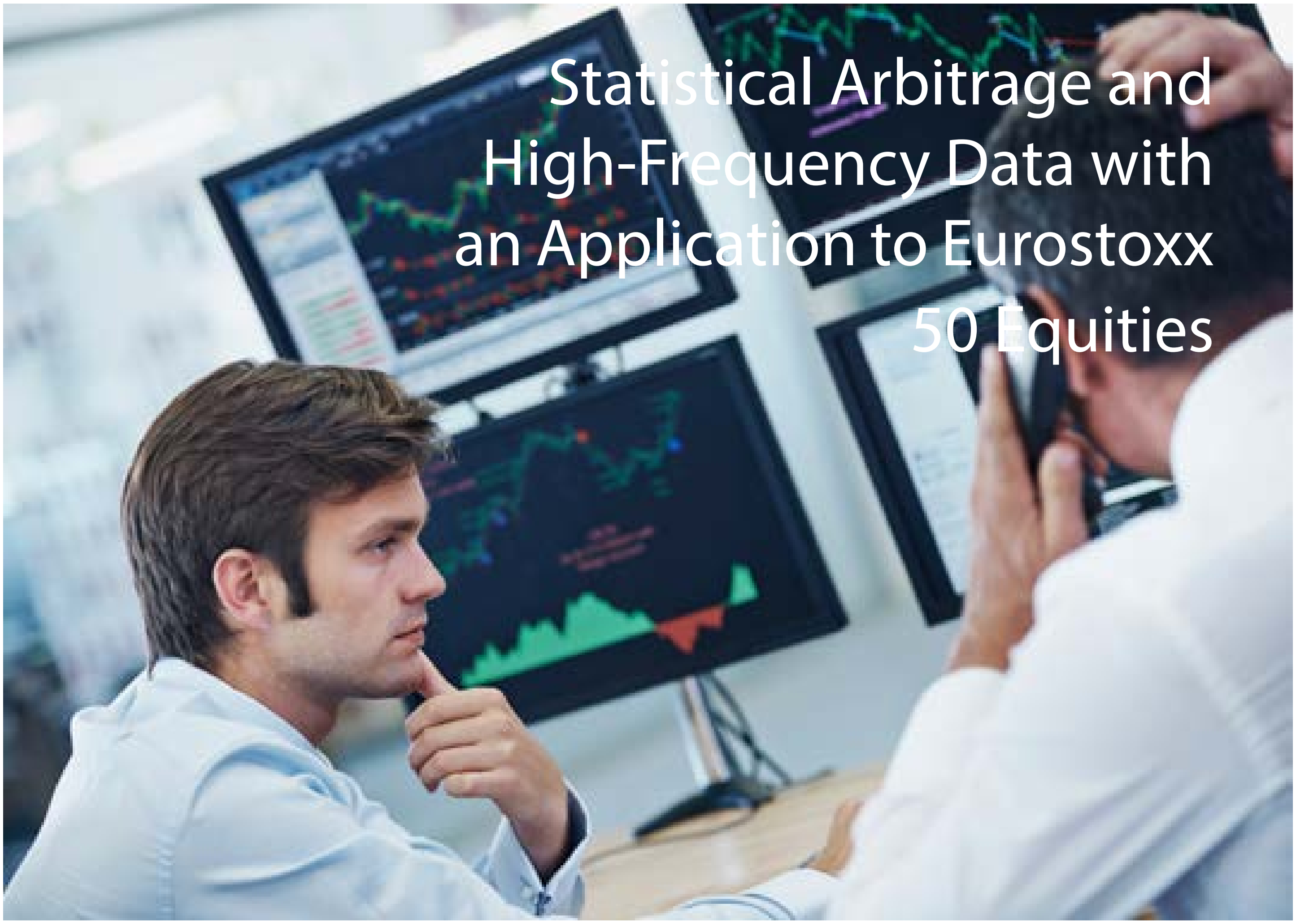
CAIA ASSOCIATION®

Alternative Investment Analyst Review

TRADING STRATEGIES

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

Christian Dunis, Gianluigi Giorgioni, Jason Laws, Jozef Rudy



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

Christian Dunis
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

Gianluigi Giorgioni
Senior Lecturer in Economics and Finance at Liverpool Business School and a member of CIBEF

Jason Laws
Reader of Finance at Liverpool Business School and a member of CIBEF

Jozef Rudy
Associate Researcher with CIBEF

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 six different frequencies, namely 5-minute, 10-minute, 20-minute, 30-minute, 60-minute, and daily sampling intervals. The high frequency data spans from July 3, 2009 to November 17, 2009; our daily data spans from January 3, 2000 to November 17, 2009.

We introduce a novel approach, which helps to enhance the performance of the basic trading strategy. The approach consists of selecting the pairs for trading based on the best in-sample information ratios and the highest in-sample t-stat of the Augmented Dickey-Fuller (ADF) test, which is applied to the residuals of the co-integrating regression using daily data. We form the portfolios of five best trading pairs and compare the performance with appropriate benchmarks.

Another improvement we introduce is the use of the high-frequency data. The advantage of incorporating the high-frequency data is higher, potentially achievable information ratios[1] compared to the use of daily closing prices and thus higher profit potential for investors (Aldridge 2009).

