Predicting the Price of EU ETS Carbon Credits: A Correlation, Principal Component and Latent Root Approach

by

Heiða Njóla Guðbrandsdóttir

Thesis
Master of Science
May 2011

Predicting the Price of EU ETS Carbon Credits:
A Correlation, Principal Component and Latent Root Approach

Heiða Njóla Guðbrandsdóttir

Thesis submitted to the School of Science and Engineering
at Reykjavík University in partial fulfillment
of the requirements for the degree of
Master of Science

May 2011

Supervisor:

Dr. Haraldur Óskar Haraldsson
Assistant Professor, Reykjavík University, Iceland

Examiner:

Dr. Hlynur Stefánsson
Assistant Professor and Director, Reykjavík Unitversity, Iceland

Predicting the Price of EU ETS Carbon Credits: A Correlation, Principal Component and Latent Root Approach
30 ECTS thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Financial Engineering

Department of Financial Engineering
School of Science and Engineering
Reykjavik University
Menntavegi 1
101 Reykjavik
Iceland

Telephone: +354 599 6200

Bibliographic information:
Heiða Njóla Guðbrandsdóttir, 2011, *Predicting the Price of EU ETS Carbon Credits. A Correlation, Principal Component and Latent Root Approach,* Master of Science Thesis, School of Science and Engineering, Reykjavik University, pp. 1-99.

Printing: Háskólaprent
Reykjavík, Iceland, May 2011

Abstract

The aim of this thesis is to examine what drives the changes in the price of carbon credits in the European Union's Emission Trading Scheme (EU ETS) and to make predictions based on these relationships. The study, which is based on the British energy market and global equity indices, starts with a large dataset which is reduced in dimension using correlation and principal component analysis. Predictions are then made by multiple linear regression, principal component regression and latent root regression. Correlation is the preferred dimension reduction technique to be followed by principal component regression. Certified emission reduction units (CERs) are shown to be the only same-day market relationship which provides useful predictions of European Union Allowance prices (EUAs), however this relationship is lost when data is lagged by one business day. No significant correlation is found between EUAs and the UK power market and the theoretical price of carbon credits; switching price, is shown to be a poor indicator of the price of carbon credits. The latent root model shows a notable performance out-of-sample, capturing the overall trend of EUAs over the prediction horizon.

Keywords: EU ETS, EUAs, Carbon Credits, Principal Component Analysis, Latent Root Regression

Útdráttur

Rannsókn þessi heitir á íslensku: *Verðþróun losunarheimilda á ETS-markaði Evrópusambandsins: Fylgni-, meginþátta- og eigingildagreining*. Markmiðið með rannsókninni er að kanna hvaða þættir hafa áhrif á verðbreytingar gengis losunarheimilda á markaði Evrópusambandsins (EU ETS) og byggja spálíkön á þeim samböndum. Rannsóknin, sem byggir á breskum orkumarkaði og alþjóðlegum hlutabréfavísitölum, notast við stórt gagnasafn sem síðan er minnkað með tveimur tölfræðilegum aðferðum: fylgnigreiningu og meginþáttagreiningu (e. Principal Component Analysis). Spálíkön eru svo byggð með margfaldri línulegri aðhvarfsgreiningu (e. Multiple Linear Regression), meginþátta-aðhvarfsgreiningu (e. Principal Component Regression) og eigingilda-aðhvarfsgreiningu (e. Latent Root Regression). Fylgnigreiningin reyndist besta aðferðin til að minnka vídd gagnanna fyrir samdægurssambönd. Svo skyldi beita meginþátta-aðhvarfsgreiningu á hið minnkaða gagnasafn. Gengi CERs hefur mikla samdægursfylgni við gengi losunarheimilda (e. EUAs), en þetta samband hverfur þegar gögnum er seinkað um einn dag. Enga fylgni mátti greina milli gengis losunarheimilda og gengis raforku á breskum markaði. Ennfremur reyndist hið fræðilega verð losunarheimilda (e. switching price) ekki góður mælikvarði á verðþróun losunarheimilda. Eigingilda-líkanið sýndi ágæta frammistöðu þegar það var prófað á nýjum gögnum, sem höfðu ekki verið notuð við afhvarfsgreiningu, þar sem líkanið nær að endurspegla heildarstefnu verðþróunar losunarheimildanna á spátímabilinu.

Efnisorð: EU ETS, losunarheimildir, græn vottorð, meginþáttagreining, eigingilda-aðhvarfsgreining

Predicting the Price of EU ETS Carbon Credits: A Correlation, Principal Component and Latent Root Approach

Heiða Njóla Guðbrandsdóttir

Thesis submitted to the School of Science and Engineering
at Reykjavík University in partial fulfillment
of the requirements for the degree of
Master of Science

May 2011

Student:

____

Heiða Njóla Guðbrandsdóttir

Supervisor:

____

Dr. Haraldur Óskar Haraldsson

Examiner:

____

Dr. Hlynur Stefánsson

Acknowledgement

I dearly thank my supervisor, Dr. Haraldur Óskar Haraldsson, Assistant Professor at the School of Science and Engineering at Reykjavik University, for his guidance during the course of this thesis and invaluable help gathering data as well as suggesting the topic. I would also like to thank my boyfriend, Hrannar Þór Hallgrímsson, for his patience and support. Last but not least I thank my parents, Guðbrandur Gimmel and Ingunn Þórðardóttir, for their helpful suggestions and my family and friends for words of encouragement.

Contents

List of Figures

Appendix

List of Tables

Appendix

List of Abbreviations

AAU	Assigned Amount Units
ACF	Autocorrelation Function
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
BIC	Bayesian Information Criterion
BL	Base Load
BLUE	Best Linear Unbiased Estimate
CDE	Carbon Dioxide Equivalent
CDS	Clean Dark Spread
CDM	Clean Development Mechanism
CER	Certified Emission Reduction Unit
CLRM	Classical Linear Regression Model
CSS	Clean Spark Spread
DS	Dark Spread
EIT	Economies in Transition
ERU	Emission Reduction Unit
EU	European Union
EU ETS	European Union Emission Trading Scheme
EUA	European Union Allowance Unit
GHG	Green House Gas
GWP	Global Warming Potential
IET	International Emissions Trading

JI	Joint Implementation
LRR	Latent Root Regression
MSE	Mean Squared Error
MWh	Megawatt Hour
NAP	National Allocation Plan
OECD	Organization for Economic Co-operation and Development
OLS	Ordinary Least-Squares
OTC	Over the Counter
PCA	Principal Component Analysis
PC	Principal Component
PCR	Principal Component Regression
PDF	Probability Density Function
PL	Peak Load
RGGI	Regional Green House Gas Initiative
RMU	Removable Unit
SS	Spark Spread
SVD	Singular Value Decomposition
UNFCCC	United Nations Framework Convention on Climate Change

1 Introduction

The increased concentration of green house gases in the atmosphere is now a generally accepted fact among scientists and politicians. More important is the consensus that the concentration is largely due to human activity and that the increased concentration is directly linked to the phenomena of global warming and climate change (Stern, 2007). Should no action be taken, the earth's average surface temperature could rise by 4°C by the year 2100, which could result in catastrophic effects.

To address this risk, countries of the world gathered for the Earth Summit in Rio de Janeiro in 1992 and agreed on an international treaty; the United Nations Framework Convention on Climate Change (UNFCCC). The aim of the treaty was to slow down green house gas emissions and the convention addressed important issues of climate change mitigation, new technology and promoting education. The treaty was however not legally binding so in 1997 an addition to the treaty was adopted, known as the Kyoto Protocol. According to the Kyoto Protocol, overall emissions will be reduced by five percent compared to 1990 levels over the period from 2008 to 2012, known as the Kyoto commitment period.

Following the Kyoto Protocol the European Union (EU) began preparing an EU carbon market to facilitate EU Member States to meet their commitments in a cost-effective way. On January 1st 2005 the European Union Emission Trading Scheme (EU ETS), the first international trading system of its kind, was launched. The EU ETS is a cap and trade system where overall emission levels are capped but members are free to buy or sell emission allowances as needed. It is currently the largest multi-country, multi-sector emission trading scheme in the world where trading in 2009 accumulated to over $ 118 billion.

Understanding the market and its key price-drivers is essential for managing large-scale investment choices as well as successfully planning existing operations, especially in the energy intensive industrial sector. As exact predictions on the market are virtually impossible, knowledge of what drives changes in the price of carbon credits is extremely valuable in constructing optimal hedging and investment strategies.

Although the theoretical foundation of carbon markets is widely acknowledged, empirical studies have only been published recently or are forthcoming. Several analyses have been carried out on the subject of carbon-price development.

Taschini & Paolella (2006) focused on the econometric modeling of the allowances. They conducted an analysis of the statistical distribution of emission trading allowances and constructed GARCH-models to address the tail behavior and heteroskedastic dynamics in the returns. Benz & Truck (2009) examined different phases of price and volatility behavior in the returns with the use of Markov switching and AR-GARCH models for stochastic modeling. The models were found to be effective in capturing short-term behavior. Daskalakis et al. (2009) compared three main markets under the EU ETS: Powernext, Nord Pool and ECX and concluded that spot prices were better approximated by Geometric Brownian motion augmented by jumps as the spot prices are likely to be characterized by jumps and non-stationarity. Uhrig-Homburg & Wagner (2009) examined the relationship between spot and futures markets in the EU ETS and concluded that futures markets lead the price discovery process of carbon credits. The above analyses are however based on data from phase I of the market environment, i.e. from 2005-2007.

Bataller et al. (2007) analyzed the effect of different factors on the price of allowances, including weather. They concluded that the most emission intensive energy sources were the principal factors in the determination of carbon prices and that only extreme temperatures could influence the prices. Alberola et al. (2008) came to a similar conclusion stating that EUA spot prices not only react to energy prices, but also to unanticipated temperature changes during colder events. In a Master’s thesis, Obermayer (2009) explored the relationship between carbon credits and German energy complex assets, including electrical power, coal, natural gas and oil. He found power to be the only significant correlation to EUAs and suggested further work on British data. Frunza et al. (2010) showed that energy, natural gas, oil, coal and equity indices acted as major factors in driving the carbon allowance prices. They then used an arbitrage pricing model via a hidden Markov chain model to predict futures prices and found the model to be effective both in and out-of-sample. Frunza et al. (2010) also briefly touched on the subject of principal component analysis. Finally Zhang & Wei (2010) summarized the main arguments of empirical studies on the EU ETS completed thus far.

To this date no analyses have been done focusing on British data, to the knowledge of this author, despite the fact that the United Kingdom is the second largest emitter of the OECD countries included in the EU ETS, after Germany (International Energy Agency, 2009). An in depth principal component analysis has also yet to be presented. Earlier studies have mainly focused on time-series models based on the EUA returns but fewer have examined market-relationships to EUA prices. A quantitative or semi-quantitative method of dimension reduction has not been examined with an application to the prediction of EUA prices.

The goal of this thesis is therefore to answer the following two questions:

1. What drives the changes in the price of emission allowances and can predictions be made based on that knowledge?
2. Considering a large-dimension dataset, what is the most effective way of reducing the dimension of the data to a manageable, yet useful size?

To address these topics and thereby the goals of this thesis a large dataset is collected, covering over thirty different variables based on British power market data, global equity indices and relevant currencies. The dimension of the data is reduced using two different methods: correlation and principal component analysis. In and out-of-sample predictions are then made using multiple linear regression, principal component regression and latent root regression.

The results show that the dimension reduction based on correlation and principal component analysis yield two completely different sets of variables. Correlation generates a dataset of equity indices along with CERs while the principal component analysis highlights dark spread, clean spark spread and WTI crude but switching price is rejected as a useful predictor of EUA returns. Dimension reduction based on correlation is the most effective in providing a strong same-day relationship able to capture over 80% of the variability of the data, but the relationship is lost when data is lagged by one business day. When testing out-of-sample none of the models is able to capture the variability of the returns, however the same-day and lag-1 day latent root models are able to capture the overall trend of EUA prices over the period of the prediction horizon.

British data do not seem to provide a useful relationship to EUA returns and British energy returns are weakly correlated to EUA returns, as opposed to German energy returns as shown by previous studies. The multiple linear regression and principal component regression prove to be useful tools for modeling if the reduced dataset is highly correlated to EUA returns.

The thesis is structured as follows: Chapter 2 gives the relevant background and development to this date of the international climate policy and the Kyoto Protocol as well as describing the fundamentals of the European Union's Emission Trading Scheme. Chapter 3 describes the theoretical framework and methods used to solve the problems presented above. Chapter 4 then examines the data used for the analysis and briefly describes the fundamental theory of the theoretical relationship between carbon price and the energy sector. In chapter 5 the main-results of the analysis are presented and in chapter 6 conclusions are drawn based on those results. The appendix then follows with more detailed results for the enthusiastic reader.

2 Background

The chapter describes the development of the international climate policy from which the European Union's Emission Trading Scheme has emerged. The first section covers the Kyoto protocol, followed by a section which explains the key-rules and structure of the EU ETS.

2.1 Climate Change and the International Climate Policy

The increased concentration of green house gases[1] (GHGs) in the atmosphere is now a generally accepted fact among scientists and politicians. More important is the consensus that the concentration is largely due to human activity and that the increased concentration is directly linked to the phenomena of global warming and climate change (Stern, 2007).

The earth's surface temperature has risen by 0.74°C on average since the dawn of the industrial revolution in the 18th and 19th century and green house gas emissions have risen by more than 30 percent (Labatt & White, 2007). A temperature increase of another 1.8 to 4 degrees is expected by the year 2100 assuming no action is taken to reverse the development. To address this risk, countries of the world gathered for the Earth Summit in Rio de Janeiro in 1992 and agreed on an international treaty; the United Nations Framework Convention on Climate Change (UNFCCC). The aim of the treaty was to take the first steps towards reversing or slowing down global warming by reducing emissions of green house gases and to define the objectives and principles of the member countries, known as the Parties, with respect to green house gas emission. The convention addressed the important issues of climate change mitigation, developing new technology, promoting education, research and information exchange ("United Nations Framework Convention on Climate Change," 2010).

With 194 Parties, the convention is close to universal membership. The Parties of the UNFCCC are differentiated into three groups; the Annex I Parties, refering to the developed nations which are members of the Organization for Economic Co-operation and Development (OECD) plus economies in transition (the EIT Parties); the Annex II Parties, including only the OECD members; and finally

[1] The UNFCCC identifies six primary green house gases: carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs) and sulphur hexafluoride (SF_6)

non-Annex I Parties which include Annex II and other parties which are mostly developing countries. Although the UNFCCC was an important step towards reducing green house gas emissions, the treaty was not legally binding. In December 1997 a number of nations attending the third session of the Conference of the Parties (COP3) adopted an addition to the treaty. The addition is known as the Kyoto Protocol (Depledge & Lamb, 2005).

The protocol recognizes that the majority of emissions stem from developed countries (Annex I) and therefore places a heavier burden on industrialized nations, imposing "common but differentiated responsibilities" ("United Nations Framework Convention on Climate Change," 2010). The Kyoto protocol contains legally binding emissions targets for 37 industrialized countries, stating that on average, Parties will reduce emissions by 5 percent against 1990 levels over the period from 2008 to 2012, also known as the Kyoto commitment period. To date 184 Parties have ratified the treaty which entered into force on February 16th 2005 when the target of 55 Parties had ratified the treaty and accounted for a minimum of 55% of 1990 emissions. To date all but one Annex I Party have ratified the treaty, the only exception being the United States of America.

In reality the developed countries are facing much more than a 5% reduction. Many of the OECD countries did not meet their non-binding emission reductions to 1990 levels by the year 2000, in fact, emissions rose compared to 1990 levels. This trend is now reversing, but in order to meet targets the required overall emission reduction is nearly 20% by the end of 2012 should no abatement attempts be made ("What is the EU doing on climate change?," 2010). To facilitate the required emissions reductions for countries with commitments under the Kyoto Protocol some means of mitigation are offered. These are known as the Kyoto flexible mechanisms: International Emissions Trading (IET), The Clean Development Mechanism (CDM) and Joint Implementation (JI).

The economic basis for international emissions trading is that the marginal cost of abating emissions differs between countries. Under IET, Annex I Parties receive (or purchase) a predetermined (desired) amount of credits, so-called assigned amount units (AAUs). One AAU gives the right to emit one ton of carbon dioxide equivalent (CDE)[2]. If desired the Parties may choose to trade their AAUs. A country with low abatement costs could therefore sell its redundant AAUs to another country where abatement costs are higher and hence increase the efficiency of the Kyoto agreement.

The clean development mechanism and joint implementation are so-called project-based mechanisms which generate emission reductions via specific projects. CDM is designed to encourage Annex I

[2] The carbon dioxide equivalent is a metric measure used as a comparison between various GHGs. The carbon dioxide equivalent of a certain GHG is based on their global warming potential (GWP). The equivalent is commonly expressed in million metric tons of carbon dioxide equivalent (MMTCDE) (Taschini & Paolella, 2006).

countries to invest in emission reduction projects in developing countries (non-Annex I countries). For every ton of CDE reduced a certified emission reduction unit (CER) is granted in return which the Annex I countries can then use to lower their overall emissions. Similarly JI enables developed countries (Annex I) to carry out joint implementation projects with other industrialized countries granting them either emission reduction units (ERUs) or specific removable units (RMUs) based on the type of project ("United Nations Framework Convention on Climate Change," 2010). Figure 2.1 shows a schematic view of the flexible mechanisms and their credits.

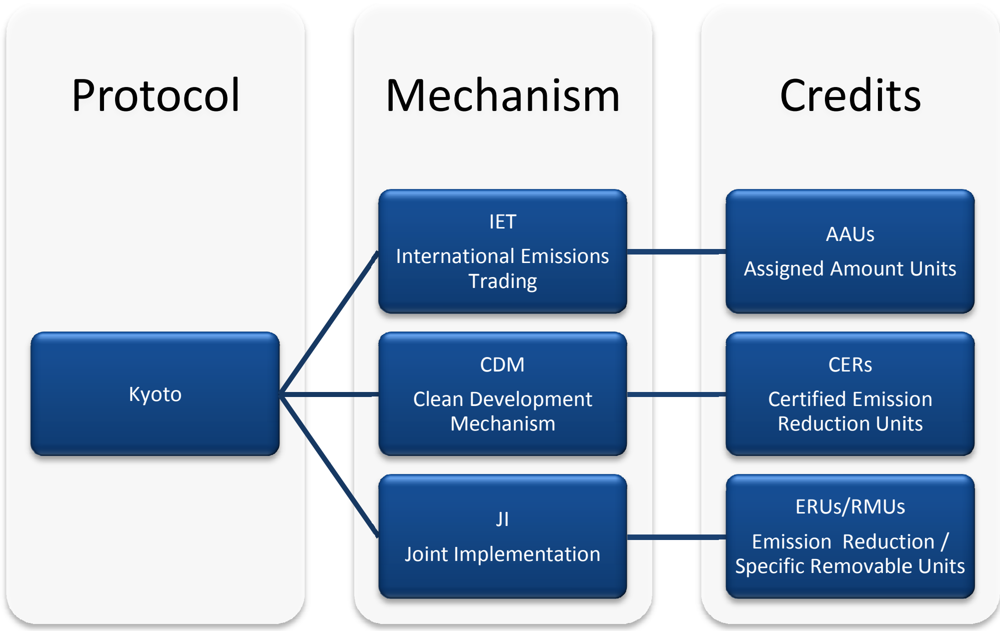

Figure 2.1: A schematic view of the mechanisms of the Kyoto Protocol.

2.2 The European Union Emission Trading Scheme

Following the Kyoto Protocol the European Union (EU) began preparing an EU carbon market to facilitate EU member states to meet their commitments in a cost-effective way. On January 1st 2005 the European Union Emission Trading Scheme (EU ETS), the first international trading system of its kind, was launched. It is currently the largest multi-country, multi-sector emission trading scheme in the world and covers over 11,000 installations in all 27 EU member states as well as the other three (non-EU) members of the European Economic Area; Iceland, Norway and Lichtenstein ("EU Press Releases: MEMO/08/35," 2008).

The EU ETS is a cap and trade system where overall emission levels are capped but members are free to buy or sell emission allowances as needed. The allowances, referred to as European Union allowance units (EUAs), are the currency of the trading scheme and are equal to 1 ton of carbon dioxide equivalent. Under the EU ETS one EUA is compatible to one AAU (refer to figure 2.1). The EU ETS is organized into trading periods called phases (see figure 2.2), where the launch of the trading scheme marks the beginning of phase I, known as the trial period. Phase II spans the five year Kyoto commitment period from 2008 to 2012, phase III runs from 2013-2020 and the fourth trading period runs from 2021-2028.

In order for a country to receive allowance units it must draw up a national allocation plan (NAP), which determines the country's overall emission levels and how the EUAs are to be distributed within the country's installations. By the end of the year, each installation reports its total emissions which can then be credited with the allocated EUAs. If the installation is long, EUAs can either be sold or the allowances banked (however banking was prohibited during phase I). When allowances are banked the owner of the allowances is in fact saving the credits to a later date when he has use for them. If, for instance, an installation meets its annual emission goal and is long EUA allowances, the allowances can be banked and used the following year or when needed. The allowances can also be sold immediately for market spot price. If the installation is short allowances it can choose to buy EUAs or participate in JI or CDM projects to meet emission reduction targets. Non-compliance results in a fine. During phases I and II the fine for non-compliance has been €30/ton, but is scheduled to be raised to over €100/ton in phase III, which is much higher than the spot price of EUAs.

The EU ETS is an artificial market, meaning that there is no actual demand for carbon credits other than the one imposed by law and regulations. The market therefore depends heavily on the quality of regulatory framework and phase I was a trial period used to get the necessary regulatory bodies up and running. From the beginning, the EU ETS has covered power stations and other combustion plants above a certain capacity threshold, oil refineries and coke ovens as well as iron and steel plants. Factories making pulp, paper, board, cement, glass, lime and bricks were also included.

A key rule in phase I was the banning of banking carbon credits to phase II. The market experienced an early crash in 2006 where price of credits plummeted to near zero after the announcement that the market would be long at the end of phase I. Most blame the national allocation plans, saying that emissions had been over-allocated and many countries therefore holding soon to be worthless carbon credits because of the rule of no banking. Some have however suggested that the fault was not over-allocation, but rather over-abatement (Ellerman & Buchner, 2008). Regardless of the reason, the market survived and despite extremely low spot prices, the futures market for phase II credits still maintained high prices indicating a common belief that the market regulations would adjust to this newfound knowledge. Although phase I is a clear example of market crash due to regulatory mistakes it also establishes that the cap and trade system is highly transparent and new information influences price formation quickly.

During phase I all allowances were allocated to installations free of charge and the redundant/missing EUAs sold or bought when needed. For phase II a marginal amount of credits was auctioned and banking to phase III allowed. Phase II coincides with the legally binding Kyoto period to reduce emissions against 1990 levels. Emissions have in fact been reduced, in part by means of abatement, but probably mostly due to the global financial crisis. Demand for energy has dropped accordingly and a price drop in the 2009 spot prices for carbon credits is evident. Trading volumes did however double from 2008 to 2009 indicating a high level of speculation and overall the market value of the transactions was higher in 2009 (Kossoy & Ambrosi, 2010).

