CHARLES UNIVERSITY IN PRAGUE

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Alice Nušlová

Pairs Trading at the Prague Stock Exchange

Bachelor thesis

Prague 2014

Author: Alice Nušlová
Supervisor: PhDr. Ladislav Krištoufek Ph.D.

Year of defense: 2014

Declaration of Authorship

I hereby proclaim that I wrote my bachelor thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used.

I grant a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, June 30, 2014

Alice Nušlová

Acknowledgments

I would like to express my gratitude to my supervisor PhDr. Ladislav Krištoufek Ph.D. for helpful advice and for allowing me to investigate a topic that I have selected independently. I am truly indebted to Mgr. et Mgr. Vojtěch Patočka whom I consulted with all programming obstacles. His invaluable assistance has given me encouragement and motivation. Next I wish to thank my parents for being patient with my studies and for supporting me throughout the academic years.
Finally, this thesis would not exist without the help of director of Kurzy.cz Mr. Antonín Foller who provided me with data on the Prague Stock Exchange.

Bibliographic entry: NUŠLOVÁ, Alice. *Pairs Trading at the Prague Stock Exchange*. Prague, 2014. Bachelor thesis, Charles University, Faculty of Social Sciences, Institute of Economic Studies. Supervisor: PhDr. Ladislav Krištoufek Ph.D.

Title: Pairs Trading at the Prague Stock Exchange
Author: Alice Nušlová
Department: Institute of Economic Studies
Supervisor: PhDr. Ladislav Krištoufek Ph.D.
Supervisor's e-mail address: kristoufek@ies-prague.org

Abstract: Since its birth in the 1980s, pairs trading has become a widely used strategy for making profits among hedge funds and institutional investors. This technique identifies pairs of securities whose historical prices show long-run relationship, and takes advantage of their short-term relative mispricing. Profit is generated due to correcting behavior of security prices as they converge towards equilibrium value of their spread. The aim of this thesis is to compare two traditional approaches to pairs trading: cointegration and sum of squared deviations between normalized historical returns, known as distance criterion, within the Prague Stock Exchange equity market. We further investigate whether the two methods, so commonly employed in the US equity market, can be applied with similar success in the PSE. Our results reveal that the strategy using distance criterion outperforms the method of cointegration in nearly every aspect considered. Nevertheless, its returns are not statistically different from zero, and in other measures the return distribution lags behind the one found in the US equity market.We conclude that with the form of trading we present here the PSE is not a suitable stock market for pairs trading technique.

JEL classification: G10, G11
Keywords: cointegration, pairs trading, pairs selection, mean reversion, statistical arbitrage, Prague Stock Exchange, sum of squared deviations, long/short position
Length of the thesis: 62,964 characters

Název práce: Pairs Trading at the Prague Stock Exchange
Autor: Alice Nušlová
Institut: Institut ekonomických studií
Vedoucí bakalářské práce: PhDr. Ladislav Krištoufek Ph.D.
E-mail vedoucího: kristoufek@ies-prague.org

Abstrakt: Od svého zrodu v 80. letech 20. století se párové obchodování stalo široce používanou investiční strategií mezi hedgovými fondy a institucionálními investory. Tato technika určí akcie, jejichž historické ceny mezi sebou vykazují dlouhodobý vztah, a následně využívá jejich krátkodobého relativního cenového vychýlení. Zisk je generován díky opravnému chování cen akcií, které konvergují k ekvilibriu tohoto vztahu. Cílem této bakalářské práce je porovnat dvě tradiční metody výběru a obchodování párů: kointegraci a součet čtverců odchylek mezi normalizovanými historickými výnosy, známý jako metoda vzdáleností, v kontextu Burzy cenných papírů Praha. Testujeme dále, zda tyto metody, zcela běžně používané na americkém akciovém trhu, mohou být podobně úspěšně aplikovány na pražské burze. Výsledky odhalují, že strategie využívající metodu vzdáleností překonává kointegrační přístup téměř ve všech srovnávacích statistikách. Přesto však její výnosy nejsou statisticky odlišné od nuly, a distribuce těchto výnosů se nevyrovná výsledkům na burze v USA zveřejněným v dřívějších analýzách. Docházíme k závěru, že při námi prezentované formě obchodování není pražská burza vhodným akciovým trhem pro techniku párového obchodování.

Klasifikace JEL: G10, G11
Klíčová slova: kointegrace, párové obchodování, výběr párů, návrat ke střední hodnotě, statistická arbitráž, Burza cenných papírů Praha, součet čtverců odchylek, dlouhá a krátká pozice
Délka práce: 62,964 znaků

Bachelor Thesis Proposal

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Author:	**Alice Nušlová**	Supervisor:	PhDr. Ladislav Krištoufek Ph.D.
E-mail:	ali.nuslova@gmail.com	E-mail:	kristoufek@ies-prague.org
Phone:	+420776593418	Phone:	line 312 (IES)
Specialization	Economic Theories	Defense Planned:	September 2014

Proposed Topic:

Cointegration in Pairs Trading

Topic Characteristics:

The objective of this thesis is to propose strategies for pairs trading using cointegration approach. The idea of pairs trading is based on fluctuations around the long-run equilibrium of the two stocks that form a pair. The quantity representing the difference in normalized prices between the two stocks is called the spread. When the value of the spread substantially deviates from its mean value, a long-short position is taken with the assumption that the spread will revert back to its equilibrium. If this is the case, the position is unwound and, consequently, profit is made. In the thesis we will focus on implementing trading strategies for various trading horizons, ranging from intraday to weekly and monthly holding periods. We will compare our results and discuss profitability of each strategy.

As a first step of the process, after providing theoretical background supporting the whole thesis, we will need to identify suitable stocks for trading pairs. There are two possibilities how to form pairs. One of them is stock fundamentals analysis, which involves looking at company's data, e.g. revenue, debt-to-equity ratio, etc. The second approach focuses on technical analysis and takes into account historical prices of stocks. The latter approach is the one we will follow along in our thesis. We will form suitable pairs by ordering the stocks with minimum-distance method of their normalized historical prices.

After identification we need to test whether the stocks in a pair are cointegrated. We will use MATLAB computing software to attest the right choice of pairs using cointegration framework (Engle and Granger (1987), Johansen (1988)). Once this is achieved, we will set the trade signals for each pair according to historical comovement of both stocks.

As a last step, the two strategies will be backtested and their functionality will then be verified on out-of-sample data. The purpose of this work, however, is not to practically prove the profitability of the strategies; it should serve as an inspiration on a possible method of trading pairs in various trading horizons, which, no matter how promising results it may yield, does not take into account the real-world trading obstacles.

Outline:

1. Introduction
2. Literature review and theoretical background
   a. Minimum-distance method for matching stocks into pairs, justification of its suitability
   b. Johansen cointegration analysis
   c. Engle-Granger cointegration test
3. Formation of pairs from stocks traded on the United States' stock exchange market
   a. Execution of pairs using minimum-distance method
   b. Verification of cointegration of pairs in MATLAB using cointegration framework (Engle and Granger (1987), Johansen (1988))
4. Empirical model
   a. Determining entering and unwinding position for each pair, separately for various trading horizons
   b. Backtesting the strategies and verifying their functionality
5. Conclusion
   a. Relevance of the strategies
   b. Comparison of different trading horizons
   c. Assessing the possible profitability of both methods

Core Bibliography:

Engle, R., & Granger, C. (1987). *Co-Integration and Error Correction: Representation, Estimation and Testing* (Sv. Vol. 55, No. 2). Econometrica.

Gatev, E., Goetzmann, W., & Rouwenhorst, K. (2006). *Pairs Trading: Performanceof a Relative-Value Arbitrage Rule.* Oxford University Press.

Johansen, S. (1988). *Statistical Analysis of Cointegration Vectors.* Copenhagen: Journal of Economic Dynamics and Control 12.

Tsay, R. (2005). *Analysis of Financial Time Series.* John Wiley & Sons, Inc.

Vidyamurthy, G. (2004). *Pairs Trading: Quantitative Methods and Analysis.* John Wiley and Sons, Inc.

Author

Supervisor

Contents

List of Tables

List of Figures

Acronyms

PSE Prague Stock Exchange

SSD sum of squared deviations

ADF augmented Dickey-Fuller

AEG augmented Engle-Granger

OLS ordinary least squares

AIC Akaike information criterion

BIC Bayesian information criterion

Introduction

"... Human beings don't like to trade against human nature, which wants to buy stocks after they go up not down."

- Nunzio Tartaglia

Pairs trading is an investment strategy that was pioneered at the New York Stock Exchange by a group of people around quantitative analyst Nunzio Tartaglia (Bookstaber (2007)). Up until 1980s, when this kind of statistical arbitrage was born, financial world was aware of returns to simultaneous buy-low and sell-high strategy. Trading in pairs, however, was a major breakthrough, for it found a way how to tie up shares together based on their long-run equilibrium value. The equilibrium value is termed the spread, and it captures the degree of mutual mispricing of one security relative to the other. Because the securities move largely together, pairs trading expects mean-reverting behavior of the spread and avails of temporary divergence from the equilibrium value. It sells the relatively overvalued security and buys the relatively undervalued one with an expectation that the mispricing will correct itself in the future. The greater the mispricing, the higher the potential return.

Three traditional methods of pairs trading are recognized: cointegration, distance and stochastic spread method. All of them have received much attention among hedge funds and institutional investors, not so in academic literature. To this time, the distance-approach study by Gatev et al. (1999) has been the only renown empirical analysis. Few more similar works exist, but evidence of performance of the strategy outside the US is scarce.

This thesis contributes to the field by analyzing methods of distance and cointegration at the Prague Stock Exchange (PSE). The distance method largely follows Gatev et al. (2006) and it is based on the sum of squared deviations between normalized historical returns. Cointegration method has its grounding in the article by Engle and Granger (1987). We compare which approach yields better results, and evaluate whether pairs trading in the PSE would be a good investment strategy in terms of risk and reward. Comparison is made with Gatev et al. (2006) in order to see how much number of shares listed, size of the equity market and other characteristics affect return distribution.

This bachelor thesis is organized as follows. Chapter 1 provides review of literature that laid cornerstones to the theory of pairs trading. It touches the topic of mean reversion, zooms in on the evolution of findings in regard to return reversals and states important contradicting explanations for short-term contrarian profits that were in the center of research around 1990. It then describes present-day trends that are driven by diminishing returns of the strategy. Chapter 2 focuses on data issues and explains distribution of sample periods. Chapter 3 summarizes theory pertaining to the cointegration method, and explains normalization of returns. Chapter 4 is devoted to methodology; it presents two approaches to pairs formation and trading that we employ in empirical research: cointegration and distance criterion. Computation of returns and measurement of investment performance follow. Chapter 5 states scenarios that could be the outcome of empirical analysis. Results of the analysis are shown in Chapter 6, along with their possible explanation, and differences in outcomes between the two methods of pairs formation are contrasted. A paragraph is devoted to the comparison of distance method with evidence from the US market. In the last part of the thesis we present conclusion.

1 Literature Review

Pairs trading investment strategy was first documented by Gatev et al. (1999), yet the concepts which it builds upon, namely reversion of stock prices and simultaneous long and short position, had been in center of focus among traders and financial mathematicians long before that.

Mean-reverting behavior of the stock market was investigated and evidenced by Poterba and Summers (1988) in a study on the NYSE and 17 foreign equity markets. The authors used variance ratio tests to conclude that over long horizons monthly stock returns exhibit negative serial correlation and as such have significant predictable components. Fama and French (1988) updated on the finding by subperiod testing which suggested that the mean reversion in 1926-1985 was largely caused by the 1926-1940 period. De Bondt and Thaler (1985) in their market-behavioral research focused on the links between mean reversion and overreaction hypothesis. The hypothesis asserted that individuals tend to overweight consistent pattern of news pointing in the same direction, resulting in systematical overshooting of stock prices. Such irrational behavior, as the empirical test confirmed, is corrected by subsequent price movement in inverse direction to its fair market value. Buying underpriced losers and selling overpriced winners yielded cumulative abnormal returns even five years into the investment period.

Later on, Jegadeesh (1990) and Lehmann (1990) conclude abnormal returns even for short-term contrarian strategies and explain them as evidence of inappropriate reaction to new information in accordance with the paper of De Bondt and Thaler (1985). These early documentations of short-term return predictability provide academical evidence for the potential of statistical arbitrage to generate significant profits, proven already by the Morgan Stanley hedge fund in the preceding years (Bookstaber (2007)). Jegadeesh (1990) buys and sells stocks on the basis of their prior-month returns and holds them for one month with a difference in abnormal returns on the extreme decile portfolios of 2.49 percent per month. Lehmann (1990) uses a shorter, weekly scale, however even such time frame for contrarian strategy results in positive profits in 90 percent of the weeks. Since changes in fundamental valuation of firms over such short intervals are improbable, the paper disfavours the association of expected security returns with the security's fundamentals.

Forces driving the price swings and contrarian profits had become a source of disagreement and contradictory theories in subsequent papers. Lo and MacKinlay (1990) disputed the overreaction explanation and attributed majority of the expected profits from contrarian investing to positive cross-correlation between securities.[1] The authors made a case by construction of a return-generating process in which returns of each security were serially independent. Still, positive expected profits from buying losers and selling winners persisted. Five years later, Jegadeesh and Titman (1995b) refuted findings of Lo and MacKinlay (1990). According to them, the analysis by Lo and MacKinlay (1990) did not relate systematic stock price over- or underreactions to contrarian profits because delayed reactions to common factors that imply lead-lag structure affected both covariances as well as cross-covariances, i.e. the two components of equation by Lo and MacKinlay (1990). They designed another equation for decomposition of contrarian profits that included a more detailed set of stock price reaction scenarios, with under and overreaction to common factors and idiosyncratic news. With their alternative decomposition, Jegadeesh and Titman (1995b) showed that delayed reactions could not be exploited by contrarian trading scheme, and supported conclusions of Lehmann (1990) and Jegadeesh (1990). Moreover, they contributed to the literature when they claimed reversal of a firm-specific component of returns as the primary source of contrarian profits.

[1]This relationship arises from asymmetrical sensitivity of stocks to new information when high return for one stock today suggests high probability of rising return for another stock the following day, resulting in a lead-lag structure.

All the cited works shared one identical conclusion: they proved predictability of stock returns, and hence gradually rendered obsolete weak-form efficient markets hypothesis advocated by Fama (1970).

It was not until few years later when Gatev et al. (1999) published the first empirical test of pairs trading strategy. Pairs trading evolved from simple contrarian principles, but while in contrarian investing stocks were not related, pairs trading used the idea of relative pricing between them. Gatev et al. (1999) employed sum of squared deviations (SSD) between normalized historical prices to form pairs and attained profits robust to modest transaction costs. In the paper he concentrated on mechanism and practical issues of the strategy, nevertheless the article also continued in the footsteps of its predecessors in that it investigated the importance of mean reversion for generating pairs trading profits. The idea behind the bootstrapping test conducted for that purpose was following: if mean reversion were the only driving force of profits, it would suffice to randomly match stocks to become profitable. The outcome shed further light on the technique because such contrarian strategy did not make money. The paper found alternative explanation which stated that profits of the strategy arise from temporary mispricing of close substitutes[2], and they are influenced by common factor exposures of stock prices.

Subsequent version of the paper (Gatev et al. (2006)) extended the testing period by five years and continued to object to certain sources of pairs trading profits claimed to be the driving force of profitability in earlier literature, in particular to unrealized bankruptcy risk and the inability of arbitrageurs to take advantage of the profits due to short-sale constraints. It further discovered diminishing returns, and, by implementing a one-day waiting rule, it disproved that contrarian trading profits cannot be realized by traders transacting at bid and ask

[2]From the notion of relative pricing, close substitutes should sell for the same price; if they do not, the Law of One Price (see for example Chen and Knez (1995)) is broken, suggesting market inefficiency and possibility to avail of relative-value arbitrage, i.e. pairs trading.

prices, a theory advocated by Jegadeesh and Titman (1995a).

In the meantime, other authors presented their advance in the field, albeit with much less attention. Vidyamurthy (2004) provided detailed discussion of cointegration approach based on Engle and Granger (1987), and Elliott et al. (2005) proposed stochastic spread method. These two methods, together with the SSD by Gatev et al. (1999), now represent main approaches that have been widely adopted by practitioners.

However, neither of the papers carried empirical results, which is not unusual when we research literature. Articles presenting practical implementation of the strategies are generally scarce. This may be explained by lack of interest among academia or likely by proprietary nature of the findings. Since Gatev et al. (2006), only one academical work, written by Broussard and Vaihekoski (2012), replicated and tested the SSD strategy on a market outside the US. Works which would empirically explore the two other leading methods are practically non-existent, or limited to university theses and working papers that have not been published by recognized journals. Broussard and Vaihekoski (2012) investigated pairs trading in Finland and found excess returns slightly higher than those in Gatev et al. (2006), which was probably caused either by thinner trading and wider bid-ask spreads in comparison with the US market, or due to different consideration of pairs: in Broussard and Vaihekoski (2012), majority of pairs are represented by multiple share classes of the same stock while Gatev et al. (2006) forms pairs by matching stocks issued by different companies.

Extant papers on pairs trading have been consistently bringing forth evidence of declining profits for traditional strategies (Gatev et al. (2006), Do and Faff (2010), Do and Faff (2012)). As a consequence, current trends attempt to explore new analytical techniques of pairs formation and trading that would halt this course, for example by more reliable identification of close substitutes or by giving correct trading signals.

The former has been researched by Huck (2009) and Do and Faff (2010), trading triggers have been investigated by Liew and Wu (2013).