Market neutral strategies are generally known for attractive investment properties, such as low exposures to equity markets and relatively low volatility (Capocci 2006), but recently the profitability of these strategies has deteriorated (Gatev et al. 2006). While Gatev et al. (2006) study goes only go back to 2002, the Hedge Fund Research Equity Market Neutral Index that started in 2003, does not show the supposed qualities for which market neutral strategies are known (i.e., steady growth and low volatility). The industry practice for market neutral hedge funds is to use a daily sampling frequency and standard co-integration techniques to find matching pairs (Gatev et al. 2006). This approach is useful 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 we can compare the results of simulated trading using intraday data on various sampling frequencies with daily data.

2. Literature Review

The following section provides a brief review of the relevant literature.

2.1. Market neutral strategies

Pair trading, a well-known technique, was developed in 1980 by a team of scientists led by Wall Street quant Nunzio Tartaglia (Gatev et al. 2006). The strategy is widely documented in the existing 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 a pair of shares is formed where the investor is long one share and short the other share. The rationale is that there is a long-term equilibrium (spread) between the share prices and the shares will 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), a spread position is established. 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 the pair. The appropriate amount of each share is expressed by the variable beta, which tells the investor the number of the shares of X he has to short/go long for each one share of Y. There are various ways of calculating beta: it can be fixed, or it can be time-varying. To make beta time-varying, we will use rolling ordinary least squares (OLS) regression, double exponential smoothing-based (DESP) model, and the Kalman filter.

2.2. 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 (Nath 2003), which looks at market neutral strategies in the U.S. fixed-income market.

2.3. Co-integration

Co-integration is a quantitative technique introduced in a seminal paper by Engle and Granger (1987) that is based on finding long-term relationships between asset prices. Given its utility, co-integration might help identify potentially related pairs of assets. However, we will consider all of the possible pairs from the same industry, not only the co-integrated ones. We do this so that we will be able to measure whether the co-integrated pairs in the in-sample period perform better in the out-of-sample period than the non-co-integrated 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 co-integrating vectors that can be found in the system. The spread between the assets is not the one with the lowest variance, as with OLS, but rather the most stable one in the long-term (Alexander 2001). According to Alexander, the Engle and Granger (1987) methodology is preferred in financial applications due to its simplicity and lower variance, which are important points to consider from a risk management perspective. Since we only deal with pairs of shares in this paper, we also prefer the simpler Engle and Granger (1987) methodology.

There are many applications of co-integration in the world of investing, for example, index replication, which exploits long-term qualities of co-integration and requires only occasional portfolio rebalancing (e.g., Dunis and Ho (2005) and Alexander and Dimitriu 2002). There are also market-neutral arbitrage strategies based on co-integration, where one enters a trade when a relationship deviates from the long-term mean and exits the trade when the spread has returned to the long-term mean. While Burgess (2003), Lin et al. (2006), and Galenko et al. (2007) refer to their strategies as high-frequency trading, they use daily closing prices among four world indexes, rather than intraday continuous or intraday minute data.

2.4. Time Adaptive Models

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

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

2.5. Hedge Funds

The pair trading technique is used primarily by hedge funds and with regard to the type of investment strategy it falls under, there is a distinct hedge fund classification bearing the name equity market neutral funds (Khandani and Lo 2007 and Capocci 2006). Hedge funds employ dynamic trading strategies that are dramatically different from the ones employed by mutual funds and this enables them to offer investors more attractive investment opportunities (Fung and Hsieh 1997 and Liang 1999).

3. The Eurostoxx 50 Index and Related Financial Data

We use 50 stocks that formed the Eurostoxx 50 index as of November 17, 2009 (Appendix F). The data downloaded from Bloomberg includes six frequencies: 5-minute, 10-minute, 20-minute, 30-minute, 60-minute data (high-frequency data), and daily prices. We refer to all the data related in the minute dataset as high-frequency for brevity's sake.

Our database of minute data spans from July 3, 2009 to November 17, 2009.[2] Intraday stock prices are not adjusted automatically by Bloomberg for dividend payments and stock splits, so we adjust them accordingly.[3] Our database only includes the prices at which the shares were traded; although we do not consider bid and ask prices, some of our recorded trades were transacted at the bid and some at the ask. We have as many as 8,000 data points when data are sampled at 5-minute intervals for a 5-month period. 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 and for the 20-minute data, we have 2,000 data points.

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

In Exhibit 1, we provide the start and the end of the in- and out-of-sample periods for all of the frequencies. For high-frequency data, the in- and out-of-sample periods are the same length. 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.[6]

We used the Bloomberg sector classification with the industry sector ticker. We divide the shares in our

	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

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

database into 10 industrial sectors: (1) basic materials, (2) communications, (3) consumer cyclical, (4) consumer non-cyclical, (5) diversified, (6) energy, (7) financial, (8) industrial, (9) technology, and (10) utilities. Note that there is only one share in each of the diversified and technology categories (Appendix E) that prevents both these shares from forming pairs.

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 of the shares in the sample. In contrast, at times with high-frequency data and for a given pair, one share has a price related to a particular minute while no price is recorded for the other share, due to no transaction having taken place in that minute. In such an event, unmatched 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). This situation presents itself only rarely, as these 50 shares are the most liquid European shares listed.

4. Methodology

In this section, we describe in detail the techniques that we use in simulated trading. First, we describe the Engle and Granger (1987) co-integration approach, then, we describe the techniques we used in order to make the beta parameter adaptive: rolling OLS, the DESP model, and the Kalman filter.

Since the Kalman filter proves to be a superior technique for the beta calculation (as shown later), only the Kalman filter is used for the calculation of the spread to obtain the final results that are presented in this paper.

4.1. Co-integration model

First, we form the corresponding pairs of shares from the same industry. Once these are formed, we evaluate whether the pairs are co-integrated in the in-sample period. We investigate whether the fact that some pairs are co-integrated helps to improve the profitability of the pairs selected. Thus, in the first stage we consider pairs that are co-integrated and pairs that are not co-integrated.

