integer towards infinity with the function ceil). 176 is the number of all the possible pairs out of 50 shares, provided that only the pairs of shares from the same industry are selected.

where $\tilde{\beta}_{t+1}$ is the prediction of the value of β_t in time period t+1, a_t the level estimated at time t and b_t the trend estimated at time t and k the number of look-ahead periods.

We optimized the α and k parameters present in Equations (3), (4), (7) and (5) respectively. Optimized values for α and k are 0.8126 and 30 respectively.

d. Time-varying parameter models with Kalman filter

The Kalman filter allows parameters vary over time and it is more optimal than rolling OLS for adaptive parameter estimation, see Dunis and Shannon (2005). Further details of the model and estimation procedure can be found in Harvey (1981) and Hamilton (1994).

The time varying beta model can be expressed by the following system of state-space equations:

$$Y_t = \beta_t X_t + \varepsilon_t \tag{8}$$

$$\beta_t = \beta_{t-1} + \eta_t \tag{9}$$

where Y_t is the dependent variable at time t, β_t is time-varying coefficient, X_t is the independent variable at time t, and ε_t and η_t are independent uncorrelated error terms. Equation (8) is known as a measurement equation and Equation (9) as the state equation, which defines beta as a simple random walk in our case. We thus use similar model to Dunis and Shannon (2005) or Burgess (1999). For the full specification of the Kalman filter model please see Appendix a.

We optimized the noise ratio, see Appendix a for the noise ratio definition. The resulting value for the noise ratio of 3.0e-7 was then used for all the remaining pairs and frequencies.

5. THE PAIR TRADING MODEL

The procedures described in this section were applied to both daily and high-frequency data. The pairs had to belong to the same industry to be considered for trading. It was the only condition in order to keep our strategy simple. This leaves us with pairs immune to industry-wide shocks.

a. Pair trading: a self-financing strategy

A pair trading strategy requires one to be long one share and short another. Pair trading is a so-called self-financing strategy, see e.g. Alexander and Dimitriu (2002), meaning that an investor can borrow the amount he wants to invest, say from a bank. Then, to be able to short a share, he deposits the borrowed amount with the financial institution as collateral and obtains borrowed shares. Thus, the only cost he has to pay is the difference between borrowing interest rates paid by the investor and lending interest rates paid by the financial institution to the investor. Subsequently, to go short a given share, the investor sells the borrowed share and

obtains cash in return. From the cash he finances his long position. On the whole, the only cost is the difference between both interest rates (paid vs. received). A more realistic is the situation where an investor does not have to borrow capital from a bank in the beginning (e.g. the case of a hedge fund that disposes of capital from investors) allows us to drop the difference in interest rates. Therefore, a short position would be wholly financed by an investor. In our first scenario the investor would be paid interest from a financial institution which lends him shares, however this interest was neglected for the ease of calculation. As our strategy proves robust and profitable, it does not affect our conclusions, it could only affect them on the positive side.

b. Spread calculation

First, we calculate the spread between the shares. The spread is calculated as

$$z_t = P_{Y_t} - \beta_t P_{X_t} \tag{10}$$

where z_t is the value of the spread at time t, P_{X_t} is the price of share X at time t, P_{Y_t} is the price of share B at time t and β_t is the adaptive coefficient beta at time t.

Beta was calculated at each time step using 3 of the methods described in the methodological part, namely the rolling OLS, the DESP model and the Kalman filter.

We did not include a constant in any of the models. Intuitively speaking, when the price of one share goes to 0, why would there be any threshold level under which the price of the second share cannot fall? Furthermore, by not including a constant, we obtain a model with fewer parameters to be estimated.

c. Entry and exit points

First we estimate the spread of the series using Equation (10). The spread is then normalized by subtracting its mean and dividing by its standard deviation. The mean and the standard deviation are calculated from the in-sample period and are then used to normalize the spread both in the in- and out-of-sample periods.

We sell (buy) the spread when it is 2 standard deviations above (below) its mean value and the position is liquidated when the spread is closer than 0.5 standard deviation to its mean. We decided to wait for 1 period before we enter into the position, to be on the safe side and make sure that the strategy is viable in practice. For instance, in case of 5-minute data, after the condition for entry has been fulfilled, we wait for 5-minutes before we enter the position.

We chose the investment to be money-neutral, thus the amounts of euros to be invested on the long and short side of the trade to be the same.[9] As the spread is

[9] Above we explained that our positions are money neutral on both sides of the trade. However in practice this is not always possible, as an investor is not able to buy share fractions. Thus, it might occur that we wish to be long 1000 euros worth of share A and short 1000 euros worth of share B. But the price of share X is 35 euros and the price of share Y is 100 euros. In this case we would need to buy 28.57 shares X and sell 10 shares Y. In the paper we simplified the issue and supposed that an investor is able to buy fractions of the shares. The reason is that one is able to get as close as one

away from its long term mean, we bet on the spread reverting to its long term mean, but we do not know whether we will gain more from our long or short position[10]. We do not assume rebalancing once we enter into the position. Therefore, after an initial entry into the position with equal amounts of euros on both sides of the trade, even when due to price movements both positions stop being money-neutral, we do not rebalance the position. Only two types of transactions are allowed by our methodology, entry into a new position, and total liquidation of the position we were in previously.

For an illustration, in Figure 1 below we show the normalized spread and the times when the positions are open. When the dotted line is equal to 1(-1), the investor is long (short) the spread.

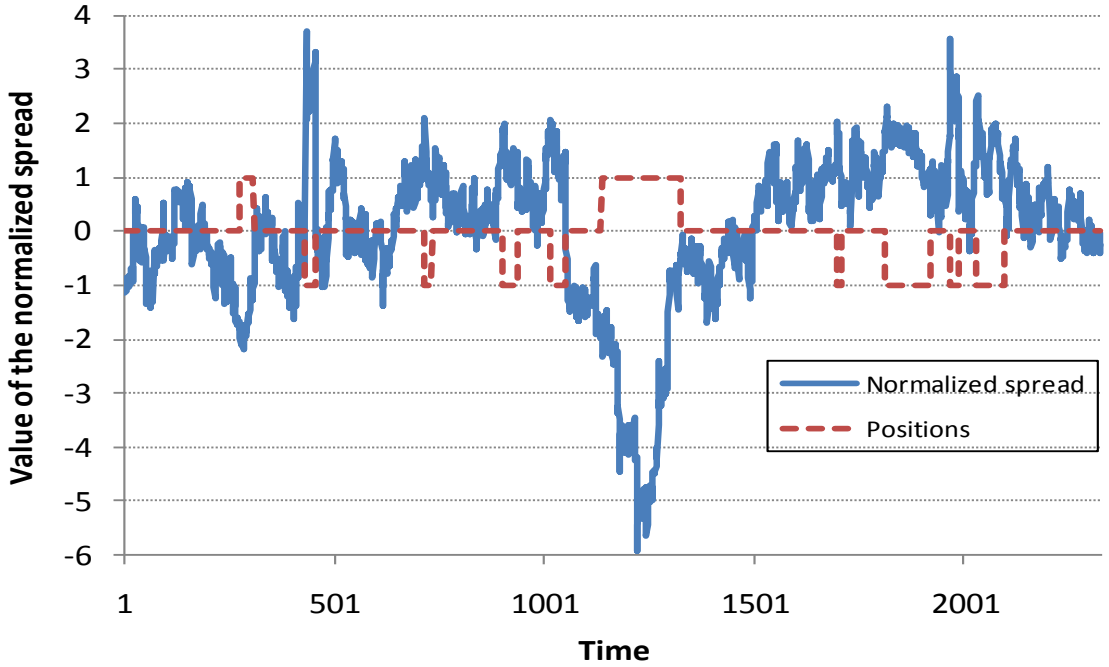


Figure 1. The normalized spread of the pair consisting of Bayer AG and Arcelor Mittal sampled a 20-minute interval

In Figure 2 we show the cumulative equity curve for the pair consisting of Bayer AG and Arcelor Mittal[11]. Note how the investment lost almost 10% around half the sample as position was entered into too soon and continued to move against the investor. Finally it reverted and recovered almost all the capital lost.

---

wishes to the money neutral position in practice. The only thing one has to do is to increase the amount of money on both sides of the trade. If in the previous example we wished to be long and short 100,000 euros, we would buy 2857 shares X and 1000 shares Y.

[10] We do not know which of the cases will occur in advance: whether the shares return to their long term equilibrium because the overvalued share falls more, the undervalued rises more, or both perform the same.