In phase III auctioning of allowances will increase year by year, starting from about 10-20 percent, meaning that installations will no longer receive all allowances for free and auctioning will be the default method of allocation, although some sectors and sub-sectors that are at risk of carbon leakage[3] will still receive free allowances. The number of free allowances to be given to industrial installations will be decided in 2011 (Kossoy & Ambrosi, 2010). Each year more allowances will be auctioned and by the end of phase IV the vast majority of all carbon allowances will be auctioned. This development partnered with the annual reductions of overall available allowances will put an upwards pressure on the price of EUAs, hence forcing companies to either pay more for the right to emit or invest in technologies and projects to reduce or offset their emissions. Phase III will include two new sectors; aviation and the aluminum industry.

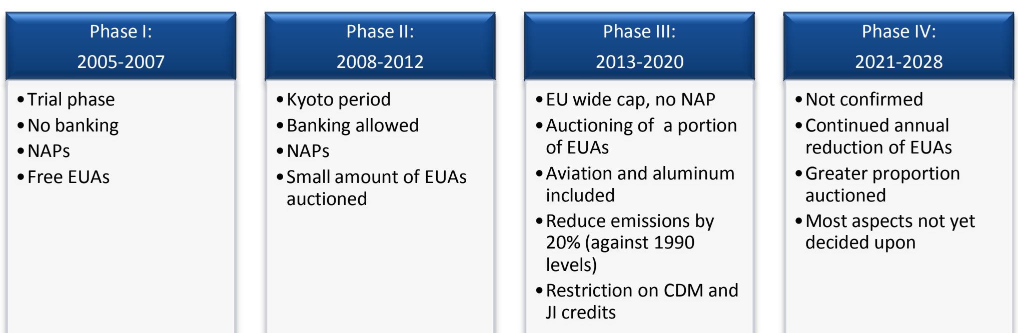

Figure 2.2: Key facts for EU ETS phases.

The EU ETS was the only mandatory emission trading scheme in the world until New Zealand's ETS was entered into force in November 2009. New Zealand's government did however announce that the full implementation of the trading scheme could be delayed unless other developed countries established similar regulations. Following the Copenhagen climate conference in 2009, where high expectations for a legally binding global agreement were not met, the sense of uncertainty of the future of the global emissions reductions effort increased.

Expectations were lower when the 16th session of the Conference of the Parties took place in Cancun, Mexico in December 2010. The climate change conference did however achieve important progress towards an international agreement after the Kyoto period by agreeing upon a "balanced package" of decisions to help countries address climate change. The package also provides the necessary

[3] Carbon leakage occurs when companies in sectors that are subject to strong international competition choose to relocate from the EU to third world countries that are more lenient on GHG emissions.

foundation for further negotiations on a comprehensive agreement, covering all major emitters ("United Nations Framework Convention on Climate Change," 2010).

Currently the EU ETS is the only framework that promises to reduce green house gases after the Kyoto period, i.e. beyond 2012.

3 Theoretical Framework

The chapter covers theory pertinent to the analysis. First a section on time series analysis, then a brief chapter about correlation and its relevance to the analysis followed by a section about principal component analysis, multiple linear regression, principal component regression and finally latent root regression. The sections on time series analysis and correlation rely on Tsay (2002), the section on principal component analysis and latent root regression is mostly based on Jolliffe (2002) and Webster, Gunst & Mason (1974) respectively and finally the section on multiple linear regression is based on Montgomery & Runger (2007) and Gujarati & Porter (2009).

3.1 Time Series Analysis

3.1.1 Returns

Consider an asset, having price P_t at time t. The one period simple gross return from time $t-1$ to t can be written as

$$1 + \mathrm{R_t} = \frac{\mathrm{P_t}}{\mathrm{P_{t-1}}} \quad \text{or} \quad \mathrm{P_t} = \mathrm{P_{t-1}}(1 + \mathrm{R_t}) \tag{3.1}$$

The simple return or simple net return, $\mathrm{R_t}$, is

$$\mathrm{R_t} = \frac{\mathrm{P_t}}{\mathrm{P_{t-1}}} - 1 = \frac{\mathrm{P_t} - \mathrm{P_{t-1}}}{\mathrm{P_{t-1}}} \tag{3.2}$$

Taking the natural logarithm of the simple gross return yields the continuously compounded return, r_t. Let p_t and p_{t-1} denote the natural logarithm of P_t and P_{t-1} respectively. The continuously compounded return or log return is therefore

$$\mathrm{r_t} = \ln(1 + \mathrm{R_t}) = \ln\left(\frac{\mathrm{P_t}}{\mathrm{P_{t-1}}}\right) = \ln(\mathrm{P_t}) - \ln(\mathrm{P_{t-1}}) = \mathrm{p_t} - \mathrm{p_{t-1}} \tag{3.3}$$

For small returns, a commonly used approximation is

$$\mathrm{R_t} \approx \mathrm{r_t} \tag{3.4}$$

3.1.2 Moments of a Random Variable

Now consider a continuous random variable, X. Let $f(x)$ be the probability density function (PDF) and let E denote the expectation operator. The ℓth moment of X is then defined as

$$m'_\ell = E(X^\ell) = \int_{-\infty}^{\infty} x^\ell f(x)dx$$

When analyzing distributions other than the normal distribution the first four moments are of particular interest. The first moment measures the central location of the distribution of X, also known as the mean or expectation of X. Provided that the integral exists, the ℓth central moment of X about the mean μ_x is defined as

$$m'_\ell = E(X - \mu_x)^\ell = \int_{-\infty}^{\infty} (x - \mu_x)^\ell f(x)dx$$

The second central moment of X, the variance, denoted by σ_x^2, measures the variability of X. The positive square root of the variance is the standard deviation, σ_x. The third central moment is a measure of symmetry or lopsidedness, with respect to the mean and the fourth central moment is a measure of the tail behavior of X. In statistics the third and fourth normalized or standardized central moments are referred to as skewness and kurtosis, denoted by $s(x)$ and $k(x)$ respectively, as a measure of asymmetry and tail thickness

$$s(x) = E\left[\frac{(X - \mu_x)^3}{\sigma_x^3}\right] \qquad k(x) = E\left[\frac{(X - \mu_x)^4}{\sigma_x^4}\right]$$

A distribution that is skewed to the right will have positive skewness but negative skewness if skewed to the left. Furthermore, the property $k(x) - 3$ is known as the excess kurtosis, since the kurtosis for a normal distribution is equal to three. A distribution with positive excess kurtosis will have heavier tails than a normal distribution.

Now consider the random sample $\{x_1, x_2, \dots, x_N\}$ of X with N observations where $n = 1,2, \dots, N$. Let $\hat{\mu}_x$, $\hat{\sigma}_x^2$, $\hat{s}(x)$ and $\hat{k}(x)$ denote the sample mean, sample variance, normalized sample skewness and normalized sample kurtosis respectively. Then

$$\hat{\mu}_x = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{3.5}$$

$$\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{n=1}^{N} (x_n - \hat{\mu}_x)^2 \tag{3.6}$$

$$\hat{s}(x) = \frac{1}{(N-1)\hat{\sigma}_x^3}\sum_{n=1}^{N} (x_n - \hat{\mu}_x)^3 \tag{3.7}$$

and
$$\hat{\mathrm{k}}(\mathrm{x}) = \frac{1}{(\mathrm{N}-1)\hat{\sigma}_{\mathrm{x}}^{4}} \sum_{\mathrm{n}=1}^{\mathrm{N}} (\mathrm{x}_{\mathrm{n}} - \hat{\mu}_{\mathrm{x}})^{4} \tag{3.8}$$

represent the unbiased sample estimators as counterparts to the population central moments.

3.1.3 Multivariate Returns and Covariance

Let $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ be a random vector consisting of p random variables. Provided that the expectations exist the mean vector and covariance matrix of $\boldsymbol{X}$ are defined as

$$E(\boldsymbol{X}) = \boldsymbol{\mu}_x = \left[E(X_1), E(X_2), \dots, E(X_p)\right]'$$

$$Cov(\boldsymbol{X}) = \boldsymbol{\Sigma} = \mathrm{E}[(\mathbf{X} - \boldsymbol{\mu}_{\mathrm{x}})(\mathbf{X} - \boldsymbol{\mu}_{\mathrm{x}})']$$

whose sample counterparts are simply vector notations of equations (3.5), (3.6), (3.7) and (3.8) (assuming that the population mean is not known). Thus, the covariance of a random variable with itself is simply the variance of the random variable. Furthermore assuming that the population mean is unknown the covariance of two random variables, X and Y is

$$Cov(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right] = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y}) \tag{3.9}$$

where $\bar{x}$ and $\bar{y}$ represent the sample means of X and Y respectively according to equation (3.5).

3.2 Correlation

Correlation is an important factor in both principal component analysis and regression introduced in later sections. It is a measurement of how two random variables move together, i.e. the strength of the linear dependence between them. Let X and Y be two random variables. The correlation between X and Y is then defined as

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E\left[(X-\mu_x)(Y-\mu_y)\right]}{\sqrt{E(X-\mu_x)^2 E(Y-\mu_y)^2}} \tag{3.10}$$

It can be shown that $-1 \leq \rho_{x,y} \leq 1$ and $\rho_{x,y} = \rho_{y,x}$. X and Y are completely uncorrelated if $\rho_{x,y} = 0$. Now consider again a sample of N observations from both X and Y. The sample correlation is then represented as

$$\hat{\rho}_{x,y} = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_{n}^{N}(x_n - \bar{x})^2 \sum_{n}^{N}(y_n - \bar{y})^2}} \tag{3.11}$$

where $\bar{x}$ and $\bar{y}$ represent the sample means of X and Y respectively according to equation (3.5).

Correlation also has a useful geometric property. Consider X and Y as two vectors in space. Let $Var(X)$ and $Var(Y)$ represent the squared lengths of the vectors and let their mutual projection in the Euclidian space be given by $Cov(X, Y)$. Let θ be the angle between the two vectors. Equation (3.10) can then be rewritten as

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \cos(\theta) \tag{3.12}$$

The correlation between X and Y can therefore be readily interpreted by examining a graphical representation. This property will be utilized further in the principal component analysis in section 5.1.3.

Another important factor when analyzing time series is autocorrelation. Consider again the return series r_t, where t is the time index and the greatest time index is T. The concept of autocorrelation is a generalized version of correlation when the linear dependence between r_t and its past values $r_{t-\ell}$ is of interest. The correlation coefficient between r_t and $r_{t-\ell}$ is denoted by ρ_ℓ and refered to as the lag-ℓ autocorrelation of r_t. Autocorrelation is further defined as

$$\rho_\ell = \frac{Cov(r_t, r_{t-\ell})}{\sqrt{Var(r_t)Var(r_{t-\ell})}} \tag{3.13}$$

whose lag- ℓ sample autocorrelation counterpart is

$$\hat{\rho}_\ell = \frac{\sum_{t=\ell+1}^{T}(r_t - \bar{r})(r_{t-\ell} - \bar{r})}{\sqrt{\sum_{t=\ell+1}^{T}(r_t - \bar{r})^2 \sum_{t=\ell+1}^{T}(r_{t-\ell} - \bar{r})^2}} \tag{3.14}$$

Iteratively calculating the lag- ℓ autocorrelation for increased values of ℓ, generates the function $\hat{\rho}_1, \hat{\rho}_2, \ldots$, called the sample autocorrelation function (ACF) of r_t which can be plotted against ℓ to capture the linear dynamic of the data to identify any serial correlations of the returns.

3.3 Principal Component Analysis

Principal component analysis (PCA) is a dimension reduction technique. Given a multidimensional data set in $\mathbb{R}^p$, the aim of PCA is to describe the data in a lower dimension using synthetic variables which are linear combinations of the original variables. The synthetic variables are also known as the principal components (PCs). Reducing the dimension of the data generally incurs a loss of information, but PCA performs this reduction in such a way as to minimize this loss.

3.3.1 Definition

Suppose that $\boldsymbol{x}$ is a vector of p centered[4] random variables whose covariance structure and correlations between the p variables is of interest. PCA is applied to $\boldsymbol{\Sigma}$, the covariance matrix of $\boldsymbol{x}$. Furthermore let $\boldsymbol{\alpha}_1$ be a vector of p constants $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p}$. The first principal component, z_1, is the linear combination of the elements of $\boldsymbol{x}$ having maximum variance,

$$z_1 = \boldsymbol{x}\boldsymbol{\alpha}_1 = x_1\alpha_{11} + x_2\alpha_{12} + \cdots + x_p\alpha_{1p} = \sum_{j=1}^{p} x_j\alpha_{1j}$$

To find the second PC, a linear combination $\boldsymbol{x}\boldsymbol{\alpha}_2$, uncorrelated with $\boldsymbol{x}\boldsymbol{\alpha}_1$, is found in the direction of maximum variance (apart from the first PC) and so on. At the kth stage of the process a linear function $\boldsymbol{x}\boldsymbol{\alpha}_k$ is found that has maximum variance subject to being uncorrelated with $\boldsymbol{x}\boldsymbol{\alpha}_1, \boldsymbol{x}\boldsymbol{\alpha}_2, \ldots, \boldsymbol{x}\boldsymbol{\alpha}_{k-1}$. The kth linear function is the kth PC. In general the kth PC can be written as

$$z_k = \boldsymbol{x}\boldsymbol{\alpha}_k = x_1\alpha_{k1} + x_2\alpha_{k2} + \cdots + x_p\alpha_{kp} = \sum_{j=1}^{p} x_j\alpha_{kj} \tag{3.15}$$

where $k < p$ is an integer. If $\boldsymbol{x}$ consists of simple returns of p assets then z_k is the return of a portfolio that assigns the weight α_{kj} to the jth asset. In order to maintain the proportional allocation assigned to the jth asset the vector $\boldsymbol{\alpha}_k$ is standardized such that $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1$. Up to p PCs can be found, but the goal of the method is to reduce the dataset so that most of the variation in the data is accounted for by the first m PCs, where $m \ll p$ is determined according to a selection criteria which will be covered in the next section.

In matrix form, equation (3.15) can be rewritten as

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A} \tag{3.16}$$

where $\boldsymbol{Z}$ is a $N \times p$ matrix whose kth column contains the kth PC, $\boldsymbol{X}$ is the centered $N \times p$ data matrix and $\boldsymbol{A}$ is an orthogonal $p \times p$ matrix whose kth column, $\boldsymbol{\alpha_k}$ is called the kth eigenvector of the covariance matrix, $\boldsymbol{\Sigma}$.

Using properties of linear combinations of random variables it can be shown that

$$\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{\Lambda} \tag{3.17}$$

where $\boldsymbol{\Lambda}$ is a diagonal $p \times p$ matrix whose kth diagonal element is λ_k; the kth eigenvalue of $\boldsymbol{\Sigma}$ also referred to as the kth latent value. The diagonal elements of $\boldsymbol{\Lambda}$ are ordered by decreasing magnitude such that $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$. Furthermore considering the kth diagonal element of $\boldsymbol{\Lambda}$ it can be shown that

[4] Centered, meaning that each column mean has been subtracted form the column values

$$\lambda_k = Var(\boldsymbol{x}\alpha_k) = Var(z_k) \tag{3.18}$$

It then follows from equations (3.17) and (3.18) that

$$\mathbf{A}'\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{\Lambda} \tag{3.19}$$

and the covariance matrix can be reconstructed from the PCs by

$$\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{\Lambda}\mathbf{A}' \tag{3.20}$$

3.3.2 Selection Criteria

When choosing the number of PCs to retain after the analysis several selection criteria exist. Three will be explained for the purpose of the analysis.

The first, and most widely used, is the cumulative proportion of variance accounted for by the eigenvalues. Again let $\boldsymbol{x}_i$ denote the ith column vector of $\boldsymbol{X}$, z_i represent the ith PC, λ_i the ith eigenvalue and $tr(\boldsymbol{\Sigma})$ denote the trace of the covariance matrix. The variance of $\boldsymbol{x}_i$ is then

$$\sum_{i=1}^{p} Var(\boldsymbol{x}_i) = tr(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(z_i) \tag{3.21}$$

The result of equation (3.21) yields the semi-quantitative selection criteria; proportion of cumulative variance of the kth PC (Tsay, 2002)

$$\frac{Var(z_k)}{\sum_{i=1}^{p} Var(\boldsymbol{x}_i)} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \tag{3.22}$$

The criterion is semi-quantitative because the final decision of the number of PCs, m, to retain is qualitative – often a rule of thumb[5].

The second method, backward elimination, involves iteratively eliminating the most dominant variable of the PC having the lowest latent value until a certain number of variables are left. The most dominant variable is the variable whose coefficient in the linear combination has the highest absolute value.

The third method, forward selection, is the opposite of backward elimination. Instead of eliminating poor candidates, the method iteratively chooses the most dominant variable of the PC having the highest latent value. The process is repeated until a sufficient number of variables has been selected.

[5] A common approach is choosing m such that the PCs account for over 80 percent of the total variability of the data.

3.3.3 Singular Value Decomposition

To physically compute the principal components further definitions are needed. The PCs can be expressed by performing singular value decomposition (SVD). The SVD of the centered $N \times p$ data matrix $\boldsymbol{X}$ is defined as

$$\boldsymbol{X} = \boldsymbol{ULA}' \tag{3.23}$$

where $\boldsymbol{U}$ is an $N \times p$ orthogonal matrix spanning the column space of $\boldsymbol{X}$ and $\boldsymbol{A}$, as defined above, spans the row space. $\boldsymbol{L}$ is a $p \times p$ diagonal matrix with diagonal entries $l_1 \geq l_2 \geq \cdots l_p \geq 0$, called the singular values of $\boldsymbol{X}$ (Hastie, Tibshirani, & Friedman, 2008; Jolliffe, 2002). The sample covariance matrix is given by $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}/(N-1)$ and hence from equation (3.23)

$$\boldsymbol{S}(N-1) = \boldsymbol{X}'\boldsymbol{X} = \boldsymbol{AL^2A}' \tag{3.24}$$

is the eigen decomposition of $\boldsymbol{X}'\boldsymbol{X}$. The columns of $\boldsymbol{A}$ are called the principal component directions or eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ and as previously mentioned the first PC has the property of having the largest sample variance of all the possible linear combinations of the columns of $\boldsymbol{X}$. By definition of singular values and eigenvalues the sample variance of the kth PC is easily found to be

$$Var(z_k) = Var(\boldsymbol{X}\alpha_k) = \frac{l_k^2}{N-1} = \lambda_k \tag{3.25}$$

and in fact $z_k = \boldsymbol{X}\alpha_k = \boldsymbol{u}_k l_k$ meaning that the columns of $\boldsymbol{UL}$ also give the PCs of $\boldsymbol{X}$. The importance of using SVD for PCA is twofold; it provides an efficient method to find the PCs as well as yielding the eigenvectors and singular values, and hence the eigenvalues and variance of the PCs for the sample covariance matrix. The scaled versions of the PC scores given by $\boldsymbol{U} = \boldsymbol{ZL}^{-1}$ are a bonus (Jolliffe, 2002).

3.4 Multiple Linear Regression

Linear regression is a statistical technique often used to capture the relationship between different variables. The variables are referred to as dependent variables and independent variables or regressors. As is implied by the name, the dependent variable is a function of the independent variables. Multiple linear regression treats the scenario where a single dependent variable is a function of multiple regressors.

3.4.1 Definition

To define the regression model let again $X = x_{i1}, x_{i2}, \ldots, x_{ip}$, $i = 1,2,\ldots,N$, be a vector of p random variables each having N observations and let Y be the dependent variable, also of N observations. The linear relationship between X and Y can be defined by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \qquad i = 1,2,\ldots N$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are constants, to be estimated, that assign weights to the regressors. The relationship is however rarely strictly linear so an error term is added

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon_i \qquad i = 1,2,\ldots,N \tag{3.26}$$

called a multiple linear regression model with p regressors, where the error term, ϵ_i, is a random error with mean zero and an unknown variance σ^2. Using matrix notation the model can be expressed as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{3.27}$$

where the first column of $\boldsymbol{X}$ is the column vector $\mathbf{1}$.

Although the model is called a linear regression model it can accommodate non-linear terms such as the interaction term $\beta_{12}X_1X_2$ by simply setting $\beta_{p+1} = \beta_{12}$ and $X_{p+1} = X_1X_2$. In fact, any regression model whose parameters are linear is a linear regression model regardless of the shape of the generated plane.

3.4.2 Least Squares Estimation of Parameters

In order to find the "best fit" to the data an estimation of the betas, denoted by $\widehat{\boldsymbol{\beta}}$, needs to be made. The method of ordinary least squares (OLS) or sum of squared errors can be used to estimate the regression parameters (betas). The goal is to minimize the least squares function L[6]

$$L = \sum_{i=1}^{N} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \tag{3.28}$$

The least squares estimator $\widehat{\boldsymbol{\beta}}$ is then found by taking the partial derivatives of L and finding the minimum

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{0} \tag{3.29}$$

Equation (3.29) yields the equations that must be solved

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \tag{3.30}$$

Multiplying both sides of equation (3.30) by $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ then gives the least squares estimate of the betas

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{3.31}$$

[6] Note that this is *not* the same matrix L as in the SVD in the PCA section.

The fitted model can then be represented as $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The difference between the actual observations and the fitted values is known as the error or residual defined by

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{3.32}$$

3.4.3 Assumptions of the Classical Model and the Gauss-Markov assumptions

Several assumptions underlying the method of least squares have to be met in order to draw inferences about the true betas of the population. The Gaussian, standard classical linear regression model (CLRM), the cornerstone of most econometric theory, makes seven assumptions:

A1. The model is linear in its parameters as discussed in section 3.4.1
A2. Regressors are independent of the error term, i.e. $Cov(X_i, e_i) = 0$
A3. The expected value of the error terms is zero, i.e. $E[e_i] = 0$
A4. The errors are homoskedastic, have constant variance, i.e. $Var(e_i) = \sigma^2$
A5. No autocorrelation between the errors, i.e. $Cov(e_i, e_j) = 0$ where $i \neq j$
A6. The number of observations, N, must be greater than the number of estimated parameters, p
A7. The given X values must not all be the same, i.e. X must have variability

Under the above assumptions the OLS estimators are both unbiased and have minimum variance among all alternative estimators. Given a model that upholds the assumptions of the CLRM, inferences can be made about the population with the added assumption that the errors are normally distributed with mean zero and variance σ^2.

When the sole purpose is parameter estimation, but not to draw conclusions about the population the estimation can be based on fewer assumptions. A best linear unbiased estimator (BLUE) is defined by the assumptions underlying the Gauss-Markov theorem[7]. An OLS estimator $\hat{\beta}$ is said to be a best linear unbiased estimator (BLUE) of β if the following holds:

A1. The model is linear
A2. The parameter estimate is unbiased
A3. It has minimum variance, i.e. it is an efficient estimator

The residuals do not need to be normally distributed nor homoskedastic for an estimator to be BLUE (Gujarati & Porter, 2009).