Huck (2009) designs another metrics for pairs selection through forecasting and multi-criteria decision methods and reports non-zero excess returns at 1% significance level for the S&P100 index. However promising, these results cannot be assessed or compared with the distance method as Huck (2009) does not account for transaction costs and uses weekly data. Do and Faff (2010) add two additional metrics to the conventional SSD, resulting in enhanced profits: frequency of zero crossings of paired stocks and industry homogeneity. The authors make us believe in the conclusion that the industry grouping is an innovative approach but its pioneeer is in fact Gatev et al. (1999). Yet, Do and Faff (2010) contribute to our knowledge of the strategy when they test it during the recent financial crisis, through which they show its substantial performance in turbulent periods. Also, the paper disproves long believed fallacy that profits are carried away by increased competition of hedge funds; rather, majority of the decline can be attributed to worsening arbitrage risks[3].

Liew and Wu (2013) criticize conventional approaches to pairs trading that we are going to use in this thesis and instead propose to apply copulas[4] between two stock returns. The association between financial assets, they claim, is rather complex to be captured by linear association such as cointegration error term or linear correlation coefficient. The paper ascertains that the copula approach results in more trading opportunities and, since it is independent of correlation and cointegration, its authors regard it to be a new alternative method. Nonetheless, its empirical grounding requires further research because the single pair it uses is by no means an indicator of success at a stock exchange as a whole.

[3]To be mentioned in Chapter 2

[4]Copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform.

The evolution of pairs trading lies in exploring new methods rather than testing the conventional ones. With markets becoming more efficient, hedge funds will seek to gain competitive advantage and to avail of arbitrage opportunities with greater or at least constant profits. Thus, pairs trading is likely to become a black box even more so than it is today, with practical outcomes and know-how hidden from public awareness.

2 Data

We use daily data for the Prague Stock Exchange over June 2008 - March 2014. Number of stocks varies between 12 and 14, and the number of pairs that could potentially be tradeable is computed as

$$P_N = \binom{N}{2} = \frac{N!}{2!(N-2)!}$$

where N is the total number of stocks eligible for trading. The highest number of pairs is 91. Our sample comprises nine periods where each is divided into two stages: period of formation and the trading period. Gatev et al. (2006) uses a 12-month time frame of daily data for identifying suitable pairs and 6 months of data for simulated trading, and rolls the time period forward by one month. In this thesis we keep the trading period at 6 months but extend the formation period to 15 months, as the method of cointegration requires a longer time scale to reliably identify matching stocks. Unlike Gatev et al. (2006), we roll periods forward by 6 months, making the trading periods follow up. The formation and trading horizons remain our time frames throughout the analysis. For sample periods and stocks included, refer to Table A.1. The number of stocks and pairs considered for pairs formation in each period is in Table A.2.

Industry-group structure of the PSE is greatly diversified, with industries comprising Integrated Telecommunications Services, Banks, Multiline Insurance, Broadcasting, Tobacco, Casinos and Gaming, Textiles and Leather Goods, Electric Utilities, Coal, Real Estate Development, Distillers and Wineries, Leisure and Recreation, Oil and Gas [5]. The PSE ranks among small equity markets, therefore constructing pairs with restriction to the same industry is not feasible. Throughout all sample periods only two stocks are from identical sector: Erste and KB. This may represent a potential pitfall because pairing up of stocks across

[5]These are the industry groups as classified by the Thomson Reuters Classification.

sectors induces a threat of industry-wide shocks that could bring about losses in the trading period. Gatev et al. (2006) shows that average excess return for unrestricted pairs is lower than for homogenous-industry strategy. We comment on this issue in Chapter 6.

We impose two restrictions on stocks to include them in the analysis:

1. Each stock is traded on every business day while listed in order to avoid illiquidity.
2. Each stock is listed throughout at least one whole formation and investment period, i.e. a 21-month frame.

Evaluation of these criteria requires forward-looking information and thus introduces a look-ahead bias, treatment of which is beyond the scope of this thesis. Criterion 1 is met by all stocks except for the ECM, which was not traded since 21 June 2011 until its delisting in 2013, although during that time it was still listed on the PSE. We keep the ECM until 21 June 2011 but screen it out from our data samples after this date. Criterion 2 is not fulfilled by Zentiva, Tatra Mountain Resort and Stock Spirits Group. The last two mentioned can not be considered for research because they went public on the PSE in the fourth quarter of 2012 and 2013, respectively, a time frame not sufficient to include it in both formation and trading period. Zentiva left the PSE before the end of the first formation period. The shares of ECM, KIT Digital, AAA Auto Group and Fortuna were either introduced or delisted in the course of the analysis, and as such are included only in certain samples. For illustration, Fortuna went public on the PSE in October 2010 and is therefore included in the next nearest pairs formation period, starting in December 2010.

Irregularities affecting stock prices such as stock splits or stock dividends do not occur. Few companies underwent a capital reduction, a type of restructuring that does not need to be taken into consideration by data adjustments because theoretically the effect on share price is

minimal. The case of Telefonica CR, however, shows that in practice it is not always so and animal spirits may affect prices. In November 2012 the company decreased shareholders' equity through reduction in face value of its shares by 13%. The decision came into force on October 17, 2010, and over the following two months the company's share price dropped by 20%. This difference in market price pre- and post-capital restructuring probably reflects sentiment-driven response of investors rather than rational behavior: first, earnings per share remained unchanged; second, increased debt-to-equity ratio of the Czech Republic subsidiary was more than compensated for by a year-on-year decrease in debt-to-equity ratio of Telefonica SA.
Telefonica CR underwent a second round of capital reduction in the last quarter of 2013 through share cancellation and repurchase. This time the decision and implementation left its share price almost unaffected, which can be attributed to two opposing tendencies: the company is devalued by the exact amount of cash disbursed to buy back shares, but also shares of the remaining shareholders represent ownership of a greater fraction of the company. In neither case do we adjust data for this kind of capital restructuring.

All calculations and pairs trading models are programmed in the software R.

3 Theory

3.1 Preliminaries

In order to understand the cointegration approach, the key concepts of time series data need to be clarified. We define here the notion of covariance stationarity, integration and cointegration, show how we can test for stationarity and illustrate on a problem of spurious regression how cointegration helps us determine whether two nonstationary series are related. We then continue to describe the method, based on Engle and Granger (1987). The last subsection explains the process of normalization necessary for executing the distance method.

Covariance stationarity is a weaker form of stationarity[6], although fully sufficient for our purposes. A time-series process y_t with finite second moment ($E(y_t^2) < \infty$) is covariance stationary if, for all values, its mean and variance are constant and independent of time ((3.1) and (3.2)) and the covariances depend only upon the distance between the two time periods, but not the time periods per se ((3.3)):

$$E(y_t) = \mu \tag{3.1}$$

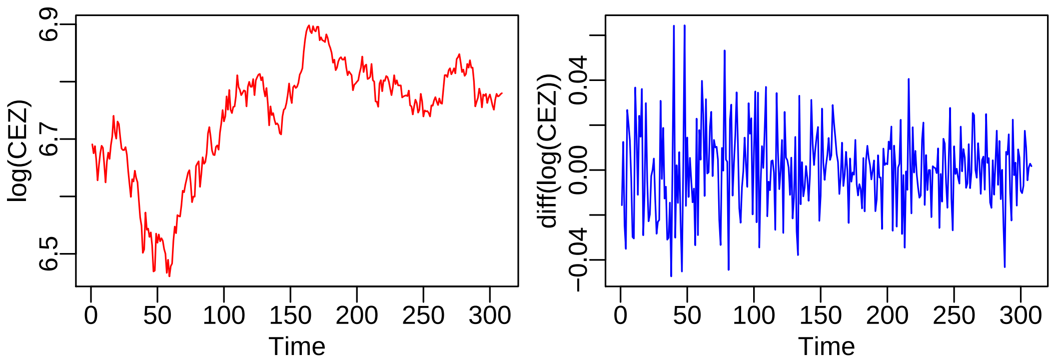
$$Var(y_t) = \sigma^2 \tag{3.2}$$

$$Cov(y_t, y_{t+i}) = \gamma_i \quad \forall\, i \geq 1 \tag{3.3}$$

For an example of nonstationary vs. stationary process we select a period dating from December 6, 2008 until March 5, 2010 to compare a plot of CEZ daily stock prices in logarithms with a plot of their first differences, which is the approximate percentage change in price:

[6] A stationary process is a stochastic process whose joint probability distribution remains unchanged when shifted in time, i.e. if we take any collection of random variables in the sequence and shift that sequence ahead i time periods, the joint probability distribution must remain stable.

Figure 3.1: Nonstationary and Stationary Stochastic Process



Process on the left-hand side in Figure 3.1 shows a trending behavior and therefore is clearly nonstationary. At a minimum, it does not fulfill condition (3.1). On the other hand, the series for the change variable shown on the right exhibits the property of mean reversion bceause it fluctuates around a constant value. By appearance we may guess that the change variable is covariance stationary.

There are many analytical ways how to determine whether a series is stationary or not. The most popular one, and the one employed by Engle and Granger (1987), is a formal test for a unit root formulated by Dickey and Fuller (1979). Three variants of the test exist, depending on the role of the constant term and the trend. The test begins with least squares estimation of AR(1) model:

$$y_t = \alpha + \rho y_{t-1} + e_t, \quad t = 1, 2, ... \tag{3.4}$$

where e_t are independent random errors with zero mean and constant variance. y_t has a unit root if, and only if, $\rho = 1$. We are interested in testing this hypothesis against the possibility that y_t is stationary, and so our alternative is one-sided $\rho < 1$[7]. We do not consider the case when $\rho > 1$ since it would imply that y_t is explosive, which is not the case for stock prices on our time scale.

[7] Practically this means $0 < \rho < 1$

A more established and convenient form for carrying out the unit root test is to subtract y_{t-1} from both sides of (3.4) and define $\theta = \rho - 1$:

$$y_t - y_{t-1} = \alpha + \rho y_{t-1} - y_{t-1} + e_t \tag{3.5}$$

$$\Delta y_t = \alpha + \theta y_{t-1} + e_t, \tag{3.6}$$

and estimate (3.6) by OLS. Then, the hypotheses are as follows:

$$H_0 : \rho = 1 \Leftrightarrow H_0 : \theta = 0$$

$$H_1 : \rho < 1 \Leftrightarrow H_1 : \theta < 0$$

If we fail to reject H_0, we conclude that the series is nonstationary and has a unit root. If we reject H_0, the series does not have a unit root. This test is known as the Dickey-Fuller test.
The asymptotic critical values for rejection of H_0 depend on the model and are taken from Davidson and MacKinnon (1993):

Table 3.1: Asymptotic Critical Values for the Dickey-Fuller Test

	Significance level		
Model	1%	5%	10%
$\Delta y_t = \theta y_{t-1} + \epsilon_t$	-2.56	-1.94	-1.62
$\Delta y_t = \alpha + \theta y_{t-1} + \epsilon_t$	-3.43	-2.86	-2.57
$\Delta y_t = \alpha + \lambda t + \theta y_{t-1} + \epsilon_t$	-3.96	-3.41	-3.13

The critical values need to be specially generated because under H_0, y_t is nonstationary, which means that its variance increases with the sample size and the t statistic does not have an approximate standard normal distribution. They are more negative than the t distribution critical values.

Under the condition $E(e_t|y_{t-1}, y_{t-2}, ..., y_0) = 0$, (3.6) is a dynamically complete model, but oftentimes finite distributed lag models such as (3.6) may suffer from dynamic form misspecification. One implication of this is that their adjacent errors are positively correlated, which violates the Gauss-Markov assumption of no serial correlation: $\forall\, t \neq s, E(e_t e_s | \mathbf{x}_t, \mathbf{x}_s) = 0$, where for our case $\mathbf{x}_t = y_{t-1}, \mathbf{x}_s = y_{s-1}$. As a consequence, estimates of the regression coefficients are inefficient and the usual OLS standard errors are biased downward, rendering the significance tests on the coefficients invalid. That is why we need to make sure we capture the full dynamic nature of the process, and include sufficient lag terms. We may allow Δy_t to follow an AR model in an extended equation of the form

$$\Delta y_t = \alpha + \lambda t + \theta y_{t-1} + \sum_{p=1}^{m} \gamma_p \Delta y_{t-p} + e_t, \tag{3.7}$$

where $|\gamma_p| < 1$. We can add p lags of Δy_t to the equation to account for the dynamics of the process, and the lag length is often dictated by the frequency of the data and the sample size. The optimal number of lags can be determined through the use of measures of the quality of a model such as Akaike information criterion (AIC) or Bayesian information criterion (BIC), but the more lags we include, the more initial observations we lose, possibly resulting in small sample power of the test. The unit root test based on (3.7) is referred to as the augmented Dickey-Fuller (ADF) test. The hypotheses remain as in the nonaugmented version, and the same critical values apply.

Referring back to Figure 3.1, we now have a formal tool to confirm that the left-hand side graph is indeed nonstationary. However, after differencing once, the series shows the property of mean reversion, and we reject the null hypothesis of a unit root. The minimum number of times a series must be differenced to make it stationary is called order of integration, and denoted $I(d)$. Therefore, the process on the right is said to be integrated of order zero, or $I(0)$, while the process on the left is integrated of order one, or $I(1)$. We observe that the concepts of stationarity and integration are tightly linked.

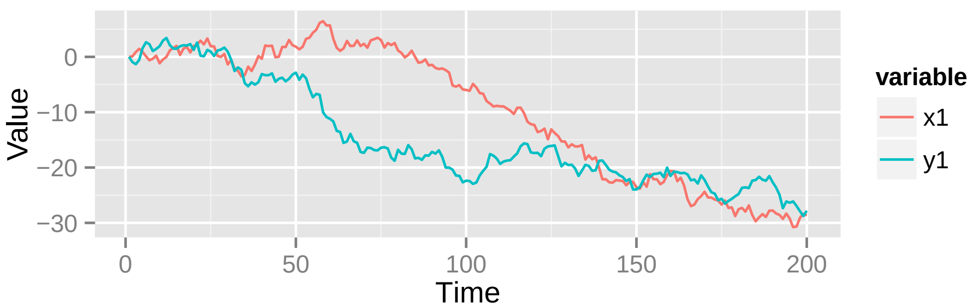
Knowing whether a time series is stationary or nonstationary is of paramount importance in regression analysis. For nonstationary series we may encounter a problem of obtaining significant regression results when in fact the series are not related. Such relationships are said to be spurious.

We demonstrate our point in a manner similar to Granger and Newbold (1974), one of the earliest studies of spurious regression phenomenon, albeit in our example we run single realization instead of Monte Carlo simulation done by Granger and Newbold (1974). We use the R software to generate two independent random walks x_t and y_t:

$$x_t = x_{t-1} + \epsilon_t, \quad t = 1, 2, ..., 200,$$
$$y_t = y_{t-1} + \upsilon_t, \quad t = 1, 2, ..., 200,$$

where ϵ_t and υ_t are i.i.d. random errors, $\epsilon_t, \upsilon_t \sim N(0, 1)$, and we set the initial values of the series to $x_0 = y_0 = 0$. Since ϵ_t and υ_t are mutually independent processes, x_t and y_t are also independent, and we should find no evidence of a relationship between them. However, when we look at Figure 3.2, we see that these entirely unrelated series have positive relationship. We continue to study this relationship in a more

Figure 3.2: Time Series of Two Random Walk Variables



analytical manner through running the OLS regression of y on x:

$$y_1 = \beta_0 + \beta_1 x_1 + u,$$

Table 3.2: OLS Regression Results

	Dependent variable:
	y1
x1	0.577^{***} (0.037)
Intercept	-9.056^{***} (0.567)
Observations	200
R^2	0.553
Residual Std. Error	6.176 (df = 198)
F Statistic	244.950^{***} (df = 1; 198)
Note:	*p<0.1; **p<0.05; ***p<0.01

summary of which is displayed in Table 3.2. For 198 degrees of freedom one can use the standard normal 5% critical value $c = 1.96$ for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. The regression yields huge t statistic on x: $t_{\hat{\beta}_1} = 0.577/0.037 \approx 15.6$, implying that the estimated slope coefficient is significantly different from 0 at any conventional significance level. Spurious regressions typically have low Durbin-Watson statistic, and this one is no different: $DW = 0.0328$. With 200 observations, one regressor and a model with intercept, the 1% critical values for one-sided DW test are $d_L = 1.664$ and $d_U = 1.684$. $DW < d_L$ is a condition for rejection of H_0 that claims errors to be uncorrelated. Thus, the errors exhibit positive serial correlation and are a source of downward-biased standard error for x_1 and the inflated t statistic. The regression model has an R-squared of 0.553, and so x_t is estimated to explain about 55.3% of the variation in the dependent variable. In any case, these results, no matter how appealing, are completely meaningless for we already explained that the processes are independent.

Using a real example from our thesis, we next choose sample period one with 13 stocks and 315 observations[8], dating from June 6, 2008 until September 5, 2009, and perform simple OLS regression for every possible pair. Except for one, all t statistics for slope coefficients are found to be statistically different from zero for two-tailed 1% level test.

[8]For sample periods and stocks included see Table A.1

In fact, the t statistic is in most cases in order of tens. Of the 78 pairs, 15 have R^2 greater than 0.85. Because it is unlikely that all stocks would relate to one another, we are again dealing with the problem of spurious regression. Later we reveal how many pairs are truly related.

Obviously, $I(1)$ variables should be used in regression analysis with great caution. Still, nonstationarity is ubiquitous in economics and financial data and it certainly is a property of stock prices, hence we need to establish a concept which would ensure that regressing one $I(1)$ variable on another will provide informative results. Such concept is cointegration.

3.2 Cointegration

We are now ready to present the definition of cointegration as introduced by Engle and Granger (1987):

The components of the vector x_t are said to be *cointegrated of order d, b*, denoted $x_t \sim CI(d,b)$, if (i) all components of x_t are $I(d)$; (ii) there exists a vector $\alpha(\neq 0)$ so that $z_t = \alpha' x_t \sim I(d-b)$, $b > 0$. The vector α is called the cointegrating vector.