[11] The pair was chosen only for an illustration of the approach. Both shares are from the same industry: Basic materials, see Appendix e. In Figure 2 the same pair of shares is shown as was the case in Figure 1.

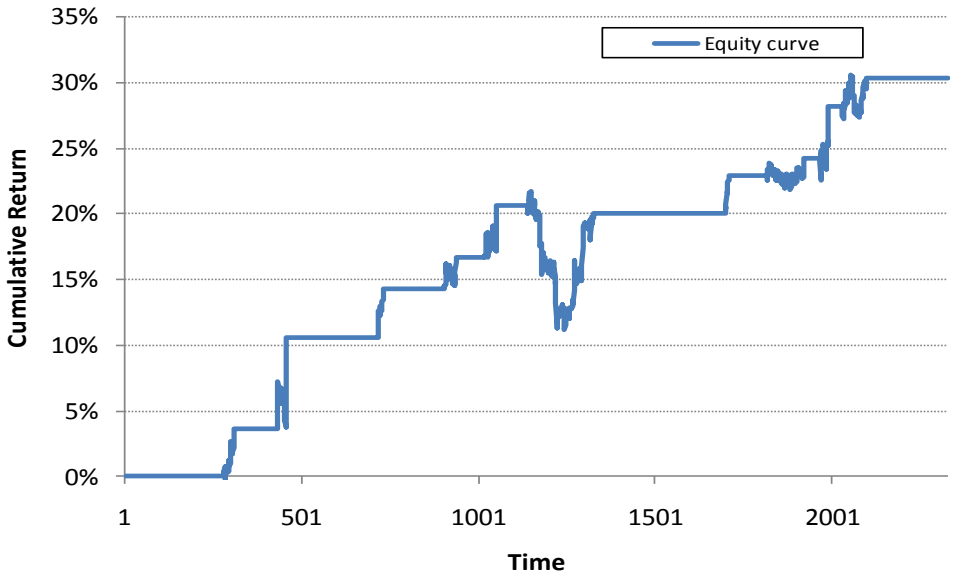


Figure 2. Cumulative equity curve in percent of the pair trading strategy applied to Bayer AG and Arcelor Mittal sampled at a 20-minute interval

In the next section we explain the different indicators calculated in the in-sample period, trying to find a connecting link with the out-of-sample information ratio and as a consequence offer a methodology for evaluating the suitability of a given pair for arbitrage trading.

d. Indicators inferred from the spread

All the indicators are calculated in the in-sample period. The objective is to find the indicators with high predictive power of the profitability of the pair in the out-of-sample period. These indicators include the t-stat from the ADF test (on the residuals of the OLS regression of the two shares), the information ratio and the half-life of mean reversion.

i. Half-life of mean reversion

The half-life of mean reversion in number of periods can be calculated as:

$$Halflife = -\frac{\ln(2)}{k} \tag{11}$$

where k is the median unbiased estimate of the strength of mean reversion from Equation (12), see Wu *et al.* (2000, p. 759) or Dias and Rocha (1999, p. 24). Intuitively speaking, it is half the average time the pair usually takes to revert back to its mean. Thus, pairs with low half-life should be preferred to high half-lives by traders.

Equation (12) is called the OU equation and can be used to calculate the speed and strength of mean reversion, see Mudchanatongsuk *et al.* (2008). The following formula is estimated on the in-sample spread:

$$dz_t = k\left(\mu - z_t\right)dt + \sigma dW_t \tag{12}$$

where μ is the long-term mean of the spread, z_t value of the spread at particular point in time, k the strength of mean reversion, σ the standard deviation and W_t the Wiener process. The higher the k, the faster the spread tends to revert to its long-term mean. Equation (12) is used indirectly in the paper, it is just the supplementary equation from which we calculate the half-life of mean reversion of the pairs.

ii. Information ratio

We decided to use the information ratio (IR), a widely used measure among practitioners which gives an idea of the quality of the strategy[12]. An annualized information ratio of 2 means that the strategy is profitable almost every month. Strategies with an information ratio around 3 are profitable almost every day, see Chan (2009). For our purpose we calculated the information ratio as:

$$Annualized\ Information\ Ratio = \frac{R}{\sigma}.\sqrt{hours\ traded\ \ per\ day.252} \tag{13}$$

where R is the average return we obtain from the strategy and σ is the standard deviation of return of the strategy. However, it is not the perfect measure and Equation (13) overestimates the true information ratio if returns are autocorrelated, see e.g. Sharpe (1994) or Alexander (2008, p. 93).

6. OUT-OF-SAMPLE PERFORMANCE AND TRADING COSTS

a. Return calculation and trading costs

The return in each period is calculated as

$$Ret_t = \ln(P_{X_t} / P_{X_{t-1}}) - \ln(P_{Y_t} / P_{Y_{t-1}}) \tag{14}$$

where P_{X_t} is the price of the share we are long in period t, $P_{X_{t-1}}$ the price of the share we are long in period t-1, P_{Y_t} the price of the share we are short in period t, and $P_{Y_{t-1}}$ the price of the share we are short in period t-1.

We consider conservative total transaction costs of 0.3% one-way in total for both shares, similar to e.g. Alexander and Dimitriu (2002). We are dealing with the 50 most liquid European shares in this paper. Transaction costs consist of 0.1%[13] of brokerage fee for each share (thus 0.2% for both shares), plus a bid-ask spread for each share (long and short) which we assume to be 0.05% (0.3% in total for both shares).

[12] IR has now become more popular among practitioners in quantitative finance than Sharpe ratio. The formula for a Sharpe ratio (SR) calculation can be found in Appendix g. Note that the only difference between IR and SR is the risk free rate in the denominator of SR.

[13] For instance Interactive Brokers charges 0.1% per transaction on XETRA market (see http://www.interactivebrokers.com/en/p.php?f=commission and http://www.interactivebrokers.com/en/accounts/fees/euroStockBundlUnbund.php?ib_entity=llc, the bundled cost structure. Last accessed 14th February 2010)

We calculate a median bid-ask spread for the whole time period investigated for 6 randomly chosen stocks sampled at a 5-minute interval. We chose 6 stocks using the same randomization procedure which we used to select 6 random pairs for the genetic optimization purposes for rolling OLS, DESP and Kalman filter. Median value of the 6 median values of the bid-ask spreads was 0.05%. The bid-ask spread at every moment was calculated as:

$$Bid / Ask\ Spread = \frac{abs(P_A - P_B)}{avg(P_A + P_B)} \tag{15}$$

where P_A is the ask price of a share at any particular moment and P_B is the bid price at the same moment.

We buy a share which depreciates significantly whilst on the other hand we sell those that appreciate significantly. Therefore in real trading it may be possible not to pay the bid-ask spread. The share that we buy is in a downtrend. The downtrend occurs because transactions are executed every time at lower prices. And the lower prices are the result of falling ask prices which get closer to (or match) bid prices, thus effectively one does not have to pay bid-ask spread and transacts at or close to the bid quote. The opposite is true for rising prices of shares.

b. Preliminary out-of-sample results

In Table 2 we present the out-of-sample information ratios excluding transaction costs for the pair trading strategy at all the frequencies we ran our simulations for. Results across all the three methods used are displayed.

Results are superior for the Kalman filter method for the most sampling frequencies. That is why we focus exclusively on this methodology in our further analysis. It is interesting to note that rolling OLS and DESP do not offer clearly better results compared to the case when beta is fixed.

AVERAGE VALUES	Fixed Beta	rolling OLS	DESP	Kalman
5-minute data	0.96	0.92	1.27	1.21
10-minute data	0.96	0.88	0.77	1.27
20-minute data	0.90	1.03	0.75	1.19
30-minute data	0.97	1.09	0.88	1.34
60-minute data	0.94	0.91	0.99	1.23
Daily data	0.49	-0.33	0.52	0.74

Table 2. Out-of-sample information ratios for the simulated pair trading strategy at different frequencies. Transaction costs have not been considered.

From Table 2 it is also clear that higher sampling frequencies offer more attractive investment characteristics then using daily data for all the methodologies.

In Figure 3 we present adaptive betas calculated using the three approaches mentioned. Both the OLS and DESP beta seem to fluctuate around the Kalman filter beta.