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

$$Y_t = \beta X_t + \varepsilon_t \qquad (1)$$

In the second step, the residuals of the OLS regression are tested for stationarity using the ADF at 95% confidence level (Said and Dickey 1984).

4.2. 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.

$$Y_t = \beta_t X_t + \varepsilon_t \qquad (2)$$

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

We optimized the length of the OLS rolling window using genetic optimization.[7] For more details on 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 randomly[8] chosen pairs[9] 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 of maximizing 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 also optimized in case of the DESP model and the Kalman filter.

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

4.3. Double Exponential-smoothing Prediction Model

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 the β_t series, we smooth it using the DESP model. The 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 is a double exponentially smoothed series, and α is 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:

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

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

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

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

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

4.4 Time-varying Parameter Models with Kalman Filter

The Kalman filter allows parameters to vary over time and it is more optimal than a rolling OLS for adaptive parameter estimation (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 use a similar model to Dunis and Shannon (2005) and Burgess (1999). For the full specification of the Kalman filter model, see Appendix A.

We optimized the noise ratio (Appendix A). The resulting value for the noise ratio of 3.0e-7 was then used for all of 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. In order to keep the strategy simple, this was the only restriction placed on our strategy. This leaves us with pairs that are immune to industry-wide shocks.

5.1. Pair Trading: a Self-financing Strategy

A pair trading strategy requires one to be long one share and short the other share. Pair trading is a so-called self-financing strategy (e.g., Alexander and Dimitriu 2002), meaning that an investor can borrow the amount of cash that he wants to invest, say from a bank. Then, to be able to short a share, he deposits the borrowed amount of cash with the financial institution as collateral and obtains borrowed shares. Thus, the only cost he has to pay is the difference between the borrowing interest rates paid by the investor and the 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 versus received). We consider a more realistic scenario in which an investor does not have to borrow capital from a bank in the beginning (e.g., the case of a hedge fund that uses capital from investors). This 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; this interest was neglected for the ease of calculation. As our strategy proves robust and profitable, it does not affect our conclusions since it biases our results downward.

5.2. Spread calculation

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

$$z_t = P_{Y_t} - \beta_t P_{X_t} \quad (10)$$

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 the three methods described in the methodological section, namely the rolling OLS, the DESP model, and the Kalman filter. We did not include a constant in any of the models; therefore, we obtain a model with fewer parameters to be estimated.

5.3. 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 for both 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 within 0.5 standard deviations of its mean. We decided to wait for 1 period before we enter into the position to be conservative and ensure that the strategy is viable in practice. For instance, in the 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 will be the same.[10] As the spread is 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.[11] We do not assume rebalancing once we enter

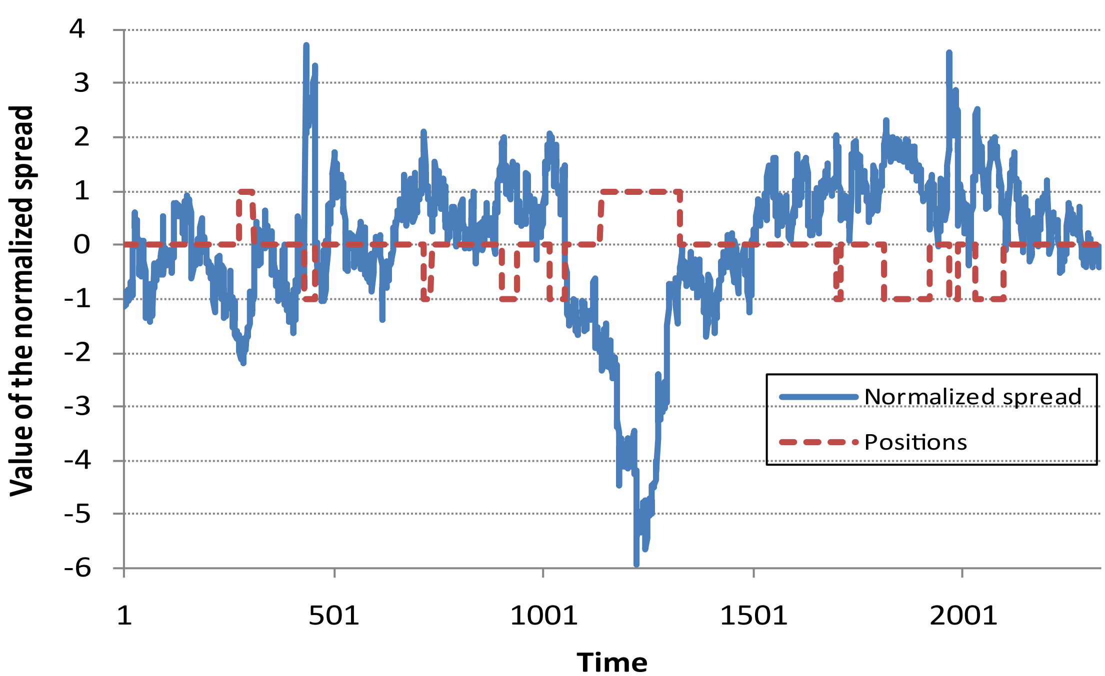


Exhibit 2 The normalized spread of the pair consisting of Bayer AG and Arcelor Mittal sampled a 20-minute interval