3.4.4 Significance Testing: The t-statistic and White's robust t-statistic

When an estimate of the betas has been found it is often desirable to verify the significance of the estimated parameters. This can be done using the t-test statistic. To test whether an individual

[7] The Gauss-Markov theorem states that given the assumptions of the CLRM, the OLS estimators in the class of unbiased linear estimators, have minimum variance and are BLUE.

regression coefficient, β_j, equals a value β_{j0} the hypothesis $H_0 = \beta_j = \beta_{j0}$ and $H_1 = \beta_j \neq \beta_{j0}$ is tested with the statistic

$$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}} \tag{3.33}$$

where C_{jj} is the diagonal element of the inverse of the covariance matrix of $\boldsymbol{X}$, $(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}$, corresponding to the estimated beta, $\hat{\beta}_j$. The null hypothesis is rejected if $|t_0| > t_{\alpha/2,n-(p+1)}$.

Significance testing is based on the assumptions of normally distributed, homoskedastic residuals. The homoskedasticity assumption is often hard to meet, rendering the parameter confidence intervals unreliable. White (1980) offered a remedy called a heteroskedasticity-consistent covariance matrix estimator to treat heteroskedastic residuals in a manner that yields reliable confidence intervals of the estimator without changing the regression coefficients. Furthermore the estimator holds regardless of the shape of the heteroskedasticity of the residuals. White's robust t-statistic, HC0, is defined as

$$HC0 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{diag}\left[e_i^2\right]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{3.34}$$

where the entries on the main diagonal of HC0 are the estimated squared standard errors of the regression coefficients. Dividing the regression coefficients by these standard errors gives a ratio that can be used to derive the p-values for hypothesis testing (Hayes & Cai, 2007).

According to Hayes & Cai (2007) a weighted version of White's t-statistic is more reliable. The HC4 statistic was introduced by Cribari-Neto, (2004) and is defined as

$$HC4 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{diag}\left[\frac{e_i^2}{(1-h_{ii})^{\delta_i}}\right]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{3.35}$$

where,

$$\delta = \min\left\{4, \frac{Nh_{ii}}{p+1}\right\} \tag{3.36}$$

where $h_{ii} = \boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'$. The h_{ii}s are the diagonal elements in the "hat" matrix $\mathbf{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and also know as leverage values[8]. The δ controls the level of discounting for the ith observation with the truncation point at 4. The p-values can be extracted from HC4 in the same way as for HC0 (Hayes & Cai, 2007).

[8] "The use of leverage adjusted residuals is based on an extensive literature on the finite-sample bias of HC0." (Hayes & Cai, 2007)

3.4.5 Model Adequacy Checking

After a model has been fitted to data there is no guarantee that the model adequately describes the relationships between the variables. It is interesting to examine how much of the variation in the data is absorbed by the error terms. There are many means to this end but four will be examined for the sake of the analysis; the coefficient of determination, R^2, the adjusted coefficient of determination, R^2_{adj}, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).

The coefficient of determination is a measure of how much variation of the dependent variable is explained by the model. The basis of R^2 lies in the concept of analysis of variance (ANOVA) which is mainly comprised of three concepts: the error sum of squares, SS_E, the regression sum of squares, SS_R and the total corrected sum of squares, SS_T. Where $SS_E = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, $SS_R = \boldsymbol{\hat{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \frac{\left(\sum_{i=1}^{N} y_i\right)^2}{N}$ and $SS_T = \boldsymbol{y}'\boldsymbol{y} - \frac{\left(\sum_{i=1}^{N} y_i\right)^2}{N}$. For a linear model the sums are connected by the relationship $SS_T = SS_R + SS_E$. Furthermore the quantity $MS_E = SS_E/(N-p)$ is called the error mean square or mean squared error. Using the above relationships the coefficient of determination is the defined as

$$R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_R}{SS_T} \tag{3.37}$$

This statistic should however be used with caution. Generally R^2 increases every time a new variable is added to the model but that does not always imply a better model. Adding new variables to the model incurs a loss of one error degree of freedom meaning that in order for the model to be superior the error sum of squares of the new model should be reduced by an amount equal to the original error mean square. To counter this tendency an adjusted version of R^2 exists, where

$$R^2_{adj} = 1 - \frac{SS_E/(N-p)}{SS_T/(N-1)} = 1 - \frac{MS_E}{SS_T/(N-1)} \tag{3.38}$$

Equation (3.38) states that the adjusted coefficient of determination will only increase if the error mean square is reduced when a new variable (increased value of p) is added to the model, therefore R^2_{adj} is a good guard against overfitting.

The AIC and BIC statistics are also measures of goodness of fit and can be said to describe the tradeoff between bias and variance. Both punish for including more parameters in the model but BIC includes a larger penalty for overfitting. The AIC and BIC are defined as follows:

$$AIC = N \log\left(\frac{SSE}{N}\right) + 2p \tag{3.39}$$

$$BIC = N \log\left(\frac{SSE}{N}\right) + p \log(N) \tag{3.40}$$

where p stands for number of variables and N is the number of observations or data points. When comparing models by their respective AIC and BIC, the model having the lowest value is considered the best model in terms of AIC and BIC.

3.5 Principal Component Regression

Now consider again the standard regression model as defined by equation (3.27) and the principal components as defined by equation (3.16). Since $\boldsymbol{A}$ is orthogonal, $\boldsymbol{X\beta}$ can be written as $\boldsymbol{XAA}^{-1}\boldsymbol{\beta}$ (assuming again that $\boldsymbol{X}$ is a centered matrix) and equation (3.27) can therefore be rewritten as

$$\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\epsilon} \tag{3.41}$$

where $\boldsymbol{\gamma} = \boldsymbol{A}^{-1}\boldsymbol{\beta}$. The predictor variables have simply been replaced by their PCs in the regression model. If predictors are near-singular or are related via multicolinearity the model explained by equation (3.41) is unsatisfactory. The high inter-correlations of the predictor variables are transformed into low variances of the PCs. These low-variance relationships can be detected by examining the latent values (eigenvalues) as indicated by equation (3.25). Principal component regression (PCR) tackles this by reducing the dataset to a subset of m PCs. Having chosen the m PCs to retain, the reduced model can be written as

$$\boldsymbol{y} = \boldsymbol{Z}_m\boldsymbol{\gamma}_m + \boldsymbol{\epsilon}_m \tag{3.42}$$

A least squares estimate can then be used to find an estimate for the new parameters according to the equation

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{A}\widehat{\boldsymbol{\gamma}} \tag{3.43}$$

Applying least squares to equation (3.43) is equivalent to finding $\widehat{\boldsymbol{\beta}}$ by applying least squares directly to equation (3.27). It is however more straight forward to find the estimate using equation (3.41). The vector $\widehat{\boldsymbol{\gamma}}$ is then defined as

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{L}^{-2}\boldsymbol{Z}'\boldsymbol{y} \tag{3.44}$$

where $\boldsymbol{L}$ is the diagonal matrix containing the eigenvalues as defined in the section on SVD.

Using the PCs as predictor variables instead of the original regressors enables the contributions of each PC to be more easily interpreted than those of the original variables. Furthermore, because the PCs are uncorrelated, the contribution and estimated parameters of each PC to the model are unaffected by the addition of other PCs to the regression whereas for the original variables both the contributions and coefficients can change dramatically when another regressor is added to the model. This is especially true in the presence of multicolinearity where the main advantage of PC regression occurs. Even if multicolinearity is not a problem, regressing on the PCs can still have advantages for computation and interpretation. It should however be noted that although interpretation of the contributions is improved by regressing on the PCs, the interpretation of the regression equation itself may be futile if the PCs do not have a clear meaning.

3.6 Latent Root Regression

In the presence of multicolinearities the OLS estimation of parameters can result in very poor estimates. For instance if there is an almost exact linear dependence among the regressors, the R^2 can be very close to one. This type of ill-conditioning of the data is referred to as near singularity. Latent Root Regression (LRR), suggested by Webster, Gunst & Mason in 1974, is a modified version of OLS which identifies near singularities, determines whether these near singularities have predictive value or not and then obtains the modified (biased) least squares estimates of the parameters having adjusted for the non-predictive near singularities using the latent roots and latent vectors of the data.

Consider again equation (3.27). Let $y_i^* = (y_i - \bar{y})/\eta$ where $\eta^2 = \sum_{i=1}^{N}(y_i - \bar{y})^2$. Define the matrix $B = [\boldsymbol{y}^*: \boldsymbol{X}]$, i.e. the $N \times p + 1$ matrix of dependent *and* independent variables. After having applied PCA on B the least squares coefficients are defined as

$$\boldsymbol{\beta} = -\eta \left(\sum_{i=0}^{p} \lambda_i^{*-1} \right)^{-1} \sum_{j=0}^{p} \alpha_{0j} \lambda_j^{-1} \boldsymbol{\alpha}_j^0 \tag{3.45}$$

with residual sum of squares

$$SSE(X_1, X_2, \dots, X_p) = \eta^2 \left(\sum_{i=0}^{p} \lambda_i^{*-1} \right)^{-1} \tag{3.46}$$

where α_{0j} is the coefficient for the dependent variable of the jth eigenvector, $\boldsymbol{\alpha}_j^0$ is a vector which contains all elements of the jth eigenvector, except the first one (the dependent variable) and $\lambda_i^* = \frac{\lambda_i}{\alpha_{0i}^2}$ (Webster et al., 1974).

As mentioned in section 3.3 the kth principal component, z_k is represented in terms of the latent vectors. Furthermore the kth latent root, corresponding to the kth latent vector, measures the spread of the N data points in the direction defined by the latent vector. If α_{0k} is nearly zero the kth principal component is nearly orthogonal to the $\boldsymbol{y}^*$ axis so if both α_{0k} and λ_k are small, the kth latent vector, α_k, reveals a non-predictive near singularity, i.e. a strong linear dependence in the independent variables only, resulting in little or no change in the dependent variable.

Now suppose that the latent vectors $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_{k-1}$ correspond to non-predictive near singularities. Setting the corresponding entries of B equal to zero the modified least squares coefficients are defined as

$$\boldsymbol{\beta}^* = -\eta \left(\sum_{i=k}^{p} \lambda_i^{*-1} \right)^{-1} \sum_{j=k}^{p} \alpha_{0j} \lambda_j^{-1} \boldsymbol{\alpha}_j^0 \tag{3.47}$$

Similarly the residual sum of squares for the modified coefficients are

$$SSE(X_1, X_2, \dots, X_p) = \eta^2 \left(\sum_{i=k}^{p} \lambda_i^{*-1} \right)^{-1} \tag{3.48}$$

Thus the coefficients have been estimated utilizing only linear combinations not having both λ_k and α_{0k} small and thus having adjusted for the effect of non-predictive near singularities (Webster et al., 1974).

4 Data

The chapter covers data relevant to the analysis. First is a section on the trends of EUA prices and trading volumes. Then a section on the suggested connection between carbon allowance prices and the energy sector and finally a section describing the complete datasets chosen for the analysis and the sources used to obtain the data.

4.1 European Union Allowance Units

Since the dawn of the EU ETS in 2005 the price changes of allowances have been very volatile. Figure 4.1 shows the evolution of the spot price and the December 2008 futures contract for EUAs. The relevant trading volumes are shown in figure 4.2.

Figure 4.1: EUA prices for phases I and II (Source: "Point Carbon," 2010).

The market experienced an early crash in April 2006 when the spot price of carbon fell by almost two thirds in a single day from over 30 €/ton to just 9 €/ton prompted by uncoordinated announcements of several countries that the 2005 emissions were well within the total cap. The imminent surplus of EUAs on the market quickly devalued the spot price of carbon. An apparent surprise to the sector analysts, the crash of April 2006 registered the lowest prices since the launch of the market in January 2005. (Jones, 2006). By the end of 2006 and during early months of 2007 EUA price had dropped to virtually zero or under €1. The price spread between the spot price and the Dec 2008 futures contracts had increased from €3-5 to over €16. Market players reported buying inexpensive EUAs for phase I compliance and banking any project-based CERs delivered in 2005-2007 to phase II. The prohibition of banking EUAs to phase II accompanied by the surplus on the market contributed to making phase I EUAs almost worthless at the close of phase I (Capoor & Ambrosi, 2007).

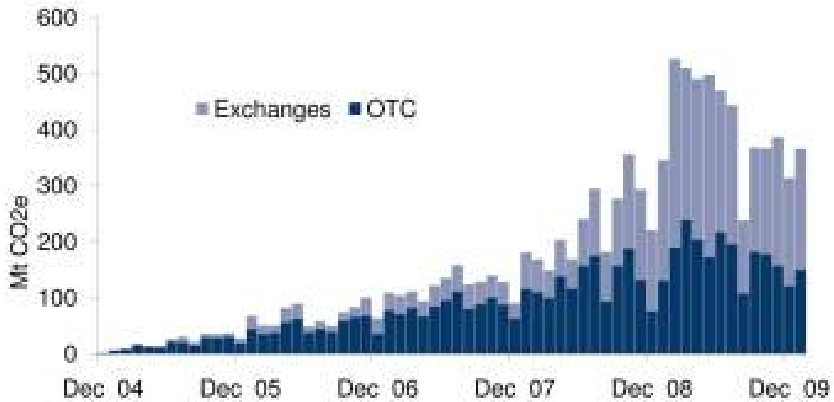

Figure 4.2: Monthly EUA trading volumes for phases I and II (Source: "Point Carbon," 2010).

Despite the severe crash of phase I EUAs the market still appeared to have faith in the EU ETS. The high prices of the December 2008 futures prices against the phase I spot price showed an increased interest in phase II as the EU announced tighter caps for the next trading period as well as allowing phase II allowances to be banked to phase III. The turmoil in 2006 and 2007 EUA prices is also a testimony of the efficiency in price formation of the carbon market. New information appears to be reflected in price quickly, suggesting market transparency.

Starting phase II, EUA prices rose and expectations grew again, but the price-increases were short-lived – carbon was hit by the recession in mid-year 2008 as economic slowdown reduced the demand for carbon. Commodity prices fell and energy intensive production of various kinds was cut back, reducing emissions. The need to purchase allowances was also reduced, since credits had been allocated for free prior to phase II when demand for commodities was strong, emissions higher and the economy was healthy. After the recession hit the European economy many struggled with raising funds. In a difficult credit environment companies who were long allowances chose to sell EUAs on the market to raise cheap cash and shore up their balance sheets causing record breaking daily and monthly trading volumes at the time, especially in the spot market. The increased supply was not met by demand hence driving lower prices for carbon well into 2009.

The companies' decision to sell EUAs was boosted by the fact that the allowances had never been paid for, but had been granted for free. Amplifying the problem was the fact that there was an overlap in the time when allowances for 2009 were issued (February 2009) and the time when 2008 allowances were to be surrendered (April 2009). Many market participants therefore chose to sell their 2008 allowances for cash knowing that they could cover any shortage with the 2009 allowances. Furthermore they could hedge their position by buying later maturity futures or options at attractive prices, in effect yielding a discount-priced loan when attaining credit was expensive or even impossible (Capoor & Ambrosi, 2009).

In February 2009 EUA prices had fallen to €8, compared to €30 nine months earlier stocking fear amongst market participants that phase II would repeat the end of phase I. However, by May 2009 the worst had passed and prices rebounded rather quickly and stabilized around €13 to €16 for the remainder of the year and well into 2010.

Given the above summary of the price history of EUAs there have been some concerns that the high volatility of carbon credits will discourage investment in emission reduction projects. The price volatility of phase I was primarily driven by regulatory mistakes as phase I, or the "learning phase" as named by the EU, ran its course. Phase II then correctly reflected the macro-economic situation in the world as commodity prices fell and demand for carbon assets was reduced accordingly. A fixed price of carbon cannot be accommodated by a cap-and-trade system such as the EU ETS due to the fact that the market should adjust to the level of scarcity of carbon. If a fixed price were the target a tax scheme would be more appropriate. While the increased proportion of auctioned allowances should help stabilize the price of carbon, given the current scheme the price will continue to reflect short and long term supply and demand.

Carbon trading volumes (refer to figure 4.2) have continued to grow as the trading scheme matures. Over-the-counter (OTC) forward contracts have been responsible for most transactions although forward contracts on exchanges have been growing for the past couple of years. The spot market has also been growing steadily. From its birth the EU ETS carbon market has grown fast, tripling its size in the first year and doubling in size between 2006 and 2007 when considering trading volumes. Despite the recession, the market value rose between 2008 and 2009 and trading volumes doubled. As shown in table 4.1 the value of the EU ETS market in 2009 is fifteen times larger than it was four years earlier as the trading done in 2009 accumulated to $ 118 billion (€85 billion).

Table 4.1: Carbon market volumes and values in 2005-2009.*

	2005	2006	2007	2008	2009
EU ETS Value [M $]**	7,908	24,436	49,065	100,526	118,474
	(99%)	(99%)	(99%)	(99%)	(96%)
EU ETS Volume [MtCO_2]**	321	1,104	2,060	3,093	6,326
	(98%)	(97%)	(98%)	(94%)	(86%)
Total Value [M $]	7,971	24,699	49,361	101,492	122,822
Total Volume [MtCO_2]	328	1,104	2,108	3,278	7,362

**Source: Capoor & Ambrosi, 2007; 2008; 2009; Kossoy & Ambrosi, 2010.*

***Numbers in brackets represent percentage of the total market value or volume respectively.*

Table 4.1 also shows overall trading volumes and value of the carbon market worldwide. In the last couple of years other allowance markets besides the EU ETS have also started to grow notably especially trading of AAUs of the Kyoto protocol as well as under the Regional Green House Gas Initiative (RGGI)[9] (Kossoy & Ambrosi, 2010) but the EU ETS is by far the largest market with market share of over 99% of total worldwide trading value during phase I and 97-98% of total market volume. Focusing on the EU ETS should therefore give a realistic view of the global market.

4.2 The Power Sector and EUAs

The power sector is one of the largest market players of the EU ETS accounting for over 70% of all emissions in 2007 and 2008 (*Report on 2008 EU Emissions Trading System emissions data*, 2009). Given its size, the power sector is the most dominant influence on the demand side and could thereby influence market dynamics.

Prior to the initialization of the EU ETS the profit or spread made by power plants could be stated as the difference between the price of electricity on the market and the price of the fuel used to generate power having considered the efficiency of the plant and disregarded operational costs. These somewhat crude relationships are referred to as dark spread (DS) and spark spread (SS). Dark spread represents the theoretical profit made when a coal-fired power plant sells a unit of electricity. Similarly, spark spread represents the profit made when a gas-fired power plant sells a unit of electricity. Dark spread and spark spread can be expressed as follows:

$$DS = P_{Electricity} - P_{Coal} \cdot \frac{1}{\rho_{Coal}} \tag{4.1}$$

$$SS = P_{Electricity} - P_{Gas} \cdot \frac{1}{\rho_{Gas}} \tag{4.2}$$

where $P_{Electricity}$ represents the price of base load electricity per megawatt hour (MWh) sold on the market, P_{Coal} is the price of coal per MWh, P_{Gas} is the price of natural gas per MWh and ρ_{Coal} and ρ_{Gas} are efficiencies of the coal-fired and gas-fired power plants respectively. The values of ρ_{Coal} and ρ_{Gas} vary from plant to plant but $\rho_{Coal} = 0{,}40$ and $\rho_{Gas} = 0{,}55$ are often used as industry-wide averages (International Energy Agency, 2005). Figure 4.3 shows historical coal, natural gas and base load electricity prices.

After the EU ETS emerged the simple economic equations above have been altered, rendering changes in the competitive environment in the European power sector. Given the price of the right to emit one

[9] The RGGI is the first mandatory, market- based effort in the USA to reduce GHG emissions. Ten participating states have capped CO_2 regionally and require the power sector to surrender tradable CO_2 allowances for each ton of emitted CO_2. The ten participating states are Conneticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Rhode Island and Vermont.

ton of CO_2, the clean dark spread (CDS) and clean spark spread (CSS) can be defined. Clean dark spread represents the dark spread having adjusted for the price of CO_2 and similarly clean spark spread represents the spark spread after adjusting for the price of CO_2. The clean dark spread and clean spark spread are defined as follows:

$$CDS = P_{Electricity} - \left[P_{Coal} \cdot \frac{1}{\rho_{Coal}} + P_{CO_2} \cdot E_{Coal}\right] \tag{4.3}$$

$$CSS = P_{Electricity} - \left[P_{Gas} \cdot \frac{1}{\rho_{Gas}} + P_{CO_2} \cdot E_{Gas}\right] \tag{4.4}$$

where P_{CO_2} stands for the price of carbon allowances per ton CO_2 and E_{Coal} and E_{Gas} are so-called emissions factors in tons CO_2 per MWh where $E_{Coal} = 0{,}86$ and $E_{Gas} = 0{,}36$[10] (*Tendances Carbone - No. 3*, 2007).

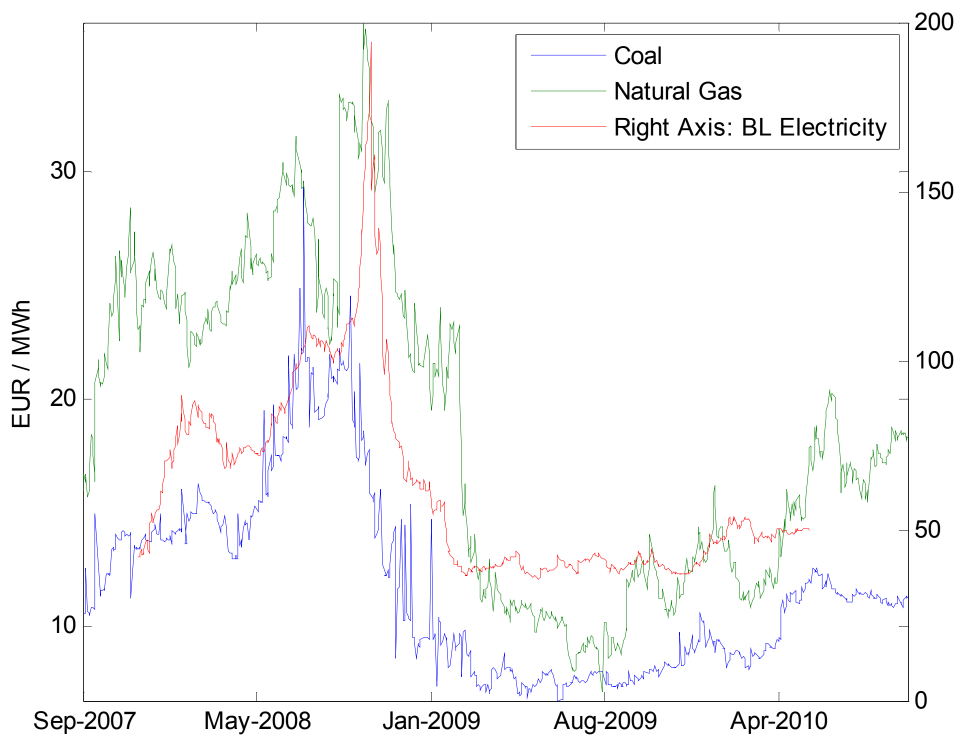

Figure 4.3: Historical natural gas, coal and base load electricity prices in Euros per megawatt hour (Source: Bloomberg).