If we concentrate on the case of $d = 1, b = 1$ and a two-component vector x_t, this tells us that in general, linear combination $z_t = \alpha_1 x_{1t} - \alpha_2 x_{2t}$ of two $I(1)$ processes x_{1t}, x_{2t} will result in an $I(1)$ process for any vector α. This means that the series may wander off far from each other without future convergence. However, for some particular $\alpha \neq 0$ the linear combination *may* yield a stationary $I(0)$ process, i.e. a process exhibiting mean reversion, constant variance and autocorrelations that depend only on the time distance between the two variables. If such α exists, then x_{1t}, x_{2t} are cointegrated, and share similar stochastic trends. Briefly, cointegration asserts that the variables have a stable long-run relationship. In this thesis, we focus on the situation when $x_t \sim CI(1,1)$ and leave the general case aside, as it is beyond the scope of the text.

In practice, the coefficient on x_{1t} is fixed at unity and α_2 is called the cointegration parameter, and we will follow this along in the rest of the thesis. The cointegration parameter tells us how many shares of x_{2t} we should buy/sell if we sell/buy one share of x_{1t}. It is the ratio in which to hold our position.

In pairs trading terminology, the linear combination z_t is termed the spread. Since the value of the spread reflects the degree of mutual mispricing between stocks, it is central to the strategy, for pairs trading attempts to avail of occasional temporary divergence from the mean value of the spread and subsequent returns to it. Because the expectation that the spread will revert back needs to be fulfilled in order to have profit potential, pairs trading heavily relies upon cointegration between stocks.

The most commonly used test of cointegration for two assets is based on a two-step procedure by Engle and Granger (1987):

1. After nonstationarity verification of x_{1t}, x_{2t}, the coefficients from the equation $x_{1t} = \alpha_2 x_{2t} + z_t$ are estimated by OLS and the residuals $\hat{z}_t$ are saved.
2. Residuals $\hat{z}_t$ are tested for stationarity using the unit root test: $\Delta\hat{z}_t = \gamma\hat{z}_{t-1} + \upsilon_t$ (or with lags of $\Delta\hat{z}_t$ to account for serial correlation)

The test is named the (augmented) Engle-Granger (AEG) test. It is basically a test of the stationarity of the least squares residuals. The null hypothesis states that the processes are not cointegrated, hence if H_0 cannot be rejected, the regression is spurious. The potential cointegration parameter α_2 needs to be estimated, which induces errors that are carried over into the second estimation through the use of residuals. Due to this, asymptotic distributions of the cointegration tests are different from those of ordinary unit root tests such as ADF, and so are their asymptotic critical values, taken from Hamilton (1994) and displayed in Table 3.3.

Pointing back to the OLS regressions for period one, majority of which were suspected to be spurious, we can now use the AEG test to verify whether this is the case. The two-step procedure yields five cointegrated pairs - a fraction of the pairs suggested to be related by the t test, and so the problem of spurious regression is proven to exist in our data. Statistics on the number of cointegrated pairs in each period is in Table A.2.

Table 3.3: Asymptotic Critical Values for the Cointegration Test

	Significance level		
Regression model	1%	5%	10%
$x_{1t} = \alpha_2 x_{2t} + z_t$	-3.39	-2.76	-2.45
$x_{1t} = \alpha_0 + \alpha_2 x_{2t} + z_t$	-3.96	-3.37	-3.07
$x_{1t} = \alpha_0 + \lambda t + \alpha_2 x_{2t} + z_t$	-3.98	-3.42	-3.13

3.3 Normalization

The distance method, proposed by Gatev et al. (2006), is less sophisticated than the method of cointegration. It relies on the process of standardization (or normalization) that accompanies both pairs formation and trading. Standardization is a conversion process when from individual data points we subtract the population mean and divide the difference by the population standard deviation. Obtained score is a dimensionless quantity that represents number of standard deviations a data point is above or below its mean. In our analysis, means and standard deviations are sample statistics since they are computed from data points in formation periods, a relatively small data sets. Therefore, the resulting score is a sample version of a standard score requiring population parameters. Hence, to take this difference into consideration, we will refer to the process as normalization.

4 Methodology

4.1 Cointegration Approach

In accordance with the cointegration definition by Engle and Granger (1987) applied to the case $CI(1,1)$, stock prices first need to fulfill the condition of nonstationarity (be $I(1)$ processes) to be considered for having a cointegrating relationship. Vidyamurthy (2004) claims that the assumption that the logarithm of stock prices is a random walk is a rather standard one. Based on a visual inspection of plots of stock prices in our sample periods and the fact that stock prices generally exhibit long periods of growth and decline, we choose equation that includes both a constant and a deterministic trend:

$$y_t = \alpha + \lambda t + \rho y_{t-1} + + \sum_{p=1}^{5} \gamma_p \Delta y_{t-p} + e_t, \quad t = 1, 2, ... \tag{4.1}$$

$$\Delta y_t = \alpha + \lambda t + \theta y_{t-1} + \sum_{p=1}^{5} \gamma_p \Delta y_{t-p} + e_t, \tag{4.2}$$

where y_t represents time series of logarithms of stock prices. We further set the maximum number of augmenting lags to five (trading week) to allow for the possibility that error terms are autocorrelated. In the thesis, optimal number of lags is determined by minimizing AIC, done automatically by the R software. Since BIC penalizes free parameters more strongly, AIC is more appropriate for the choice of lag terms, as in the trade-off between the danger of losing observations and that of autocorrelation, the latter is more severe. The maximum data points we can lose for each series is five, a negligible amount compared to the number of data points we have available, and so the test power will not be affected.

For the ADF unit root test based on (4.2), the null and alternative hypotheses are $H_0 : \theta = 0$ and $H_1 : \theta < 0$, respectively. Under the null,

y_t has a unit root while under the alternative, y_t is trend-stationary. We perform the one-tailed test on a standard significance level of 5% with 309-317 data points, which means that the asymptotic critical value $c = -3.41$ applies (see Table 3.1). H_0 is rejected if a t statistic on y_{t-1}, $t_{\hat{\theta}}$, is lower than the critical value: $t_{\hat{\theta}} < c$. Strictly speaking, trend-stationary variables are not $I(1)$ processes, and hence all stocks identified as such are excluded from the analysis. Had we not taken into account the linear trend, we would have mistakenly identified the trend-stationary processes as nonstationary.

Carrying on to conduct the AEG test, we need to build up an OLS model. The series we use have nonzero mean, so a constant term is necessary. After the preceding step we are left only with nonstationary stocks, resulting in the model

$$y_t = \mu + \beta x_t + e_t, \tag{4.3}$$

i.e. without a linear time trend, where y_t, x_t are time series of logarithms of stock prices. Estimating (4.3) by OLS gives residuals (the spread) of the form $\hat{e}_t + \hat{\mu} = y_t - \hat{\beta} x_t$. The OLS estimator $\hat{\beta}$ is the potential cointegration coefficient that we will use in trading periods as a ratio in which to buy and sell stocks in case the pair is cointegrated, $\hat{\mu}$ is the equilibrium value, and $\hat{e}_t$ is a time series with zero mean. The linear relationship on the right-hand side of the equation is also termed the equilibrium relationship. In order to eliminate the risk of serial correlation, five extra lag terms are added to the equation based on which we test for cointegration:

$$\Delta e_t = \theta e_{t-1} + \sum_{p=1}^{5} \gamma_p \Delta e_{t-p} + u_t \tag{4.4}$$

As in equation (4.2), number of lags in equation (4.4) is optimized by AIC, and the same null and alternative hypotheses as for the ADF test apply for the AEG test. The cointegration-test asymptotic critical value for a model with a constant is $c = -3.37$, found in Table 3.3. Cointegrated pairs are ones for which $t_{\hat{\theta}} < c$. While in testing of stock prices for unit root the ideal state was not to reject H_0, in the AEG test of residuals the aim is the opposite and we wish to achieve rejection of the null hypothesis in as many cases as possible. Having the fewest stationary stocks (the fewest stocks for which unit root is rejected to be concrete) and the most cointegrated pairs potentially increases the number of trades and thus may lead to more statistically indicative results.

Aside from cointegration, we require $\hat{\beta}$ to be positive. Trading with $\hat{\beta} < 0$ is not a contrarian investment strategy as the positions are both long or both short. All pairs that satisfy $\hat{\beta} > 0$ and are found to be cointegrated are eligible for trading. The remaining pairs are not considered throughout the rest of the analysis.

Once we have obtained the cointegrated pairs, we need to establish the spread as a necessary metric for trading. The estimators $\hat{\beta}$ and $\hat{\mu}$ of cointegrated pairs are carried over from the formation period, in which they were estimated, to the trading period where they form the trading spread between the same stocks: $\hat{e}_t^T + \hat{\mu}^F = y_t^T - \hat{\beta}^F x_t^T$, where F and T denote whether the values were sourced in formation or trading period. Neither the intercept $\hat{\mu}$ nor the cointegration parameter $\hat{\beta}$ can be derived from OLS model estimated in the trading period itself because that would introduce a look-ahead bias; in the trading period we attempt to simulate real conditions and we cannot use data that are not available at any moment during the trading period.

4.2 Distance Approach

In a manner similar to Gatev et al. (2006), pairs are selected on the basis of minimized SSD between standardized historical returns of two stocks. The methodology proceeds in the following steps:

(1) In formation period, we compute daily returns for each stock as a logarithm of a ratio of stock price to its value on the preceding day:

$$r_{it} = \log\left(\frac{P_{it}}{P_{it-1}}\right) \tag{4.5}$$

This has an interpretation of approximate percentage return (standardly used in the literature).

(2) Returns are normalized by subtracting their sample mean and dividing by their sample standard deviation:

$$r_{it}^{Z} = \frac{r_{it} - E(r_i)}{\sigma_i^r} \tag{4.6}$$

(3) The returns r_{it}^{Z} are added by cumulative summation, and such series, denoted as r_i^*, are then used in the equation of the SSD:

$$SSD = \sum_{t=1}^{N} (r_{it}^* - r_{jt}^*)^2 \quad \forall\, i \neq j \tag{4.7}$$

(4) Values obtained from equation (4.7) are ordered and five pairs with the lowest SSD are chosen for trading. This method is consistent with selection based on criterion of maximum correlation between cumulative returns. The remaining pairs are not considered throughout the rest of the analysis.

In the distance method, the spread reflects degree of relative mispricing between stocks as it did in the cointegration approach, although its computation lies on other grounds. In this case, the spread is the difference between cumulatively summed normalized returns of stocks that form a pair. Normalization is done such that the mean and standard deviation computed from percentage returns in the formation period are carried over to the trading period and what we normalize are trading period returns: $r_{it}^{Z} = \frac{r_{it}^{T} - E(r_i^F)}{\sigma_i^F}$, with F and T signifying in which period the values were obtained. These series are added by cumulative summation, and the spread is represented by the difference between the summed series. The risk of look-ahead bias precludes us from using mean and standard deviation computed right in the trading period.

4.3 Trading Rules

Both cointegration and distance method use oscillations about the equilibrium value of the spread as a way how to assess when to open a position. When the spread crosses a preestablished threshold, it means that the degree of mutual mispricing between two stocks has deviated far from its historical mean, which is a signal for entering the trade. The steps how in each method these thresholds are generated follow the same principle: they are based on standard deviations calculated from the formation-period values of the spread.

In the cointegration method, the long-run equilibrium value of the linear combination $y_t - \hat{\beta}x_t$ is $\hat{\mu}$. The standard deviation of this combination is computed from the residual series: $\Delta = \sqrt{Var(\hat{\mu} + \hat{e}_t)}$. We put on the trade on a distance of 2Δ in either direction from the mean $\hat{\mu}$, and unwind the position upon reversion to it, i.e. when the spread of the pair crosses the value $\hat{\mu}$. More specifically, we buy (long) the spread[9] when, at time t, the relative mispricing reaches deviation of -2Δ from the mean, and sell (short) the spread at time t+i:

$$y_t - \hat{\beta}x_t = \hat{\mu} - 2\Delta \tag{4.8}$$

$$y_{t+i} - \hat{\beta}x_{t+i} = \hat{\mu} \tag{4.9}$$

Conversely, we sell (short) the spread when, at time t, it is $+2\Delta$ above

[9] To buy the spread means that we buy the relatively underpriced stock x_t and sell the relatively overpriced y_t

the mean, and buy (long) the spread at time t+i:

$$y_t - \hat{\beta} x_t = \hat{\mu} + 2\Delta \tag{4.10}$$

$$y_{t+i} - \hat{\beta} x_{t+i} = \hat{\mu} \tag{4.11}$$

We refer to $\hat{\mu}$ as the closing threshold. The profit on the trade is the incremental change in the spread, 2Δ.

In the distance method, the equilibrium value of the spread is computed as mean of the spread in formation period. From the same spread we obtain standard deviation. As in the cointegration method, trading signals are set at two standard deviations from the equilibrium, and position is closed upon mean reversion:

$$r_{it}^T - r_{jt}^T = E(r_{it}^F - r_{jt}^F) \pm 2\sqrt{Var(r_{it}^F - r_{jt}^F)}, \quad i \neq j \tag{4.12}$$

$$r_{it}^T - r_{jt}^T = E(r_{it}^F - r_{jt}^F), \quad i \neq j, \tag{4.13}$$

where F and T denote whether the values were taken from formation or trading period.

In both methods pairs can have multiple cash flows during the trading period, or they may have none in case the prices never diverge by more than two historical standard deviations. If a pair remains open on the last day of the trading period the position is liquidated regardless of convergence.

Contrarian investment strategies, including pairs trading, are subject to bid-ask bounce (Harris (2002), Jegadeesh (1990), Jegadeesh and Titman (1995a)) that inflates computed profits and therefore needs to be treated in pursuance of more realistic profit estimation. On equity markets we may see two different prices of stock: the price quoted for an immediate sale - bid, and that for an immediate purchase - ask. This discrepancy is called bid-ask spread and is common for almost all kinds of assets. However, it often happens that sellers sell near an ask price and buyers near a bid price, a phenomenon termed as bid-ask bounce.

In pairs trading, it is likely that winners' price reflects an ask quote and losers' price a bid quote, making the divergence of prices appear larger than it actually is. Upon convergence, the converse is true and the winners are traded at a bid quote (losers analogically), which further contributes to an upward bias in profits. To minimize the effect of the bias, we initiate a trading position one day after receiving a signal of divergence and liquidate it on the day following the reversion to the mean value of the spread. This waiting rule should provide a more realistic return estimation because it makes provision for potential difficulties and time delays investors encounter when executing a trade. Gatev et al. (2006) and Broussard and Vaihekoski (2012) use the drop in excess returns resulting from the one-day gap as an estimate of transaction costs, and we do likewise.

4.4 Computation of Returns and Performance

To clear out any upward bias in profits and get a more accurate estimation of returns arising solely from pairs trading strategy, we assume that in times when no positions are opened we earn zero return on capital.

For computation of portfolio returns in the cointegration method we follow Vidyamurthy (2004):

$$[\log(P^L_{t+i}) - \log(P^L_t)] - \beta[\log(P^S_{t+i}) - \log(P^S_t)] \tag{4.14}$$

from which, after rearranging, we get

$$[\log(P^L_{t+i}) - \beta \log(P^S_{t+i})] - [\log(P^L_t) - \beta \log(P^S_t)], \tag{4.15}$$

where β is the cointegration coefficient and L, S represent long and short position, respectively. The equations (4.14) and (4.15) refer to the case when we long the spread. During trading period, it may occasionally happen that a stock in a pair that was once bought as relatively underpriced will in the same trading period become relatively overpriced (or vice versa), and hence the spread will be once bought and the next time shorted. When the spread is shorted, we trade the stocks in the same ratio but now the profit computation follows the equation

$$-[\log(P^S_{t+i}) - \log(P^S_t)] + \beta[\log(P^L_{t+i}) - \log(P^L_t)] \tag{4.16}$$

Due to the way the trading signals are generated, holding both long and short position in the spread of identical pair on the same days is not attainable.

Portfolio returns generated by the distance method are calculated as daily weighted percentage returns for long and short position, and are going to be executed according to the formula

$$r_{pt} = w_{1t}r^L_t - w_{2t}r^S_t \tag{4.17}$$

where r^L_t and r^S_t are daily returns for the positions, and w_{1t}, w_{2t} represent their daily weights. From Broussard and Vaihekoski (2012), this formula gives basically the same result as the one in Gatev et al. (2006): $r_{pt} = \frac{\sum_{i\in p} w_{it}r_{it}}{\sum_{i\in p} w_{it}}$ if the weights are adjusted accordingly. Our weights initially start at one to take into account that the strategy is dollar-neutral[10], and are marked-to-market daily based on changes in the value of stocks: $w_{it} = w_{it-1}(1 + r_{it-1})$, which implies that we compute r_{pt} as the daily reinvested payoffs. Had we not adjusted the weights daily but rather assumed them to be one at all times, the strategy would have incurred tremendous transaction costs due to daily buying and selling orders executed in an effort to keep the weights constant. This is not necessary nor would the results generated this way be any indicative of potential real-world profits.

Because the initial investment in a trade sums to zero, r_{pt} from equation (4.17) can be interpreted as daily excess return. To compute profits

[10]Each position has the same absolute dollar amount at t=0, and thus starts as a net zero position, without any investment.

for one pair throughout the entire holding period, we simply sum equation (4.17) across all days on which the position is held, and obtain its cumulative total excess return.