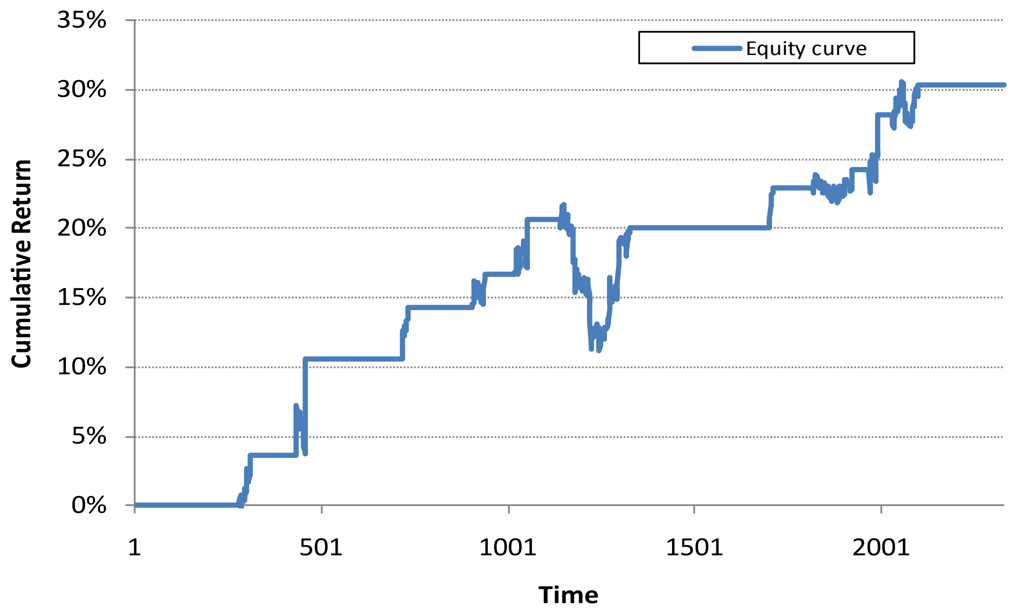


Exhibit 3 Cumulative equity curve in percent of the pair trading strategy applied to Bayer AG and Arcelor Mittal sampled at a 20-minute interval

into the position. Therefore, after an initial entry into the position, with equal amounts of euros on both sides of the trade, we do not rebalance the position, even when the positions cease being money-neutral due to price movements. Only two types of transactions are allowed by our methodology: entry into a new position and total liquidation of the position.

For an illustration, in Exhibit 2 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.

In Exhibit 3, we show the cumulative equity curve for the pair consisting of Bayer AG and Arcelor Mittal.[12] Note how the investment lost almost 10% around the mid-point of the sample period, as the position was entered into too soon and continued to move against the investor. Finally it reverted and recovered almost all of the capital lost.

In the next section, we explain the different indicators calculated in the in-sample period, in search of a link with the out-of-sample information ratio. As a result of our analysis, we offer a methodology for evaluating the suitability of a given pair for arbitrage trading.

5.4. 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.

5.4.1. Half-life of mean reversion

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

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

where is the median unbiased estimate of the strength of mean reversion from Equation 12 (Wu et al. 2000 and Dias and Rocha 1999). Intuitively, it is half of the average time the pair usually takes to revert back to its mean. Thus, traders should prefer pairs with a low half-life to those with a high half-life.

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

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

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

5.4.2. Information ratio

We decided to use the information ratio (IR), a widely applied measure among practitioners, as it gives an idea of the quality of the strategy.[13] 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 (Chan 2009). For our purpose, we calculated the information ratio as:

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

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 a perfect measure and Equation 13 overestimates the true information ratio, if returns are autocorrelated (e.g., Sharpe 1994 or Alexander 2008).

6. Out-Of-Sample Performance And Trading Costs

6.1. Return calculation and trading costs

The return in each period is calculated as:

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

where P_{X_t} is the price of the share we are long in period t; P_{Y_t} is the price of the share we are long in period t-1; is the price of the share we are short in period t; and $P_{Y_{t-1}}$ is 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 Alexander and Dimitriu (2002), for example. We are dealing with the 50 most liquid European shares in this paper. Transaction costs consist of 0.1%[14] 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% (0.1% for both shares).

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

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

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 shares that depreciates significantly, while we sell shares 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. The lower prices are the result of falling ask prices, which get closer to (or match) bid prices, so effectively one does not have to pay bid-ask spread and will transact at or close to the bid quote. The opposite is true when prices of shares are rising.

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

Exhibit 4 Out-of-sample information ratios for the simulated pair trading strategy at different frequencies. Transaction costs have not been considered