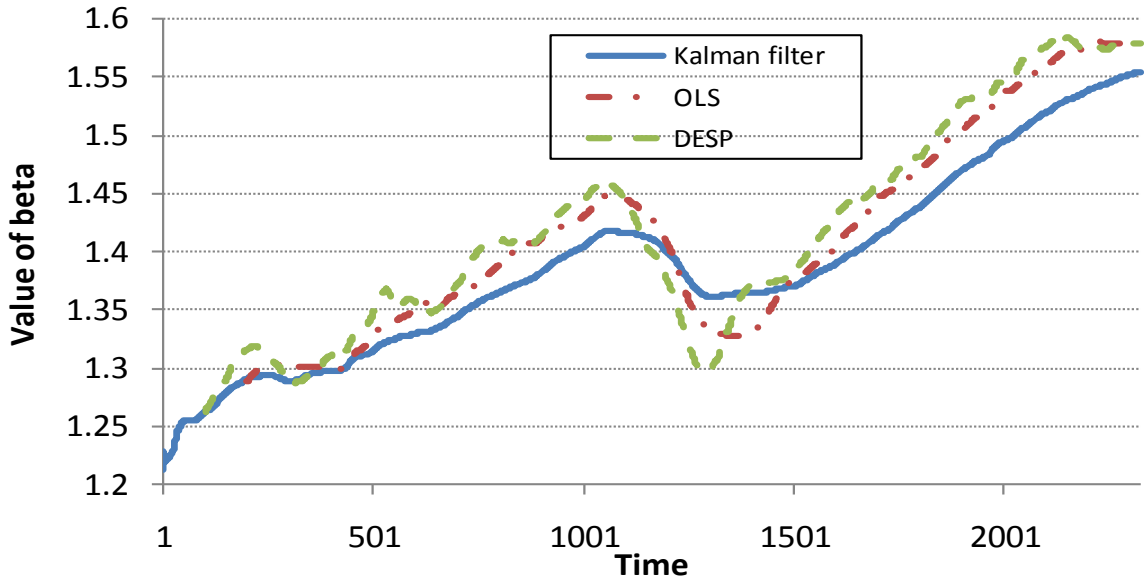


Figure 3. Various betas calculated for the Bayer AG and Arcelor Mittal pair sampled at a 20-minute interval

In Figure 4 we show the distribution of the information ratios including transaction costs for the 20-minute sampling frequency with the Kalman filter used for the beta calculation.

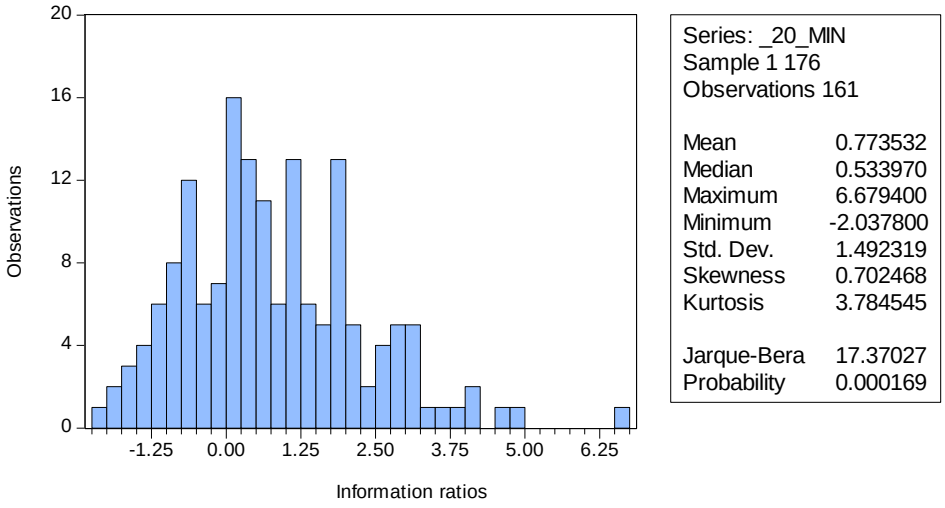


Figure 4.Distribution of information ratios for a 20-minute sampling frequency. One-way transaction costs of 0.4% have been considered.

From the above figure it is clear that an average pair trading is profitable and that pairs are mainly situated to the right of 0.

We also present the distribution of information ratios for daily data to be able to investigate more closely the difference between higher and lower sampling frequencies, see Figure 5.

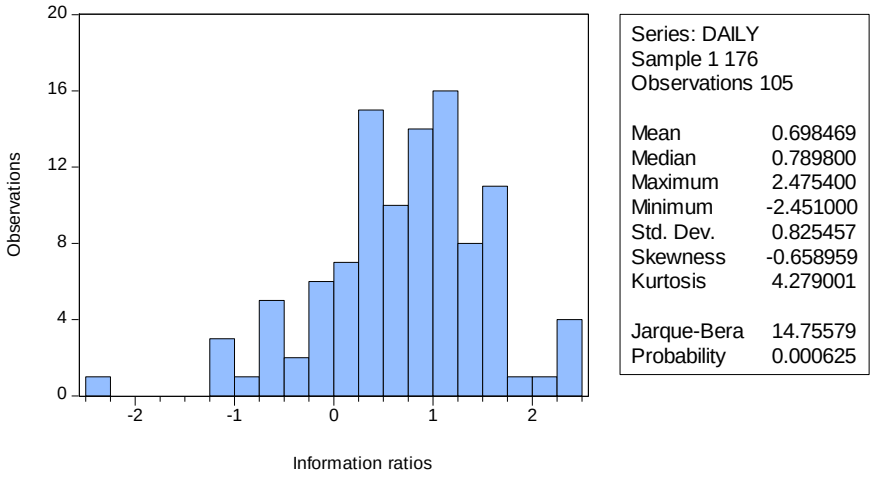


Figure 5. Information ratios of the pairs using daily sampling frequency. One-way transaction costs of 0.4% have been considered.

One important thing to consider is the lower amount of observations. The out-of-sample period for daily data included has only 229 data points (see Table 1) when compared to Figure 4 and the pairs that did not record any transaction were excluded.

Again, as in Figure 4, majority of the information ratios is situated in positive territory. Distributions of information ratios for other sampling frequencies can be seen in Appendices b-e.

The summary statistics for all trading frequencies can be seen in Table 3 below. The main difference between the daily data and high-frequency data is the maximum drawdown and maximum drawdown duration, see Magdon-Ismail (2004). Both these measures are of primary importance to investors. The maximum drawdown defines the total percentage loss experienced by the strategy before it starts "winning" again. In other words, it is the maximum negative distance between the local maximum and subsequent local minimum measured on an equity curve and gives a good measure of the downside risk for the investor (see Appendix g).

AVERAGE VALUES	5-minute	10-minute	20-minute	30-minute	60-minute	Average HF	Daily
Information ratio (ex TC)	1.21	1.27	1.19	1.34	1.23	1.25	0.74
Information ratio (incl. TC)	0.26	0.64	0.77	0.97	0.97	0.72	0.70
Return (ex TC)	16.03%	17.58%	17.12%	20.25%	18.71%	17.94%	19.55%
Return (incl. TC)	1.92%	7.83%	10.33%	14.08%	14.08%	9.65%	18.62%
Volatility (ex TC)	17.55%	18.51%	18.57%	19.35%	19.57%	18.71%	29.57%
Positions taken	49	34	24	21	17	29	3
Maximum drawdown (ex TC)	4.09%	4.25%	4.07%	4.07%	4.08%	4.11%	13.61%
Maximum drawdown duration (ex TC)	5	10	10	20	19	13	79

Table 3. The out-of-sample annualized trading statistics for pair trading strategy with the Kalman filter used for the beta calculation

On the other hand, maximum drawdown duration is expressed as the number of days since the drawdown has begun until the equity curve returns to the same percentage gain as before. Both these measures are important for the psychology of investors, because when the strategy is experiencing a drawdown, investors might start questioning the strategy itself.

Both statistics are significantly higher for daily data than for any higher frequency. The maximum drawdown for daily data is 13.61%, whereas it is 4.11% for high-frequency data. The maximum drawdown duration ranges from 5 to 20 days for the high-frequency data and is as much as 79 days for the daily data.

Information ratios (excluding trading costs) are slightly higher for high-frequency data as has already been shown in Table 2. But when trading costs are considered, high-frequency data are more affected than daily data, as one would expect (due to the higher number of transaction). For instance the information ratio for the 5-minute data drops from an attractive 1.21 to only 0.26 when the trading costs are considered. Average information ratio of the pairs sampled at high-frequencies is 0.72, very similar to the daily sampling frequency (0.70). However, we consider very conservative trading costs which penalize high-frequency data too much, and the information ratio achievable in the real trading might be considerably higher.

c. Further investigations

We further analyze our results below and address some interesting issues from an investment perspective.

i. Relationship between the in-sample t-stats and the out-of-sample information ratios

We examine whether the in-sample cointegration of a given trading pair implies better out-of-sample performance. One can logically assume that a higher stationarity of the residual from the cointegration equation implies a higher confidence that the pair will revert to its mean. Thus we would expect a significant positive correlation between the t-stat of the ADF test on the OLS residuals and the out-of-sample information ratio. We perform this analysis only on daily data. We deal with intraday data later.