In order to make the two analyzed methods directly comparable, we compute the distance-method returns once again by using formulas (4.15) and (4.16)[11]. β in this case is the ratio of stocks constructed to obtain market-neutral portfolio[12]. For the sake of clarity, we denote this ratio as k and use β in its classical financial meaning, i.e. as a measure of systematic risk of a stock in comparison to the market as a whole. To determine k, we first regress return series of portfolio stocks on market returns:

$$r_1 = \alpha_1 + \beta_1 r_m + \upsilon_1 \tag{4.18}$$

$$r_2 = \alpha_2 + \beta_2 r_m + \upsilon_2, \tag{4.19}$$

where r_m is the PX index return, r_1, r_2 are stock returns and β_1, β_2 are measures of systematic risk. Lets assume that in our pairs portfolio we buy one unit of stock 1 and sell k units of stock 2; then the return is $r_p = r_1 - kr_2$. After plugging in for r_1, r_2 from (4.18) and (4.19) we receive

$$r_p = \alpha_1 + \beta_1 r_m + \upsilon_1 - k(\alpha_2 + \beta_2 r_m + \upsilon_2) \tag{4.20}$$

$$r_p = \alpha_1 - k\alpha_2 + (\beta_1 - k\beta_2) r_m + \upsilon_1 - k\upsilon_2 \tag{4.21}$$

For this porfolio to be market-neutral, the correlation of its returns and the market returns must be 0: $\beta_1 - k\beta_2 = 0$. Hence, $k = \frac{\beta_1}{\beta_2}$: for each unit of stock 1 that we long/short, we short/long k units of stock 2.

As a measure of the risk-adjusted investment performance we choose

[11] This is not a part of the included R code.

[12] Market-neutral portfolio is a portfolio whose return is uncorrelated with the market return. Regardless of whether the market is bullish or bearish, market-neutral strategy performs in a steady manner and with lower volatility

information ratio designed by Treynor and Black (1973):

$$IR = \frac{E(R_p - R_b)}{\sqrt{Var(R_p - R_b)}} \tag{4.22}$$

where R_p is the portfolio return, R_b is the benchmark return, which is in our case the PX equity index, the numerator represents the expected active return and the denominator is the standard deviation of the active return, or tracking error. The information ratio is a commonly employed performance measure among hedge funds because it uses appropriate benchmark which eliminates market risk, showing only risk taken form active management. As such, it shows value added relative to this benchmark. The infamous and still very popular Sharpe ratio (Sharpe (1966)) has been criticized for using the risk-free rate because it places all managers on the same playing field irrespective of style.

Opinions on what level of information ratio should be regarded as satisfying are not consistent. According to Grinold and Kahn (2000), the information ratio is analogous to a normal bell-shaped curve with IR=0 as the mean of the distribution. Generally speaking, a figure of 0.5 reflects a good performance, 0.75 very good and 1.00 outstanding.

5 Scenario Analysis

The proposed scenarios take into account specifics of the PSE that do not pertain to the US and Finnish equity markets, such as size, industry diversity, liquidity, efficiency and volumes traded. Although these factors did not alter our methodology used from what is outlined in Gatev et al. (2006) or Broussard and Vaihekoski (2012), they are likely to affect the ability to form pairs and generate profits. We expect three possible outcomes:

Scenario 1: Due to the absence of pairs that would belong to the same sector, industry shocks will cause that no stocks will be identified as cointegrated, and using the cointegration method for trading will not be feasible.

In the distance method, trading five pairs with the smallest SSDs will lead to situations when pairs wander-off from each other and positions remain open longer than our preestablished threshold, basically changing the strategy to buy-and-hold one. Positions in these trades will need to be closed regardless of convergence, most of them will yield negative returns and average excess return will be negative as well. Using evidence from the Finnish stock market presented by Broussard and Vaihekoski (2012), who trade a pair on average 23 days, we set our threshold for average days per trade at twice this value to make allowance for the fact that multiple share classes may have stronger mean-reverting behavior. In the distance-method Scenario 1, we expect average return per trade to be negative and average days per trade to be above 46:

$$\begin{aligned} \textit{S1:} \quad \bar{r} &< 0 \\ \overline{days} &> 46 \end{aligned}$$

Scenario 2: The cointegration method will reveal certain cointegrated pairs. However, due to industry diversity, sector shocks will break the cointegrating relationship in trading periods, resulting in non-convergence and losses.

Pairs selected in the distance method will take long time to converge once the trade is entered, averaging beyond the threshold set in Scenario 1. Nevertheless most traded pairs will gradually move toward the equilibrium value of their spread, and the average excess return will be positive:

$$\textit{S2:} \quad \bar{r} > 0$$
$$\overline{days} > 46$$

Scenario 3: Because the PSE is a small equity market, where stocks are generally more interrelated, and it is considerably influenced by events in Germany and within the European Union, sentiment of investors will be similar regarding all the stocks. In each sample period there will be at least 10% of cointegrated pairs. Industry heterogeneity will be causing sector shocks, affect stock prices and swing the spreads, nevertheless the spreads will revert back to their mean values in time horizons averaging at most 30 days for the cointegration approach. The optimal number of average days per trade was determined based on averages from the formation period, where we obtained the value of 20 days. In trading period we need to take into account the fact that the cointegration coefficient β was taken from the formation period, and so we should expect the average days per trade to be inflated in trading samples. The executed trades will lead to positive returns.

In Scenario 3, we expect the outcome of the distance method to be average days per trade lower than 46, and if trades need to be closed on the last day of trading period, their returns will mostly be positive.

They will generate positive average excess return:

$$\textit{S3:} \quad \bar{r} > 0$$
$$\overline{days} < 46$$

6 Empirical Results and Discussion

6.1 Cointegration-Based Trading

Table A.2 shows the number of stationary stocks and cointegrated pairs. In the cointegration method, the necessary condition of nonstationarity is met by nearly every stock. In four sample periods, the ADF test leads to the rejection of unit root in the share price of Unipetrol. On this account, Unipetrol has to be detracted from the periods concerned.

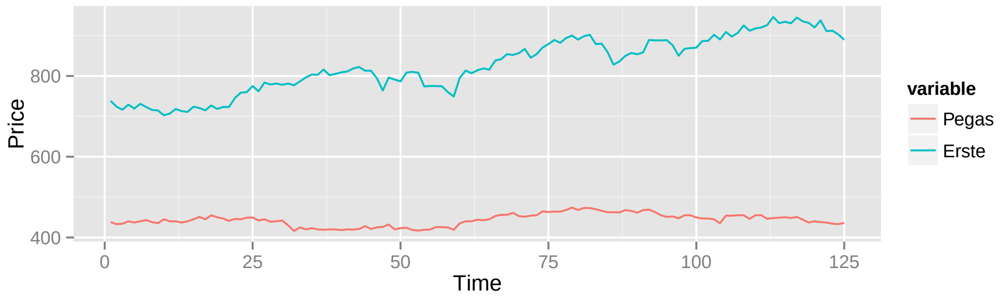
Cointegrated pairs with positive cointegration coefficient β are scarce, reaching a maximum 11.5% of all pairs only in one sample period, with total average across all samples 4.14% of the number of pairs. Ten cointegrated pairs are taken out because they have $\beta < 0$. The sample period nine does not contain any cointegrated pair, and is not suitable for trading. As a result, we do not reject Scenario 1.

Scenario 2 cannot be rejected in any sample period apart from sample nine because we are able to identify cointegrated pairs in every one of them. However, once traded, these pairs do not seem to be related anymore. For illustration, in Figure 6.1 we plot price series in absolute values[13] of the shares of Erste and Pegas in trading period three. Trading signal is generated at time 35 and the position is liquidated at a loss of -4.63% on the last day of the period. Not surprisingly, cointegration test in the trading period reveals that this pair is no longer cointegrated. Investigation of possible causes that could stand behind the boom in share price of Erste shows that in the middle of 2010, banking sector, one of the sectors that was worst affected by the debt crisis, starts picking up. The positive sentiment is brought about by rising profits of European banks and declining risk aversion of investors. In October, Erste presents press report for the 3rd quarter of 2010 with

[13]The cointegration was investigated in logarithms of prices. We plot here absolute values for clearer depiction of the price development.

increased Y-o-Y net profit, better interest margin and lower operating costs. The rise in the banking industry that does not impact Pegas, the textile producer, is an example of how industry shock can break cointegrating relationship.

Figure 6.1: Time Series of Prices of Erste and Pegas in Trading Sample 3



The long-short position in Erste and Pegas from period three represents one of the 82% trades that need to be closed on the last day of trading as the spread diverges without subsequent mean reversion. Pursuant increasing evidence that cointegration between selected stocks is no longer in place in trading periods, we decide to analyze cointegration in trading samples as well. Cointegration tests verify that virtually none of the pairs that we selected would not qualify for long-run relationship anymore. Research of historical events in industries and companies concerned confirms that industry heterogeneity is a cause of loss among many pairs that do not converge to the equilibrium until the end of the trading period. Due to that, Scenario 2 is further substantiated.

One implication of the industry shocks and broken cointegrating relationships is evidenced by histogram of percentage returns, shown in Figure A.1. The distribution statistics are displayed in Table A.3. Returns are measured per trade, and trades occur on time scale of a 6-month trading period. The strategy yields negative average return -5.17%, nearly 50% of the trades end up in loss and the distribution of returns shows high downside risk, also referred to as 'fat-tail' risk: its kurtosis is extremely positive and its skewness large and negative. The positive excess kurtosis (leptokurtosis) of 8.023 means that the distribution has significant mass concentrated in outliers, and so extreme events are more probable. This characteristic would not be so worrying if skewness were positive. However, its value of -2.579 along with the shape of the histogram tell us that the strategy has high percentage of returns concentrated around zero but also large negative returns with smaller probability. The combination of leptokurtic and negatively skewed distribution results in large downside tail, which is not a strategy any investor would undertake, as strategies should exhibit no skewness and relative platykurtosis. In fact, such negative statistics are rarely seen among hedge funds.

Based on the described properties of return distribution it is reasonable to expect large minimum and modest maximum in returns. The minimum is -0.896: during one single trade we lost 89.6% of the initial value of the investment. Maximum is 13.5%. Sample standard deviation 0.218 as a measure of dispersion confirms high volatility of returns and therefore riskiness of our strategy. Most of the time, trades will end up in the range between -26.97% and 16.63%. For example the S&P 500, a common benchmark for large-cap funds, had a standard deviation of 0.217 in May 2011. Nonetheless, since we are not merely holding an index but have a developed strategy, we would expect to receive reward for this strategy and the risk undertaken. Its excess return relative to a benchmark index is captured by average information ratio. It is 0.0697, which implies that the technique does not outperform the PX index, and perceived by the metrics of Grinold and Kahn (2000), it is considered poor.

Table A.3 presents comparison between strategy that delays the opening of the pairs position by one day with rule that opens a position at the end of the day that prices diverge by more than 2 historical standard deviations. With one day of waiting, the trading yields a negative average return of -5.17%. Contrary to what is expected from the theory of bid-ask bounce, end-of-the-day opening of the position gives slightly more negative return: -5.27%. A possible explanation is that the trading signal is generated immediately after crossing of the 2-deviation trading threshold; however, that does not guarantee the spread will start mean-reverting from that exact instance. Oftentimes, it does not, and hence higher value of the spread on the day following the opening signal raises returns. On the closing day, delaying liquidation by one day should again increase the spread and returns. In Figure 6.2 for instance, the first trade, at time 8, deviates further behind the selling threshold for another week before it crosses the threshold back on the way towards its mean[14]. Without waiting, the return on the trade is 0.446; with a one-day lag, it jumps to 0.543[15].

Unfortunately, the waiting rule as an approximation of transaction costs fails to deliver meaningful results. Rather, it puts into question whether it would not be more convenient to delay the trading as a part of the strategy. Nevertheless, despite the seeming advantage of one-day lag, it is with high probability nothing else but another consequence of non-cointegration. Hence, instead of waiting for one day before trade execution, one should refrain from trading the pair completely. We find support for this theory in other measures that are in favour of the no-waiting approach: minimum and maximum are shifted to the right on the return axis, skewness is less negative and excess kurtosis less leptokurtic.

With average days per trade of ≈ 58, we can reject Scenario 3 that the average would be at most 30 days. This large number only confirms that once industry shocks occur, prices of paired stocks do not maintain the cointegrating relationship.

[14]This trade is taken from the distance method for the sake of illustration.
[15]Computed by using (3.16)

6.2 Distance-Based Trading

Histogram of returns per trade is shown in Figure A.2, return distribution statistics in Table A.4. Average excess return employing formula (4.17). i.e. value-weighted approach, is 2.927% calculated with a one-day lag. The distance method recognizes the discrepancy between waiting vs. immediate execution of trade as we expect it from the theory of bid-ask bounce; with the latter approach, returns rise to 3.44%. We can therefore assume the one-day lag outcome to be an estimate of returns after inclusion of transaction costs. If it were not for the low information ratio, it would be considered a good performance. But the $IR = 0.105$ is close to zero, and so we obtained no active return on investment. Considering the shortselling constraints and costs, it would be preferable to hold the PX index.

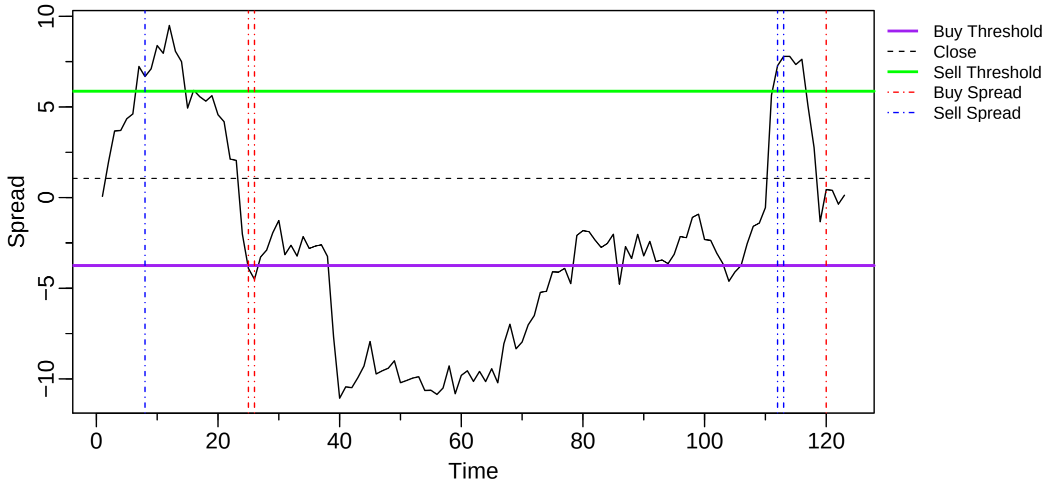
Based on the standard deviation of 0.165, majority of the trades should fall in the range between -13.57% and 19.43%. While we got positive average return, only 32.7% of trades are round-trips, average number of days the pairs are traded is 53.88 and 38% of trades end up in loss, which indicates that long-run relationship often does not hold. Out of the trades liquidated at the end of the period, majority have negative returns as stated in Scenario 1. But Scenario 1 also states that average return is negative, so we reject it. Scenario 2 correctly predicts that the average number of days per trade will be over 46. At the same time, it claims positive average return. We do not reject it. Almost half of the trades which close at the end of the period are profitable, suggesting that the comovement of stocks happens on a larger time scale but the lockstep in the development of returns is persistent. Hence, it is true that most traded pairs converge toward the spread's equilibrium. Finally, we reject Scenario 3 because the average time per trade is not lower than 46 days.

The histogram of returns shows some evidence that the underlying disribution is normally distributed. By making a Q-Q plot we prove that our distribution has heavier tails than the normal because the Q-Q plot does not form a straight diagonal line but its ends are bent to the shape of S. The points are arched and refer to negative skewness, which is all confirmed in the excess return distribution in A.4. These properties are not optimal as they induce a risk of extreme adverse events. Moreover, they hamper our ability to test hypotheses. Return data have a unimodal distribution, but they are not entirely symmetric and the sample comprises of 52 points. The ideal conditions of normality and large sample, required for a one-sample t test for the mean, cannot be fulfilled. Still, we decide to include t statistic and conduct the t test, results of which have only orientational character. We want to test whether the average excess return is zero against an alternative that the return is positive: $H_0 : r = 0$, $H_1 : r > 0$. The t statistic is computed as $t = \frac{\bar{r}-0}{se(\bar{r})} = \frac{0.0293-0}{0.023} \approx 1.277$. The 5% critical value for 51 degrees of freedom is $c = 1.675$. Because $t_r < c$, we find that the average return is not statistically greater than zero.

To give an example on how the pairs are traded, in Figure 6.2 we illustrate the spread between normalized returns of CETV and AAA in the trading period seven. The spread is formed such that when it is sold, we sell shares of CETV and buy shares of AAA, and vice versa. Horizontal trading thresholds indicate the boundaries above/below which values of the spread deviate by more than two historical standard deviations. The dashed horizontal line informs when the position should be liquidated, and vertical lines represent buy and sell signals for the spread. During the period, the position is opened three times: twice as a short position in the spread (at times 8 and 113) and once as a long position (at time 26). All trades have positive returns.

Distance-based trading does not have in the PSE as good performance as is presented by Gatev et al. (2006), who trades pairs in the S&P 500. We attribute this to greater number of potential pairs in the US market that allows to select pairs from the same industry and hence

Figure 6.2: Spread Between CETV and AAA with Thresholds and Trading Triggers



achieve a more stable relationship. Our surmise is confirmed by the percentage of mixed-sector pairs in the S&P 500 analysis, which is mere 20%. The industry heterogeneity puts the PSE in a great disadvantage, reflected in almost all compared values. The US market yields nearly twice as many trades per pair (1.16 vs 2.02) and its pairs are more likely to be traded at least once in the 6-month trading period. The time per trade is shorter, which could again be caused by tighter comovement in stocks, or by greater trading volume and hence better liquidity and efficiency in general. It is calculated as average time pairs are open in months divided by average number of trades per pair: $\frac{3.75}{2.02} = 1.86$ months for the S&P 500 vs. 2.23 months for the PSE. Regarding returns, the principal difference is that the t statistic in empirical analysis of Gatev et al. (2006) leads to strong rejection of hypothesis of zero returns, while in our case zero returns cannot be rejected even at 10% significance level[16]. The skewness is also in favor of the US market, with 0.34 vs -0.344. Range of attained returns is over 84% in the PSE, a great volatility, against the range 27% in the S&P 500. The single measure in which the trading in the PSE outperformed the other one is

[16] The 10% critical value for a one-tailed t test with 51 df is 1.298.

excess kurtosis; still, since the minimum and maximum of returns is not spread out far from zero in the S&P, it does not present serious problem in terms of outliers. Comparison measures are in Table A.6.