Setting equation (4.3) equal to (4.4) and solving for P_{CO_2} the switching price of carbon can be derived:

[10] Hence gas-fired power plants are both more efficient than their coal-fired counterparts and pollute less.

$$P_{Switch} = \frac{\frac{P_{Coal}}{\rho_{Coal}} - \frac{P_{Gas}}{\rho_{Gas}}}{E_{Gas} - E_{Coal}} \tag{4.5}$$

Switching price is a fictional price representing equilibrium between clean dark spread and clean spark spread. It can be interpreted as the price of carbon where it becomes equally profitable to generate electricity by burning coal and by burning natural gas. It therefore represents the price of carbon above which it becomes more profitable in the short term to switch from coal fired production to the cleaner natural gas and below which it is more profitable to switch from natural gas to coal (*Tendances Carbone - No. 3*, 2007). Figure 4.4 shows the historical development of dark and spark spreads as well as clean dark and spark spreads.

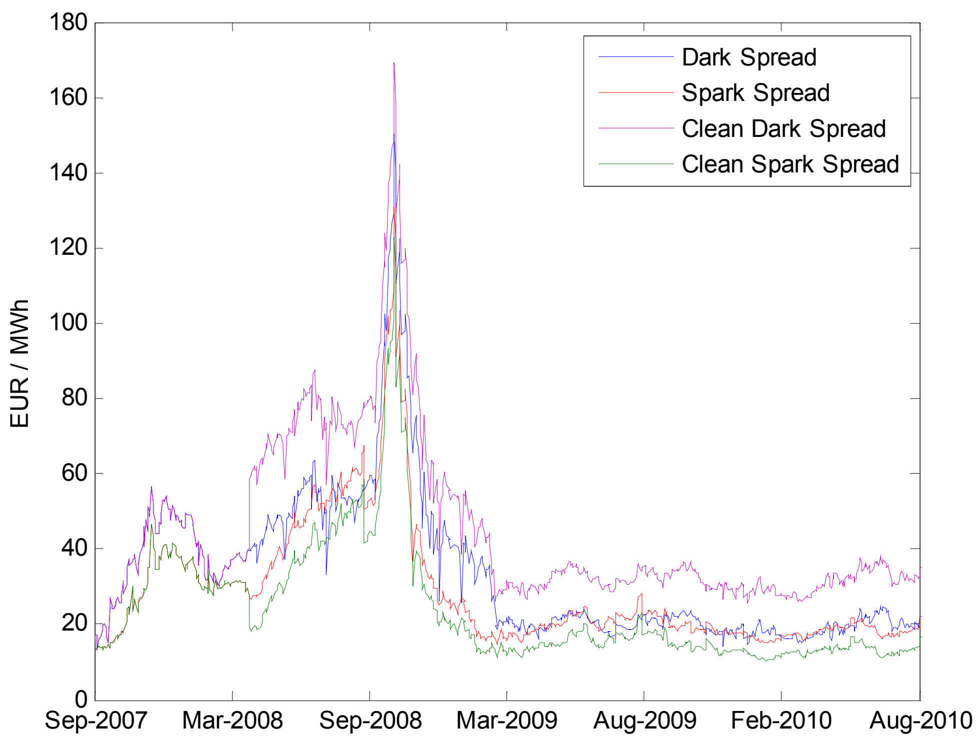

Figure 4.4: Historical dark spread, spark spread, clean dark spread and clean spark spread prices in Euros per megawatt hour (Source: Bloomberg).

By the theory of supply and demand, the demand for carbon allowances should decrease when more electric power producers choose to burn natural gas creating downward pressure on the price of carbon until it stabilizes around or above the implicit switching price. An assumption of switching price theory is stable electricity prices, but in reality they are not as can clearly be seen in figure 4.3. Many electricity producers simply stabilize their clean dark spreads and clean spark spreads by passing on the cost of carbon to consumers through higher energy prices (Obermayer, 2009).

Examining figure 4.5 below it is clear that the actual price of carbon does not follow the theoretical price, indicating that other factors than the fundamental power sector pricing theory heavily influence the price. Although carbon prices are considered volatile the switching price, based on the two commodities coal and natural gas, becomes even more unstable and does not appear to capture the price dynamics of EUAs.

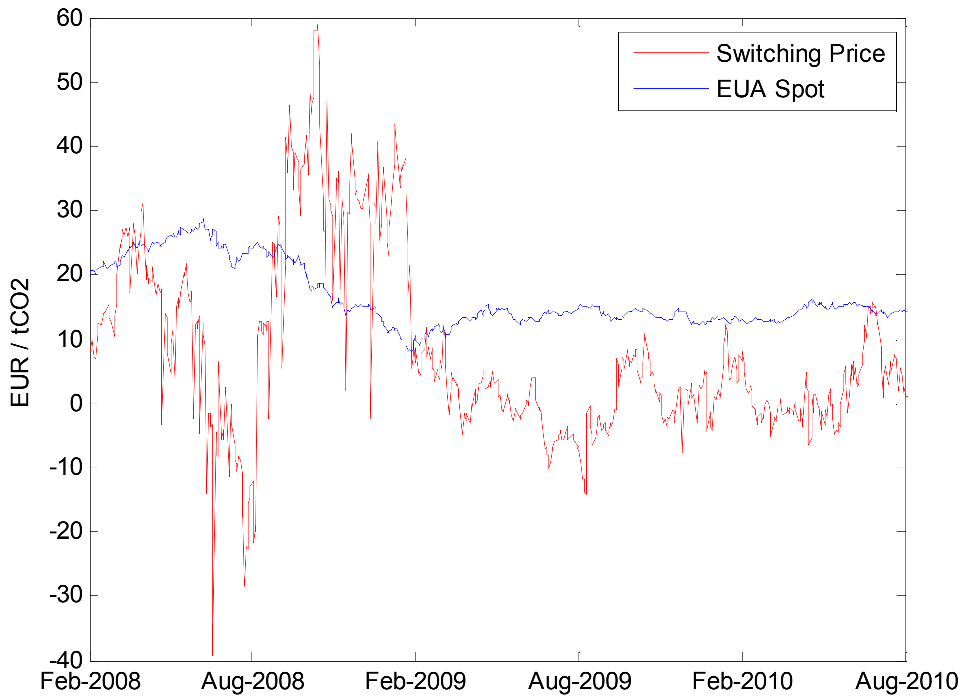

Figure 4.5: Switching price versus EUA spot price in Euros per ton CO_2 (Source: Bloomberg).

4.3 Data Description

The analysis aims at testing a large dataset of various time series using dimension reduction techniques to gain a-priory knowledge for regression modeling. It is known that traders targeting carbon closely follow energy, as well as commodities such as brent crude oil, gasoil, natural gas, coal and sometimes also equity indices. Table 4.2 shows the chosen data for the analysis, their source as well as the period available. The table also includes the derived time series described in equations (4.1)-(4.5). Since the aluminum industry will be included in phase III of the EU ETS, LME aluminum prices are included in the database. A weather index is also included since other analyses have concluded that weather may be a factor in EUA prices as mentioned in the introduction. Two datasets are then defined; the training data set, which spans the period from September 2008 to August 2009, and the test data set, which spans the one year period from October 2009 to October 2010.

Table 4.2: The data chosen for the analysis (Source: Bloomberg).

Time Series:	Exchange:	Ticker:	Quoted unit:	Data:
EUA Spot	Bluenext	PNXCSPOT Index	EUR/Metric ton	24.6.2005-22.10.2010
		PNXCSPT2 Index		
CER futures	ICE ECX	CARZ0 Comdty	EUR/Metric ton	14.3.2008-22.10.2010
WTI crude futures		EN1 Comdty		3.2.2006-22.10.2010
Gasoil futures		QS1 Comdty		4.1.2005-22.10.2010
Natural gas futures		FN1 Comdty	p/therms	4.1.2005-22.10.2010
Electricity Base Load		AT1 Comdty	GBP/MWh	4.1.2005-22.10.2010
Electricity Base Load		AT2 Comdty	GBP/MWh	4.1.2005-22.10.2010
Electricity Peak Load		AI1 Comdty	GBP/MWh	4.1.2005-22.10.2010
Electricity Peak Load		AI2 Comdty	GBP/MWh	4.1.2005-22.10.2010
Brent crude oil		CO1 Comdty		4.1.2005-22.10.2010
Coal		XWZ0 Comdty	USD/Metric ton	10.9.2007-22.10.2010
Aluminum Primary	LME	LMAHDS03 LME Cmdty	USD/ton	4.1.2005-22.10.2010
Aluminum Alloy	LME	LMAADS03 LME Cmdty	USD/ton	4.1.2005-22.10.2010
Dark Spread*			EUR/MWh	11.09.2007-22.10.2010
Clean Dark Spread*			EUR/MWh	11.09.2007-22.10.2010
Spark Spread*			EUR/MWh	4.1.2005-22.10.2010
Clean Spark Spread*			EUR/MWh	4.1.2005-22.10.2010
Switching Price*			EUR/Metric ton	11.09.2007-22.10.2010
Weather		UGHIE	Index	7.2.2008-22.10.2010
Nasdaq 100 Index		NDX Index	Index	4.1.2005-22.10.2010
Standard & Poor's 500 Index		SPX Index	Index	4.1.2005-22.10.2010
FTSE 100 Index		UKX Index	Index	4.1.2005-22.10.2010
Deutscher Aktien Index		DAX Index	Index	4.1.2005-22.10.2010
Compagnie des Agents de Change		CAC Index	Index	4.1.2005-22.10.2010
Amsterdam Stock Exchange		AEX Index	Index	4.1.2005-22.10.2010
Portuguese Stock Index		PSI20 Index	Index	4.1.2005-22.10.2010
USD/GBP		USDGBP	Currency	4.1.2005-22.10.2010
USD/EUR		USDEUR	Currency	4.1.2005-22.10.2010
EUR/GBP		EURGBP	Currency	4.1.2005-22.10.2010

**Calculated from other time series according to equations (4.1)-(4.5). "AT1 Comdty" is used for $P_{Electricity}$.*

5 Results

The chapter describes the results of the study. The first section is a statistical analysis of the data, including the first four moments, key correlations and the principal component analysis. The second section shows in-sample regression results, including base-line regressions, principal component regression and latent root regression. The third section tests the regression models on out-of-sample data and finally, the fourth section provides a summary of the results.

5.1 Statistical Analysis of phase II data

5.1.1 Data Distribution

The distribution of EUA returns is shown in table 5.1. The mean of the returns is negative over the duration of the period in question. The standard deviation is high relative to equity indices, indicating high volatility, and the positive skewness indicates that the distribution is skewed to the right, meaning that the right tail of the distribution is somewhat longer. The distribution has heavier tails than a normal distribution since the excess kurtosis is close to one.

Table 5.1: The first four moments of European Union Allowance Units for the training data set.

	Mean	Standard deviation	Skewness	Kurtosis	Excess kurtosis
EUA	-0.0015	0.0319	0.0913	3.9285	0.9285

A complete table, showing the first four moments for all variables used in the analysis is shown in table 1 of the appendix. Examining table 1 shows that switching price dominates the variability of the data with standard deviation of 1.9620. The next highest standard deviation is 0.0952 for coal. The variability of switching price is therefore over twenty times larger than the next highest variability.

A graphical representation of the results of table 5.1 are shown in figure 5.1. The histogram (left) shows the actual EUA returns with a fitted normal curve. The positive skew is not obvious but the heavier tails are more apparent. Plotting the sample quantiles of EUAs (right) versus theoretical quantiles from a normal distribution also shows that the returns are not normally distributed. The slight s-shaped curve formed by the EUA returns indicates heavier tails.

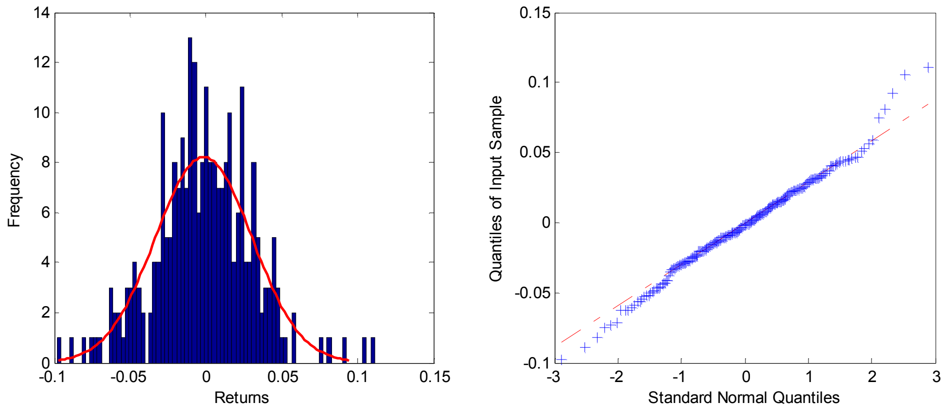

Figure 5.1: A histogram of EUA spot returns with a fitted normal PDF (left) and a qq-plot of the sample quantiles of EUAs shown in blue versus theoretical quantiles from a normal distribution, shown in red (right).

5.1.2 Correlation Analysis

Correlations of the sample data were analyzed for same-day as well as lagged values. Lagged values were 1 business day, 2 business days, 1 business week and 1 business month. Correlations were low for values lagged by one business day and onward. Top ten correlations to EUA returns are shown in table 5.2.

Table 5.2: Top ten correlations to EUA returns by lag-value, in order of decreasing value.

Same-day		Lag-1 business day		Lag-2 business days		Lag-1 business week	
EUA	1.00	EUA	1.00	EUA	1.00	EUA	1.00
CER	0.91	CER	0.17	SPX	-0.19	SS	0.09
CAC	0.40	NDX	0.17	NDX	-0.18	Weather	-0.09
PSI20	0.40	SPX	0.15	Brent crude	-0.17	PSI20	-0.09
DAX	0.38	Weather	-0.14	CER	-0.15	CSS	0.08
Gasoil	0.37	EUR/GBP	0.09	UKX	-0.15	Brent crude	-0.08
UKX	0.37	Gasoil	-0.09	AEX	-0.15	UKX	-0.08
WTI crude	0.36	WTI crude	0.08	WTI	-0.14	Electr. PL 2	0.07
Brent crude	0.32	Coal	0.07	Gasoil	-0.14	DS	0.07
CDS	0.32	Natural gas	0.07	Natural gas	-0.13	CDS	0.07
AL Primary	0.27	USD/GBP	0.07	DAX	-0.12	CAC	-0.07

The same-day correlations are strongest, especially the correlation between EUAs and CERs, reaching over 90 percent. The next three highest correlations are all equity indices. Moving to lag-1 business day, the correlations have fallen substantially. CERs, which formerly had the greatest correlation to EUAs now show low correlation. Reaching the lag of one business week all correlations are weak.

Tables 2-6 of the appendix show the complete top ten correlations and inter-correlations for the phase II sample data for same-day, lag-1 business day, lag-2 business days, lag-1 business week and lag-1 business month.

The autocorrelation of EUAs was also examined. As shown in figure 5.2 no substantial autocorrelation was found. Most points are within the confidence limit and no pattern is visible. The autocorrelation of the remainder of the training data set was examined using a visual inspection as seen in the figure below. No substantial autocorrelation was found.

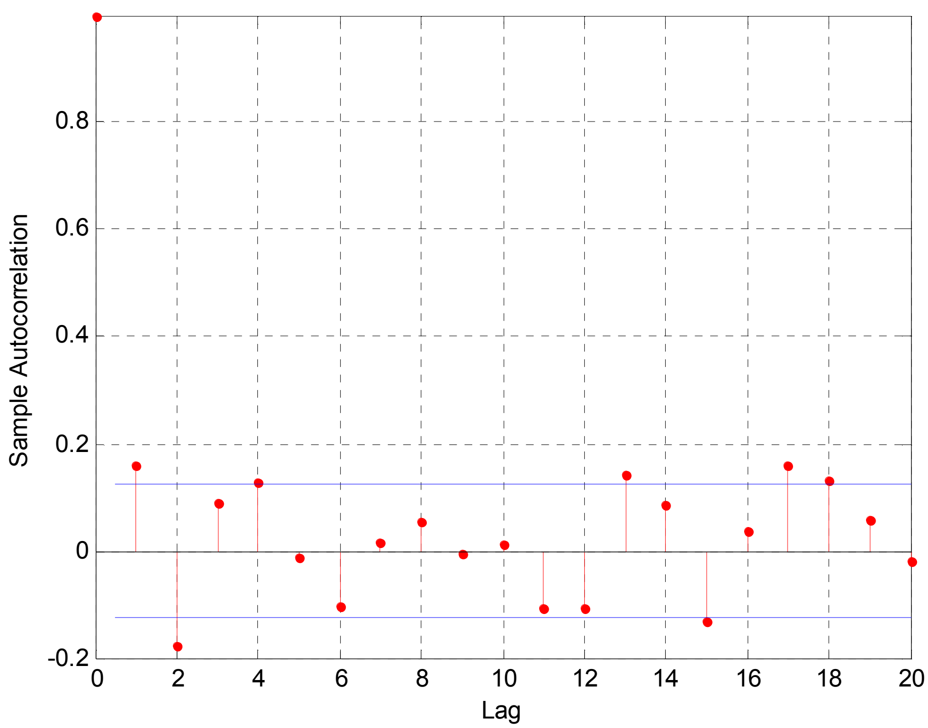

Figure 5.2: Autocorrelation of EUA returns from zero to twenty lags. The red dots indicate the value of the correlation and the blue lines represent the confidence level of two standard deviations or approximately 95%.

Based on correlations the top five variables of each lag were chosen as input variables into a baseline regression model, i.e. the top five correlations to EUA returns of each column in table 5.2. Regression results are explained in section 5.2.1.

5.1.3 Principal Component Analysis of phase II data

A principal component analysis was performed on the training data set. However, switching price had to be excluded from the analysis. As mentioned in section 5.1.1, the standard deviation of switching price was very large, dominating all other variables. This became apparent in the PCA since switching price, if included, was responsible for over 99% of the variability of the first PC. Regressing using only switching price did not yield reliable results and it was therefore eliminated from the PCA and regression analysis.

Table 5.3 shows results for the PCA excluding switching price. The dominant variable in each PC is the variable having the highest coefficient in the linear combination which defines the PC, i.e. the largest contributor to that particular PC. The table then shows the corresponding latent value and the cumulative proportion of total variance.

Table 5.3: PCA results.

Principal Component	Dominant Variable	Coefficient	Latent Value	Cumulative %Variance
1	DS	0.608044	0.020094	0.450114
2	CSS	0.583827	0.007582	0.619960
3	WTI crude	0.435572	0.006313	0.761384
4	Natural gas	0.556801	0.003033	0.829334
5	WTI crude	0.554578	0.001611	0.865421
6	Weather	0.779301	0.001039	0.888703
7	Weather	0.489392	0.000833	0.907371
8	Electricity PL AI2	0.479754	0.000738	0.923899
9	CER	0.622906	0.000705	0.939685
10	AL Alloy	0.523707	0.000561	0.952254
11	Gasoil	0.596883	0.000490	0.963231
12	Electricity PL AI1	0.448033	0.000399	0.972169
13	Brent crude	0.643844	0.000215	0.976974
14	Electricity PL AI1	0.623862	0.000204	0.981537
15	Electricity BL AT2	0.741246	0.000184	0.985655
16	USD/GBP	0.571120	0.000126	0.988488
17	DAX	0.631926	0.000113	0.991011
18	AL Primary	0.585925	0.000108	0.993432
19	PSI20	0.652388	0.000084	0.995315
20	EUR/GBP	0.499297	0.000068	0.996848
21	UKX	0.463253	0.000037	0.997679
22	UKX	0.531602	0.000033	0.998419
23	NDX	0.558473	0.000029	0.999073
24	CAC	0.819821	0.000021	0.999548
25	Electricity BL AT1	0.719882	0.000017	0.999934
26	SS	0.670489	0.000003	0.999998
27	EUR/GBP	0.583759	0.000000	1.000000

Principal component 1 (PC1) accounted for over 45% of the variation of the data. The next three PCs accounted for approximately 38%. Looking at the first four PCs over 80% of the variation of the data was accounted for. The most dominant variables in the first five PCs were therefore dark spread, clean spark spread, WTI crude and natural gas. These variables were used as input to the principal component regression which will be covered in section 5.2.2.

A graphical representation of the data with respect to the first two PCs is shown on a biplot in figure 5.3. The horizontal axis represents principal component 1. Dark spread is the most dominant variable with respect to this axis and also highly correlated to PC1 as the cosine of the angle between the PC direction and the eigenvector gives the correlation between the two as discussed in section 3.2. In terms of the first two PCs the biplot confirms the results of table 5.3.

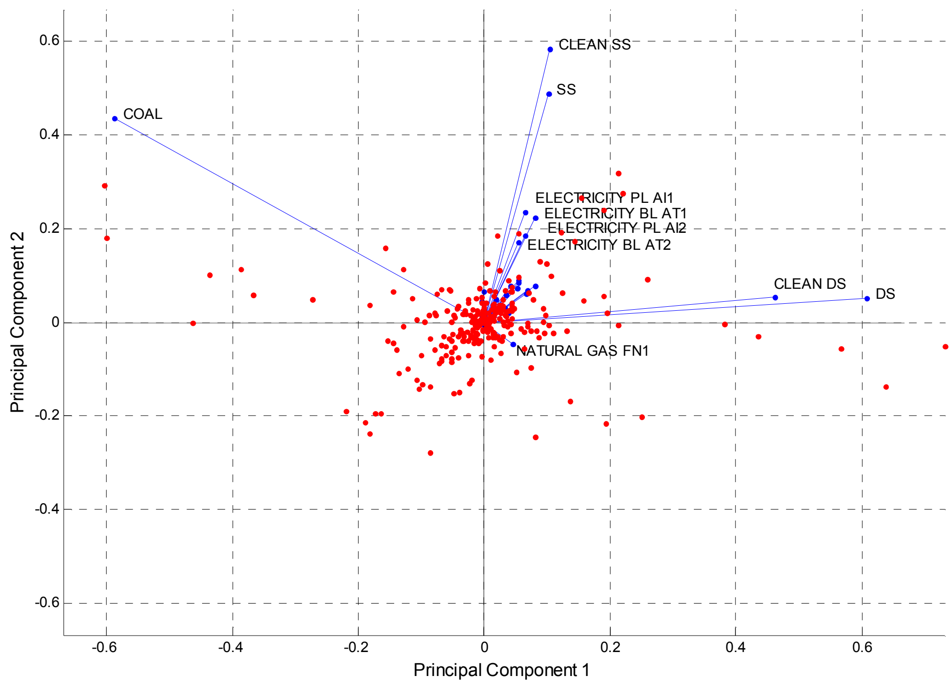

Figure 5.3: Biplot showing graphical PCA results for the first two PCs.