We bootstrap (with replacement) the pairs consisting of information ratios and t-stats[14]. The t-stat is obtained from the coefficient of the ADF test of the cointegrating equation. After bootstrapping (with replacement) the correlation coefficient 5,000 times at a 95% confidence interval, we obtain a lower/upper limits for the coefficients shown in Table 4 below.

[14] Our objective is to analyze the relation between the t-stat and the information ratio for all the pairs. Instead of calculating a point estimate of a correlation coefficient, we prefer to calculate the confidence intervals of a true correlation coefficient. We perform bootstrapping with replacement, the standard computer-intensive technique used in statistical inference to find confidence intervals of an estimated variable, see e.g. Efron, B. and Tibshirani, R. J. (1993) An Introduction to the Bootstrap, Chapman & Hall, New York.. It is a quantitative process in which we randomly repeat the selection of data (we repeat it for 5,000 times). Some samples might contain the same item more than once (hence the bootstrapping with replacement), whereas others may not be included at all. The process provides a new set of samples which is then used to calculate the unbiased confidence intervals for the true correlation coefficient. Bootstrapping in our case is a simple process of creating 5000 random samples from the original data set in such a way, that the corresponding pairs are selected 176 times from an original data set to form each of 5000 samples.

in-sample t-stats vs. oos information ratio	5-minute	10-minute	20-minute	30-minute	60-minute	Daily
LOWER	0.04	-0.05	-0.18	-0.22	-0.26	-0.18
UPPER	0.32	0.23	0.13	0.10	0.09	0.22

Table 4. 95% confidence intervals of the correlation coefficients between t-stats generated in the in-sample period and the out-of-sample information ratios

The in-sample t-stat seems to have certain predictive power for the out-of-sample information ratio, although not for all the frequencies. The only frequencies for which the t-stat works are data sampled 5- and 10-intervals. For all the other frequencies the centre of the distribution is either very close to 0 (20-minute and daily data) or slightly negative (30-minute and 60-minute data). For instance, 95% confidence intervals for daily data are almost perfectly centred around 0 (-0.18 and 0.22), implying that the true correlation coefficient might be 0.

ii. Relationship between t-stats for different high-frequencies and pairs

In this paper we have various sampling frequencies defined as high frequency. Those are data sampled at 5-, 10-, 20-, 30- and 60-minute intervals. In this section we investigate whether there is a certain structure in their t-stats which could help us reduce the dimensionality of higher frequencies. This would enable us to pick only one higher frequency representative of all the intervals for further analysis.

To do that we apply principal component analysis (PCA) to all the high-frequency pairs, see Jollife (1986) for the most comprehensive reference of PCA. PCA is a statistical technique which tries to find linear combinations of the original assets accounting for the highest possible variance of the total variance of the data set. If there is a strong common behaviour of the assets, in our case the t-stats across different pairs and frequencies, just a few first principal components should suffice to explain the behaviour of the entire data set.

As the first step to obtain the data suitable as an input to PCA we form the matrix of t-stats from the ADF test. Each row of the matrix contains t-stats for different pairs (we have 176 rows, the same amount as the number of pairs) and each column contains t-stats for these pairs sampled at different frequencies (thus we have 5 columns, one for 5-, 10-, 20-, 30- and 60-minute interval). The matrix is normalized across the columns by subtracting the mean and dividing by the standard deviation of each column. In this way, we obtain a matrix with mean 0 and unit variance in each column.

The covariance of such a normalized matrix serves as an input for a principal components analysis. The first principal component explains over 97.9% of the variation in the data, confirming that there is a clear structure in the dataset. This means that trading pairs have similar t-stats across all the frequencies (in other words columns of the original matrix are similar).

This finding is further reinforced by comparing variances between t-stats. From the original matrix of t-stats, we calculate variances for each frequency. We obtain 5 variances between the pairs (1 for each high frequency), which all vary around 0.58, quite a high variance for t-stats when considering that t-stats range from 0.18 to 2.83. Then we compute the variance of the t-stats for each of the 176 pairs across

the 5 frequencies. These are much smaller in magnitude, the maximum variance being just around 0.14. Thus, the fact that variances between different frequencies are small when considering each of the 176 pairs, but variances between the pairs are high further demonstrates that t-stats tend to be similar across all the frequencies for any particular pair.

As a conclusion we summarize that once a pair has been found to be cointegrated (in any time interval higher than the daily data) it tends to be cointegrated across all the frequencies. Hence we only need to look at one frequency.

iii. Does cointegration in daily data imply higher frequency cointegration?

We just demonstrated that there is a clear structure in the high-frequency dataset of the t-stats. The conclusion was that it is sufficient to consider only one higher frequency (here we decide for 5-minute data) as a representative for all the high-frequencies. In this section we investigate the relationship between the t-stats for daily data (computed from 1st January 2009 to 9th September 2009 for daily data) and the t-stats for 5-minute data (computed from the out-of-sample period for 5-minute data, i.e. 10th September 2009 to 17th November 2009).

We perform bootstrapping (with replacement) to obtain confidence intervals of the true correlation coefficient. The dataset is bootstrapped 5,000 times and a 95% confidence intervals are -0.03/0.33.

The boundaries of the confidence intervals imply that there is a possible relation between the variables. The true correlation coefficient is probably somewhere around 0.15 (in the centre of the confidence intervals mentioned above). Thus, cointegration found in daily data implies that the spread should be stationary for trading purposes in the high-frequency domain.

iv. Does in-sample information ratio and the half-life of the mean reversion indicate what the out-of-sample information ratio will be?

We showed above that there is a relationship between the profitability of the strategy and the stationarity of the spread computed from the t-stat of the ADF test. Here we try to find additional in-sample indicators (by looking at the in-sample information ratio and the half-life of mean reversion) of the out-of-sample profitability (measured by the information ratio) of the pair.

We follow the same bootstrapping procedure we already performed in the previous sections to estimate the confidence intervals. That is, bootstrapping is performed 5,000 times (with replacement) as in other cases.

In Table 5 below we show the bootstrapped correlation coefficients among the in- and out-of-sample information ratios (not taking into account transaction costs) across all frequencies.

in-sample vs. oos information ratio	5-minute	10-minute	20-minute	30-minute	60-minute	Daily
LOWER	-0.02	0.10	-0.09	-0.26	-0.16	0.07
UPPER	0.31	0.42	0.26	0.07	0.15	0.32

Table 5. 95% confidence intervals of the correlation coefficients between information ratios generated in the in- and out-of-sample periods

The confidence bounds indicate that the in-sample information ratio can predict the out-of-sample information ratio to a certain extent. Whereas in Table 4 the t-stat only worked for 5- and 10-minute data, the information ratio works for data sampled 5-, 10-, 20-minute and daily intervals. On the other hand, the in-sample information ratio does not work well for 30- and 60-minute data. We assume that the relationship should be positive whereas for 30- and 60-minute data the centre between the confidence bounds is negative and close to 0, respectively. Overall, the average lower/upper interval across all the frequencies presented is -0.06/0.26.

Next we perform a bootstrapping of the pairs consisting of the in-sample half-life of mean reversion and the out-of-sample information ratio.

We show the 95% confidence interval bounds of the true correlation coefficient in Table 6 below. As we would expect, the lower the half-life is, the higher the information ratio of the pair. The extent of the dependence is slightly lower than the one presented in Table 5. The average lower/upper interval across all the frequencies presented is -0.20/0.06. So we find that there is negative relation between the half-life of mean reversion and subsequent out-of-sample information ratio.

half-life vs. oos information ratio	5-minute	10-minute	20-minute	30-minute	60-minute	Daily
LOWER	-0.18	-0.25	-0.24	-0.19	-0.15	-0.19
UPPER	0.08	-0.01	0.00	0.08	0.13	0.08

Table 6. 95% confidence intervals of the correlation coefficients between the in-sample half-life of mean reversion and the out-of-sample information ratios

Thus the 2 indicators presented here seem to have certain predictive power as to the out-of-sample information ratio of the trading pair.

7. A DIVERSIFIED PAIR TRADING STRATEGY

Standalone results of trading the pairs individually are quite attractive as shown in Table 3 but here we try to improve them using the findings from the previous section. We use the indicators mentioned just above to select the 5 best pairs for trading and present the results in what follows.