Overall, the S&P market offers more trading opportunities and non-zero returns with shorter time to achieve them, and if conditions in the S&P persist until today, we can consider it superior for employing pairs trading strategy. Analysis of present-day suitability of the US equity market for pairs trading is beyond the scope of this text.

6.3 Comparison of Cointegration and Distance Method

Table 6.1 summarizes trading differences between the two analyzed methods. The distance method yields nearly twice as many trades and has almost two times higher fraction of round-trips - positions that revert to the mean in the course of the trading period[17]. This measure alone is already a sign of its superior performance, since mean reversion of the spread ceteris paribus generates higher returns than convergence that does not reach the closing threshold. Both strategies have their trades open for average time of 2.59 months (distance) and 2.78 months (cointegration), which is far beyond Broussard and Vaihekoski (2012) (23 trading days) and even high above Gatev et al. (2006) (1.86 months). Distance method has a lower fraction of distinct pairs, which is another indicator that its outcome might be better than that of the cointegration method, as it indicates that the same pairs are chosen in more samples and therefore have a more stable mean-reverting relationship.

The use of similar methodology, which is in both cases based on the concepts of spread divergence and reversion to the equilibrium value, leads to the selection of the same most frequent pair, made up of KB (banks) and VIG (multiline insurance). Despite industry diver-

[17]From the original definition in stock trading context, round-trip means purchase/sale and subsequent sale/purchase of securities, regardless of whether it occurs during or at the end of a period. We change slightly the meaning to fit our purposes

Table 6.1: Trading Statistics: One-Day Waiting Rule

	Cointegration	Distance
Total number of trades*	28	52
Total number of round-trips*	5	17
Total number of nondistinct pairs*	30	45
Total number of distinct pairs*	24	32
Average number of trades opened per trading period	4.11	5.78
Average number of days per trade	58.0357	53.8846
Stocks most frequently paired up (# of nondistinct pairs)	KB (11), AAA (8), VIG (5)	Erste (13), KB (12), VIG (9)
Most frequent pair (# periods)	KB+VIG (4)	KB+VIG (5)

*across all periods; round-trip = trade which is closed before the end of the trading period

sity among all stocks (except for Erste and KB), the methods achieved to find a pair in which shares are as close in the type of industry as they can get. Contrary to the past, when they were separate entities, banks and multiline insurance companies are nowadays becoming more and more interrelated as banks offer insurances as investment opportunities. The institutional reposition of finance has played a major role in bringing the industries together and we have evidence of this in our analysis. In the distance method, the three most frequent pairs are KB and VIG (5 periods), Erste and VIG (3), and KB and Erste (3). In the cointegration method, the pairs are much more distinct.

Another meaningful statistic are shares that were most frequently paired up. Again, banking and insurance are in the first places. The influence of these industries on other companies listed on the PSE is economically straightforward - during economic booms, the demand for loans, banking and insurance services boosts earnings of these institutions. Along with that, lending standards are looser and the consumer demand for products and services of companies increases. Hence, the prosperity of banks moves together with the prosperity of companies. In a recession, banks compensate for a riskier environment by tightening their lending conditions and consumer demand is weak, leading to the comovement with other sectors in opposite direction.

This is about all what the two methods share in common. Table A.5 and Figure A.3 reveal major differences between return distributions, where returns for the two methods are calculated along the same metric based on equations (3.15) and (3.16). The results are in favour of the distance method in all aspects except for the average information ratio. The distance-based return distribution is leptokurtic, but far less than in the cointegration method, and the positive skewness means that if extreme events occur, they bring large positive returns. The average return of the distance method reaches 3.72%, i.e. it is by 8.89% higher, and it is achieved with lower volatility. Maximum return is 84.3%, nothing compared to 13.5% in the cointegration-based trading. Even though the distance method has not beaten the PX index held on exactly the same days, in terms of other metrics it has performed considerably better than the method of cointegration.

Conclusion

The purpose of this thesis was to investigate pairs trading strategies of cointegration and minimum distance in the Prague Stock Exchange. The choice of the topic was motivated by scarce published empirical research and by evidence of declining profitability of pairs trading in the US market, presumably caused by worsening arbitrage risk and increased efficiency. Since the PSE is a small equity market, chances were that pairs trading profits might not be arbitraged away as quickly, and that events within the European Union and our most influential neighbor, Germany, could affect stocks similarly, regardless of industry. On the other hand, conditions in the market such as the low number of stocks and industry diversity indicated that the success of the strategy may not be so straightforward.

We aspired to analyze how the individual methods would perform, which one would be a better investment strategy and if pairs trading would yield more satisfactory results in the US equity market than in the PSE.

The cointegration method identified on average 4.76% of pairs to be cointegrated. This relationship, however, did not last until trading periods and 46% of trades had negative return. The method failed in all measured statistics, including maximum loss, volatility, information ratio and average return. Investigation of causes showed that industry diversity is a crucial factor when it comes to cointegration, because industry shocks were the reason why the spread between pairs of stocks increased and did not revert to its equilibrium value. The distance method, returns of which were computed along the formula proposed by Gatev et al. (2006), revealed almost a bell-shaped distribution of returns and average return not statistically different from zero. When compared to pairs trading results in the S&P 500, presented by Gatev et al. (2006), the method did not generate as many trading opportunities, had greater risk, and the mean reversion occured in longer time span. We claim this to be the result of higher number of potential pairs in the S&P 500, which ensures that pairs with a more stable comovement of prices will be selected for trading.

When we compared the two approaches, we found few intriguing similarities. The pair that was most frequently selected in both was formed by the shares of KB and VIG. We attribute this to the narrow link between banking and insurance sectors as these institutions nowadays frequently offer similar services. Moreover, shares that were most frequently paired up also belonged to banking and insurance industries (KB, Erste, VIG). One possible explanation is positive correlation between prosperity in these industries and prosperity of companies, arising from economic cycles. The comparison further showed that the distance method greatly outperformed the method of cointegration in terms of trading opportunities, the share of trades with negative return and consistency of pairs choice. We can state that the distance method is a more appropriate technique in the context of the PSE. Nevertheless, the selection of strategy is a Sophie's choice, for in both cases it would be better to simply hold the PX index.

This thesis is a unique evidence of pairs trading in the PSE. Still, it omits some considerations that, if implemented, would considerably improve the meaningfulness of results. The recommendation for future research includes employing risk control measures such as stop-loss at 20% of position value or maximum holding period, incorporating short-selling costs and expanding the number of sample periods. These could notably alter the return distributions and would reflect more realistic estimation of performance. The improvement towards better results is, however, questionable.

References

Bookstaber, R. M. 2007. *A demon of our own design.* Hoboken, N.J.: J. Wiley.

Broussard, J. & M. Vaihekoski. 2012. Profitability of pairs trading strategy in an illiquid market with multiple share classes. *Journal of International Financial Markets, Institutions and Money* vol. 22.issue 5, pp. 1188–1201. URL: `http://linkinghub.elsevier.com/retrieve/pii/S1042443112000583`.

Chen, Z. & P. Knez. 1995. Measurement of market integration and arbitrage. *Review of Financial Studies* vol. 8.issue 2. URL: `EBSCOdatabase`.

Davidson, R. & J. G. MacKinnon. 1993. *Estimation and inference in econometrics.* Oxford University Press. ISBN: 0195060113.

De Bondt, W. & R. Thaler. 1985. Does the stock market overreact? *Journal of Finance* vol. 40.issue 3, pp. 793–805. URL: `EBSCOdatabase`.

Dickey, D. A. & W. A. Fuller. 1979. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association* vol. 74.issue 366, pp. 427–431. URL: `EBSCOdatabase`.

Do, B. & R. W. Faff. 2010. Does simple pairs trading still work? *Financial Analysts Journal* vol. 66.issue 4, pp. 83–95. URL: `EBSCOdatabase`.

— . 2012. Are pairs trading profits robust to trading costs? *The Journal of Financial Research* 35.2, pp. 261–287. URL: `http://ssrn.com/abstract=1707125`.

Elliott, R. J., J. V. D. Hoek, & W. P. Malcolm. 2005. Pairs trading. *Quantitative Finance* vol. 5.issue 3, pp. 271–276. URL: `http://www.tandfonline.com/doi/abs/10.1080/14697680500149370`.

Engle, R. F. & C. W. Granger. 1987. Cointegration and error correction - representation, estimation, and testing. *Econometrica* 55.2, pp. 251–276. ISSN: 0012-9682. DOI: 10.2307/1913236. URL: <GotoISI>://WOS:A1987G612400001.

Fama, E. F. 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance* vol. 25.issue 2, pp. 383–417. URL: `EBSCOdatabase`.

Fama, E. F. & K. R. French. 1988. Permanent and temporary components of stock prices. *Journal of Political Economy* vol. 96.issue 2, pp. 246–273. URL: `EBSCOdatabase`.

Gatev, E., W. N. Goetzmann, & K. G. Rouwenhorst. 1999. Pairs trading: performance of a relative value arbitrage rule. *Yale School of Management Working Papers* ysm3, pp. 1–34.

— . 2006. Pairs trading: performance of a relative value arbitrage rule. *Review of Financial Studies* vol. 19.issue 3, pp. 797–827. URL: EBSCOdatabase.

Granger, C. & P. Newbold. 1974. Spurious regressions in econometrics. *Journal of Econometrics* vol. 2.issue 2, pp. 111–120. URL: EBSCOdatabase.

Grinold, R. C. & R. N. Kahn. 2000. *Active portfolio management: a quantitative approach for producing superior returns and controlling risk.* Vol. 2. McGraw-Hill.

Hamilton, J. D. 1994. *Time series analysis.* Vol. 2. Princeton University Press.

Harris, L. 2002. *Trading and exchanges: market microstructure for practitioners.* Financial Management Association Survey and Synthesis Series. Oxford University Press, USA. ISBN: 9780199792702.

Huck, N. 2009. Pairs selection and outranking. *European Journal of Operational Research* vol. 196.issue 2, pp. 819–825. URL: http://linkinghub.elsevier.com/retrieve/pii/S0377221708003160.

Jegadeesh, N. 1990. Evidence of predictable behavior of security returns. *The Journal of Finance* 45.3, pp. 881–898. ISSN: 00221082. URL: http://www.jstor.org/stable/2328797.

Jegadeesh, N. & S. Titman. 1995a. Short-horizon return reversals and the bid-ask spread. *Journal of Financial Intermediation* vol. 4.issue 2, pp. 116–132. URL: ScienceDirect.

— . 1995b. Overreaction, delayed reaction, and contrarian profits. *The Review of Financial Studies* 8.4, pp. 973–993. URL: http://www.jstor.org/stable/2962296.

Lehmann, B. 1990. Fads, martingales, and market efficiency. *Quarterly Journal of Economics* vol. 105.issue 1, pp. 1–28. URL: EBSCOdatabase.

Liew, R. Q. & Y. Wu. 2013. Pairs trading. *Journal of Derivatives & Hedge Funds* vol. 19.issue 1, pp. 12–30. URL: EBSCOdatabase.

Lo, A. & A. MacKinlay. 1990. When are contrarian profits due to stock market overreaction? *Review of Financial Studies* vol. 3.issue 2, pp. 175–205. URL: EBSCOdatabase.

Poterba, J. & J. Summers. 1988. Mean reversion in stock prices. *Journal of Financial Economics* vol. 22.issue 1, pp. 27–59. URL: EBSCOdatabase.

Sharpe, W. F. 1966. Security prices, risk, and maximal gains from diversification. *Journal of Finance* vol. 21.issue 4, pp. 743–744. URL: EBSCOdatabase.

Treynor, J. L. & F. Black. 1973. How to use security analysis to improve portfolio selection. *The Journal of Business* 46.1, pp. 66–86. URL: http://EconPapers.repec.org/RePEc:ucp:jnlbus:v:46:y:1973:i:1:p:66-86.

Vidyamurthy, G. 2004. *Pairs trading: quantitative methods and analysis*. J. Wiley.

Appendix

Table A.1: Sample Periods

Sample	Formation Period (D/M/Y)	Points	Trading Period (D/M/Y)	Points	Stocks
1	6/6/2008-5/9/2009	315	6/9/2009-5/3/2010	123	Erste, KB, CEZ, Telefonica, Unipetrol, CETV, Orco, Philip Morris, Pegas, NWR, VIG, AAA, ECM
2	6/12/2008-5/3/2010	309	6/3/2010-5/9/2010	127	
3	6/6/2009-5/9/2010	314	6/9/2010-5/3/2011	125	
4	6/12/2009-5/3/2011	313	6/3/2011-5/9/2011	128	Erste, KB, CEZ, Telefonica, Unipetrol, CETV, Orco, Philip Morris, Pegas, NWR, VIG, AAA
5	6/6/2010-5/9/2011	316	6/9/2011-5/3/2012	126	Erste, KB, CEZ, Telefonica, Unipetrol, CETV, Orco, Philip Morris, Pegas, NWR, VIG, AAA, Kit Digital
6	6/12/2010-5/3/2012	317	6/3/2012-5/9/2012	127	Erste, KB, CEZ, Telefonica, Unipetrol, CETV, Orco, Philip Morris, Pegas, NWR, VIG, AAA, Kit Digital, Fortuna
7	6/6/2011-5/9/2012	317	6/9/2012-5/3/2013	123	Erste, KB, CEZ, Telefonica, Unipetrol, CETV, Orco, Philip Morris, Pegas, NWR, VIG, AAA, Fortuna
8	6/12/2011-5/3/2013	314	6/3/2013-5/9/2013	127	Erste, KB, CEZ, Telefonica, Unipetrol, CETV, Orco, Philip Morris, Pegas, NWR, VIG, Fortuna
9	6/6/2012-5/3/2014	314	6/9/2013-5/3/2014	123	

Table A.2: Formation and Trading Statistics

Period	Stocks	Pairs	Stationary	Cointegrated	Trades, COIN	Trades, DIST
1	13	78	0	5	3	4
2	13	78	1	9	6	6
3	13	78	0	1	1	7
4	12	66	0	1	2	6
5	13	78	0	2	4	8
6	14	91	0	9	8	3
7	13	78	1	2	3	7
8	12	66	1	1	1	5
9	12	66	1	0	-	6

Stocks = original number of stocks, without detractions due to stationarity; Pairs = original number of pairs, without detractions due to stationarity or $\beta < 0$; Stationary = stocks with unit-root rejection; Cointegrated = cointegrated pairs with positive beta; Trades, COIN = number of trades generated using the cointegration method; Trades, DIST = number of trades generated using the distance method

Figure A.1: Histogram of Cointegration-Trading Returns: One-Day Lag

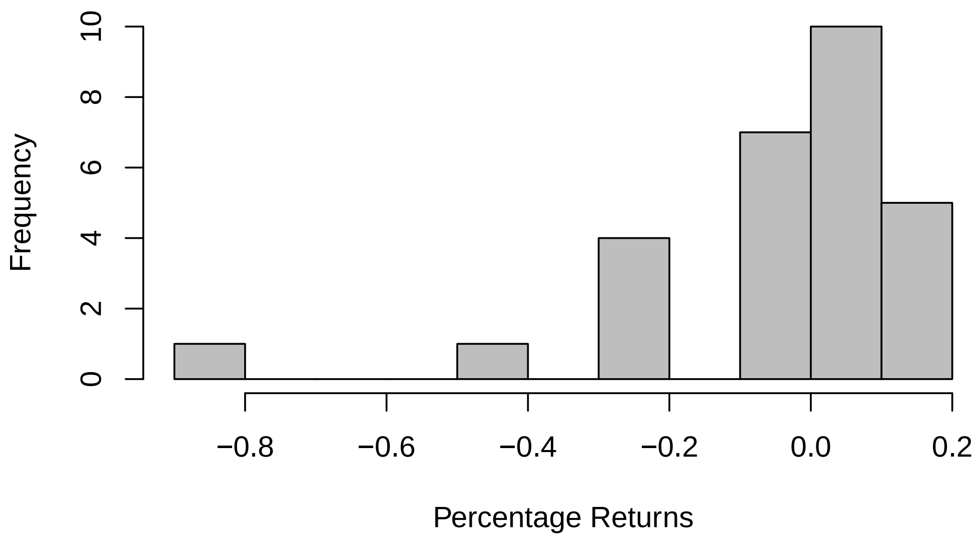


Table A.3: Return Distribution for Cointegration Trading: One-Day Lag vs. End-of-the-Day Execution

	One Day Waiting Rule	No Waiting
Average excess return*	-0.0517	-0.0527
Standard deviation	0.218	0.221
Average information ratio	0.0697	0.0607
Excess return distribution		
Median	0.0153	-0.004
Skewness	-2.579	-2.390
Excess kurtosis	8.023	6.975
Minimum	-0.896	-0.888
Maximum	0.135	0.141

* Per trade in a 6-month trading period

Figure A.2: Histogram of Distance-Trading Returns: Value-Weighted Approach, One-Day Lag