An interesting result is revealed in figure 5.3. The eigenvectors, apart from coal, seem to fall into two clusters near the two principal components. The first cluster consists of clean dark spread and dark spread. The second cluster groups spark spread and clean spark spread. The first two PCs therefore represent two different types of power plants; coal fired power plants versus the more environmentally friendly gas fired power plants. Zooming in on the figure reveals that both brent crude and WTI crude also lie near PC1, indicating variability in a similar direction as dark spread, high correlation with PC1. Also interesting is the fact that the returns for both peak load (PL) and base load (BL) electricity lie in between the first two PCs, although closer to the direction set by PC2, which is consistent with the fact that the UK generates more electricity from gas fired power plants than from coal fired power plants. Of UK's total electricity supplied in 2010, 103.2 TWh were generated from coal but 171.5

TWh from gas. (MacLeay, 2011)[11]. Looking at these two fuel sources, coal fired plants generate 37% of the energy created using these two fuel sources and natural gas covers the remaining 63%. Referring again to figure 5.3 this proportion is somewhat reflected as the energy returns cluster lies closer to PC2, the direction which was previously linked to gas fired power plants than to PC1, the direction previously linked to coal fired power plants.

Figure 5.4 shows principal component 1 against principal component 3. PC3 contains equity indices as well as materials such as aluminum, WTI crude, brent crude, gasoil and natural gas. PC2 lies straight into the page. Heavier oils and refined products follow PC1, which is consistent with the earlier definition of PC1, lead by dark spread and clean spark spread as discussed above.

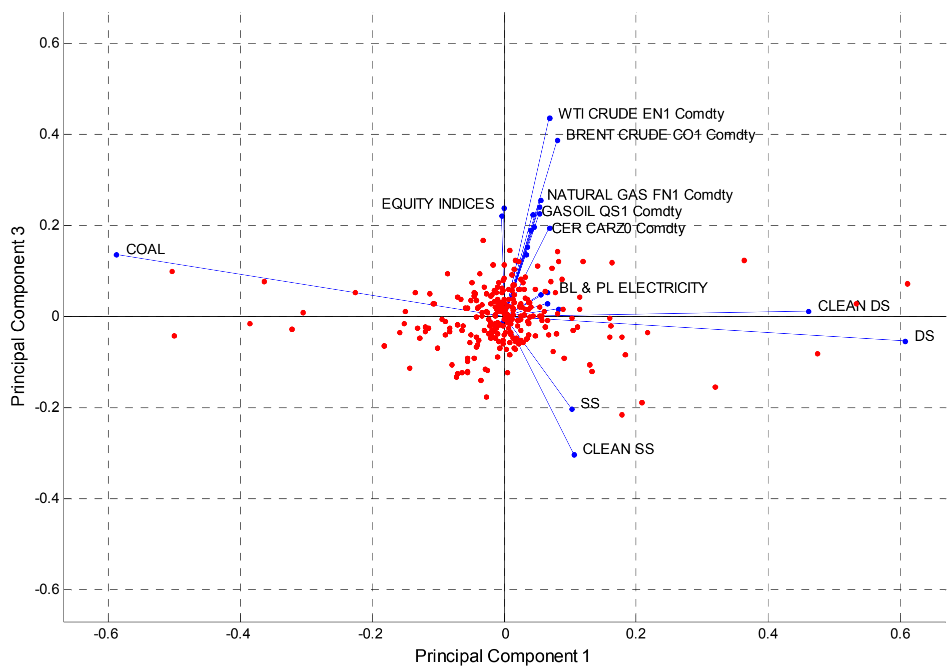

Figure 5.4: Biplot showing graphical PCA results for the first and third PC.

PCA was used to select variables for regression by dimension reduction based on the selection criteria explained in section 3.3.2. Variables chosen using simple PCA were the most dominant variables in the top five PCs of table 5.3 accounting for over 80% of the cumulative proportion of variance of the total dataset.

Five variables were also chosen using backward elimination. Table 5.4 shows the variables retained after the backward elimination process as well as their coefficients, latent values and cumulative proportion of variance.

[11] In 2010 47,3% of UK power was generated from gas, 28,4% from coal, 15,6% by nuclear plants, 6,9% from renewable sources and 1,8% from other sources of energy (MacLeay, 2011).

Table 5.4: PCA results using backward elimination.

Principal Component	Dominant Variable	Coefficient	Latent Value	Cumulative %Variance
1	WTI crude	0.7321	0.0152	0.6461
2	Coal	0.9024	0.0041	0.8213
3	DS	0.9588	0.0021	0.9115
4	CSS	0.6569	0.0012	0.9609
5	Weather	0.9845	0.0009	1.0000

Table 7 of the appendix shows the order in which the variables were eliminated.

The forward selection process was repeated until five variables were obtained. Table 5.5 shows the order of which variables were selected using the forward selection method.

Table 5.5: The order of variables generated by forward selection PCA.

Principal Component	Dominant Variable	Coefficient	Latent Value	Proportion of Variance*
1	DS	0.6080	0.0201	0.4501
2	Coal	0.7687	0.0129	0.3509
3	CDS	0.5127	0.0089	0.3195
4	CSS	0.5841	0.0074	0.3164
5	WTI crude	0.4326	0.0068	0.3440

**Note that this is not the cumulative percentage of variance, but the proportion of variance for the PC in each iteration.*

Table 8 of the appendix shows the cumulative proportion of variance based on the above results.

Examining the results of the dimension reductions shown in the above tables for simple PCA, backward elimination PCA and forward selection PCA, reveals that three variables are included in all cases, i.e. dark spread, clean spark spread and WTI crude. The fourth and fifth variables are then natural gas, coal, weather or clean dark spread depending on the type of dimension reduction method used. Of the variables listed above, the top five variables chosen using correlation as the dimension reduction technique shown in table 5.2 only included weather.

5.2 Predicting Phase II Prices

5.2.1 Base-Line Regressions

As a mean of comparison the base-line regressions were set using correlation as the dimension reduction technique. The top five correlations were used in each case as previously discussed in section 5.1.2. Regressions were done on same-day data as well as data lagged by one day. The same-day results are shown in table 5.6.

Table 5.6: Base-line regression results for same-day training data.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-6.7119 E-06	-0.0082	0.9935	-0.0082	0.9935
CER	0.8874	30.5127	4.2736 E-86	27.3495	5.6866 E-77
CAC	0.0446	0.5004	0.6172	0.4133	0.6797
PSI20	0.0286	0.3746	0.7083	0.3322	0.7400
DAX	0.0132	0.1780	0.8588	0.1296	0.8970
Gasoil	0.0534	1.7734	0.0774	1.8723	0.0623
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	1.7134 E-04	83.50%	83.17%	-2,161	-2,062

The only variable whose adjusted p-value (HC4) was within the confidence interval of 95% (p-value below 0.05) was CER. The regression was therefore repeated using only CERs plus a constant as regressors. The results are shown in table 5.7 below.

Table 5.7: Base-line regression results for same-day training data after adjustment.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0015	-0.0785	0.9375	-0.0789	0.9372
CER	0.9296	34.8001	1.9622E-98	33.8877	4.9931E-96
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	1.7668E-04	82.72%	82.65%	-2,150	-2,050

There seems to be a strong same-day relationship between EUA returns and CERs as the model captures over 80% of the variability of EUAs. The residuals of the model were stationary but did not pass a normality test and a qq-plot showed heavy set tails, indicating higher probability of extreme events. They also showed some degree of heteroskedasticity. The adjusted t-statistic (HC4) is however heteroskedasticity consistent. No signs of autocorrelation could be detected. Figure 5.5 below shows the results of table 5.7 graphically. Since the EUA returns are not strictly normal and some number of outliers is present in the heavy tails, the above results were compared to the results of a robust regression. The robust regression yielded similar results and the influence of extremities therefore rejected.

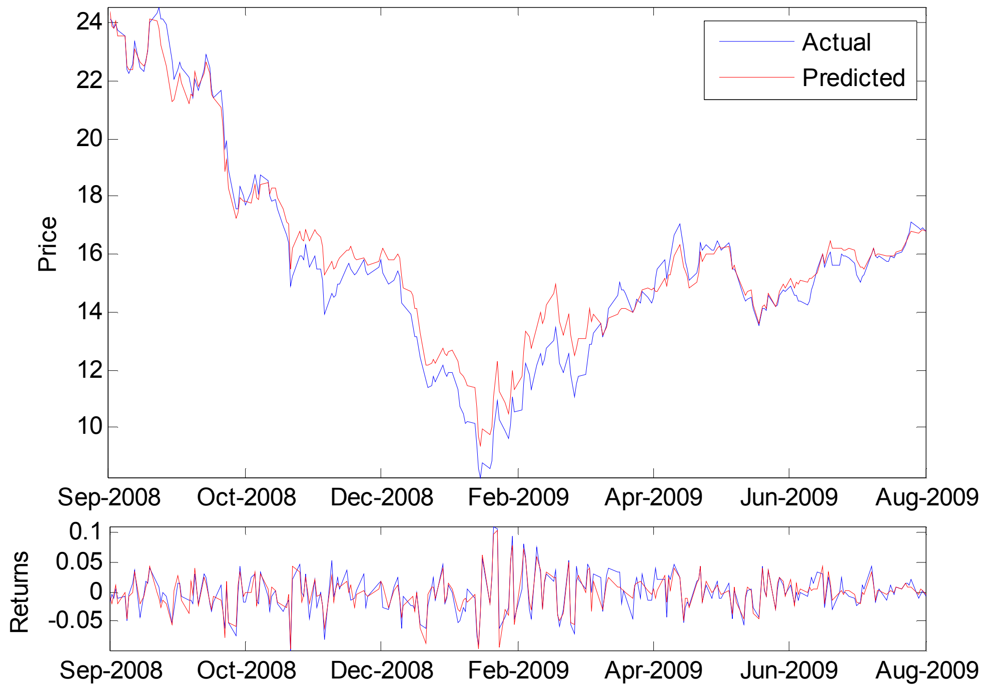

Figure 5.5: The actual vs. predicted same-day price and returns for EUAs using the base-line model.

Regressing on data which had been lagged by one business day yielded the result shown in table 5.8.

Table 5.8: Base-line regression results for lag-1 day training data.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0016	-0.8109	0.4182	-0.7847	0.4334
CER	0.1695	2.5940	0.0101	1.7320	0.0845
NDX	0.3566	1.5105	0.1322	1.4405	0.1510
SPX	-0.2239	-0.9536	0.3412	-0.8435	0.3997
Weather	-0.1510	-2.3477	0.0197	-2.0236	0.0441
EUR/GBP	0.3317	1.2980	0.1955	1.1452	0.2532
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	9.6027 E-04	7.84%	5.98%	-1,714	-1,616

The only variable whose p-value (HC4) was within the confidence interval of 95% (p-value below 0.05) was weather. The regression was therefore repeated using only weather plus a constant as regressors. The results are shown in table 5.9 below.

Table 5.9: Base-line regression results for lag-1 day training data after adjustment.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0017	-0.8385	0.4026	-0.8400	0.4017
Weather	-0.1502	-2.3048	0.0220	-2.1340	0.0338
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	1.0043 E-03	2.06%	1.68%	-1,699	-1,600

Residuals of the model were normally distributed, stationary and no autocorrelation was detected. As before, heteroskedasticity was present when residuals were investigated as a function of time. Figure 5.6 shows the results of table 5.9 graphically.

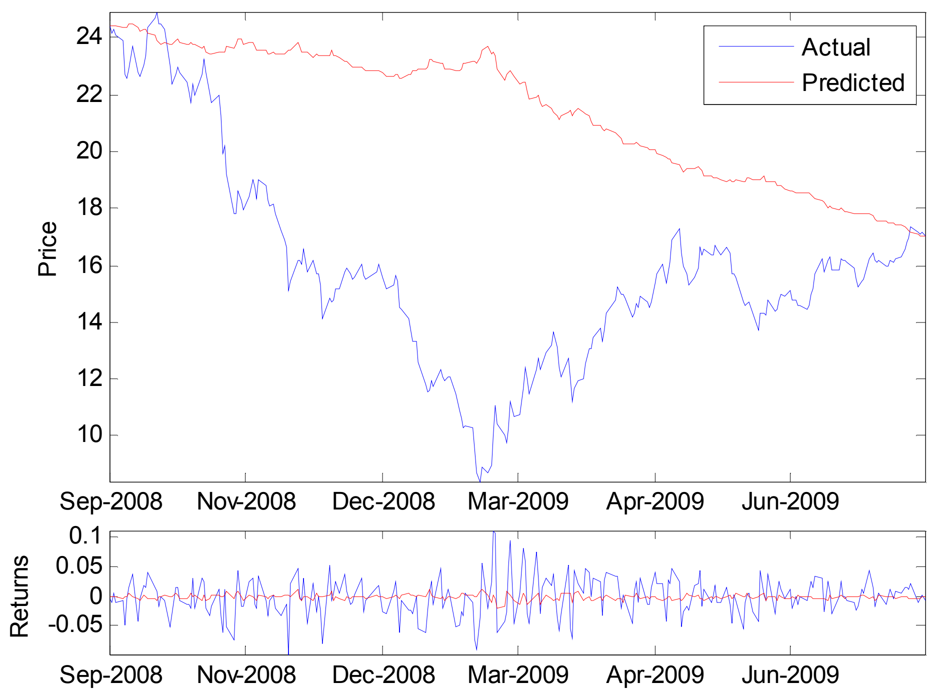

Figure 5.6: The actual vs. predicted lag-1 day price and returns for EUAs using the base-line model.

As shown in the figure the relationship previously present has completely disappeared when moving to data lagged by one business day. Regressions were also done for lagged values of one business week and one business month. The relationship between EUAs and the regressors continued to deteriorate as the lagged interval increased and no predictive value was found to be present in the data. These results will therefore be omitted for all methods.

5.2.2 Principal Component Regression

Using the three dimension reduction procedures mentioned in section 5.1.3 principal component regression (PCR) was used to model same-day relationships and relationships lagged by one day between EUA returns and the chosen regressors. Table 5.10 shows the results of the regression performed on the variables chosen using simple PCA.

Table 5.10: Simple PCA regression results for same-day training data.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0015	-0.8124	0.4174	-0.7949	0.4274
DS	-0.0630	-3.1224	0.0020	-2.6867	0.0077
CSS	0.1281	4.2957	2.4951E-05	3.4906	0.0006
WTI crude	0.1946	4.6840	4.6201E-06	3.4992	0.0006
Natural gas	0.0077	0.1222	0.9029	0.0923	0.9265
Goodness of fit	MSE	R^2	R^2_{adj}	AIC	BIC
	8.6177 E-04	16.71%	15.38%	-1,748	-1,649

The variables whose p-values (HC4) were within the confidence interval of 95% (p-value below 0.05) are dark spread (DS), clean spark spread (CSS) and WTI crude. The regression was therefore repeated using these variables. The results are shown in table 5.11.

Table 5.11: Simple PCA regression results for same-day training data after adjustment.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0015	-0.8133	0.4168	-0.8014	0.4237
DS	-0.0606	-2.9983	0.0030	-2.5925	0.0101
CSS	-0.1446	-4.5275	9.2257E-06	-3.7328	0.0002
WTI crude	0.1878	4.5163	9.6882E-06	3.4548	0.0006
Goodness of fit	MSE	R^2	R^2_{adj}	AIC	BIC
	8.5968 E-04	16.58%	15.58%	-1,748	-1,649

The same day relationship based on variable selection using simple PCA does not outperform the simple base-line model. The base-line model gives much stronger results in all goodness of fit parameters. The mean squared error is higher, R^2 and R^2_{adj}, are over 70 percentiles lower and AIC and BIC show higher numerical values, indicating a lower quality model. The residuals did however uphold all the necessary criteria; they were normally distributed, stationary and no autocorrelation was present. A slight heteroskedasticity was again visible, but using the adjusted t-statistic and its relevant p-value the results became statistically reliable. Figure 5.7 shows the results of table 5.11 graphically.

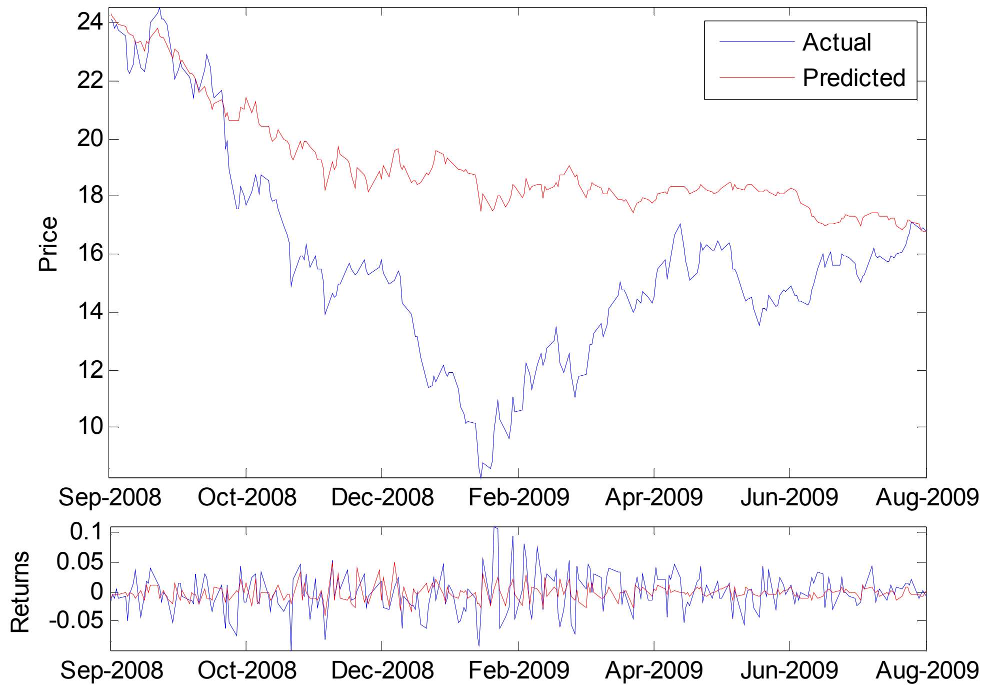

Figure 5.7: The actual vs. predicted same-day price and returns for EUAs using the simple PCA model.

The results for data lagged by one day are shown in table 9 in the appendix. As in the earlier case, no relationship was visible after moving to data lagged by one business day.

Table 5.12 shows results of the regression based on variables chosen by backward elimination.

Table 5.12: PCA using backward elimination regression results for same-day training data.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0015	-0.8285	0.4082	-0.8003	0.4243
WTI crude	0.0428	2.9233	0.0038	2.0107	0.0454
Coal	-0.0501	-1.7834	0.0757	-1.5189	0.1301
DS	0.2474	6.3124	1.2507E-09	5.1552	5.1650E-07
CSS	0.0674	1.2728	0.2043	0.8516	0.3953
Weather	-0.1883	-3.1644	0.0017	-2.1710	0.0309
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	8.2851 E-04	20.24%	18.64%	-1,760	-1,660

The variables whose p-values (HC4) were within the confidence interval of 95% (p-value below 0.05) are WTI crude, dark spread (DS) and weather. The regression was therefore repeated using these variables. The results for the adjusted regression are shown in table 5.13.

Table 5.13: PCA using backward elimination regression results for same-day training data after adjustment.

Variable	β	t-stat	p-value	HC4	p $_{HC4}$-value
Constant	-0.0015	-0.8169	0.4147	-0.8002	0.4243
WTI crude	0.0757	3.6626	0.0003	3.1041	0.0021
DS	-0.2288	-5.6386	4.6038 E-08	-4.4760	1.1547 E-05
Weather	0.1626	2.7072	0.0073	1.7969	0.0736
Goodness of fit	MSE 8.5217 E-04	R^2 17.31%	R^2_{adj} 16.32%	AIC -1,750	BIC -1,651

Residuals were normally distributed, stationary and heteroskedastic. No autocorrelation was detected. Again comparing to the base-line model of section 5.2.1 the base-line model outperforms the regression based on backward elimination shown in the table above. The mean squared error (MSE) of the base-line model is lower than the MSE of the regression based on backward elimination, R^2 and R^2_{adj} are higher and both AIC and BIC are lower, all indicating a stronger model set by the base-line regression. Figure 5.8 shows the results of table 5.13 graphically.

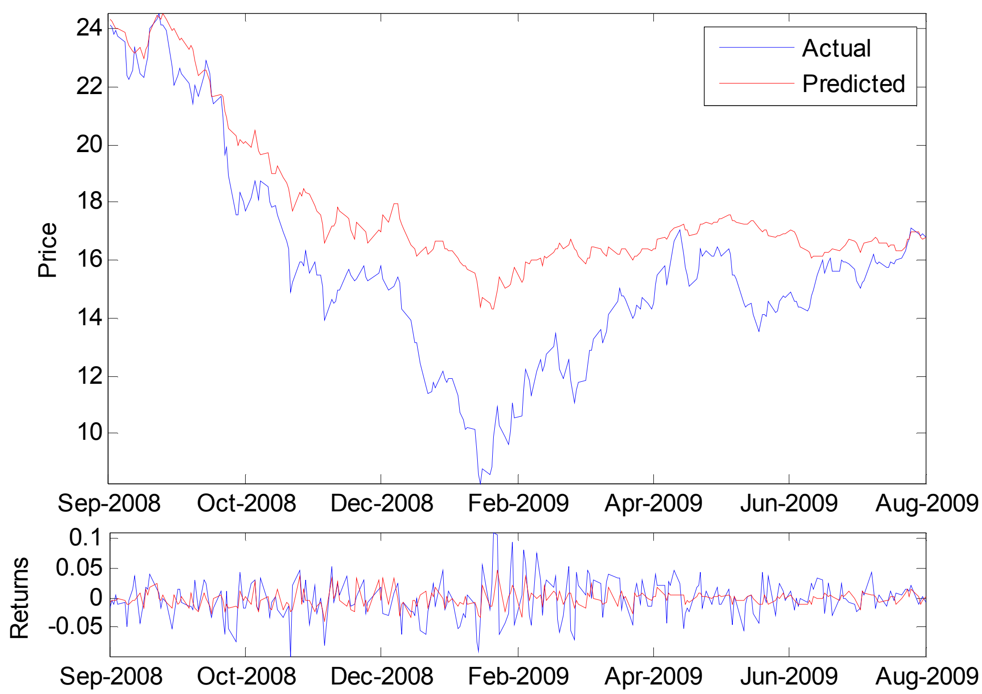

Figure 5.8: The actual vs. predicted same-day price and returns for EUAs using the backward elimination PCA model.

The results for data lagged by one day are shown in table 10 in the appendix. No relationship was visible after moving to data lagged by one business day.

Table 5.14 shows results of the regression based on variables chosen by forward selection.

Table 5.14: PCA using forward selection regression results for same-day training data.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0015	-1.3658	0.1732	-1.2659	0.2067
DS	-0.0467	-5.9145	1.0935 E-08	-2.5616	0.0110
Coal	-0.0363	-2.1600	0.0317	-1.1611	0.2467
CDS	0.2673	11.4291	0.0000	8.1374	1.9318
CSS	-0.0997	-3.3394	0.0010	-2.0184	0.0446
WTI crude	2.6300	20.4457	0.0000	9.7707	0.0000
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	3.0487 E-04	70.65%	70.06%	-2,014	-1,915

The variables whose p-values (HC4) were within the confidence interval of 95% (p-value below 0.05) are dark spread (DS), clean spark spread (CSS) and WTI crude. The regression was therefore repeated using these variables. The results are shown in table 5.15.