First, we present the results of using each indicator individually. Results of selecting 5 pairs based on the best in-sample information ratios are shown in Table 7 below.

AVERAGE VALUES	5-minute	10-minute	20-minute	30-minute	60-minute	Average HF	Daily
Information ratio IN-SAMPLE (incl. TC)	5.65	6.21	6.57	6.81	6.77	6.40	0.90
Information ratio (ex TC)	3.22	9.31	3.44	3.92	1.27	4.23	1.39
Information ratio (incl. TC)	2.27	7.71	2.58	2.88	0.75	3.24	1.32
Return (incl. TC)	21.14%	33.63%	15.16%	13.63%	5.27%	17.77%	18.50%
Volatility (ex TC)	9.30%	4.36%	5.88%	4.73%	7.02%	6.26%	14.03%
Maximum drawdown (ex TC)	3.02%	0.78%	1.19%	1.49%	1.42%	1.58%	4.26%
Maximum drawdown duration (ex TC)	7	17	18	33	34	22	55

Table 7. The out-of-sample information ratios for 5 selected pairs based on the best in-sample information ratios

Information ratios improve for pairs sampled at the high-frequency and daily intervals. The improvement is the most noticeable for pairs sampled at the high-frequency intervals, when the average information ratio net of trading costs for the high-frequency data improves from 0.72 as in Table 3 to 3.24. The information ratio for daily data improves as well (from 0.7 to 1.32). Almost all the information ratios for the pairs sampled at the high-frequency intervals are above 2, a truly attractive result for the strategy.

Maximum drawdown and maximum drawdown duration favour the pairs sampled at the high-frequency intervals as well. The average maximum drawdown for the pairs sampled at the high-frequency intervals is 1.58%, much less than the drawdown for the pairs sampled at a daily interval (4.26%). The maximum drawdown duration is 22 days on average for the high-frequency data, and 55 days for the daily data.

In Table 8 below we show trading results based on the half-life of the mean reversion as an indicator. Thus, 5 pairs with the lowest half-life of the mean reversion were selected to form the portfolio.

AVERAGE VALUES	5-minute	10-minute	20-minute	30-minute	60-minute	Average HF	Daily
Information ratio IN-SAMPLE (incl. TC)	0.58	1.33	4.35	4.42	5.51	3.24	0.46
Information ratio (ex TC)	1.59	6.42	4.35	1.34	0.59	2.86	0.57
Information ratio (incl. TC)	-3.32	0.34	0.10	-0.85	-0.04	-0.75	0.50
Return (incl. TC)	-18.27%	1.25%	0.26%	-2.40%	-0.26%	-3.88%	6.71%
Volatility (ex TC)	5.50%	3.63%	2.63%	2.83%	7.01%	4.32%	13.43%
Maximum drawdown (ex TC)	0.81%	0.92%	0.93%	1.36%	1.84%	1.17%	3.07%
Maximum drawdown duration (ex TC)	3	7	16	29	34	18	57

Table 8. The out-of-sample trading statistics for 5 pairs selected based on the best in-sample half-life of mean reversion

The information ratios net of trading costs are not attractive, with 0.50 being the highest and -3.32 being the lowest. The average information ratio for the pairs sampled at the high-frequency interval is -0.75, which means that the average pair is not profitable. The information ratio of the pairs sampled at a daily interval is 0.5, which is profitable, but worse than the basic case shown in Table 3. Thus we decide not to take the half-life of mean reversion into consideration as a prospective indicator of the future profitability of the pair.

In Table 9 below we show the results of using the in-sample t-stats of the ADF test of the cointegrating regression as the indicator of the out-of-sample information ratios.

AVERAGE VALUES	5-minute	10-minute	20-minute	30-minute	60-minute	Average HF	Daily
Information ratio IN-SAMPLE (incl. TC)	2.16	2.37	3.32	3.39	3.71	2.99	0.38
Information ratio (ex TC)	12.05	6.13	1.47	-0.22	1.28	4.14	-0.05
Information ratio (incl. TC)	5.60	2.47	-1.18	-0.90	0.15	1.23	-0.08
Return (incl. TC)	13.53%	6.49%	-3.38%	-6.49%	0.69%	2.17%	-1.50%
Volatility (ex TC)	2.42%	2.62%	2.88%	7.23%	4.56%	3.94%	18.82%
Maximum drawdown (ex TC)	0.52%	0.57%	1.21%	1.19%	1.23%	0.94%	3.64%
Maximum drawdown duration (ex TC)	3	7	18	35	39	20	74

Table 9. The out-of-sample trading statistics for 5 pairs selected based on the best in-sample t-stats of the ADF test

Focusing on the information ratios after transaction costs, they are worse than when the in-sample information ratio was used as an indicator. The out-of-sample information ratio after deduction of transaction costs is higher using the t-stats than using the in-sample information ratio only for a 5-minute data. For all the other frequencies, the in-sample information ratio is a better indicator.

In Table 10 below we present the results of using the t-stat of the ADF test for daily data (from 1st January 2009 to 9th September 2009) as an indicator of the out-of-sample information ratio of the pairs sampled at the high-frequency intervals. Average information ratio for all the high-frequency trading pairs is around 3, which makes it the second best indicator after the in-sample information ratio.

AVERAGE VALUES	5-minute	10-minute	20-minute	30-minute	60-minute	Average HF
Information ratio IN-SAMPLE (incl. TC)	1.08	1.25	0.75	1.18	1.34	1.12
Information ratio (ex TC)	6.86	8.95	4.62	3.73	2.40	5.31
Information ratio (incl. TC)	2.12	5.40	3.12	2.64	1.75	3.01
Return (incl. TC)	7.65%	18.96%	15.89%	16.28%	12.55%	14.27%
Volatility (ex TC)	3.61%	3.51%	5.09%	6.16%	7.15%	5.11%
Maximum drawdown (ex TC)	0.79%	0.66%	0.92%	1.03%	1.43%	0.97%
Maximum drawdown duration (ex TC)	4	5	5	10	13	7

Table 10. The out-of-sample trading statistics for selected 5 pairs based on the best in-sample t-stats of the ADF test for daily data

We also include an equally weighted combination of the indicators. We use the formula below:

$$Combined\_ranking = \frac{R_1 + R_2}{2} \tag{16}$$

where R_1 and R_2 and are the rankings based on the in-sample information ratio and the in-sample t-stat of the series sampled at a daily interval. In other words, we assign a ranking from 1 to 176 to each pair of shares based on the 2 indicators mentioned just above. Then we calculate the average ranking for each trading pair and reorder them based on the new ranking values. Finally we form the portfolio of the first 5 trading pairs.

The trading results of the combined ratio are presented in the Table 12 below.

AVERAGE VALUES	5-minute	10-minute	20-minute	30-minute	60-minute	Average HF	Daily
Information ratio IN-SAMPLE (incl. TC)	1.12	-0.81	-0.25	-0.04	0.99	0.20	0.20
Information ratio (ex TC)	0.73	3.43	4.11	6.61	8.01	4.58	0.43
Information ratio (incl. TC)	-0.52	1.75	2.92	5.25	6.78	3.24	0.35
Return (incl. TC)	-6.03%	9.42%	18.01%	26.92%	35.11%	16.69%	5.00%
Volatility (ex TC)	11.67%	5.40%	6.17%	5.13%	5.18%	6.71%	14.15%
Maximum drawdown (ex TC)	3.58%	1.11%	1.00%	1.05%	1.53%	1.65%	5.21%
Maximum drawdown duration (ex TC)	1,783	855	257	157	54	21	169

Table 11. The out-of-sample trading statistics for 5 best pairs selected based on combined ratio calculated according to Equation (16)

The average information ratio for the pairs sampled at the high-frequency intervals is 3.24. Unfortunately, the pair trading strategy using daily data only achieves information ratio of 0.35 after transaction costs, which is even worse than the original, unoptimized case.

We also combine the t-stat of the ADF test for a given high-frequency and the information and obtain attractive results. Although the average information ratio net of trading costs for the trading pairs sampled at the high-frequency intervals is higher than was the case in Table 3 (thus when no indicator was used), the information ratios for the 20- and 60-minute sampling frequencies are negative and thus results are not consistent across all the high-frequency intervals. This in our opinion disqualifies the usage of this indicator for predicting the future profitability of the pairs.