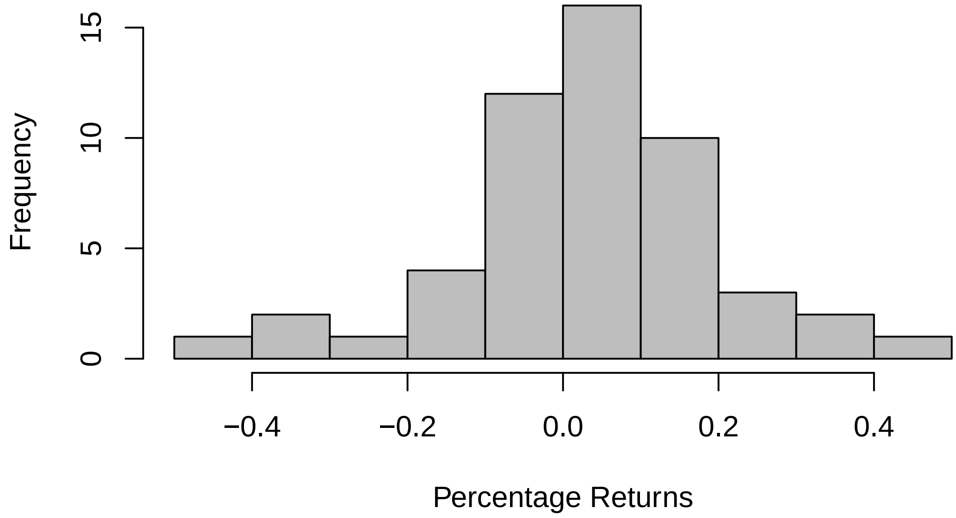


Table A.4: Return Distribution for Distance Trading: One-Day Lag vs. End-of-the-Day Execution; Value-Weighted Approach

	One Day Waiting Rule	No Waiting
Average excess return*	0.0293	0.0344
Standard deviation	0.165	0.159
Standard error	0.023	0.022
t-statistic	1.277	1.565
Average information ratio	0.1051	0.0875
Excess return distribution		
Median	0.0385	0.0241
Skewness	-0.344	-0.512
Excess kurtosis	1.439	1.055
Minimum	-0.409	-0.393
Maximum	0.431	0.368

* Per trade in a 6-month trading period

Figure A.3: Comparison of Return Histograms: One-Day Lag

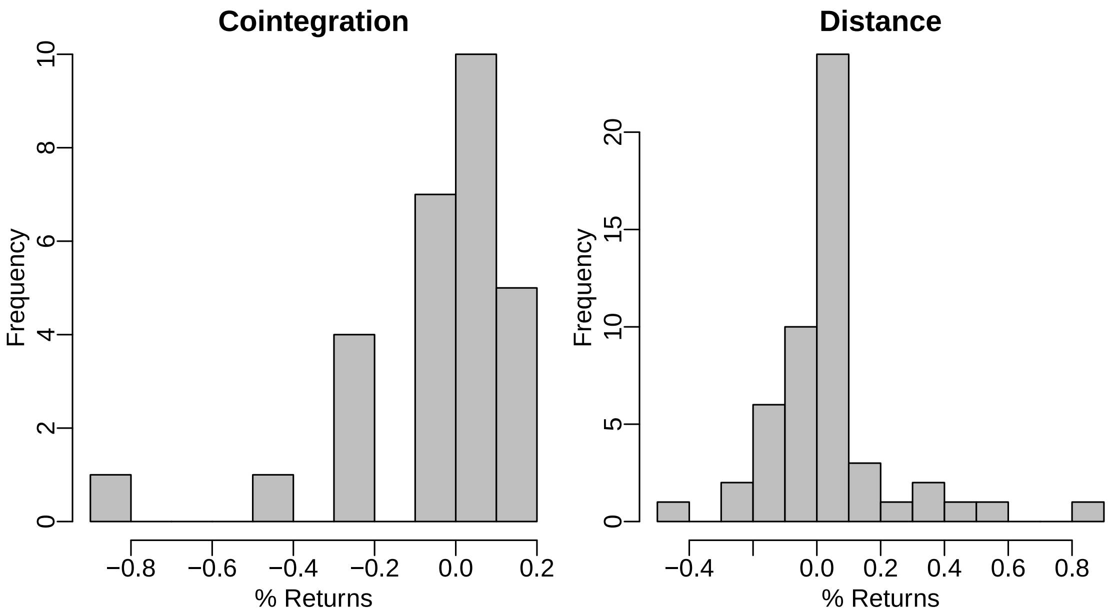


Table A.5: Comparison of Return Distributions for Cointegration and Distance Methods: One-Day Waiting Rule

	Cointegration	Distance
Average excess return	-0.0517	0.0372
Standard deviation	0.218	0.210
Average information ratio	0.0697	-0.0897
Excess return distribution		
Median	0.0153	0.0313
Skewness	-2.579	1.328
Excess kurtosis	8.023	5.114
Minimum	-0.896	-0.472
Maximum	0.135	0.843
Share of negative observations	46.4%	36.5%

Table A.6: Comparison of Distance Method with Analysis by Gatev et al. (2006)

		S&P500-Market Analysis by Gatev et al. (2006)*
Average number of pairs traded per 6-month period	4.56	4.81
Average time pairs are open in months	2.59	3.75
Average number of trades per pair	1.16	2.02
Average number of months per trade	2.23	1.86
t statistic	1.277	6.26
Excess return distribution		
Skewness	-0.344	0.34
Excess kurtosis	1.439	10.64
Minimum	-0.409	-0.126
Maximum	0.431	0.144
Observations with excess return <0	38%	35%