Table 5.15: PCA using forward selection regression results for same-day training data after adjustment.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0015	-0.8133	0.4168	-0.8014	0.4237
DS	-0.0606	-2.9983	0.0030	-2.5925	0.0101
CSS	-0.1446	-4.5275	9.2257 E-06	-3.7328	0.0002
WTI crude	0.1878	4.5163	9.6882 E-06	3.4548	0.0006
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	8.5968 E-04	16.58%	15.58%	-1,748	-1,649

Residuals were normally distributed, stationary and heteroskedastic. No autocorrelation was detected. By eliminating the variables which were not statistically significant the coefficient of determination fell by over 50 percentiles, the mean squared error increased and the AIC and BIC increased in value, all indicating a lower quality model. The results of table 5.15 therefore do not outperform the base-line model of section 5.2.1. The results are shown graphically in figure 5.9 below.

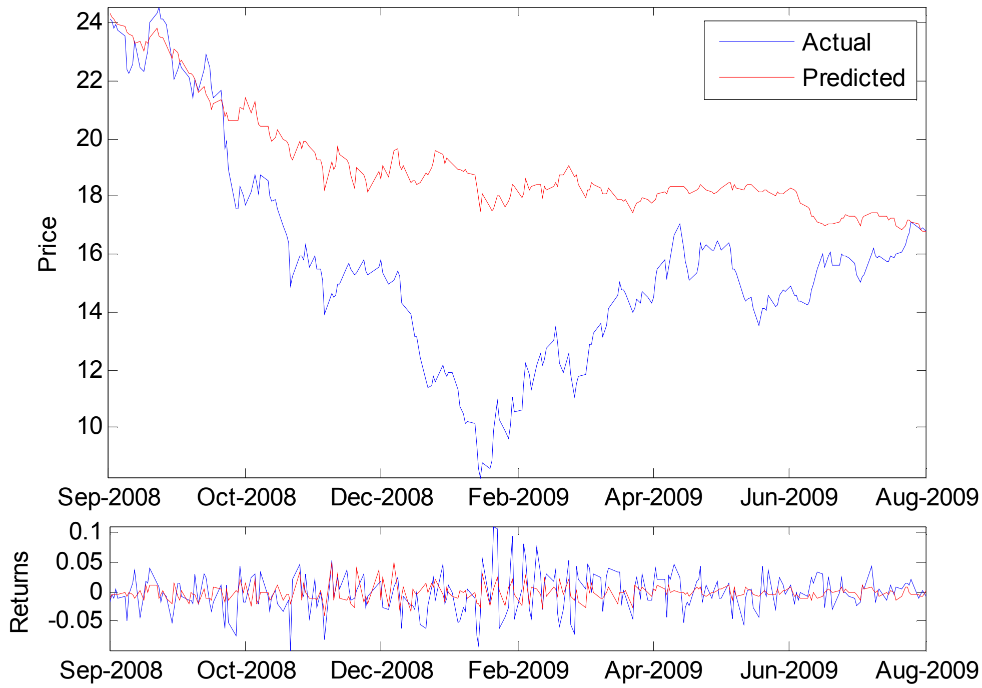

Figure 5.9: The actual vs. predicted lag same-day price and returns for EUAs using the forward selection PCA model.

The results for data lagged by one day are shown in table 11 in the appendix. No relationship was visible after moving to data lagged by one business day.

5.2.3 Biased estimators using Latent Root Regression

As a final comparison to the base-line model, latent root regression (LRR) was performed on same-day training data and training data lagged by one day. Components were considered non-predictive near-singularities if the relevant eigenvalue was smaller than 0.05 and the coefficient in the linear combination was smaller than 0.10. Non-predictive near singularities were iteratively eliminated and near singularities passing the above criteria retained. The result of the latent root regression for same-day data is shown in table 5.16 and a graphical representation is shown in figure 5.10.

None of the parameters in table 5.16 were statistically significant at the 95% confidence interval which is no surprise as the parameters are purposefully biased. The latent root model does however not exceed the standard set by the base-line model and should therefore be rejected.

Table 5.16: Latent root regression results for same-day training data.

PC No.	β	t-stat	p-value	HC4	p_{HC4}-value
3	-0.0043	-8.6554	6.6613 E-16	-0.0697	0.9445
6	0.0025	1.1682	0.2439	0.0086	0.9932
7	0.0063	5.7089	3.2765 E-08	0.0585	0.9534
8	0.0031	2.2743	0.0238	0.0167	0.9867
19	0.0667	47.3027	3.3661 E-125	0.4882	0.6259
20	0.2378	155.1403	1.3049 E-246	1.4327	0.1532
21	0.0264	48.4720	1.5095 E-127	0.3007	0.7639
22	0.0395	59.0875	5.8850 E-147	0.3436	0.7315
24	0.0471	86.9719	3.4909 E-186	0.6832	0.4951
26	0.1283	60.9882	4.1173 E-150	0.6135	0.5401
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	9.7693 E-04	6.03%	2.58%	-1,758	-1,722

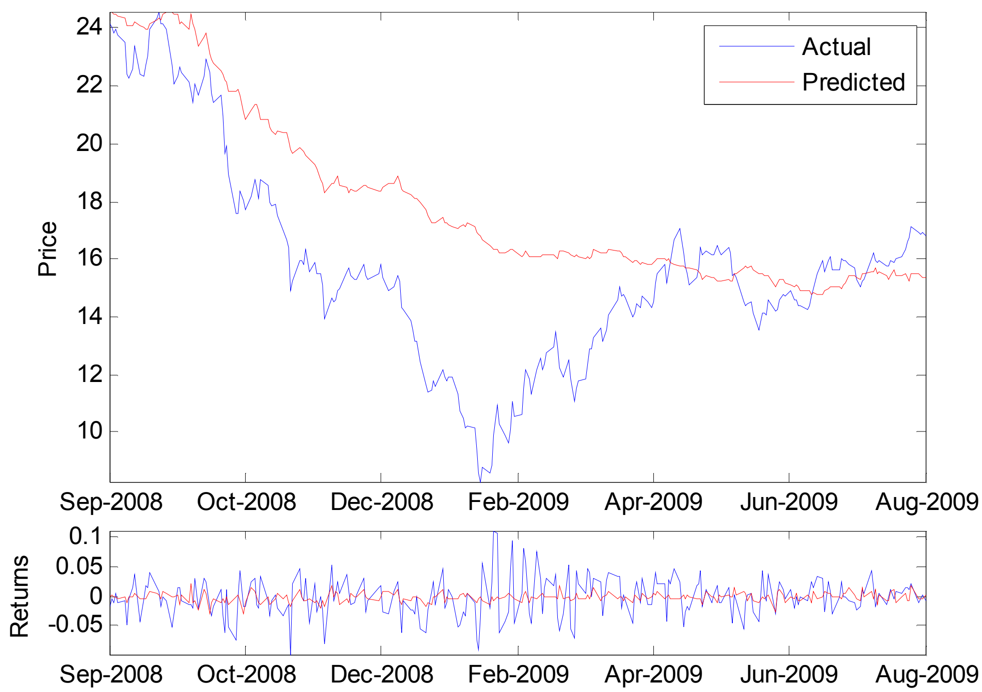

Figure 5.10: The actual vs. predicted same-day price and returns for EUAs using the latent root model.

Regressing on data which had been lagged by one business day yielded the results shown in table 5.17 and a graphical representation of the results is shown in figure 5.11 below.

Table 5.17: Latent root regression results for lag-1 day training data.

PC No.	β	t-stat	p-value	HC4	p_{HC4}-value
6	0.1416	0.4336	0.0000	1.3732	0.1709
7	-0.1185	-0.3660	0.0000	-1.2019	0.2305
10	-0.1319	-0.7739	0.0000	-2.0235	0.0441
11	0.2010	1.0088	0.0000	2.0476	0.0416
12	0.0262	0.1081	0.0000	0.2562	0.7980
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	1.1690 E-03	5.67%	4.15%	-1,710	-1,692

Two of the parameters of table 5.17 were statistically significant at the 95% confidence interval. Again it should be kept in mind that these parameter estimates are biased and therefore not BLUE. The latent root model for data lagged by one business day exceeds the standard set by the adjusted base-line model for data lagged by one business day and yields similar results as the original unadjusted base-line model shown in table 5.8.

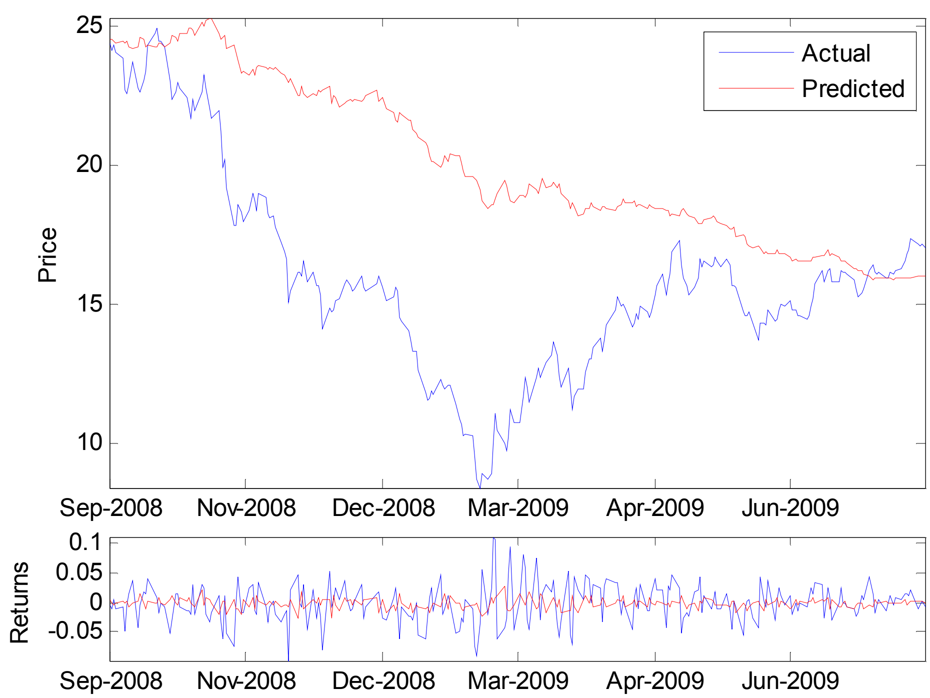

Figure 5.11: The actual vs. predicted lag-1 day price and returns for EUAs using the latent root model.

5.3 Testing Out-of-Sample

5.3.1 Base-Line Regression Models

Based on the results in sections 5.2.1 - 5.2.3 the base-line models performed best. To test their ability, out-of-sample predictions were made using the models presented in table 5.7 and table 5.9 on predefined test-data spanning one year from October 2009 to October 2010. The out-of-sample results for same-day predictions of the base-line model are shown in table 5.18 and a graphical representation is shown in figure 5.12.

Table 5.18: Out-of-sample regression results for same-day test data for the base-line model.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0015	-.09705	0.3327	-1.7736	0.0000
CER	0.9296	11.2515	0.0000	0.3183	0.2188
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	3.9635 E-05	88.47%	88.42%	-2,755	-2,748

The model's performance holds out-of-sample, however, based on figure 5.12, the model performs well on a short horizon during the first two to three months of the test-period. After that the model seems to overestimate negative returns and underestimate positive returns.

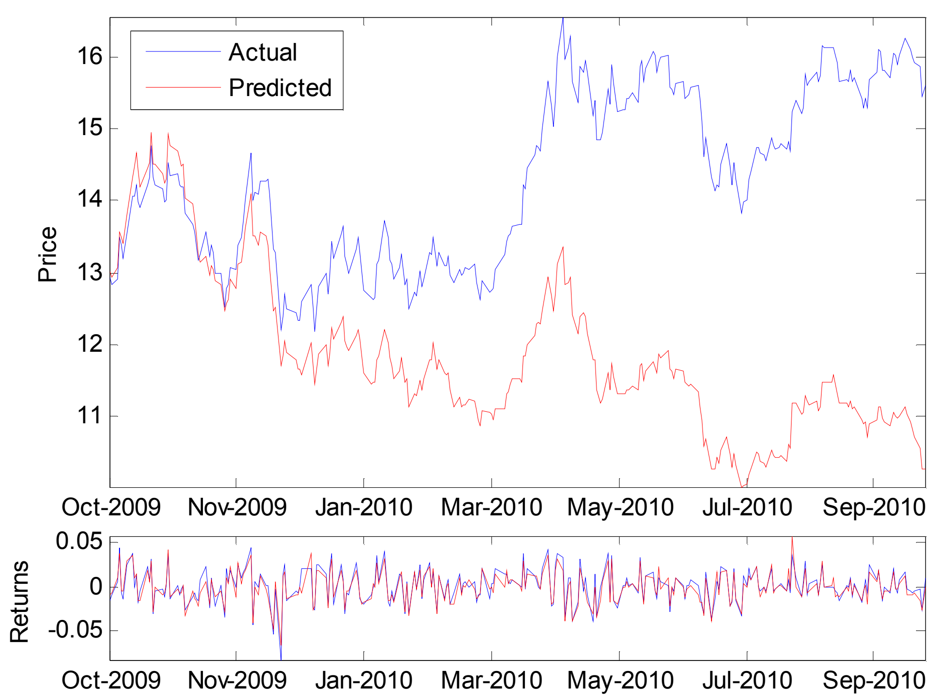

Figure 5.12: The actual vs. predicted same-day price and returns for out-of-sample EUAs using the base-line model.

The out-of-sample results for test data lagged by one day for the base-line model are shown in table 5.19.

Table 5.19: Out-of-sample regression results for lag-1 day test data for the base-line model.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0017	-0.6766	0.4992	-1.4434	0.1501
Weather	-0.1502	-1.2893	0.1984	-2.1558	0.0320
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	3.6781 E-04	4.38%	4.03%	-2,141	-2,134

All goodness of fit parameters show improvement, indicating the validity of the model itself. Examining figure 5.13 reveals that the model neither captures the trends of the EUA price nor the direction of its development. While the price of EUAs was slowly trending upwards, the predicted price steadily decreased for the duration of the test-period.

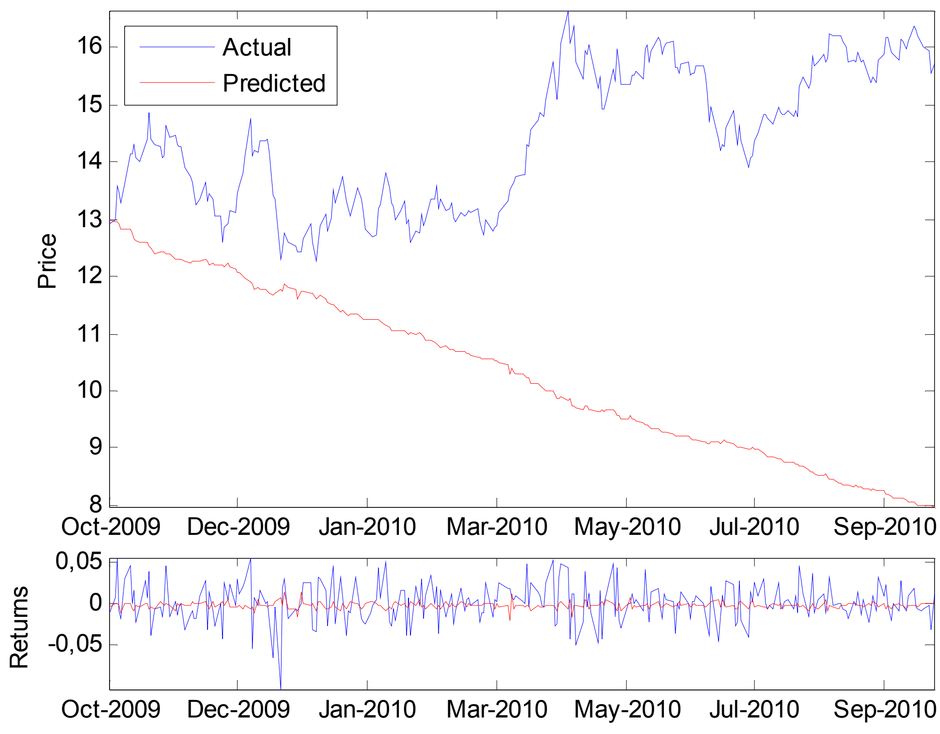

Figure 5.13: The actual vs. predicted lag-1 day price and returns for out-of-sample EUAs using the base-line model.

5.3.2 Principal Component Regression Models

Although the PCA models of section 5.2.2 did not outperform the base-line models, set by using correlation as the dimension reduction technique, their performance out-of-sample should be examined. Table 5.20 shows the out-of-sample results for same-day test data using the simple PCA model.

Table 5.20: Out-of-sample regression results for same-day test data for simple PCA.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0015	-0.7139	0.4759	-1.3034	0.1935
DS	-0.0606	-1.5237	0.1288	-2.2795	0.0234
CSS	-0.1446	-2.9385	0.0036	-3.5799	0.0004
WTI crude	0.1878	1.5922	0.1125	2.8809	0.0043
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	3.5277 E-04	14.86%	13.90%	-2,158	-2,144

Again all goodness of fit parameters show a deterioration, apart from AIC and BIC which have improved slightly. Examining figure 5.14 again shows that the model is incapable of following the trends of actual EUA prices.

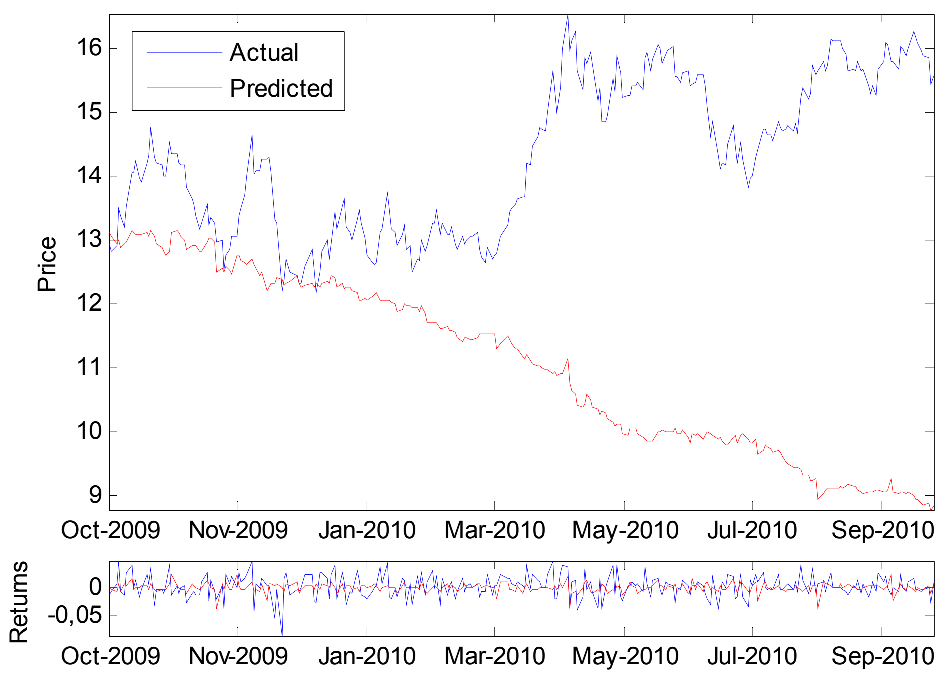

Figure 5.14: The actual vs. predicted same-day price and returns for out-of-sample EUAs using the simple PCA.

The results for data lagged by one day are shown in table 12 in the appendix. No relationship was visible after moving to data lagged by one business day.

The backward elimination PCA model was tested and results are shown in table 5.21. All goodness of fit parameters have deteriorated and the model does not capture the overall trends of the EUA price as shown in figure 5.15. As with the out-of-sample testing of the base-line model, the backward elimination PCA model seems to capture the price changes to some extent on the short horizon, during the first two or three months, but then fails to model the actual price trend.

Table 5.21: Out-of-sample regression results for same-day test data for PCA using backward elimination.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0015	-0.6598	0.5099	-1.0347	0.3017
WTI crude	0.0757	0.5933	0.5535	0.9348	0.3507
DS	-0.2288	-5.3209	0.0000	-4.5553	0.0000
Weather	0.1626	1.5361	0.1257	2.0086	0.0456
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	5.5658 E-04	22.67%	21.80%	-2,034	-2,020

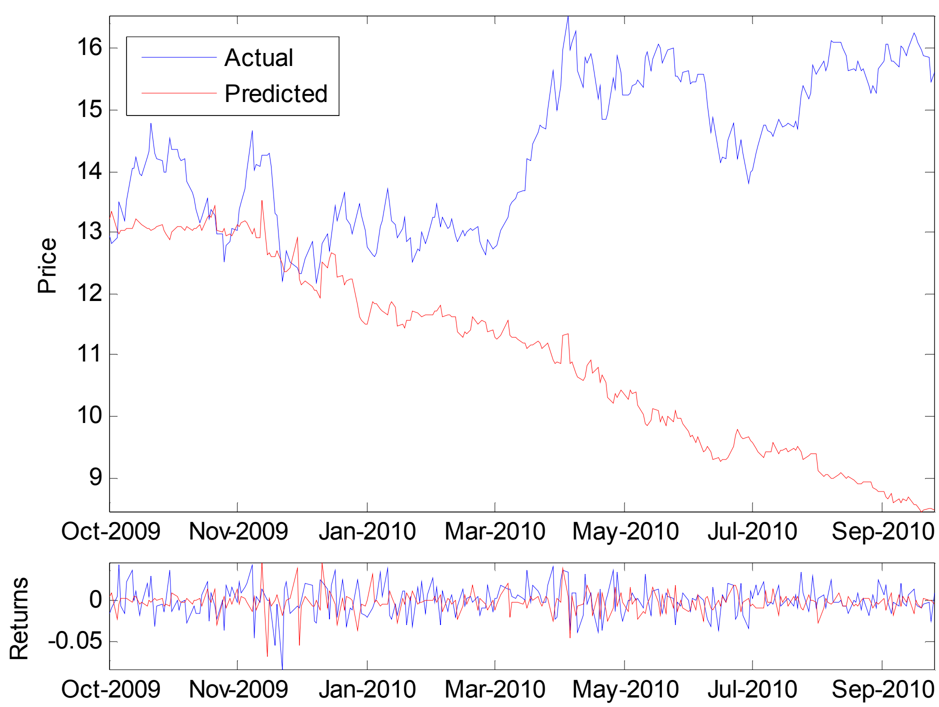

Figure 5.15: The actual vs. predicted same-day price and returns for out-of-sample EUAs using the backward elimination PCA.

The results for data lagged by one day are shown in table 13 in the appendix. No relationship was visible after moving to data lagged by one business day.

The forward selection PCA model was tested and results are shown in table 5.22. The model performs similarly out-of-sample as in-sample, but does not capture the overall trends of the EUA price as shown in figure 5.16. As with the out-of-sample testing of the base-line model, the forward selection PCA model seems to capture the price changes to some extent on the short horizon, during the first two or three months, but then fails to model the actual price trend.