AVERAGE VALUES	5-minute	10-minute	20-minute	30-minute	60-minute	Average HF	Daily
Information ratio IN-SAMPLE (incl. TC)	0.96	2.21	1.25	4.87	4.08	2.67	0.51
Information ratio (ex TC)	3.02	15.80	-0.05	2.03	-0.12	4.14	0.46
Information ratio (incl. TC)	1.30	7.60	-0.58	0.92	-0.52	1.74	0.43
Return (incl. TC)	7.61%	7.92%	-3.91%	4.33%	-4.49%	2.29%	6.81%
Volatility (ex TC)	5.87%	1.04%	6.78%	4.68%	8.63%	5.40%	15.96%
Maximum drawdown (ex TC)	0.71%	0.92%	1.81%	1.74%	1.85%	1.41%	5.65%
Maximum drawdown duration (ex TC)	4	8	19	38	61	26	40

Table 12. The out-of-sample trading statistics for 5 best pairs selected based on the combined ratio of the in-sample t-stat of the ADF test and the in-sample information ratio

To summarize, on the one hand we were able to improve the information ratios net of trading costs for daily data from around 0.7 as in Table 3 to 1.3 as in Table 7 using the in-sample information ratio as an indicator of the future profitability of the pairs.

On the other hand, 3 different indicators heavily improved the attractiveness of the results for the pairs sampled at the high-frequency intervals. We were able to increase the out-of-sample information ratio from 0.72 as in Table 3 (the average out-of-sample information ratio for all the 176 pairs sampled at the high-frequency intervals) to around 3, using the in-sample information ratio, the t-stat of the ADF test of the series sampled at a daily interval and a combination of the two (see Table 7, Table 10 and Table 11).

Below we compare the results of the pair trading strategy at both frequencies (an average of all the high-frequency intervals and a daily one) with the appropriate benchmarks. In practice, one would choose only one high-frequency interval to trade, but here we look at an average, which represents all the frequencies for reasons of presentation. In fact, pairs sampled at all the high-frequency intervals are

attractive for trading purposes when the in-sample information ratio is used as the indicator of the future profitability. Due to homogeneity we also use the in-sample information ratio as the indicator for the pairs sampled at daily interval.

In Table 13 below we present a comparison of our pair trading strategy sampled at a daily interval with the results of buy and hold strategy of the the Eurostoxx 50 index and Market Neutral Index (HFRXEMN Index in Bloomberg). The results span from 1st January 2009 to 17th November 2009, the out-of-sample period for our pairs sampled at a daily interval.

AVERAGE VALUES	Market neutral index	Eurostoxx 50	Daily Strategy
Information Ratio (incl. TC)	-1.04	0.54	1.32
Return (incl. TC)	-4.56%	15.34%	18.50%
Volatility (incl. TC)	4.36%	28.62%	14.03%
Maximum drawdown (ex TC)	6.20%	33.34%	4.26%
Maximum drawdown duration (ex TC)	188	44	55

Table 13. Annualized trading statistics compared in the out-of-sample period for the pair trading strategy sampled at daily interval, with the in-sample information ratio used as the indicator of the future profitability of the strategy

The strategy outperforms its primary benchmark, the Market neutral index both on the absolute and risk-adjusted basis. While the market neutral index lost money during the period, our strategy was profitable without showing excessive volatility relative to the return. It also outperformed the corresponding market index, the Eurostoxx 50 index.

In Table 14 below we compare the results of the average high-frequency pair trading strategy with the appropriate benchmarks in the period from 10th September 2009 to 17th November 2009. The information ratio of 3.24 of the pair trading strategy is considerably higher than any of the two indices. Thus, using high-frequency sampling frequency seems to offer significant improvement of the investment characteristics of the pair trading strategy. It offers a comparable absolute return to the one achieved by the Eurostoxx 50 index, with significantly lower volatility.

AVERAGE VALUES	Market neutral index	Eurostoxx 50	HF Strategy
Information Ratio (incl. TC)	0.90	0.78	3.24
Return (incl. TC)	3.55%	16.40%	17.77%
Volatility (incl. TC)	3.96%	21.10%	6.26%
Maximum drawdown (ex TC)	1.64%	8.31%	1.58%
Maximum drawdown duration (ex TC)	19	11	22

Table 14. Annualized trading statistics compared in the out-of-sample period for pair trading strategy sampled at the high-frequency interval, with the in-sample information ratio used as the indicator of the future profitability of the strategy

8. CONCLUDING REMARKS

In this article we apply a pair trading strategy to the constituent shares of the Eurostoxx 50 index. We implement a basic long-short trading strategy which is used to trade shares sampled at 6 different frequencies, namely data sampled at 5-minute, 10-minute, 20-minute, 30-minute, 60-minute and daily intervals.

First, we divide shares into industry groups and form pairs of shares that belong to the same industry. The Kalman filter approach is used to calculate an adaptive beta for each pair.

Subsequently, we calculate the spread between the shares and simulate trading activity based on 2 simple trading rules. We enter the position (long or short) whenever the spread is more than 2 standard deviations away from its long-term mean. All positions are liquidated when the spread returns to its long-term mean (defined as its distance being lower than 0.5 standard deviations from the long-term mean), that is, technically, when it reverts towards the long-term mean.

As such, standalone pair trading results are not very attractive. That is why we introduce a novel approach to select the best pairs for trading based on the in-sample information ratio of the series, the in-sample t-stat of the ADF test of the series sampled at a daily interval and a combination of the two, as these are shown to be good indicators of the out-of-sample profitability of the pair.

We then build a diversified pair trading portfolio based on the 5 trading pairs with the best in-sample indicator value. Our diversified approach is able to produce information ratios of over 3 for a high frequency sampling interval (an average across all the high-frequency intervals considered), and 1.3 for a daily sampling frequency using the in-sample information ratio as an indicator. This is a very attractive result when compared to the performance of the Eurostoxx 50 index and the index of Market Neutral Hedge Funds with information ratios lower than 1 during the review period. It also shows how useful the combination of the high-frequency data and the concept of cointegration can be for quantitative fund management.

APPENDICES

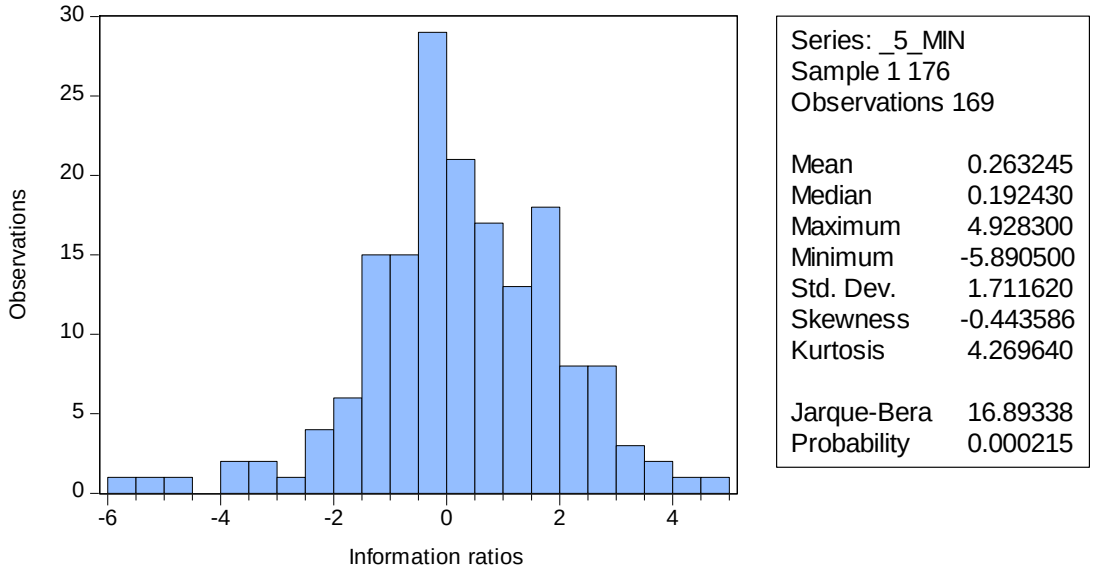
a. Kalman filter estimation procedure

The full specification of the model:

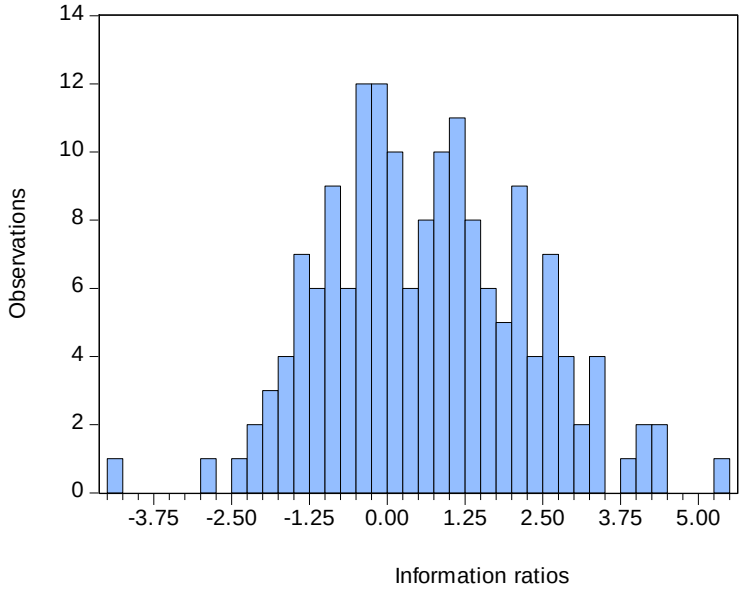
$$\begin{aligned} \beta_{t|t-1} &= \beta_t \\ v_t &= Y_t - X_t\beta_t \\ F_t &= X_t P_t X'_t + H \\ \beta_{t+1} &= \beta_t + P_t X'_t \frac{v_t}{F_t} \\ P_{t+1} &= P_t - P_t X'_t X_t P_t \frac{1}{F_t} + Q \end{aligned} \tag{17}$$

The parameters that need to be set in advance are H and Q, which could be defined as the error terms of the process. Their values in isolation are not important. The most important parameter of the Kalman filter procedure is the noise ratio, which is defined as $noiseRatio = \frac{Q}{H}$. The higher the ratio, the more adaptive beta, the lower the ratio, the less adaptive beta. Thus, if we used extremely low value of noise ratio, e.g. 10^{-10}, the beta would be fixed along the dataset. Also, it is important to correctly initialize the value of beta, as in the second equation, $v_{t+1} = Y_t - X_t\beta_t$, there is no way of knowing what β_t will be at the first step. Thus, we have set β_1 to be $\beta_1 = \frac{Y_1}{X_1}$, thus the initial error term being 0.

b. Distribution of information ratios for a 5-minute sampling frequency. Kalman filter was used for the beta calculation

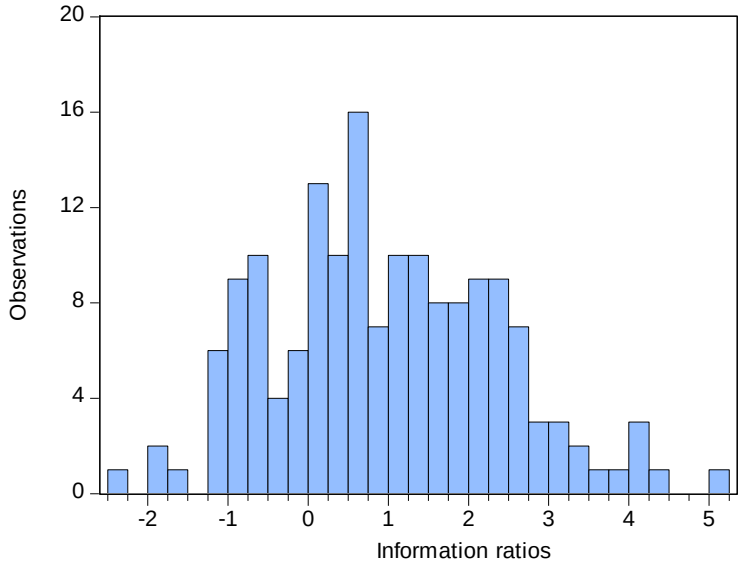


c. Distribution of information ratios for a 10-minute sampling frequency. Kalman filter was used for the beta calculation



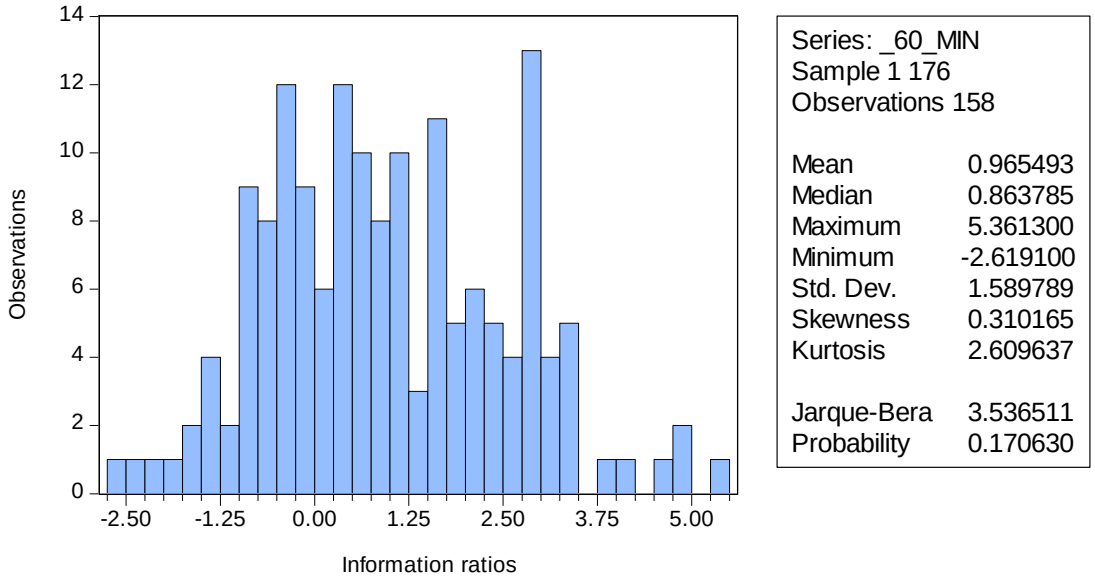
Series: _10_MIN	
Sample 1 176	
Observations 164	
Mean	0.643384
Median	0.544800
Maximum	5.275800
Minimum	-4.279000
Std. Dev.	1.614723
Skewness	0.195924
Kurtosis	2.957080
Jarque-Bera	1.061808
Probability	0.588073

d. Distribution of information ratios for a 30-minute sampling frequency. Kalman filter was used for the beta calculation



Series: _30_MIN	
Sample 1 176	
Observations 161	
Mean	0.973128
Median	0.887350
Maximum	5.056900
Minimum	-2.392900
Std. Dev.	1.399748
Skewness	0.246636
Kurtosis	2.788876
Jarque-Bera	1.931273
Probability	0.380741

e. Distribution of information ratios for a 60-minute sampling frequency. Kalman filter was used for the beta calculation



f. Constituent stocks of Eurostoxx 50 index which we used to form the pairs

Number	Company name	Bloomberg Ticker	Industrial sector
1	Air Liquide SA	AI FP Equity	Basic Materials
2	ArcelorMittal	MT NA Equity	Basic Materials
3	BASF SE	BAS GY Equity	Basic Materials
4	Bayer AG	BAYN GY Equity	Basic Materials
5	Deutsche Telekom AG	DTE GY Equity	Communications
6	France Telecom SA	FTE FP Equity	Communications
7	Nokia OYJ	NOK1V FH Equity	Communications
8	Telecom Italia SpA	TIT IM Equity	Communications
9	Telefonica SA	TEF SQ Equity	Communications
10	Vivendi SA	VIV FP Equity	Communications
11	Daimler AG	DAI GY Equity	Consumer, Cyclical
12	Volkswagen AG	VOW GY Equity	Consumer, Cyclical
13	Anheuser-Busch InBev NV	ABI BB Equity	Consumer, Non-cyclical
14	Carrefour SA	CA FP Equity	Consumer, Non-cyclical
15	Groupe Danone SA	BN FP Equity	Consumer, Non-cyclical
16	L'Oreal SA	OR FP Equity	Consumer, Non-cyclical
17	Sanofi-Aventis SA	SAN FP Equity	Consumer, Non-cyclical
18	Unilever NV	UNA NA Equity	Consumer, Non-cyclical
19	LVMH Moet Hennessy Louis Vuitton SA	MC FP Equity	Diversified
20	ENI SpA	ENI IM Equity	Energy
21	Repsol YPF SA	REP SQ Equity	Energy
22	Total SA	FP FP Equity	Energy
23	Aegon NV	AGN NA Equity	Financial
24	Allianz SE	ALV GY Equity	Financial
25	AXA SA	CS FP Equity	Financial
26	Banco Santander SA	SAN SQ Equity	Financial
27	Banco Bilbao Vizcaya Argentaria SA	BBVA SQ Equity	Financial
28	BNP Paribas	BNP FP Equity	Financial
29	Credit Agricole SA	ACA FP Equity	Financial
30	Deutsche Bank AG	DBK GY Equity	Financial
31	Deutsche Boerse AG	DB1 GY Equity	Financial
32	Assicurazioni Generali SpA	G IM Equity	Financial
33	ING Groep NV	INGA NA Equity	Financial
34	Intesa Sanpaolo SpA	ISP IM Equity	Financial
35	Muenchener Rueckversicherungs AG	MUV2 GY Equity	Financial
36	Societe Generale	GLE FP Equity	Financial
37	UniCredit SpA	UCG IM Equity	Financial
38	Alstom SA	ALO FP Equity	Industrial
39	CRH PLC	CRH ID Equity	Industrial
40	Koninklijke Philips Electronics NV	PHIA NA Equity	Industrial
41	Cie de Saint-Gobain	SGO FP Equity	Industrial
42	Schneider Electric SA	SU FP Equity	Industrial
43	Siemens AG	SIE GY Equity	Industrial
44	Vinci SA	DG FP Equity	Industrial
45	SAP AG	SAP GY Equity	Technology
46	E.ON AG	EOAN GY Equity	Utilities
47	Enel SpA	ENEL IM Equity	Utilities
48	GDF Suez	GSZ FP Equity	Utilities
49	Iberdrola SA	IBE SQ Equity	Utilities
50	RWE AG	RWE GY Equity	Utilities