*measures for top 5 pairs

```
IS=lapply( paste("IS", 1:9, sep=''), get )
OOS=lapply( paste("OOS", 1:9, sep=''), get )

### COINTEGRATION METHOD ###

# FORMATION
# Unit root test for stationarity
mylist.names=c("ADF1", "ADF2", "ADF3", "ADF4", "ADF5", "ADF6", "ADF7", "ADF8", "ADF9")
ADF=sapply(mylist.names,function(x) NULL)
for(k in 1:length(IS)){
  for(i in 1:(ncol(IS[[k]])-1)){
    AA=ur.df(IS[[k]][,i+1], type=c("trend"), lags=5, selectlags=c("AIC"))
    ADF[[k]][i]=attributes(AA)$teststat[1]
  }
}

max.len=max(sapply(ADF, length))
corrected.list=lapply(ADF, function(x) {c(x, rep(0, max.len - length(x)))})
ADF=do.call(rbind, corrected.list)

# Removing stock for which unit root hypothesis was rejected
for(k in 1:nrow(ADF)){
  for(i in 1:ncol(ADF)){
    if(ADF[k,i]<(-3.41)){
      IS[[k]][,which(ADF[k,]<(-3.41))+1]=NULL
      OOS[[k]][,which(ADF[k,]<(-3.41))+1]=NULL
    }
  }}

# Residual vectors
resid1=matrix(nrow=nrow(IS1),ncol=sum(1:(ncol(IS1)-2)))
resid2=matrix(nrow=nrow(IS2),ncol=sum(1:(ncol(IS2)-2)))
resid3=matrix(nrow=nrow(IS3),ncol=sum(1:(ncol(IS3)-2)))
resid4=matrix(nrow=nrow(IS4),ncol=sum(1:(ncol(IS4)-2)))
resid5=matrix(nrow=nrow(IS5),ncol=sum(1:(ncol(IS5)-2)))
resid6=matrix(nrow=nrow(IS6),ncol=sum(1:(ncol(IS6)-2)))
resid7=matrix(nrow=nrow(IS7),ncol=sum(1:(ncol(IS7)-2)))
resid8=matrix(nrow=nrow(IS8),ncol=sum(1:(ncol(IS8)-2)))
resid9=matrix(nrow=nrow(IS9),ncol=sum(1:(ncol(IS9)-2)))
resid=lapply( paste("resid", 1:9, sep=''), get )
mju=matrix(nrow=length(IS), ncol=sum(1:(max(sapply(IS,ncol))-2)))

for(m in 1:length(IS)){
  ind=1
  for(j in 1:(ncol(IS[[m]])-1)){
    for(i in (j+1):(ncol(IS[[m]])-1)){
      if((i!=ncol(IS[[m]]))&((i!=ncol(IS[[m]])-1)|(j!=ncol(IS[[m]])-1))){
        resid[[m]][,ind]=residuals(lm(IS[[m]][,j+1]~IS[[m]][,i+1]))
        mju[m,ind]=lm(IS[[m]][,j+1]~IS[[m]][,i+1])$coeff[1]
        ind=ind+1
      }
    }}}

# Cointegration test: AEG on residual vectors
mylist.names=c("TRIADF1", "TRIADF2", "TRIADF3", "TRIADF4", "TRIADF5", "TRIADF6", "TRIADF7", "TRIADF8",
      "TRIADF9")
TRIADF=sapply(mylist.names,function(x) NULL)
for(m in 1:length(IS)){
  for(i in 1:sum(1:(ncol(IS[[m]])-2))){
    TRI=ur.df(resid[[m]][,i], type=c("drift"), lags=5, selectlags=c("AIC"))
    TRIADF[[m]][i]=attributes(TRI)$teststat[1]
  }
}

# Finding number of cointegrated pairs
max.len=max(sapply(TRIADF, length))
```

```
corrected.list=lapply(TRIADF, function(x) {c(x, rep(0, max.len - length(x)))})
TRIADF=do.call(rbind, corrected.list)

icount=NULL
for(m in 1:length(IS)){
  icount[[m]]=0
  for(i in 1:ncol(TRIADF)){
    if(TRIADF[m,i]<(-3.37)){
      icount[[m]]=icount[[m]]+1
    }
  }}
icount

# Removing sample periods with no cointegrated pairs
ind=0
for(m in 1:length(icount)){
  if(icount[m]==0){
    IS[[m-ind]]=NULL
    OOS[[m-ind]]=NULL
    TRIADF=TRIADF[-m+ind,]
    ind=ind+1
  }
}
icount=icount[icount!=0]

# Cointegrated pairs, ordered from lowest t-statistic
ndx=NULL
for(m in 1:length(IS)){
  ndx[[m]]=order(TRIADF[m,])[1:icount[[m]]]
}

# TRADING
# Cointegration coefficient (ratio in which to trade stocks)
beta=matrix(nrow=length(IS), ncol=sum(1:(max(sapply(IS,ncol))-2)))
for(m in 1:length(IS)){
  ind=1
  for(j in 1:(ncol(IS[[m]])-1)){
    for(i in (j+1):(ncol(IS[[m]])-1)){
      if((i!=ncol(IS[[m]]))&((i!=ncol(IS[[m]])-1)|(j!=ncol(IS[[m]])-1))){
        beta[m,ind]=lm(IS[[m]][,j+1]~IS[[m]][,i+1])$coeff[2]
        ind=ind+1
      }
    }}}

# Formation of trading-period spreads
regoos1=matrix(nrow=nrow(OOS[[1]]),ncol=sum(1:(ncol(OOS[[1]])-2)))
regoos2=matrix(nrow=nrow(OOS[[2]]),ncol=sum(1:(ncol(OOS[[2]])-2)))
regoos3=matrix(nrow=nrow(OOS[[3]]),ncol=sum(1:(ncol(OOS[[3]])-2)))
regoos4=matrix(nrow=nrow(OOS[[4]]),ncol=sum(1:(ncol(OOS[[4]])-2)))
regoos5=matrix(nrow=nrow(OOS[[5]]),ncol=sum(1:(ncol(OOS[[5]])-2)))
regoos6=matrix(nrow=nrow(OOS[[6]]),ncol=sum(1:(ncol(OOS[[6]])-2)))
regoos7=matrix(nrow=nrow(OOS[[7]]),ncol=sum(1:(ncol(OOS[[7]])-2)))
regoos8=matrix(nrow=nrow(OOS[[8]]),ncol=sum(1:(ncol(OOS[[8]])-2)))
regoos=lapply( paste("regoos", 1:8, sep=''), get )

for(m in 1:length(regoos)){
  ind=1
  for(j in 1:(ncol(OOS[[m]])-1)){
    for(i in (j+1):(ncol(OOS[[m]])-1)){
      if((i!=ncol(OOS[[m]]))&((i!=ncol(OOS[[m]])-1)|(j!=ncol(OOS[[m]])-1))){
        regoos[[m]][,ind]=OOS[[m]][,j+1]-beta[m,ind]*OOS[[m]][,i+1]
        ind=ind+1
      }
    }}
  regoos[[m]]=regoos[[m]][,ndx[[m]]]
  regoos[[m]]=as.matrix(regoos[[m]])}
```

```
# Means and standard deviations
model1=matrix(nrow=nrow(IS1),ncol=sum(1:(ncol(IS1)-2)))
model2=matrix(nrow=nrow(IS2),ncol=sum(1:(ncol(IS2)-2)))
model3=matrix(nrow=nrow(IS3),ncol=sum(1:(ncol(IS3)-2)))
model4=matrix(nrow=nrow(IS4),ncol=sum(1:(ncol(IS4)-2)))
model5=matrix(nrow=nrow(IS5),ncol=sum(1:(ncol(IS5)-2)))
model6=matrix(nrow=nrow(IS6),ncol=sum(1:(ncol(IS6)-2)))
model7=matrix(nrow=nrow(IS7),ncol=sum(1:(ncol(IS7)-2)))
model8=matrix(nrow=nrow(IS8),ncol=sum(1:(ncol(IS8)-2)))
model=lapply(paste("model", 1:8, sep=''), get)

for(m in 1:length(model)){
  ind=1
  for(j in 1:sum(1:(ncol(IS[[m]])-2))){
    for(k in 1:nrow(resid[[m]])){
      model[[m]][k,j]=resid[[m]][k,j] +mju[m,j]
      ind=ind+1
    }}
  model[[m]]=model[[m]][,ndx[[m]]]
  model[[m]]=as.matrix(model[[m]])}

mylist.names=c("mean1", "mean2", "mean3", "mean4", "mean5", "mean6", "mean7", "mean8")
mean=sapply(mylist.names,function(x) NULL)
mylist.names=c("sd1", "sd2", "sd3", "sd4", "sd5", "sd6", "sd7", "sd8")
sd=sapply(mylist.names,function(x) NULL)

for(m in 1:length(model)){
  mean[[m]]=colMeans(model[[m]])
  sd[[m]]=colStdevs(model[[m]])
}

max.len=max(sapply(mean, length))
corrected.list=lapply(mean, function(x) {c(x, rep(0, max.len - length(x)))})
mean=do.call(rbind, corrected.list)
max.len=max(sapply(sd, length))
corrected.list=lapply(sd, function(x) {c(x, rep(0, max.len - length(x)))})
sd=do.call(rbind, corrected.list)

# Trading signals
a=c(1,1,1,1)
b=c(-2,0,2,0)
Thresholds=array(rep(NA), dim=c(nrow(mean),ncol(mean),length(a)))
for(k in 1:length(a)){
  for(i in 1:nrow(mean)){
    for(j in 1:ncol(mean)){
      Thresholds[i,j,k]=a[k]*mean[i,j]+b[k]*sd[i,j]
    }
  }}

# Points of entry/exit: finding rows where spreads are below or above trading signals
iakcie=NULL
rtime=NULL
for(m in 1:length(regoos)){
  iakcie[m]=ncol(regoos[[m]]); rtime[m]=nrow(regoos[[m]])
}
indBuyL=NULL; indSellL=NULL; indSellS=NULL; indBuyS=NULL

for(m in 1:length(regoos)){
  indBuyL[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  indBuyL[[m]][is.na(indBuyL[[m]])]=0
  indSellL[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  indSellL[[m]][is.na(indSellL[[m]])]=0
  indSellS[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  indSellS[[m]][is.na(indSellS[[m]])]=0
  indBuyS[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  indBuyS[[m]][is.na(indBuyS[[m]])]=0
}
```

```
for(m in 1:length(regoos)){
  for(i in 1:iakcie[m]){
    ind1=1;ind2=1;ind3=1;ind4=1
    for(j in 1:rtime[m]){
      if(regoos[[m]][j,i]<=Thresholds[m,i,1]){indBuyL[[m]][ind1,i]=j
                                               ind1=ind1+1}
      if(regoos[[m]][j,i]>=Thresholds[m,i,2]) {indSellL[[m]][ind2,i]=j
                                                ind2=ind2+1}
      if(regoos[[m]][j,i]>=Thresholds[m,i,3]) {indSellS[[m]][ind3,i]=j
                                                ind3=ind3+1}
      if(regoos[[m]][j,i]<=Thresholds[m,i,4]) {indBuyS[[m]][ind4,i]=j
                                                ind4=ind4+1}
    }
  }}

### Calculating profits
## In each time
# Long and short spreads
profitbottomup1=matrix(nrow=nrow(OOS[[1]])-1,ncol=ncol(regoos[[1]]))
profitbottomup2=matrix(nrow=nrow(OOS[[2]])-1,ncol=ncol(regoos[[2]]))
profitbottomup3=matrix(nrow=nrow(OOS[[3]])-1,ncol=ncol(regoos[[3]]))
profitbottomup4=matrix(nrow=nrow(OOS[[4]])-1,ncol=ncol(regoos[[4]]))
profitbottomup5=matrix(nrow=nrow(OOS[[5]])-1,ncol=ncol(regoos[[5]]))
profitbottomup6=matrix(nrow=nrow(OOS[[6]])-1,ncol=ncol(regoos[[6]]))
profitbottomup7=matrix(nrow=nrow(OOS[[7]])-1,ncol=ncol(regoos[[7]]))
profitbottomup8=matrix(nrow=nrow(OOS[[8]])-1,ncol=ncol(regoos[[8]]))
profitbottomup=lapply( paste("profitbottomup", 1:8, sep=''), get )

profitupbottom1=matrix(nrow=nrow(OOS[[1]])-1,ncol=ncol(regoos[[1]]))
profitupbottom2=matrix(nrow=nrow(OOS[[2]])-1,ncol=ncol(regoos[[2]]))
profitupbottom3=matrix(nrow=nrow(OOS[[3]])-1,ncol=ncol(regoos[[3]]))
profitupbottom4=matrix(nrow=nrow(OOS[[4]])-1,ncol=ncol(regoos[[4]]))
profitupbottom5=matrix(nrow=nrow(OOS[[5]])-1,ncol=ncol(regoos[[5]]))
profitupbottom6=matrix(nrow=nrow(OOS[[6]])-1,ncol=ncol(regoos[[6]]))
profitupbottom7=matrix(nrow=nrow(OOS[[7]])-1,ncol=ncol(regoos[[7]]))
profitupbottom8=matrix(nrow=nrow(OOS[[8]])-1,ncol=ncol(regoos[[8]]))
profitupbottom=lapply( paste("profitupbottom", 1:8, sep=''), get )

for(m in 1:length(profitbottomup)){
  for(i in 1:ncol(regoos[[m]])){
    profitbottomup[[m]][,i]=diff(regoos[[m]][,i])
    profitupbottom[[m]][,i]=-diff(regoos[[m]][,i])
  }
  null=rep(0, length=ncol(regoos[[m]]))
  profitbottomup[[m]]=rbind(null, profitbottomup[[m]])
  profitupbottom[[m]]=rbind(null, profitupbottom[[m]])
}


## For each particular trade
PXOOS=lapply( paste("PXOOS", 1:length(IS), sep=''), get )
# Long
buyL=NULL
sellL=NULL
profitL=NULL
profitPXL=NULL
for(m in 1:length(regoos)){
  buyL[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  sellL[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  profitL[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  profitPXL[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
}

for(m in 1:length(regoos)){
  for(i in 1:iakcie[m]){
    ind=1
```

```
    isellL=0
    for(j in 1:rtime[m]){
      if(indBuyL[[m]][j,i]>isellL){
        for(k in 1:rtime[m]){
          if((indSellL[[m]][k,i]>=indBuyL[[m]][j,i])&(indBuyL[[m]][j,i]!=rtime[m])){
            ibuyL=indBuyL[[m]][j,i]+1
            isellL=indSellL[[m]][k,i]+1
            buyL[[m]][ind,i]=indBuyL[[m]][j,i]+1
            sellL[[m]][ind,i]=indSellL[[m]][k,i]+1
            profitL[[m]][ind,i]=sum(profitbottomup[[m]][(buyL[[m]][ind,i]+1):sellL[[m]][ind,i],i])
            profitPXL[[m]][ind,i]=sum(PXOOS[[m]][(buyL[[m]][ind,i]+1):sellL[[m]][ind,i],3])
            ind=ind+1
            break
          }
          if((k==rtime[m])&(indBuyL[[m]][j,i]!=rtime[m])){
            ibuyL=indBuyL[[m]][j,i]+1
            isellL=rtime[m]
            buyL[[m]][ind,i]=indBuyL[[m]][j,i]+1
            sellL[[m]][ind,i]=rtime[m]
            profitL[[m]][ind,i]=sum(profitbottomup[[m]][(buyL[[m]][ind,i]+1):sellL[[m]][ind,i],i])
            profitPXL[[m]][ind,i]=sum(PXOOS[[m]][(buyL[[m]][ind,i]+1):sellL[[m]][ind,i],3])
            ind=ind+1
          }
        }
      }}}
  buyL[[m]][is.na(buyL[[m]])]=0
  sellL[[m]][is.na(sellL[[m]])]=0
  buyL[[m]]=as.matrix(buyL[[m]])
  sellL[[m]]=as.matrix(sellL[[m]])
}

# Information ratio, long
IRL=array(rep(NA), dim=c(4,max(sapply(buyL, ncol)),length(buyL)))
for(m in 1:length(regoos)){
  for(i in 1:ncol(buyL[[m]])){
    for(k in 1:nrow(buyL[[m]])){
      if(buyL[[m]][k,i]!=0){
        IRL[k,i,m]=mean(profitbottomup[[m]][(buyL[[m]][k,i]+1):sellL[[m]][k,i],i]-PXOOS[[m]][(buyL[[m
            ]][k,i]+1):sellL[[m]][k,i],3])/sd(profitbottomup[[m]][(buyL[[m]][k,i]+1):sellL[[m]][k,i],
            i]-PXOOS[[m]][(buyL[[m]][k,i]+1):sellL[[m]][k,i],3])
      }
    }}}
IRL

# Short
sellS=NULL
buyS=NULL
profitS=NULL
profitPXS=NULL
for(m in 1:length(regoos)){
  sellS[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  buyS[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  profitS[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  profitPXS[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
}

for(m in 1:length(regoos)){
  for(i in 1:iakcie[m]){
    ind=1
    ibuyS=0
    for(j in 1:rtime[m]){
      if(indSellS[[m]][j,i]>ibuyS){
        for(k in 1:rtime[m]){
          if((indBuyS[[m]][k,i]>=indSellS[[m]][j,i])&(indSellS[[m]][j,i]!=rtime[m])){
            isellS=indSellS[[m]][j,i]+1
            ibuyS=indBuyS[[m]][k,i]+1
            sellS[[m]][ind,i]=indSellS[[m]][j,i]+1
```

```
            buyS[[m]][ind,i]=indBuyS[[m]][k,i]+1
            profitS[[m]][ind,i]=sum(profitupbottom[[m]][(sellS[[m]][ind,i]+1):buyS[[m]][ind,i],i])
            profitPXS[[m]][ind,i]=sum(PXOOS[[m]][(sellS[[m]][ind,i]+1):buyS[[m]][ind,i],3])
            ind=ind+1
            break
          }
          if((k==rtime[m])&(indSellS[[m]][j,i]!=rtime[m])){
            isellS=indSellS[[m]][j,i]+1
            ibuyS=rtime[m]
            sellS[[m]][ind,i]=indSellS[[m]][j,i]+1
            buyS[[m]][ind,i]=ibuyS
            profitS[[m]][ind,i]=sum(profitupbottom[[m]][(sellS[[m]][ind,i]+1):buyS[[m]][ind,i],i])
            profitPXS[[m]][ind,i]=sum(PXOOS[[m]][(sellS[[m]][ind,i]+1):buyS[[m]][ind,i],3])
            ind=ind+1
          }
        }
      }}}
  sellS[[m]][is.na(sellS[[m]])]=0
  buyS[[m]][is.na(buyS[[m]])]=0
  sellS[[m]]=as.matrix(sellS[[m]])
  buyS[[m]]=as.matrix(buyS[[m]])
}

# Information ratio, short
IRS=array(rep(NA), dim=c(4,max(sapply(sellS, ncol)),length(sellS)))
for(m in 1:length(regoos)){
  for(i in 1:ncol(sellS[[m]])){
    for(k in 1:nrow(sellS[[m]])){
      if(sellS[[m]][k,i]!=0){
        IRS[k,i,m]=mean(profitupbottom[[m]][(sellS[[m]][k,i]+1):buyS[[m]][k,i],i]-PXOOS[[m]][(sellS[[m]][k,i]+1):buyS[[m]][k,i],3])/sd(profitupbottom[[m]][(sellS[[m]][k,i]+1):buyS[[m]][k,i],i]-PXOOS[[m]][(sellS[[m]][k,i]+1):buyS[[m]][k,i],3])
      }
    }}}
IRS

### DISTANCE METHOD ###

# Logarithms of prices
IS=lapply(paste("IS", 1:9, sep=''), get)
OOS=lapply(paste("OOS", 1:9, sep=''), get)

# Absolute values of prices
PIS=lapply(paste("PIS", 1:9, sep=''), get)
POOS=lapply(paste("POOS", 1:9, sep=''), get)

ISa=list(); OOSa=list(); PISa=list(); POOSa=list()
for(m in 1:length(IS)){
  ISa[[m]]=IS[[m]][,-1]
  OOSa[[m]]=OOS[[m]][,-1]
  PISa[[m]]=PIS[[m]][,-1]
  POOSa[[m]]=POOS[[m]][,-1]
}

# FORMATION
ret1=matrix(nrow=nrow(IS[[1]])-1,ncol=ncol(ISa[[1]]))
ret2=matrix(nrow=nrow(IS[[2]])-1,ncol=ncol(ISa[[2]]))
ret3=matrix(nrow=nrow(IS[[3]])-1,ncol=ncol(ISa[[3]]))
ret4=matrix(nrow=nrow(IS[[4]])-1,ncol=ncol(ISa[[4]]))
ret5=matrix(nrow=nrow(IS[[5]])-1,ncol=ncol(ISa[[5]]))
ret6=matrix(nrow=nrow(IS[[6]])-1,ncol=ncol(ISa[[6]]))
ret7=matrix(nrow=nrow(IS[[7]])-1,ncol=ncol(ISa[[7]]))
ret8=matrix(nrow=nrow(IS[[8]])-1,ncol=ncol(ISa[[8]]))
ret9=matrix(nrow=nrow(IS[[9]])-1,ncol=ncol(ISa[[9]]))
ret=lapply( paste("ret", 1:9, sep=''), get )

for(m in 1:length(ISa)){
```

```
  for(i in 1:ncol(ISa[[m]])){
    ret[[m]][,i]=diff(ISa[[m]][,i])
  }
  null=rep(0, length=ncol(ISa[[m]]))
  ret[[m]]=rbind(null, ret[[m]])
}

mylist.names=c("meanD1", "meanD2", "meanD3", "meanD4", "meanD5", "meanD6", "meanD7", "meanD8")
meanD=sapply(mylist.names,function(x) NULL)
mylist.names=c("sdD1", "sdD2", "sdD3", "sdD4", "sdD5", "sdD6", "sdD7", "sdD8")
sdD=sapply(mylist.names,function(x) NULL)

for(m in 1:length(ret)){
  meanD[[m]]=colMeans(ret[[m]])
  sdD[[m]]=colStdevs(ret[[m]])
}

# Cumulative sum of standardized returns
retN1=matrix(nrow=nrow(ret[[1]]),ncol=ncol(ret[[1]]))
retN2=matrix(nrow=nrow(ret[[2]]),ncol=ncol(ret[[2]]))
retN3=matrix(nrow=nrow(ret[[3]]),ncol=ncol(ret[[3]]))
retN4=matrix(nrow=nrow(ret[[4]]),ncol=ncol(ret[[4]]))
retN5=matrix(nrow=nrow(ret[[5]]),ncol=ncol(ret[[5]]))
retN6=matrix(nrow=nrow(ret[[6]]),ncol=ncol(ret[[6]]))
retN7=matrix(nrow=nrow(ret[[7]]),ncol=ncol(ret[[7]]))
retN8=matrix(nrow=nrow(ret[[8]]),ncol=ncol(ret[[8]]))
retN9=matrix(nrow=nrow(ret[[9]]),ncol=ncol(ret[[9]]))

retN=lapply( paste("retN", 1:9, sep=''), get )
for(m in 1:length(ret)){
  for(i in 1:ncol(ret[[m]])){
    for(k in 1:nrow(ret[[m]])){
      retN[[m]][k,i]=(ret[[m]][k,i]-meanD[[m]][i])/sdD[[m]][i]
    }
    retN[[m]][,i]=cumsum(retN[[m]][,i])