Table 5.22: Out-of-sample regression results for same-day test data for forward selection PCA.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0015	-0.7139	0.4759	-1.3034	0.1935
DS	-0.0606	-1.5237	0.1288	-2.2795	0.0234
CSS	-0.1446	-2.9385	0.0036	-3.5799	0.0004
WTI crude	0.1878	1.5922	0.1125	2.8809	0.0043
Goodness of fit	MSE	R^2	R^2_{adj}	AIC	BIC
	3.5277 E-04	14.86%	13.90%	-2,158	-2,144

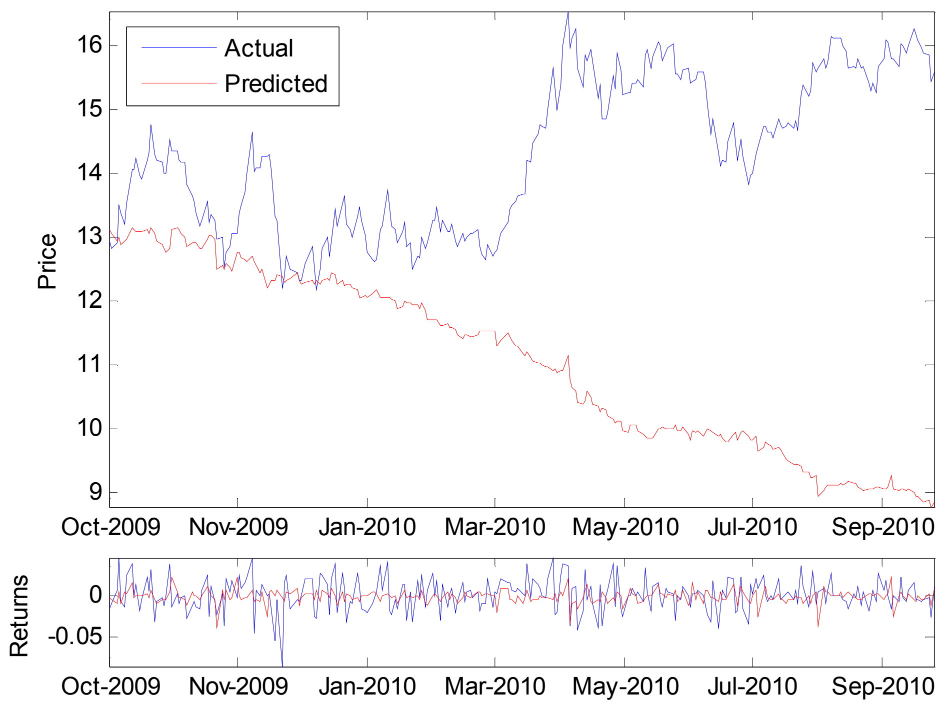

Figure 5.16: The actual vs. predicted same-day price and returns for out-of-sample EUAs using the forward selection PCA.

The results for data lagged by one day are shown in table 14 in the appendix. The model tested stronger out-of-sample, however none of the parameters, apart from one, were significant at the 95% confidence interval.

5.3.3 Latent Root Regression Models

Lastly the latent root models were tested on out-of-sample data. The results are shown in table 5.23. Compared to the in-sample results the coefficients of determination have improved as well as both the AIC and BIC, which show better performance. Of all the same-day out-of-sample models tested the latent root model seems to perform the best in capturing the macro-trend of the EUAs, although the model seems to underestimate both positive returns and negative returns as shown in figure 5.17.

Table 5.23:Out-of-sample regression results for same-day test data for the latent root model.

PC No.	β	t-stat	p-value	HC4	p_{HC4}-value
3	-0.0043	-0.0342	0.0007	-0.0535	0.9574
6	0.0025	0.0036	0.7184	0.0040	0.9968
7	0.0063	0.0286	0.0046	0.0187	0.9850
8	0.0031	0.0091	0.3648	0.0056	0.9956
19	0.0667	0.3849	0.0000	0.5656	0.5721
20	0.2378	1.2432	0.0000	2.0823	0.0383
21	0.0264	0.0799	0.0000	0.0876	0.9303
22	0.0395	0.1593	0.0000	0.1643	0.8696
24	0.0471	0.7068	0.0000	0.7972	0.4260
26	0.1283	0.2668	0.0000	0.3358	0.7373
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	3.5121 E-04	8.69%	5.55%	-2,154	-2,118

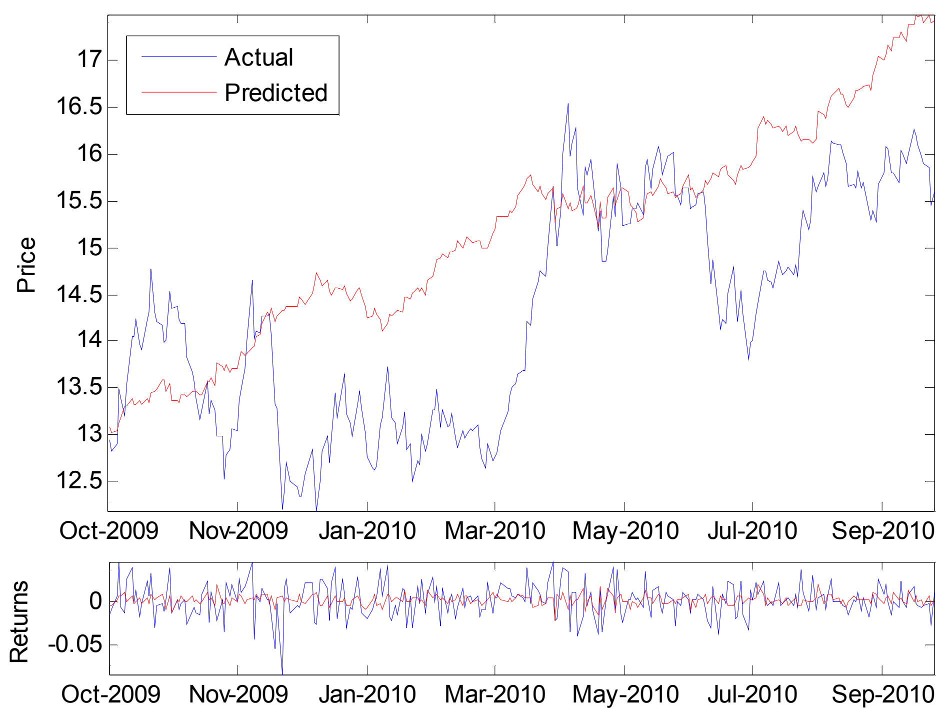

Figure 5.17: The actual vs. predicted same-day price and returns for out-of-sample EUAs using the latent root model.

Repeating the test for data lagged by one business day yielded the results shown in table 5.24. Goodness of fit parameters have again improved for the out-of-sample data. Depsite the low coefficient of determination the model captures the overall trend in the EUA price, which other lag-1 day out-of-sample models failed to do. As in the case of same-day data of the latent root model, the lag-1 day model underestimates both positive and negative returns, but manages to capture the price trend over the whole test period.

Table 5.24: Out-of-sample regression results for lag-1 day test data for the latent root model.

PC No.	β	t-stat	p-value	HC4	p_{HC4}-value
6	0.1416	21.5767	0.0000	1.0543	0.2927
7	-0.1185	-18.6673	0.0000	-0.8104	0.4184
10	-0.1319	-34.2287	0.0000	-1.6342	0.1034
11	0.2010	75.6918	0.0000	2.9787	0.0032
12	0.0262	4.8604	0.2004	0.2001	0.8416
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	3.6904E -04	6.62%	5.22%	-2,137	-2,119

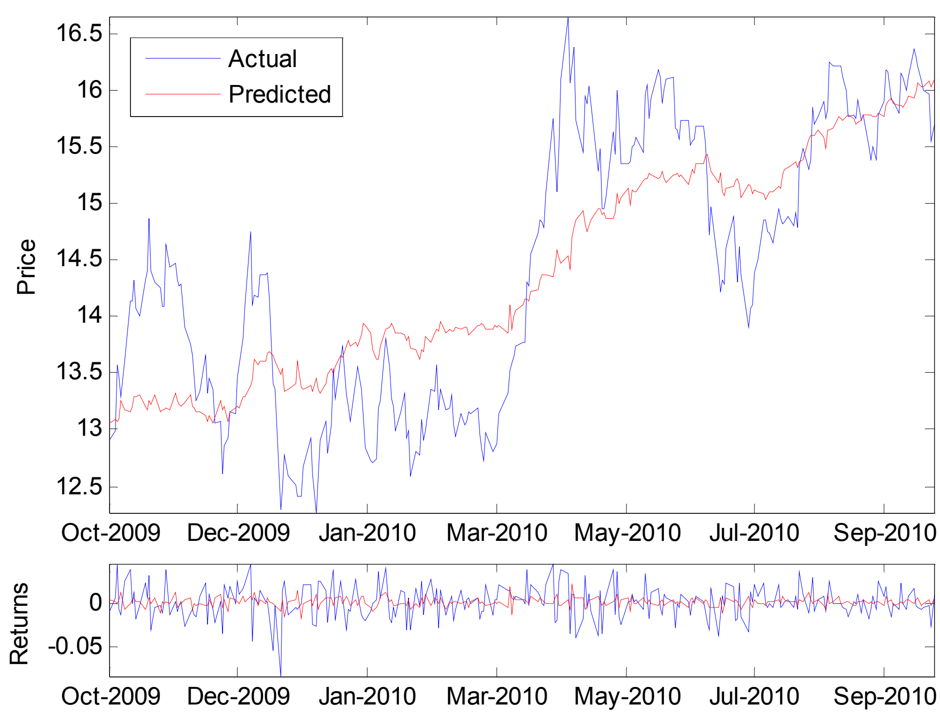

Figure 5.18: The actual vs. predicted lag-1 day price and returns for out-of-sample EUAs using the latent root model.

5.4 Summary

The chapter covered three topics: a statistical analysis, including a principal component analysis, in-sample predictions of EUA allowance prices and out-of-sample predictions.

The statistical analysis showed that EUA returns are right skewed and have heavy tails, indicating non-normal behavior. No autocorrelation was found in the returns. The first four moments were analyzed for all the data and most notable was the extremely high standard deviation of the returns of switching price. Switching price was therefore eliminated from the analysis. A correlation analysis was carried out and the top ten correlations examined for same-day, lag-1 business day, lag-2 business days and lag-1 business week. All correlations based on lagged data were found to be weak, but same-day correlations highlighted CERs and equity indices.

The principal component analysis indicated a clear division between the first three principal components. The first PC was linked to coal-fired power plants and heavier oils. The second PC was linked to the more environmentally friendly gas-fired power plants and finally the third PC contained equity indices. Another interesting result was the fact that power prices fell between the first two principal components, capturing the actual coal to natural gas proportion used for power generation in the United Kingdom. After performing dimension reduction using three different PCA-techniques the variables highlighted by PCA were dark spread, clean spark spread and WTI crude with the addition of natural gas, coal, weather and clean dark spread depending on the type of selection method used.

In-sample predictions showed that correlation was the best dimension reduction technique since the base-line model performed best on same-day data as well as day-ahead predictions. The latent root models provided similar results, however the beta-estimates were not statistically significant at the 95% confidence interval as the estimates were purposefully biased.

Out-of-sample predictions confirmed the predictions made in-sample. The base-line model continued to perform best for same-day data as well as day-ahead predictions. Finally the latent root models were able to capture the overall price trends of the EUA allowances but they were not able to capture the subtle price changes and volatility of the returns.

6 Conclusion

In this thesis market relationships have been examined to establish whether they act as key driving forces on the development of carbon prices within the EU ETS and if predictions can be made based on these relationships. The study, which is based on the British energy market, global equity indices and three relevant currencies, also focuses on which dimension reduction technique is likely to be useful in selecting the reduced set of variables.

Correlation is the preferred dimension reduction technique to be followed by principal component regression. CERs are shown to be the only market relationship which provided useful predictions of EUA prices however this relationship is lost when data is lagged by one business day.

The study shows that dimension reduction, based on correlation analysis and principal component analysis, generates two completely different sets of reduced data sets. Correlation generates a dataset of equity indices along with CERs, which are highly correlated to EUAs, but the principal component analysis highlights dark spread, clean spark spread and WTI crude along with natural gas, weather or clean dark spread, depending on the type of selection criteria used, as the relevant variables.

The reduced dataset generated by the correlation analysis on same-day data is shown to be a useful input for a multiple linear regression model. Same-day correlations, highlighting CERs, provide a dataset of good predictors of EUA price development but all correlations are low when data is lagged by one business day and the regression model fails to provide useful predictions. No significant correlation is found between EUA returns and electricity returns. Earlier studies have however found a high correlation to German power returns.

The principal component analysis shows three main clusters on the first three principal component axes. The first principal component is found to represent coal-fired power plants, or in other words, energy stemming from fuel sources yielding high CO_2 emissions. The second principal component is shown to represent “cleaner” energy, mainly gas-fired power plants. The third principal component then groups equity indices. A graphical representation of the principal components is shown to reflect the proportion of energy sources used to generate power in the United Kingdom. Predictions based on the dataset generated by the PCA do however not capture the volatility nor the overall trend of the EUA returns.

Latent root regression performs similarly as principal component regression when applied to in-sample data. However the latent root approach appears to capture the overall EUA trend when tested out-of-sample but it does not manage to capture the volatility of the returns.

The theoretical carbon price or switching price is rejected as a useful indicator of the price development of EUAs as have previous studies. Leading carbon analysts, such as Point Carbon and bulletins such as Tendances Carbone do however track the development of clean dark spread, clean spark spread and switching price daily, perhaps reflecting the market's view that the EU ETS is still a young and immature market which might slowly start to reflect these theoretical prices in the future.

Further research could be done by repeating this study on EU-wide data, focusing on correlations and principal component regression or independent component regression as the regressors often show high intercorrelations. The greatest price fluctuations have been prompted by exterior announcements or events, most recently by the nuclear crisis after the devastating tsunami in Japan. Only days later, the price of carbon had risen by over 10%. A model able to account for jumps would provide an interesting study. Another interesting topic would be to conduct further research on the EUA-CER spread. In light of the high volatility of the market and the difficulty in providing useful predictions an analysis of hedging strategies is also called for.

The EU ETS is a young and immature market and the volatility is increased by changes in the regulatory framework. As time goes by, and hopefully as a global climate agreement is reached, this turmoil is likely to subside as the rules and regulations of the market become clearer. The result: a more stable market and hopefully more reliable predictions.

7 Bibliography

Alberola, E., Chvevallier, J., & Cheze, B. (2008). Price drivers and structural breaks in European carbon prices 2005–2007. *Energy Policy*, *36*(2), 787-797. doi: 10.1016/j.enpol.2007.10.029.

Bataller, M. M., Tornero, Á. P., & Micó, E. V. i. (2007). CO2 Prices, Energy and Weather. *The Energy Journal*, *28*(3), 72-92.

Benz, E., & Truck, S. (2009). Modeling the price dynamics of CO2 emission allowances. *Energy Economics*, *31*(1), 4-15. doi: 10.1016/j.eneco.2008.07.003.

Capoor, K., & Ambrosi, P. (2007). *State and Trends of the Carbon Market 2007* (pp. 1-19). Washington D.C.

Capoor, K., & Ambrosi, P. (2008). *State and Trends of the Carbon Market 2008* (pp. 1-24). Washington D.C.

Capoor, K., & Ambrosi, P. (2009). *State and Trends of the Carbon Market 2009* (pp. 1-20). Washington D.C.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, *45*(2), 215-233. doi: 10.1016/S0167-9473(02)00366-3.

Daskalakis, G., Psychoyios, D., & Markellos, R. N. (2009). Modeling CO2 emission allowance prices and derivatives: Evidence from the European trading scheme. *Journal of Banking & Finance*, *33*(7), 1230-1241. doi: 10.1016/j.jbankfin.2009.01.001.

Depledge, J., & Lamb, R. (Eds.). (2005). *Caring for Climate: A guide to the Climate Change Convention and the Kyoto Protocol* (20th ed., pp. 1-42). Bonn, Germany: UNFCCC. Retrieved from http://unfccc.int/resource/docs/publications/caring2005_en.pdf.

Ellerman, A. D., & Buchner, B. K. (2008). Over-Allocation or Abatement? A Preliminary Analysis of the EU ETS Based on the 2005–06 Emissions Data. *Environmental and Resource Economics*, *41*(2), 267-287. Springer Netherlands. doi: 10.1007/s10640-008-9191-2.

EU Press Releases: MEMO/08/35. (2008). *The European Union*. Retrieved October 18, 2010, from http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/08/35&format=HTML &aged=0&language=EN&guiLanguage=en.

Frunza, M.-C., Guegan, D., & Lassoudière, A. (2010). Dynamic factor analysis of carbon allowances prices: From classic Arbitrage Pricing Theory to Switching Regimes. Retrieved September 25, 2010, from http://halshs.archives-ouvertes.fr/halshs-00505145/.

Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics*. (A. E. Hilbert, Ed.) (5th ed., pp. 55-106,315-522). New York: McGraw-Hill.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data mining, Inference, and Prediction* (2nd ed., pp. 61-80, 534-536). New York: Springer.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*(4), 709-722.

International Energy Agency. (2005). *Projected Costs of Generating Electricity* (pp. 1-34). Paris. Retrieved from http://www.iea.org/textbase/nppdf/free/2005/ElecCost.PDF.

International Energy Agency. (2009). *CO2 Emissions from Fuel Combustion 2009* (pp. 120-167). OECD Publishing. doi: 10.1787/co2_fuel-2009-en-fr.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer.

Jones, C. (2006). What now for carbon markets? Recovering from the April 2006 price crash. *PennWell Corporation*. Retrieved December 30, 2010, from http://www.powergenworldwide.com/index/display/articledisplay/261545/articles/cogeneration-and-on-site-power-production/volume-7/issue-4/features/what-now-for-carbon-markets-recovering-from-the-april-2006-price-crash.html.

Kossoy, A., & Ambrosi, P. (2010). *State and Trends of the Carbon Market 2010* (pp. 1-75). Washington D.C.

Labatt, S., & White, R. R. (2007). *Carbon Finance*. Hoboken, New Jersey: John Wiley & Sons, Inc.

MacLeay, I. (2011). *Energy Statistics* (pp. 1-7). London. Retrieved from http://www.ifandp.com/wp-content/uploads/2011/03/Emergy-Stats-310311.pdf.

Montgomery, D. C., & Runger, G. C. (2007). *Applied Statistics and Probability for Engineers* (4th ed., pp. 391-499). Hoboken, New Jersey: John Wiley & Sons, Inc.

Obermayer, J. (2009). *An analysis of the fundamental price drivers of EU ETS carbon credits*. KTH. Retrieved September 23, 2010, from http://www.math.kth.se/matstat/seminarier/reports/M-exjobb09/090907b.pdf.

Point Carbon. (2010). *Thomson Reuters*. Retrieved December 30, 2010, from http://www.pointcarbon.com/.

Report on 2008 EU Emissions Trading System emissions data. (2009). *Department of Energy & Climate Change* (pp. 1-28). Retrieved from http://www.decc.gov.uk/assets/decc/what we do/global climate change and energy/tackling climate change/emissions trading/eu_ets/publications/1_20090924140921_e_@@_euetsreport2008.pdf.

Stern, N. (2007). *The Economics of Climate Change: The Stern Review* (1st ed., pp. 3-24). New York: Cambridge University Press.

Taschini, L., & Paolella, M. S. (2006). An Econometric Analysis of Emission Trading Allowances. *Journal of Banking and Finance, Vol. 32, No. 10, 2008*. SSRN. doi: 10.2139/ssrn.947055.

Tendances Carbone - No. 3. (2007). (pp. 1-6).

Tsay, R. S. (2002). *Analysis of Financial Time Series* (pp. 1-78, 335-340). New York: John Wiley & Sons, Inc.

Uhrig-Homburg, M., & Wagner, M. W. (2009). Futures Price Dynamics of CO2 Emission Certificates - An Empirical Analysis. *Journal of Derivatives*, *17*(2), 73-88. SSRN. doi: 10.2139/ssrn.941167.

United Nations Framework Convention on Climate Change. (2010). *UNFCCC*. Retrieved October 14, 2010, from http://unfccc.int/2860.php.

Webster, J. T., Gunst, R. F., & Mason, R. L. (1974). Latent Root Regression Analysis. *Technometrics*, *16*(4), 513-522. The American Society for Quality Control and The American Statistical Association.

What is the EU doing on climate change? (2010). *The European Union*. Retrieved October 18, 2010, from http://ec.europa.eu/environment/climat/kyoto.htm.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817-838.

Zhang, Y.-J., & Wei, Y.-M. (2010). An overview of current research on EU ETS: Evidence from its operating mechanism and economic effect. *Applied Energy*, *87*(6), 1804-1814. doi: 10.1016/j.apenergy.2009.12.019.

Appendix

Table 1: First four moments of phase II sample data.

	Mean	Standard deviation	Skewness	Kurtosis	Excess kurtosis
EUA	-0.0015	0.0319	0.0913	3.9285	0.9285
CER futures	-0.0018	0.0312	-0.0912	4.7843	1.7843
WTI crude futures	-0.0050	0.0458	0.4428	4.1368	1.1368
Gasoil futures	-0.0032	0.0310	0.4733	3.9092	0.9092
Natural gas futures	-0.0033	0.0376	0.6240	5.8469	2.8469
Electricity Base Load	-0.0033	0.0273	-0.0565	5.7926	2.7926
Electricity Base Load	-0.0034	0.0276	-0.0632	8.1262	5.1262
Electricity Peak Load	-0.0010	0.0308	-0.4753	9.2520	6.2520
Electricity Peak Load	-0.0016	0.0295	0.2421	8.3863	5.3863
Brent crude	-0.0001	0.0400	0.2190	3.7114	0.7114
Weather	-0.0007	0.0305	0.2652	4.5582	1.5582
NDX	0.0004	0.0277	0.2567	5.9748	2.9748
SPX	0.0003	0.0281	0.1282	5.3269	2.3269
UKX	-0.0007	0.0242	0.2054	6.0631	3.0631
DAX	-0.0002	0.0261	0.5859	6.3961	3.3961
CAC	-0.1150	0.0267	0.4488	6.0640	3.0640
AEX	-0.1150	0.0283	0.1752	5.5179	2.5179
PSI20	0.0002	0.0198	0.2030	9.1735	6.1735
Aluminum Alloy	-0.0005	0.0224	0.2938	5.1011	2.1011
Aluminum Primary	-0.0003	0.0214	-0.0836	2.9701	-0.0299
Coal	-0.0003	0.0952	1.1273	15.7951	12.7951
Dark Spread	-0.0005	0.0883	1.5595	15.4367	12.4367
Clean Dark Spread	-0.0009	0.0668	0.9899	13.1282	10.1282
Spark Spread	-0.0002	0.0492	0.3956	5.4259	2.4259
Clean Spark Spread	-0.0002	0.0605	0.4664	5.0254	2.0254
Switching Price	0.2042	1.9620	0.1887	34.6467	31.6467
USD/GBP	0.0005	0.0102	0.4608	4.6862	1.6862
USD/EUR	0.0002	0.0091	0.0247	4.9677	1.9677
EUR/GBP	0.0004	0.0076	-0.2420	7.5537	4.5537

Table 2: Same-day top ten correlations of phase II sample data.