g. Calculation of the trading statistics

Annualised Return	$R^A = 252 * \frac{1}{N} \sum_{t=1}^{N} R_t$ *with* R_t *being the daily return*
Annualised Volatility	$\sigma^A = \sqrt{252} * \sqrt{\frac{1}{N-1} * \sum_{t=1}^{N} \left(R_t - \overline{R}\right)^2}$
Information Ratio	$IR = \frac{R^A}{\sigma^A}$
Maximum Drawdown	*Maximum negative value of* $\sum \left(R_t\right)$ *over the period* $MD = \underset{i=1,\cdots,t;t=1,\cdots,N}{Min} \left(\sum_{j=i}^{t} R_j \right)$
Information Ratio	$SR = \frac{R^A - R_F}{\sigma^A}$ *, where* R_F *is the risk free rate.*

References

Aldridge, I. (2009) *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems,* John Wiley & Sons, Inc., New Jersey.

Alexander, C. (2001) *Market Models: A Guide to Financial Data Analysis,* John Wiley & Sons Ltd., Chichester.

Alexander, C. (2008) *Market Risk Analysis: Practical Financial Econometrics,* John Wiley & Sons, Ltd., Chichester.

Alexander, C. and Dimitriu, A. (2002) The Cointegration Alpha: Enhanced Index Tracking and Long-Short Equity Market Neutral Strategies. *SSRN eLibrary*,

Bentz, Y. (2003) Quantitative Equity Investment Management with Time-Varying Factor Sensitivities. In Dunis, C., Laws, J. And Naïm, P. [eds.] *Applied Quantitative Methods for Trading and Investment.* John Wiley & Sons, Chichester, 213-237.

Brooks, R. D., Faff, R. W. and Mckenzie, M. D. (1998) Time-Varying Beta Risk of Australian Industry Portfolios: A Comparison of Modelling Techniques. *Australian Journal of Management,* 23, 1-22.

Burgess, A. N. (1999) A Computational Methodology for Modelling the Dynamics of Statistical Arbitrage,London Business School, PhD Thesis

Burgess, A. N. (2003) Using Cointegration to Hedge and Trade International Equities. In Dunis, C., Laws, J. And Naïm, P. [eds.] *Applied Quantitative Methods for Trading and Investment.* John Wiley & Sons, Chichester, 41-69.

Capocci, D. P. (2006) The Neutrality of Market Neutral Funds. *Global Finance Journal, June 2005,* 17, 2, 309-333.

Chan, E. (2009) *Quantitative Trading: How to Build Your Own Algorithmic Trading Business,* John Wiley & Sons, Inc., New Jersey.

Choudhry, T. and Wu, H. (2009) Forecasting the Weekly Time-Varying Beta of Uk Firms: Garch Models Vs. Kalman Filter Method. *The European Journal of Finance,* 15, 4, 437-444.

Conn, A. R., Gould, N. I. M. and Toint, P. L. (1991) A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds. *SIAM Journal on Numerical Analysis,* 28, 2, 545–572.

Dias, M. A. G. and Rocha, K. (1999) Petroleum Concessions with Extendible Options: Investment Timing and Value Using Mean Reversion and Jump Processes for Oil Prices. *Institute for Applied Economic Research Working Paper No. 620*

Dunis, C. L. and Ho, R. (2005) Cointegration Portfolios of European Equities for Index Tracking and Market Neutral Strategies. *Journal of Asset Management,* 6, 1, 33-52.

Dunis, C. L. and Shannon, G. (2005) Emerging Markets of South-East and Central Asia: Do They Still Offer a Diversification Benefit? *Journal of Asset Management,* 6, 3, 168-190.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap,* Chapman & Hall, New York.

Enders, W. and Granger, C. W. J. (1998) Unit-Root Tests and Asymmetric Adjustment with an Example Using the Term Structure of Interest Rates. *Journal of Business & Economic Statistics,* 16, 3, 304-311.

Engle, R. F. and Granger, C. W. J. (1987) Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica,* 55, 2, 251-76.

Fung, W. and Hsieh, D. A. (1997) Empirical Characteristics of Dynamic Trading Strategies: The Case of Hedge Funds *The Review of Financial Studies,* 10, 2, 275-302.

Galenko, A., Popova, E. and Popova, I. (2007) Trading in the Presence of Cointegration. *SSRN eLibrary*,

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

Giraldo Gomez, N. (2005) Beta and Var Prediction for Stock Portfolios Using Kalman's Filter and Garch Models. *Cuadernos de Administración*, 18, 29, 103-120.

Goldberg, D. E. (1989) *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison-Wesley.

Hamilton, J. (1994) *Time Series Analysis*, Princeton University Press, Princeton.

Harvey, A. C. (1981) *Time Series Models*, Philip Allan Publishers, Oxford.

Johansen, S. (1988) Statistical Analysis of Cointegration Vectors. *Journal of Economic Dynamics and Control*, 12, 231-254.

Jollife, I. T. (1986) *Principal Component Analysis*, Springer-Verlag, New York.

Khandani, A. E. and Lo, A. W. (2007) What Happened to the Quants in August 2007? *Journal of Investment Management*, 5, 4, 5-54.

Laviola, J. (2003a) Double Exponential Smoothing: An Alternative to Kalman Filter-Based Predictive Tracking. In *Proceedings of In Proceedings of the Immersive Projection Technology and Virtual Environments*. ACM Press, 199-206.

Laviola, J. (2003b) An Experiment Comparing Double Exponential Smoothing and Kalman Filter-Based Predictive Tracking Algorithms. In *Proceedings of In Proceedings of Virtual Reality*. 283-284.

Liang, B. (1999) On the Performance of Hedge Funds. *Financial Analysts Journal*, 55, 4, 72-85.

Lin, Y.-X., Mccrae, M. and Gulati, C. (2006) Loss Protection in Pairs Trading through Minimum Profit Bounds: A Cointegration Approach. *Journal of Applied Mathematics and Decision Sciences*, vol. 2006, 1-14.

Magdon-Ismail, M. (2004) Maximum Drawdown. *Risk Magazine*, 17, 10, 99-102.

Mudchanatongsuk, S., Primbs, J. A. and Wong, W. (2008) Optimal Pairs Trading: A Stochastic Control Approach. In *Proceedings of In Proceedings of the American Control Conference*. Seattle, 1035-1039.

Nath, P. (2003) High Frequency Pairs Trading with U.S. Treasury Securities: Risks and Rewards for Hedge Funds. *SSRN eLibrary*,

Said, E. S. and Dickey, A. D. (1984) Testing for Unit Roots in Autoregressive Moving-Average Models of Unknown Orders. *Biometrika*, 71, 3, 599-607.

Sharpe, W. F. (1994) The Sharpe Ratio. *Journal of Portfolio Management*, 21, 1, 49-58.

Vidyamurthy, G. (2004) *Pairs Trading - Quantitative Methods and Analysis*, John Wiley & Sons, Inc., New Jersey.

Wu, Y., Balvers, R. J. and Gilliland, E. (2000) Mean Reversion across National Stock Markets and Parametric Contrarian Investment Strategies. *The Journal of Finance*, 55, 2, 745-772.