  }}

# Sum of squared deviations of normalized returns
devN=matrix(nrow=length(retN), ncol=sum(1:(max(sapply(retN, ncol))-1)))
for(m in 1:length(retN)){
  ind=1
  for(j in 1:(ncol(retN[[m]])-1)){
    for(i in (j+1):ncol(retN[[m]])){
      devN[m,ind]=sum((retN[[m]][,j]-retN[[m]][,i])^2)
      ind=ind+1
    }
  }}

# Pairs with minimized distance, ordered from the lowest SSD
bdx=NULL
for(m in 1:nrow(devN)){
  bdx[[m]]=order(devN[m,])[1:5]
  bdx[[m]]=as.matrix(bdx[[m]])
}

# TRADING
# Retrieving stocks from pairs
stock1=array(rep(0), dim=c(1,max(sapply(bdx, length)),length(ISa)))
stock2=array(rep(0), dim=c(1,max(sapply(bdx, length)),length(ISa)))
for(m in 1:length(ISa)){
  for(n in 1:length(bdx[[m]])){
    ind=1
    for(j in 1:(ncol(ISa[[m]])-1)){
      for(i in (j+1):ncol(ISa[[m]])){
        if(bdx[[m]][n]==ind){
          stock1[1,n,m]=j
          stock2[1,n,m]=i
```

```
      }
      ind=ind+1
    }
  }}}

# Standardization of OOS returns
retoos1=matrix(nrow=nrow(OOSa[[1]])-1,ncol=ncol(OOSa[[1]]))
retoos2=matrix(nrow=nrow(OOSa[[2]])-1,ncol=ncol(OOSa[[2]]))
retoos3=matrix(nrow=nrow(OOSa[[3]])-1,ncol=ncol(OOSa[[3]]))
retoos4=matrix(nrow=nrow(OOSa[[4]])-1,ncol=ncol(OOSa[[4]]))
retoos5=matrix(nrow=nrow(OOSa[[5]])-1,ncol=ncol(OOSa[[5]]))
retoos6=matrix(nrow=nrow(OOSa[[6]])-1,ncol=ncol(OOSa[[6]]))
retoos7=matrix(nrow=nrow(OOSa[[7]])-1,ncol=ncol(OOSa[[7]]))
retoos8=matrix(nrow=nrow(OOSa[[8]])-1,ncol=ncol(OOSa[[8]]))
retoos9=matrix(nrow=nrow(OOSa[[9]])-1,ncol=ncol(OOSa[[9]]))
retoos=lapply( paste("retoos", 1:9, sep=''), get )

retoosw1=matrix(nrow=nrow(OOSa[[1]])-1,ncol=ncol(OOSa[[1]]))
retoosw2=matrix(nrow=nrow(OOSa[[2]])-1,ncol=ncol(OOSa[[2]]))
retoosw3=matrix(nrow=nrow(OOSa[[3]])-1,ncol=ncol(OOSa[[3]]))
retoosw4=matrix(nrow=nrow(OOSa[[4]])-1,ncol=ncol(OOSa[[4]]))
retoosw5=matrix(nrow=nrow(OOSa[[5]])-1,ncol=ncol(OOSa[[5]]))
retoosw6=matrix(nrow=nrow(OOSa[[6]])-1,ncol=ncol(OOSa[[6]]))
retoosw7=matrix(nrow=nrow(OOSa[[7]])-1,ncol=ncol(OOSa[[7]]))
retoosw8=matrix(nrow=nrow(OOSa[[8]])-1,ncol=ncol(OOSa[[8]]))
retoosw9=matrix(nrow=nrow(OOSa[[9]])-1,ncol=ncol(OOSa[[9]]))
retoosw=lapply( paste("retoosw", 1:9, sep=''), get )

for(m in 1:length(OOSa)){
  for(i in 1:ncol(OOSa[[m]])){
    retoos[[m]][,i]=diff(OOSa[[m]][,i])
  }
  null=rep(0, length=ncol(OOSa[[m]]))
  retoos[[m]]=rbind(null, retoos[[m]])
  retoosw[[m]]=retoos[[m]]+1
}

OOSN1=matrix(nrow=nrow(OOSa[[1]]),ncol=ncol(OOSa[[1]]))
OOSN2=matrix(nrow=nrow(OOSa[[2]]),ncol=ncol(OOSa[[2]]))
OOSN3=matrix(nrow=nrow(OOSa[[3]]),ncol=ncol(OOSa[[3]]))
OOSN4=matrix(nrow=nrow(OOSa[[4]]),ncol=ncol(OOSa[[4]]))
OOSN5=matrix(nrow=nrow(OOSa[[5]]),ncol=ncol(OOSa[[5]]))
OOSN6=matrix(nrow=nrow(OOSa[[6]]),ncol=ncol(OOSa[[6]]))
OOSN7=matrix(nrow=nrow(OOSa[[7]]),ncol=ncol(OOSa[[7]]))
OOSN8=matrix(nrow=nrow(OOSa[[8]]),ncol=ncol(OOSa[[8]]))
OOSN9=matrix(nrow=nrow(OOSa[[9]]),ncol=ncol(OOSa[[9]]))
OOSN=lapply( paste("OOSN", 1:9, sep=''), get )

for(m in 1:length(OOSa)){
  for(i in 1:ncol(OOSa[[m]])){
    for(k in 1:nrow(OOSa[[m]])){
      OOSN[[m]][k,i]=(retoos[[m]][k,i]-meanD[[m]][i])/sdD[[m]][i]
    }
    OOSN[[m]][,i]=cumsum(OOSN[[m]][,i])
  }}

# For chosen pairs, difference between normalized cumulative returns
dif1=matrix(nrow=nrow(OOSN[[1]]),ncol=nrow(bdx[[1]]))
dif2=matrix(nrow=nrow(OOSN[[2]]),ncol=nrow(bdx[[2]]))
dif3=matrix(nrow=nrow(OOSN[[3]]),ncol=nrow(bdx[[3]]))
dif4=matrix(nrow=nrow(OOSN[[4]]),ncol=nrow(bdx[[4]]))
dif5=matrix(nrow=nrow(OOSN[[5]]),ncol=nrow(bdx[[5]]))
dif6=matrix(nrow=nrow(OOSN[[6]]),ncol=nrow(bdx[[6]]))
dif7=matrix(nrow=nrow(OOSN[[7]]),ncol=nrow(bdx[[7]]))
dif8=matrix(nrow=nrow(OOSN[[8]]),ncol=nrow(bdx[[8]]))
dif9=matrix(nrow=nrow(OOSN[[9]]),ncol=nrow(bdx[[9]]))
dif=lapply( paste("dif", 1:9, sep=''), get )
```

```
for(m in 1:length(OOSN)){
  for(n in 1:length(stock1[1,,m])){
    if(stock1[1,n,m]!=0){
      dif[[m]][,n]=OOSN[[m]][,stock1[1,n,m]]-OOSN[[m]][,stock2[1,n,m]]
    }
  }
  dif[[m]]=as.matrix(dif[[m]])
}

# Trading signals
meanN=array(rep(NA), dim=c(1,ncol(stock1),length(retN)))
sdN=array(rep(NA), dim=c(1,ncol(stock1),length(retN)))
for(m in 1:length(retN)){
  for(n in 1:ncol(stock1)){
    if(stock1[1,n,m]!=0){
      meanN[1,n,m]=mean(retN[[m]][,stock1[1,n,m]]-retN[[m]][,stock2[1,n,m]])
      sdN[1,n,m]=sd(retN[[m]][,stock1[1,n,m]]-retN[[m]][,stock2[1,n,m]])
    }
  }}

a=c(1,1,1,1)
b=c(-2,0,2,0)
Thresholds=array(rep(NA), dim=c(length(OOSN),ncol(meanN),length(b)))
for(k in 1:length(b)){
  for(i in 1:length(OOSN)){
    for(j in 1:ncol(meanN)){
      Thresholds[i,j,k]=a[k]*meanN[1,j,i]+b[k]*sdN[1,j,i]
    }
  }}

# Points of entry/exit: finding rows where spreads are below or above trading signals
iakcie=NULL
rtime=NULL
for(m in 1:length(dif)){
  iakcie[m]=ncol(dif[[m]]); rtime[m]=nrow(dif[[m]])
}
indBuyL=NULL; indSellL=NULL; indSellS=NULL; indBuyS=NULL

for(m in 1:length(dif)){
  indBuyL[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  indBuyL[[m]][is.na(indBuyL[[m]])]=0
  indSellL[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  indSellL[[m]][is.na(indSellL[[m]])]=0
  indSellS[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  indSellS[[m]][is.na(indSellS[[m]])]=0
  indBuyS[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  indBuyS[[m]][is.na(indBuyS[[m]])]=0
}

for(m in 1:length(dif)){
  for(i in 1:iakcie[m]){
    ind1=1;ind2=1;ind3=1;ind4=1
    for(j in 1:rtime[m]){
      if(dif[[m]][j,i]<=Thresholds[m,i,1]){indBuyL[[m]][ind1,i]=j
                                          ind1=ind1+1}
      if(dif[[m]][j,i]>=Thresholds[m,i,2]) {indSellL[[m]][ind2,i]=j
                                           ind2=ind2+1}
      if(dif[[m]][j,i]>=Thresholds[m,i,3]) {indSellS[[m]][ind3,i]=j
                                           ind3=ind3+1}
      if(dif[[m]][j,i]<=Thresholds[m,i,4]) {indBuyS[[m]][ind4,i]=j
                                           ind4=ind4+1}
    }
  }}

### Calculating profits
PXOOS=lapply( paste("PXOOS", 1:length(IS), sep=''), get )
```

```
## Long ##
buyLD=NULL
sellLD=NULL
for(m in 1:length(dif)){
  buyLD[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
  sellLD[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
}

for(m in 1:length(dif)){
  for(i in 1:iakcie[m]){
    ind=1
    isellL=0
    for(j in 1:rtime[m]){
      if(indBuyL[[m]][j,i]>isellL){
        for(k in 1:rtime[m]){
          if((indSellL[[m]][k,i]>=indBuyL[[m]][j,i])&(indBuyL[[m]][j,i]!=rtime[m])){
            ibuyL=indBuyL[[m]][j,i]+1
            isellL=indSellL[[m]][k,i]+1
            buyLD[[m]][ind,i]=indBuyL[[m]][j,i]+1
            sellLD[[m]][ind,i]=indSellL[[m]][k,i]+1
            ind=ind+1
            break
          }
          if ((k==rtime[m])&(indBuyL[[m]][j,i]!=rtime[m])){
            ibuyL=indBuyL[[m]][j,i]+1
            isellL=rtime[m]
            buyLD[[m]][ind,i]=indBuyL[[m]][j,i]+1
            sellLD[[m]][ind,i]=rtime[m]
            ind=ind+1
          }
        }
      }}}
  buyLD[[m]][is.na(buyLD[[m]])]=0
  buyLD[[m]]=as.matrix(buyLD[[m]])
  sellLD[[m]][is.na(sellLD[[m]])]=0
  sellLD[[m]]=as.matrix(sellLD[[m]])
}

# Returns for each pair traded
weight11=matrix(nrow=nrow(retoos[[1]]),ncol=10)
weight12=matrix(nrow=nrow(retoos[[2]]),ncol=10)
weight13=matrix(nrow=nrow(retoos[[3]]),ncol=10)
weight14=matrix(nrow=nrow(retoos[[4]]),ncol=10)
weight15=matrix(nrow=nrow(retoos[[5]]),ncol=10)
weight16=matrix(nrow=nrow(retoos[[6]]),ncol=10)
weight17=matrix(nrow=nrow(retoos[[7]]),ncol=10)
weight18=matrix(nrow=nrow(retoos[[8]]),ncol=10)
weight19=matrix(nrow=nrow(retoos[[9]]),ncol=10)
weight1=lapply(paste("weight1", 1:9, sep=''), get)

weight21=matrix(nrow=nrow(retoos[[1]]),ncol=10)
weight22=matrix(nrow=nrow(retoos[[2]]),ncol=10)
weight23=matrix(nrow=nrow(retoos[[3]]),ncol=10)
weight24=matrix(nrow=nrow(retoos[[4]]),ncol=10)
weight25=matrix(nrow=nrow(retoos[[5]]),ncol=10)
weight26=matrix(nrow=nrow(retoos[[6]]),ncol=10)
weight27=matrix(nrow=nrow(retoos[[7]]),ncol=10)
weight28=matrix(nrow=nrow(retoos[[8]]),ncol=10)
weight29=matrix(nrow=nrow(retoos[[9]]),ncol=10)
weight2=lapply(paste("weight2", 1:9, sep=''), get)

for(m in 1:length(POOSa)){ind=1
  for(i in 1:ncol(buyLD[[m]])){
   for(j in 1:nrow(buyLD[[m]])){
   if((buyLD[[m]][j,i]!=0)&(buyLD[[m]][j,i]!=sellLD[[m]][j,i])&(buyLD[[m]][j,i]+1!=sellLD[[m]][j,i])&(
        buyLD[[m]][j,i]+2<sellLD[[m]][j,i])){
```

```
      weight1[[m]][1:((sellLD[[m]][j,i])-(buyLD[[m]][j,i]+1)),ind]=cumprod(retoosw[[m]][(buyLD[[m]][j,i]+2):(sellLD[[m]][j,i]),stock1[1,i,m]])
      weight2[[m]][1:((sellLD[[m]][j,i])-(buyLD[[m]][j,i]+1)),ind]=cumprod(retoosw[[m]][(buyLD[[m]][j,i]+2):(sellLD[[m]][j,i]),stock2[1,i,m]])
      ind=ind+1}else{
   if((buyLD[[m]][j,i]!=0)&(buyLD[[m]][j,i]==sellLD[[m]][j,i])){
      weight1[[m]][1,ind]=0
      weight2[[m]][1,ind]=0
      ind=ind+1}else{
   if((buyLD[[m]][j,i]!=0)&(buyLD[[m]][j,i]+1==sellLD[[m]][j,i])){
      weightS1[[m]][1,ind]=0
      weightS2[[m]][1,ind]=0
      ind=ind+1} }}
     }}
        unity=rep(1, length=ncol(weight1[[m]]))
        weight1[[m]]=rbind(unity, weight1[[m]])
        weight2[[m]]=rbind(unity, weight2[[m]])}

retL1=matrix(nrow=nrow(retoos[[1]]),ncol=4)
retL2=matrix(nrow=nrow(retoos[[2]]),ncol=4)
retL3=matrix(nrow=nrow(retoos[[3]]),ncol=4)
retL4=matrix(nrow=nrow(retoos[[4]]),ncol=4)
retL5=matrix(nrow=nrow(retoos[[5]]),ncol=4)
retL6=matrix(nrow=nrow(retoos[[6]]),ncol=4)
retL7=matrix(nrow=nrow(retoos[[7]]),ncol=4)
retL8=matrix(nrow=nrow(retoos[[8]]),ncol=4)
retL9=matrix(nrow=nrow(retoos[[9]]),ncol=4)
retL=lapply(paste("retL", 1:9, sep=''), get)

for(m in 1:length(buyLD)){ind=1
                          for(i in 1:ncol(buyLD[[m]])){
                            for(k in 1:nrow(buyLD[[m]])){
                              if(buyLD[[m]][k,i]!=0){
                                for(j in 1:((sellLD[[m]][k,i])-(buyLD[[m]][k,i]))){
                                  retL[[m]][j,ind]=weight1[[m]][j,ind]*retoos[[m]][j+buyLD[[m]][k,i],stock1[1,i,m]]-weight2[[m]][j,ind]*retoos[[m]][j+buyLD[[m]][k,i],stock2[1,i,m]]
                                }}else{break}
                              ind=ind+1
                            }}
                          retL[[m]][is.na(retL[[m]])]=0}

profitLD=matrix(ncol=max(sapply(buyLD,nnzero)), nrow=length(OOSa))
for(m in 1:length(OOSa)){
  for(i in 1:ncol(retL[[m]])){
    profitLD[m,i]=sum(retL[[m]][,i])
  }}

# Information ratio, long
IRLD=matrix(nrow=length(OOSa), ncol=max(sapply(buyLD,nnzero)))
for(m in 1:length(OOSa)){
  ind=1
  for(n in 1:ncol(buyLD[[m]])){
    for(i in 1:nrow(buyLD[[m]])){
      if(buyLD[[m]][i,n]!=0){
        IRLD[m,ind]=mean(retL[[m]][1:(sellLD[[m]][i,n]-buyLD[[m]][i,n]),ind]-PXOOS[[m]][(buyLD[[m]][i,n]+1):sellLD[[m]][i,n],3])/sd(retL[[m]][1:(sellLD[[m]][i,n]-buyLD[[m]][i,n]),ind]-PXOOS[[m]][(buyLD[[m]][i,n]+1):sellLD[[m]][i,n],3])
        ind=ind+1
      }else{break}
    }}}

## Short ##
sellSD=NULL
buySD=NULL
for(m in 1:length(dif)){
  sellSD[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
```

```
  buySD[[m]]=matrix(nrow=rtime[m],ncol=iakcie[m])
}

for(m in 1:length(dif)){
  for(i in 1:iakcie[m]){
    ind=1
    ibuyS=0
    for(j in 1:rtime[m]){
      if((indSellS[[m]][j,i]>ibuyS)&(indSellS[[m]][j,i]!=rtime[m])){
        for(k in 1:rtime[m]){
          if(indBuyS[[m]][k,i]>=indSellS[[m]][j,i]){
            isellS=indSellS[[m]][j,i]+1
            ibuyS=indBuyS[[m]][k,i]+1
            sellSD[[m]][ind,i]=indSellS[[m]][j,i]+1
            buySD[[m]][ind,i]=indBuyS[[m]][k,i]+1
            ind=ind+1
            break
          }
          if((k==rtime[m])&(indSellS[[m]][j,i]!=rtime[m])){
            isellS=indSellS[[m]][j,i]+1
            ibuyS=rtime[m]
            sellSD[[m]][ind,i]=indSellS[[m]][j,i]+1
            buySD[[m]][ind,i]=ibuyS
            ind=ind+1
          }
        }
      }}}
  sellSD[[m]][is.na(sellSD[[m]])]=0
  sellSD[[m]]=as.matrix(sellSD[[m]])
  buySD[[m]][is.na(buySD[[m]])]=0
  buySD[[m]]=as.matrix(buySD[[m]])
}

# Returns for each pair traded
weightS11=matrix(nrow=nrow(retoos[[1]]),ncol=10)
weightS12=matrix(nrow=nrow(retoos[[2]]),ncol=10)
weightS13=matrix(nrow=nrow(retoos[[3]]),ncol=10)
weightS14=matrix(nrow=nrow(retoos[[4]]),ncol=10)
weightS15=matrix(nrow=nrow(retoos[[5]]),ncol=10)
weightS16=matrix(nrow=nrow(retoos[[6]]),ncol=10)
weightS17=matrix(nrow=nrow(retoos[[7]]),ncol=10)
weightS18=matrix(nrow=nrow(retoos[[8]]),ncol=10)
weightS19=matrix(nrow=nrow(retoos[[9]]),ncol=10)
weightS1=lapply(paste("weightS1", 1:9, sep=''), get)

weightS21=matrix(nrow=nrow(retoos[[1]]),ncol=10)
weightS22=matrix(nrow=nrow(retoos[[2]]),ncol=10)
weightS23=matrix(nrow=nrow(retoos[[3]]),ncol=10)
weightS24=matrix(nrow=nrow(retoos[[4]]),ncol=10)
weightS25=matrix(nrow=nrow(retoos[[5]]),ncol=10)
weightS26=matrix(nrow=nrow(retoos[[6]]),ncol=10)
weightS27=matrix(nrow=nrow(retoos[[7]]),ncol=10)
weightS28=matrix(nrow=nrow(retoos[[8]]),ncol=10)
weightS29=matrix(nrow=nrow(retoos[[9]]),ncol=10)
weightS2=lapply(paste("weightS2", 1:9, sep=''), get)

for(m in 1:length(POOSa)){ind=1
 for(i in 1:ncol(sellSD[[m]])){
  for(j in 1:nrow(sellSD[[m]])){
  if((sellSD[[m]][j,i]!=0)&(sellSD[[m]][j,i]!=buySD[[m]][j,i])&(sellSD[[m]][j,i]+1!=buySD[[m]][j,i])&(
      sellSD[[m]][j,i]+2<buySD[[m]][j,i])){
     weightS1[[m]][1:((buySD[[m]][j,i])-(sellSD[[m]][j,i]+1)),ind]=cumprod(retoosw[[m]][(sellSD[[m]][j
          ,i]+2):(buySD[[m]][j,i]),stock1[1,i,m]])
     weightS2[[m]][1:((buySD[[m]][j,i])-(sellSD[[m]][j,i]+1)),ind]=cumprod(retoosw[[m]][(sellSD[[m]][j
          ,i]+2):(buySD[[m]][j,i]),stock2[1,i,m]])
   ind=ind+1}else{
   if((sellSD[[m]][j,i]!=0)&(sellSD[[m]][j,i]==buySD[[m]][j,i])){
```

```r
      weightS1[[m]][1,ind]=0
      weightS2[[m]][1,ind]=0
    ind=ind+1}else{
    if((sellSD[[m]][j,i]!=0)&(sellSD[[m]][j,i]+1==buySD[[m]][j,i])){
      weightS1[[m]][1,ind]=0
      weightS2[[m]][1,ind]=0
  ind=ind+1} }}
  }}
      unity=rep(1, length=ncol(weightS1[[m]]))
      weightS1[[m]]=rbind(unity, weightS1[[m]])
      weightS2[[m]]=rbind(unity, weightS2[[m]])}

retS1=matrix(nrow=nrow(retoos[[1]]),ncol=max(sapply(sellSD, nnzero)))
retS2=matrix(nrow=nrow(retoos[[2]]),ncol=ncol(sellSD[[2]]))
retS3=matrix(nrow=nrow(retoos[[3]]),ncol=ncol(sellSD[[3]]))
retS4=matrix(nrow=nrow(retoos[[4]]),ncol=ncol(sellSD[[4]]))
retS5=matrix(nrow=nrow(retoos[[5]]),ncol=ncol(sellSD[[5]]))
retS6=matrix(nrow=nrow(retoos[[6]]),ncol=ncol(sellSD[[6]]))
retS7=matrix(nrow=nrow(retoos[[7]]),ncol=ncol(sellSD[[7]]))
retS8=matrix(nrow=nrow(retoos[[8]]),ncol=ncol(sellSD[[8]]))
retS9=matrix(nrow=nrow(retoos[[9]]),ncol=ncol(sellSD[[9]]))
retS=lapply(paste("retS", 1:9, sep=''), get)

for(m in 1:length(sellSD)){
  ind=1
  for(i in 1:ncol(sellSD[[m]])){
    for(k in 1:nrow(sellSD[[m]])){
      if(sellSD[[m]][k,i]!=0){
        for(j in 1:((buySD[[m]][k,i])-(sellSD[[m]][k,i]))){
          retS[[m]][j,ind]=-(weightS1[[m]][j,ind]*retoos[[m]][j+sellSD[[m]][k,i],stock1[1,i,m]]-weightS2[[m]][j,ind]*retoos[[m]][j+sellSD[[m]][k,i],stock2[1,i,m]])
        }}else{break}
      ind=ind+1
    }}
  retS[[m]][is.na(retS[[m]])]=0}

profitSD=matrix(ncol=max(sapply(sellSD,nnzero)), nrow=length(OOSa))
for(m in 1:length(OOSa)){
  for(i in 1:ncol(retS[[m]])){
    profitSD[m,i]=sum(retS[[m]][,i])
  }}

# Information ratio, short
IRSD=matrix(nrow=length(OOSa), ncol=max(sapply(sellSD,nnzero)))
for(m in 1:length(OOSa)){
  ind=1
  for(n in 1:ncol(sellSD[[m]])){
    for(i in 1:nrow(sellSD[[m]])){
      if(sellSD[[m]][i,n]!=0){
        IRSD[m,ind]=mean(retS[[m]][1:(buySD[[m]][i,n]-sellSD[[m]][i,n]),ind]-PXOOS[[m]][(sellSD[[m]][i,n]+1):buySD[[m]][i,n],3])/sd(retS[[m]][1:(buySD[[m]][i,n]-sellSD[[m]][i,n]),ind]-PXOOS[[m]][(sellSD[[m]][i,n]+1):buySD[[m]][i,n],3])
        ind=ind+1
      }else{break}
    }}}
```

D:/RStudio/1daylag.R