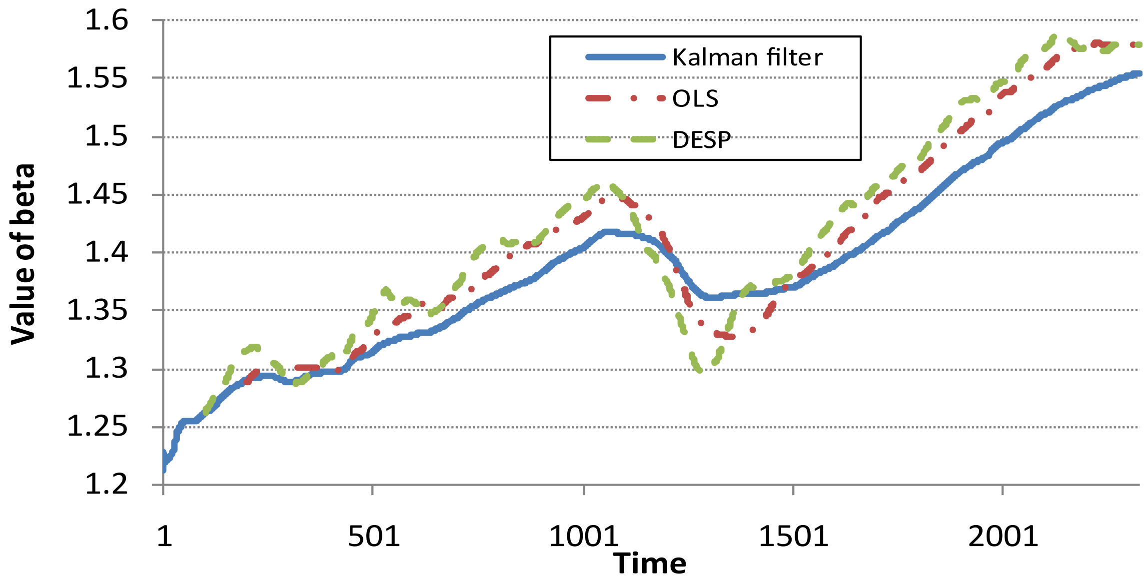


Exhibit 5 Various betas calculated for the Bayer AG and Arcelor Mittal pair sampled at a 20-minute interval

6.2. Preliminary out-of-sample results

In Exhibit 4, we present the out-of-sample information ratios, excluding transaction costs for the pair trading strategy at all of the frequencies. Results across all the three methods that we have used are displayed.

Results are superior for the Kalman filter method for most sampling frequencies. For this reason, 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 than is the case where beta is fixed.

From Exhibit 4 it is also clear that higher sampling frequencies offer more attractive investment characteristics than daily data for all of the methodologies.

In Exhibit 5, we present adaptive betas that have been calculated using the three approaches mentioned previously. Both the OLS and DESP betas seem to fluctuate around the Kalman filter beta.

In Exhibit 6, we show the distribution of the information ratios, including transaction costs for the 20-minute sampling frequency with the Kalman filter that was used for the beta calculation.

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

We also present the distribution of information ratios for daily data in order to investigate the difference between higher and lower sampling frequencies more closely (Exhibit 7).

Again, as shown in Exhibit 6, the majority of the information ratios are positive. Distributions of information ratios for other sampling frequencies can be seen in Appendices B-E.

The summary statistics for all trading frequencies are provided in Exhibit 8. The main difference between the daily data and high-frequency data is the maximum drawdown and maximum drawdown duration (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 it provides a good measure of the downside risk for the investor (Appendix G).

The maximum drawdown duration is expressed as the number of days from the start of the drawdown until the equity curve returns to the same percentage gain as before. Both of these measures are important from a psychological standpoint, because investors might start questioning the strategy when it is experiencing a drawdown.

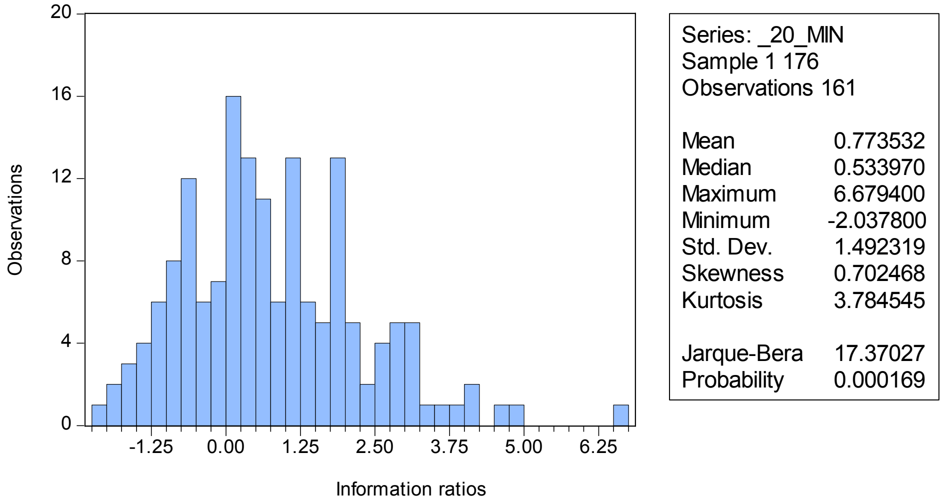


Exhibit 6 Distribution of information ratios for a 20-minute sampling frequency. One-way transaction costs of 0.4% have been considered

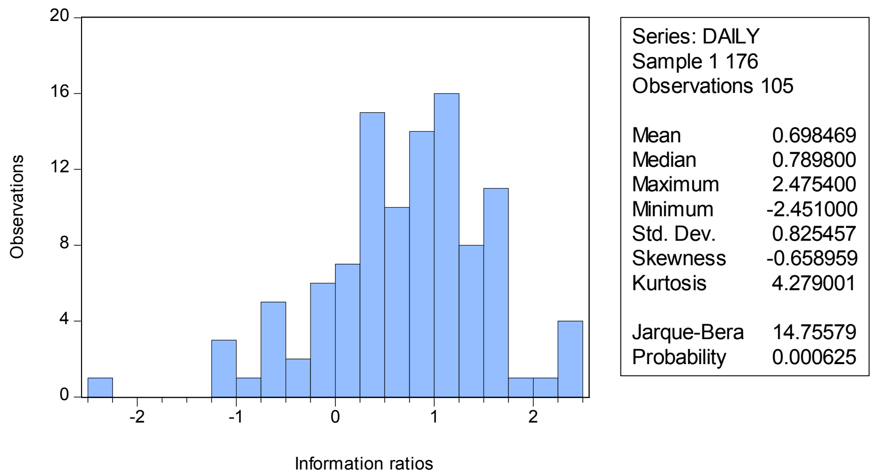


Exhibit 7 Information ratios of the pairs using daily sampling frequency. One-way transaction costs of 0.4% have been considered

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 was previously shown in Exhibit 4. When trading costs are considered, high-frequency data is more affected than daily data, due to the higher number of transactions. For instance, the information ratio for the 5-minute data drops from an attractive 1.21 to mere 0.26 when the trading costs are considered. The 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 excessively penalizes high-frequency data, so the information ratio achievable in real trading might be considerably higher.

6.3. Further investigations

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

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

We examine whether the in-sample co-integration of a given trading pair implies better out-of-sample performance. One can logically assume that a higher order of stationarity of the residual from the co-integration equation implies a higher level of confidence

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

Exhibit 8 The out-of-sample annualized trading statistics for pair trading strategy with the Kalman filter used for the beta calculation

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

Exhibit 9 95% confidence intervals of the correlation coefficients between t-stats generated in the in-sample period and the out-of-sample information ratios

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 on daily data only. We deal with intraday data later.
We bootstrap (with replacement) the pairs consisting of information ratios and t-stats.[15] The t-stat is obtained from the coefficient of the ADF test of the co-integrating 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 Exhibit 9.

The in-sample t-stat seems to have certain predictive power for the out-of-sample information ratio, although not for all of the frequencies. The only frequencies for which the t-stat works are data sampled at 5- and 10-minute intervals. For all the other frequencies, the center 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 centered around 0 (-0.18 and 0.22), implying that the true correlation coefficient might be 0.

6.3.2. 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 that could help us to 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 (Jollife 1986). PCA is a statistical technique that 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 behavior of the assets, (in our case, this would be the t-stats across different pairs and frequencies), just a few first principal components should suffice to explain the behavior 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 PCA. 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, the 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 is only around 0.14. The fact that the variances between different frequencies are small when considering each of the 176 pairs, but the variances between the pairs are high further demonstrates that t-stats tend to be similar across all of the frequencies for any given pair.

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

6.3.3. Does co-integration in daily data imply higher frequency co-integration?

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 of the high-frequencies. In this section, we investigate the relationship between the t-stats for daily data (computed from January 1, 2009 to September 9, 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., September 10, 2009 to November 17, 2009).

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

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

6.3.4. Information Ratio: In-Sample and Out of Sample

Do the 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 that we previously described in order to estimate the confidence intervals; the bootstrapping is performed 5,000 times,with replacement, as in other cases.

In Exhibit 10, we show the bootstrapped correlation coefficients between the in- and out-of-sample information ratios (not taking into account transaction costs) across all frequencies.

The confidence bounds indicate that the in-sample information ratio can predict the out-of-sample information ratio to a certain extent. Whereas in Exhibit 9 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 center 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.

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

Exhibit 10 95% confidence intervals of the correlation coefficients between information ratios generated in the in- and out-of-sample periods

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

Exhibit 11 95% confidence intervals of the correlation coefficients between the in-sample half-life of mean reversion and 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)	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

Exhibit 12 The out-of-sample information ratios for 5 selected pairs based on the best in-sample information ratios

---

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 Exhibit 11. 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 Exhibit 10. The average lower/upper interval across all the frequencies presented is -0.20/0.06. So we find that there is a negative relationship between the half-life of mean reversion and subsequent out-of-sample information ratio.

Thus, the two 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 Exhibit 8, but here we try to improve them using the findings from the previous section. We use the indicators mentioned above to select the five best pairs for trading and present the results.

First, we present the results of using each indicator individually. The results of selecting five pairs based on the best in-sample information ratios are shown in Exhibit 12.

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 Exhibit 8 to 3.24. The information ratio for daily data improves as well (from 0.7 to 1.32). Most of the information ratios for the pairs sampled at the high-frequency intervals are above 2.

The maximum drawdown and maximum drawdown duration favour the pairs that are 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 Exhibit 13, we show trading results based on using the half-life of mean reversion as an indicator. Thus, five pairs with the lowest half-life of mean reversion were selected to form the portfolio.

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 Exhibit 8. Thus, we decide not to consider the half-life of mean reversion as a prospective indicator of the future profitability of the pair.

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

Exhibit 13 The out-of-sample trading statistics for 5 pairs selected based on the best in-sample half-life of mean reversion

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

Exhibit 14 The out-of-sample trading statistics for 5 pairs selected based on the best in-sample t-stats of the ADF test

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

Exhibit 15 The out-of-sample trading statistics for 5 pairs selected, based on the best in-sample t-stats of the ADF test for daily data

In Exhibit 14, we show the results of using the in-sample t-stats of the ADF test of the co-integrating regression as the indicator of the out-of-sample information ratios.

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 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 Exhibit 15, we present the results of using the t-stat of the ADF test for daily data (from January 1, 2009 to September 9, 2009) as an indicator of the out-of-sample information ratio of the pairs sampled at the high-frequency intervals. The average information ratio for all of the high-frequency trading pairs is around 3, which makes it the second best indicator after the in-sample information ratio.

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

$$Combined\_ranking = \frac{R_1 + R_2}{2} \qquad (16)$$

where R_1 and R_2 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 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 five trading pairs.

The trading results of the combined ratio are presented in the Exhibit 16.

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

Exhibit 16 The out-of-sample trading statistics for 5 best pairs selected, based on combined ratio calculated according to Equation 16

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

Exhibit 17 Out-of-Sample Results

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

Exhibit 18 In-Sample Annualized Trading Statistics

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 an information ratio of 0.35 after transaction costs, which is worse than the original, unoptimized case.

We also combine the t-stat of the ADF test for a given high-frequency and information ratio 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 Exhibit 8 (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 of the high-frequency intervals. This, in our opinion, disqualifies the usage of this indicator for predicting the future profitability of the pairs.

Futher note that Exhibit 16 displays 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, we were able to improve the information ratios net of trading costs for daily data from around 0.7 as in Exhibit 8, to 1.3 as in Exhibit 12, using the in-sample information ratio as an indicator of the future profitability of the pairs.

Furthermore, we found that three 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 Exhibit 8 (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 and the t-stat of the ADF test of the series sampled at a daily interval, and a combination of the two (Exhibit 12, 15, and 16).

Below, we compare the results of the pair trading strategy at both frequencies (an average of all of 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 of the frequencies for reasons of presentation. In fact, pairs sampled at all

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

Exhibit 19 Out-of-Sample Annualized Trading Statistics

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 Exhibit 17, we present a comparison of our pair trading strategy sampled at a daily interval with the results of a buy and hold strategy of the the Eurostoxx 50 index and the HFR Equity Market Neutral Index. The results span January 1, 2009 to November 17, 2009, with the out-of-sample period for our pairs sampled at a daily interval.

The Exhibit displays the 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 HFR Equity Market Neutral Index, both on an absolute and risk-adjusted basis. While the Equity Market Neutral Index lost money during the period, our strategy was profitable without showing excessive volatility relative to its return. It also outperformed the corresponding market index, the Eurostoxx 50 index.

In Exhibit 19, we compare the results of the average high-frequency pair trading strategy with the appropriate benchmarks in the period from September 10, 2009 to November 17, 2009. The information ratio of 3.24 of the pair trading strategy is considerably higher than that of the two indices. Thus, using high-frequency sampling 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 a significantly lower volatility.

Exhibit 19 displays, 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 six 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 two 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 less 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 five 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 HFR Equity Market Neutral Index, with information ratios lower than 1 during the review period. It also illustrates how useful the combination of 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:

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

Exhibit A

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 and the lower the ratio, the less adaptive beta. Thus, if we used an extremely low value for the 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.

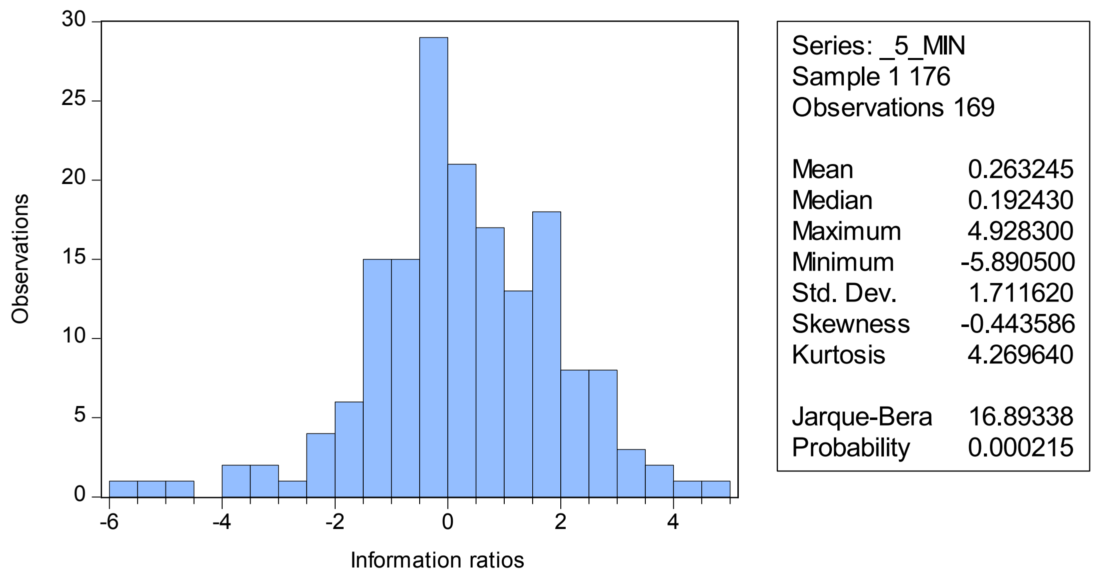


Exhibit B Distribution of information ratios for a 5-minute sampling frequency

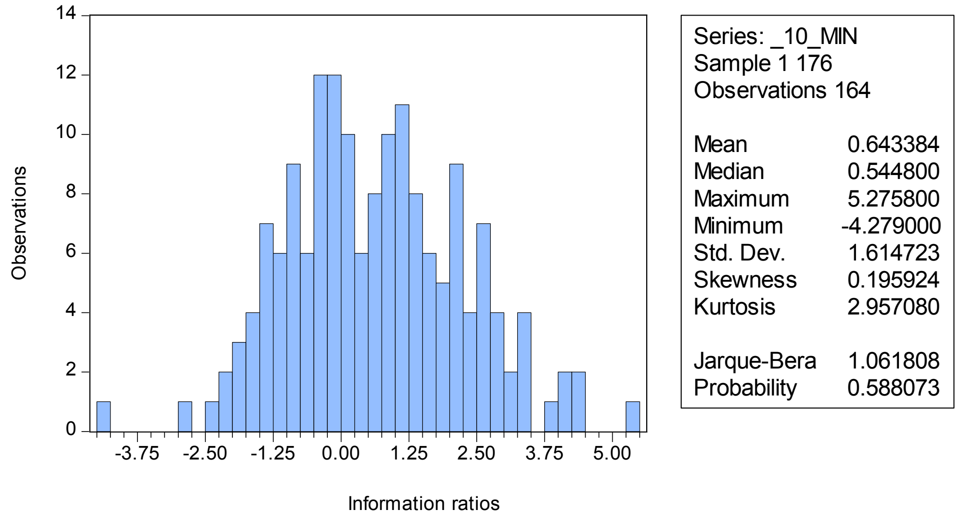


Exhibit C Distribution of information ratios for a 10-minute sampling frequency

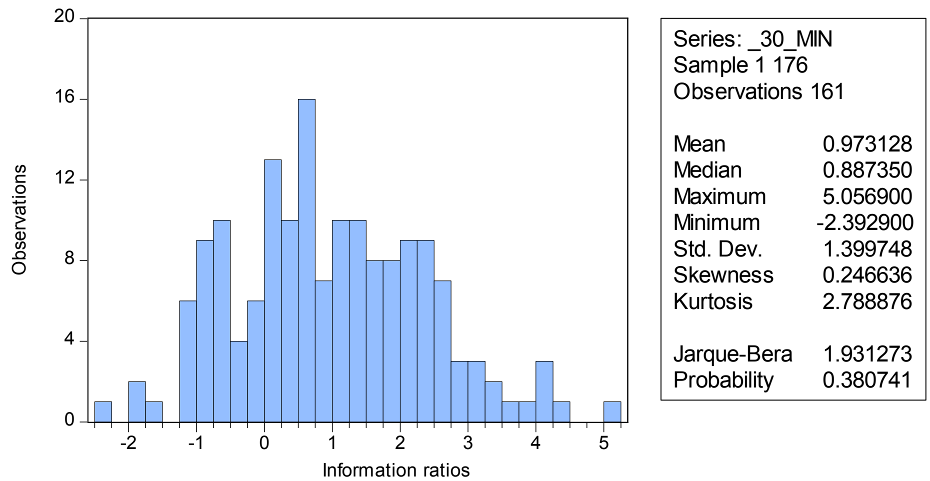


Exhibit D Distribution of information ratios for a 30-minute sampling frequency

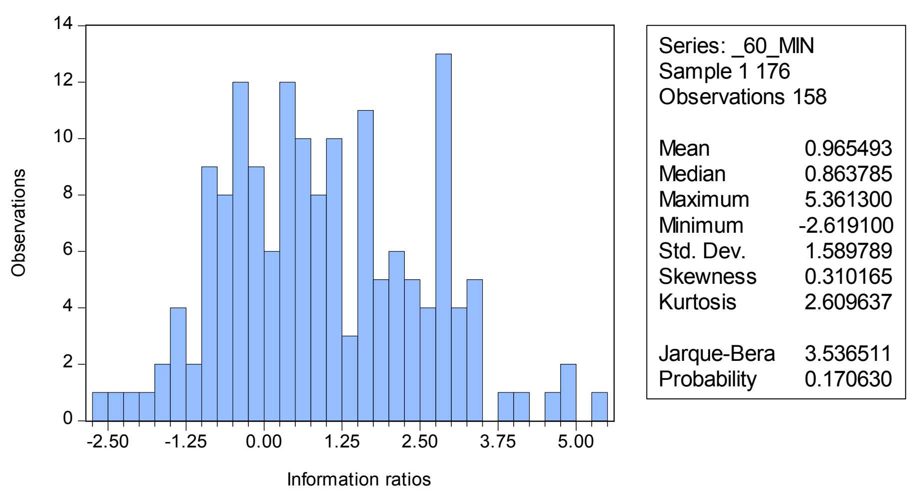


Exhibit E Distribution of information ratios for a 60-minute sampling frequency

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

Exhibit F Constituent stocks of Eurostoxx 50 index which were used to form the pairs Source: Bloomberg

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* $\sum(R_t)$ *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.*

Exhibit G. Calculation of the trading statistics

Endnotes

Sources for all figures are based on author's calculations, unless otherwise noted.

1 The information ratio is calculated as the ratio of annualized return to annualized standard deviation.

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 9:34:58 and the subsequent one at 9:35:01, the former transaction would be recorded as of 9:35. We download the data from Bloomberg, which only stores the last 100 business days of intraday data. We downloaded the data on November 17, 2009 and that is why our intraday data span from July 3, 2009.

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 Some shares do not date back as far as January 3, 2000, and as a consequence the pairs that they formed contain lower amount of data points In particular, four shares do not date back to January 3, 2000 (Anheuser-Busch starts on November 30, 2000, Credit Agricole S.A. starts on December 1, 2001, Deutsche Boerse AG starts on February 5, 2001 and GDF Suez starts on July 7, 2005).

6 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 September 10, 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).

7 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.
8 We only optimized the parameters for 6 pairs, due to the length of the genetic optimization process.

9 The 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.

10 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 1,000 euros worth of share A and short 1,000 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 of X and sell 10 shares of 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 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 2,857 shares X and 1,000 shares Y.

11 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.

12 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.

13 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.

14 For instance Interactive Brokers charges 0.1% per transaction on XETRA market (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 February 14, 2010)

15 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 (e.g., Efron, B. and Tibshirani, R. J. (1993)). 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 5,000 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.

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.

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.

Author Bios

Christian Dunis is Professor of Banking and Finance at Liverpool Business School andDirector 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.

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