	EUA	CER	WTI Crude	Gasoil	Brent crude	UKX	DAX	CAC	PSI20	AL Primary	CDS
EUA	1.00	0.91	0.36	0.37	0.32	0.37	0.38	0.40	0.40	0.27	0.32
CER	0.91	1.00	0.35	0.34	0.32	0.36	0.34	0.36	0.36	0.26	0.25
WTI crude	0.36	0.35	1.00	0.46	0.87	0.51	0.46	0.50	0.39	0.41	0.19
Gasoil	0.37	0.34	0.46	1.00	0.57	0.45	0.40	0.42	0.42	0.41	0.30
Brent crude	0.32	0.32	0.87	0.57	1.00	0.56	0.49	0.54	0.47	0.46	0.26
UKX	0.37	0.36	0.51	0.45	0.56	1.00	0.87	0.95	0.83	0.45	0.24
DAX	0.38	0.34	0.46	0.40	0.49	0.87	1.00	0.90	0.74	0.50	0.28
CAC	0.40	0.36	0.50	0.42	0.54	0.95	0.90	1.00	0.83	0.48	0.28
PSI20	0.40	0.36	0.39	0.42	0.47	0.83	0.74	0.83	1.00	0.39	0.24
AL Primary	0.27	0.26	0.41	0.41	0.46	0.45	0.50	0.48	0.39	1.00	0.18
CDS	0.32	0.25	0.19	0.30	0.26	0.24	0.28	0.28	0.24	0.18	1.00

Table 3: Lag-1 business day top ten correlations of phase II sample data.

	EUA	CER	WTI crude	Gasoil	Nat. gas	Wea-ther	NDX	SPX	Coal	USD/GBP	EUR/GBP
EUA	1.00	0.17	0.08	-0.09	0.07	-0.14	0.17	0.15	0.07	0.07	0.09
CER	0.17	1.00	0.35	0.34	0.20	0.09	0.22	0.26	-0.06	0.02	0.00
WTI crude	0.08	0.35	1.00	0.46	0.14	0.01	0.37	0.42	-0.09	0.00	0.10
Gasoil	-0.09	0.34	0.46	1.00	0.15	-0.01	0.19	0.23	-0.16	-0.12	0.01
Natural gas	0.07	0.20	0.14	0.15	1.00	-0.08	0.10	0.13	0.00	0.00	0.10
Weather	-0.14	0.09	0.01	-0.01	-0.08	1.00	-0.07	-0.06	0.02	-0.19	-0.05
NDX	0.17	0.22	0.37	0.19	0.10	-0.07	1.00	0.95	0.10	0.02	0.03
SPX	0.15	0.26	0.42	0.23	0.13	-0.06	0.95	1.00	0.11	0.01	0.03
Coal	0.07	-0.06	-0.09	-0.16	0.00	0.02	0.10	0.11	1.00	-0.12	-0.11
USD/GBP	0.07	0.02	0.00	-0.12	0.00	-0.19	0.02	0.01	-0.12	1.00	0.50
EUR/GBP	0.09	0.00	0.10	0.01	0.10	-0.05	0.03	0.03	-0.11	0.50	1.00

Table 4: Lag-2 business days top ten correlations of phase II sample data.

	EUA	CER	WTI	Gasoil	Nat. gas	Brent crude	NDX	SPX	UKX	DAX	AEX
EUA	1.00	-0.15	-0.14	-0.14	-0.13	-0.17	-0.18	-0.19	-0.15	-0.12	-0.15
CER	-0.15	1.00	0.35	0.34	0.21	0.32	0.22	0.26	0.36	0.34	0.35
WTI	-0.14	0.35	1.00	0.46	0.14	0.87	0.37	0.42	0.51	0.46	0.49
Gasoil	-0.14	0.34	0.46	1.00	0.14	0.57	0.19	0.23	0.45	0.40	0.45
Natural gas	-0.13	0.21	0.14	0.14	1.00	0.20	0.10	0.13	0.17	0.14	0.18
Brent crude	-0.17	0.32	0.87	0.57	0.20	1.00	0.42	0.45	0.57	0.50	0.54
NDX	-0.18	0.22	0.37	0.19	0.10	0.42	1.00	0.95	0.54	0.63	0.57
SPX	-0.19	0.26	0.42	0.23	0.13	0.45	0.95	1.00	0.60	0.68	0.62
UKX	-0.15	0.36	0.51	0.45	0.17	0.57	0.54	0.60	1.00	0.87	0.94
DAX	-0.12	0.34	0.46	0.40	0.14	0.50	0.63	0.68	0.87	1.00	0.87
AEX	-0.15	0.35	0.49	0.45	0.18	0.54	0.57	0.62	0.94	0.87	1.00

Table5: Lag-1 business week top ten correlations of phase II sample data.

	EUA	Electr. PL 2	Brent crude	Wea-ther	UKX	CAC	PSI20	DS	CDS	SS	CSS
EUA	1.00	0.07	-0.08	-0.09	-0.08	-0.07	-0.09	0.07	0.07	0.09	0.08
Electr. PL 2	0.07	1.00	0.23	0.00	0.28	0.26	0.25	0.31	0.34	0.39	0.33
Brent crude	-0.08	0.23	1.00	-0.03	0.57	0.54	0.47	0.23	0.26	0.02	-0.04
Weather	-0.09	0.00	-0.03	1.00	-0.02	0.00	0.03	-0.03	-0.01	0.05	0.03
UKX	-0.08	0.28	0.57	-0.02	1.00	0.95	0.83	0.19	0.24	0.10	0.04
CAC	-0.07	0.26	0.54	0.00	0.95	1.00	0.83	0.22	0.27	0.11	0.05
PSI20	-0.09	0.25	0.47	0.03	0.83	0.83	1.00	0.19	0.24	0.11	0.06
DS	0.07	0.31	0.23	-0.03	0.19	0.22	0.19	1.00	0.99	0.29	0.25
CDS	0.07	0.34	0.26	-0.01	0.24	0.27	0.24	0.99	1.00	0.30	0.23
SS	0.09	0.39	0.02	0.05	0.10	0.11	0.11	0.29	0.30	1.00	0.98
CSS	0.08	0.33	-0.04	0.03	0.04	0.05	0.06	0.25	0.23	0.98	1.00

Table.6: Lag-1 business month top ten correlations of phase II sample data.

	EUA	WTI crude	Nat. gas	Electr. BL 2	Electr. PL 1	DAX	DS	CDS	USD/ GBP	USD/ EUR	EUR/ GBP
EUA	1.00	0.04	-0.06	-0.09	-0.03	0.04	-0.08	-0.07	0.11	0.07	0.06
WTI crude	0.04	1.00	0.14	0.13	0.14	0.46	0.14	0.19	0.01	-0.09	0.10
Natural gas	-0.06	0.14	1.00	0.42	0.42	0.15	0.26	0.28	0.00	-0.11	0.12
Electr. BL 2	-0.09	0.13	0.42	1.00	0.65	0.20	0.31	0.34	0.00	-0.04	0.04
Electr. PL1	-0.03	0.14	0.42	0.65	1.00	0.16	0.35	0.39	-0.07	-0.06	-0.03
DAX	0.04	0.46	0.15	0.20	0.16	1.00	0.23	0.28	-0.03	-0.08	0.05
DS	-0.08	0.14	0.26	0.31	0.35	0.23	1.00	0.99	-0.01	-0.06	0.06
CDS	-0.07	0.19	0.28	0.34	0.39	0.28	0.99	1.00	0.01	-0.06	0.09
USD/GBP	0.11	0.01	0.00	0.00	-0.07	-0.03	-0.01	0.01	1.00	0.68	0.51
USD/EUR	0.07	-0.09	-0.11	-0.04	-0.06	-0.08	-0.06	-0.06	0.68	1.00	-0.29
EUR/GBP	0.06	0.10	0.12	0.04	-0.03	0.05	0.06	0.09	0.51	-0.29	1.00

Table 7: Order of variables for backward elimination PCA.

Principal Component	Dominant Variable	Coefficient	Latent Value
1	WTI crude	0.7321	0.0152
2	Coal	0.9024	0.0041
3	DS	0.9588	0.0021
4	CSS	0.6569	0.0012
5	Weather	0.9845	0.0009
6	CER	0.8455	0.0007
7	Gasoil	0.8109	0.0006
8	NDX	0.6227	0.0005
9	Electricity PL AI2	0.7254	0.0005
10	AL Alloy	0.8928	0.0004
11	Natural gas	0.6648	0.0003
12	DAX	0.7688	0.0002
13	Electricity BL AT2	0.7354	0.0002
14	Brent crude	0.4817	0.0002
15	Electricity PL AI1	0.4831	0.0002
16	PSI20	0.6951	0.0001
17	AL Primary	0.5963	9.82E-05
18	USD/GBP	0.9380	8.02E-05
19	AEX	0.6665	6.98E-05
20	UKX	0.7832	3.32E-05
21	SPX	0.6781	2.97E-05
22	USD/EUR	0.6388	2.48E-05
23	Electricity BL AT1	0.5773	2.26E-05
24	CAC	0.7396	2.08E-05
25	CDS	0.6116	1.42E-05
26	SS	0.6705	2.85E-06
27	EUR/GBP	0.5838	1.00E-07

Table 8: PCA performed on the variables shown in table 5.5.

Principal Component	Dominant Variable	Coefficient	Latent Value	Cumulative %Variance
1	Coal	0.6307	0.0192	0.7101
2	CSS	0.8661	0.0042	0.8666
3	WTI crude	0.9088	0.0022	0.9476
4	Coal	0.5688	0.0013	0.9973
5	CDS	0.7946	0.0001	1.0000

Regressing on data which had been lagged by one business day yielded the result shown in table 9.

Table 9: Simple PCA regression results for lag-1 day training data.

Variable	β	t-stat	p-value	HC4	p $_{HC4}$-value
Constant	-0.0014	-0.7131	0.4765	-0.7075	0.4799
DS	0.0138	0.6288	0.5300	0.6593	0.5103
CSS	0.0263	0.8106	0.4184	0.7220	0.4710
WTI crude	0.0586	1.2954	0.1964	1.1217	0.2631
Natural gas	-0.0765	-1.1213	0.2632	-1.2381	0.2168
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	1.0214 E-03	1.58%	0.00%	-1,698	-1,599

None of the betas was statistically significant at the 95% confidence level. Comparing to the base-line model of section 5.2.1 again shows that the regression based on simple PCA did not match or exceed the base-line model. Dimension reduction based on simple PCA is therefore rejected. Figure 1 shows the results graphically. The model clearly does not capture the changes in returns from day to day.

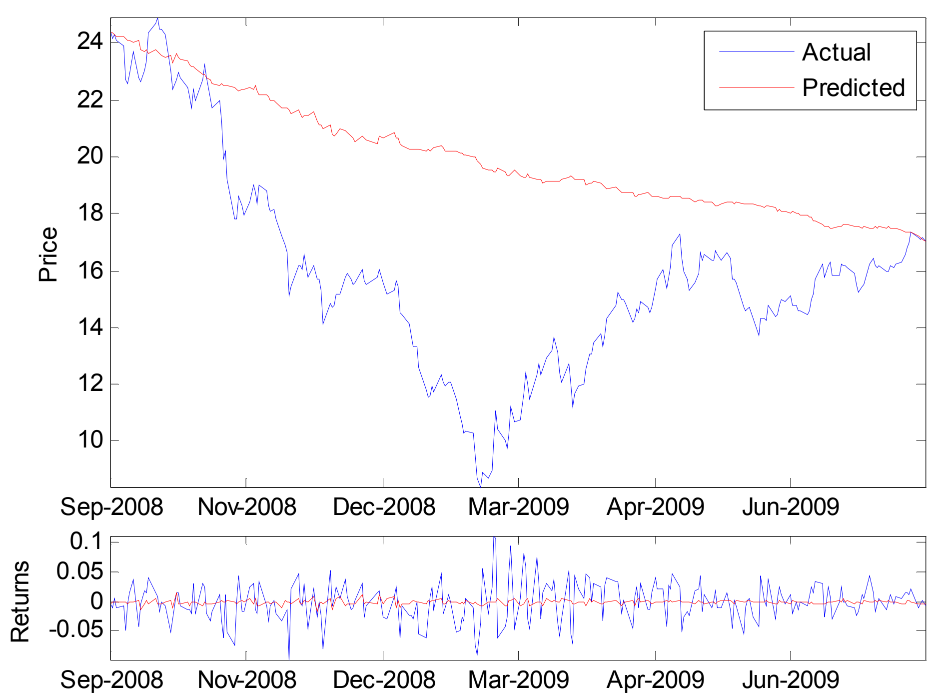

Figure 1: The actual vs. predicted lag-1 day price and returns for EUAs using the simple PCA model.

Regressing on data which had been lagged by one business day yielded the result shown in table 10.

Table 10: PCA using backward elimination regression results for lag-1 day training data.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0014	-0.7188	0.4730	-0.7089	0.4791
WTI crude	-0.0150	-0.9315	0.3525	-0.9663	0.3348
Coal	-0.0095	-0.3083	0.7581	-0.3033	0.7619
DS	0.0685	1.5861	0.1140	1.4795	0.1403
CSS	0.0488	0.8362	0.4038	0.6451	0.5194
Weather	0.1445	2.2051	0.0284	1.9673	0.0503
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	1.0050 E-03	3.52%	1.57%	-1,703	-1,604

No variable was statistically significant according to adjusted p-values. Residuals were stationary, but not normally distributed. No autocorrelation was present. Figure 2 shows the results of table 10 graphically.

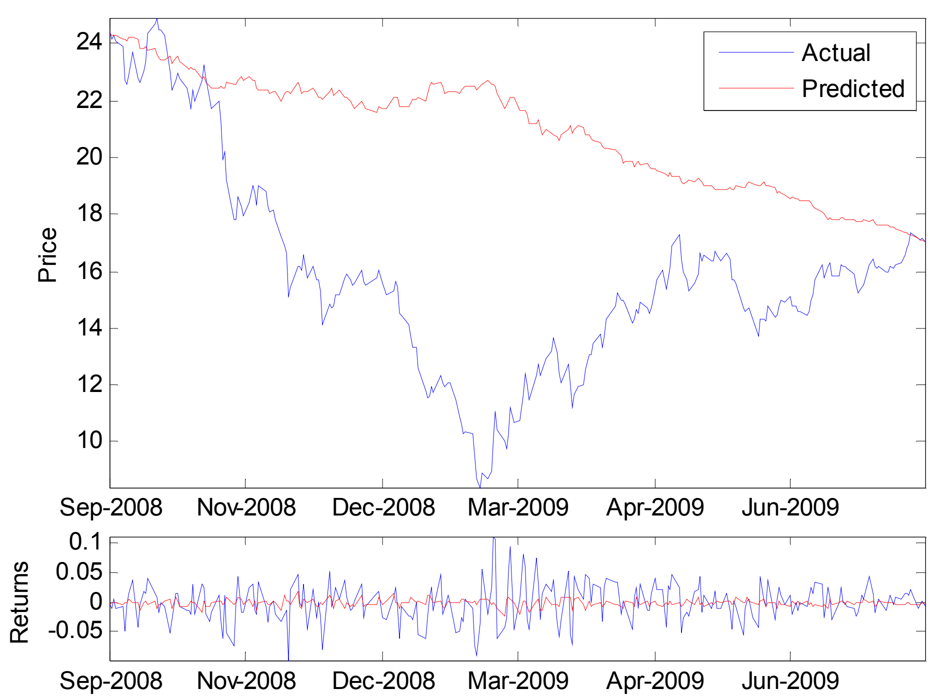

Figure 2: The actual vs. predicted lag-1 day price and returns for EUAs using the backward elimination PCA model.

Regressing on data which had been lagged by one business day yielded the result shown in the table below.

Table 11: PCA using forward selection regression results for lag-1 day training data.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0014	-0.7181	0.4734	-0.7028	0.4828
DS	0.0113	0.7871	0.4320	0.6669	0.5055
Coal	-0.0028	-0.0917	0.9270	-0.0929	0.9261
CDS	0.0753	1.7701	0.0779	1.5764	0.1162
CSS	-0.0289	-0.5324	0.5949	-0.3806	0.7038
WTI crude	0.4946	2.1151	0.0354	1.5067	0.1331
Goodness of fit	MSE 1.0074 E-03	R^2 3.32%	R^2_{adj} 1.37%	AIC -1,703	BIC -1,604

No variable was statistically significant at the 95% confidence interval according to adjusted p-values. Residuals were stationary, but not normally distributed. The qq-plot revealed heavy tails indicating an increased probability of extreme events. No autocorrelation was present. The lag-1 forward selection PCA model matched the performance set by the base-line model of table 5.9. Figure 3 shows the results of table 11 graphically.

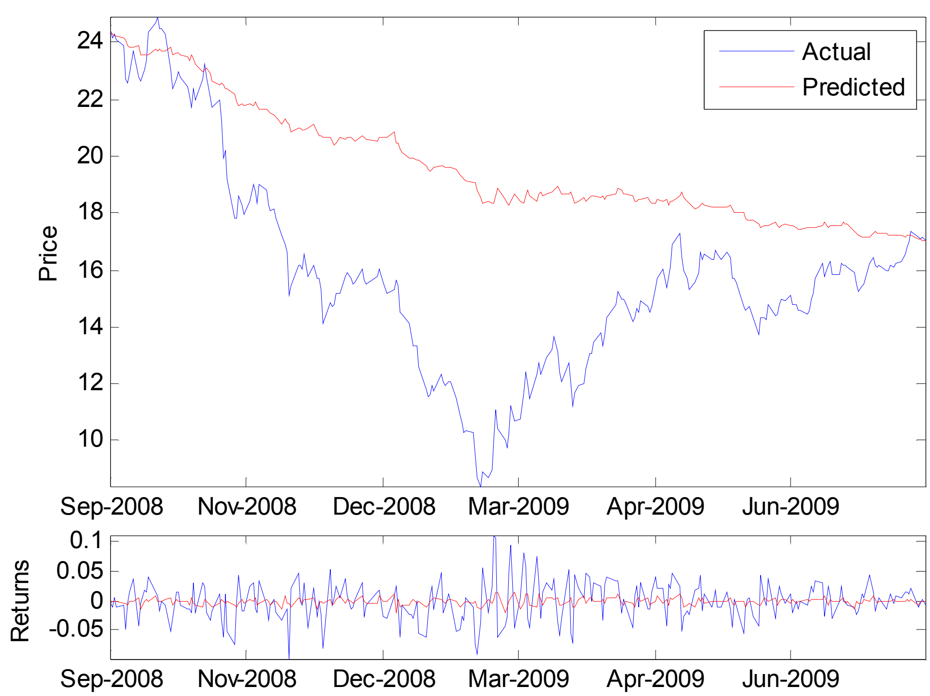

Figure 3: The actual vs. predicted lag-1 day price and returns for EUAs using the forward selection PCA model.

Testing the simple PCA model using test data lagged by one day yielded the results shown in table 12 shown below.

Table 12: Out-of-sample regression results for lag-1 day test data for simple PCA.

Variable	β	t-stat	p-value	HC4	p $_{HC4}$-value
Constant	-0.0014	-0.6358	0.5255	-1.2440	0.2146
DS	0.0138	0.2418	0.8091	0.3807	0.7037
CSS	0.0263	0.3585	0.7203	0.6671	0.5053
WTI crude	0.0586	0.4718	0.6374	0.8708	0.3846
Natural gas	-0.0765	-.06389	0.5235	-1.0032	0.3167
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	3.4469 E-04	4.32%	2.88%	-2,156	-2,138

The model shows a slight improvement in this case compared to the training dataset, but as before, the performance does not capture day to day price changes nor the overall trend of the actual EUA price. The results are shown graphically in figure 4.

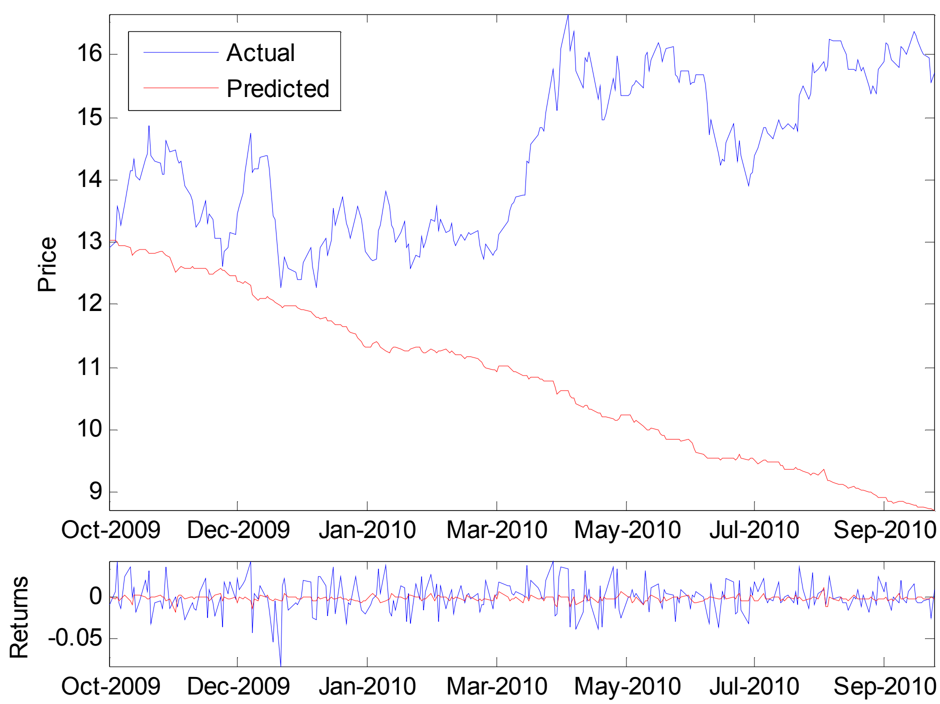

Figure 4: The actual vs. predicted lag-1 day price and returns for out-of-sample EUAs using the simple PCA.

Repeating the test for data lagged by one business day yielded the results shown in table 13. Goodness of fit parameters have all improved, but the model's performance does not capture the day to day price changes nor the overall trend of EUA prices in either case as shown in figure 5.

Table 13: Out-of-sample regression results for lag-1 day test data for backward elimination PCA.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0014	-0.7177	0.4735	-1.1688	0.2435
WTI crude	-0.0150	-0.1363	0.8917	-0.2109	0.8332
Coal	-0.0095	-0.0991	0.9211	-0.1174	0.9066
DS	0.0685	1.2988	0.1951	1.8473	0.0658
CSS	0.0488	1.0573	0.2913	1.7261	0.0855
Weather	0.1445	1.5890	0.1133	2.1746	0.0305
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	3.8602 E-04	7.94%	6.20%	-2,124	-2,102

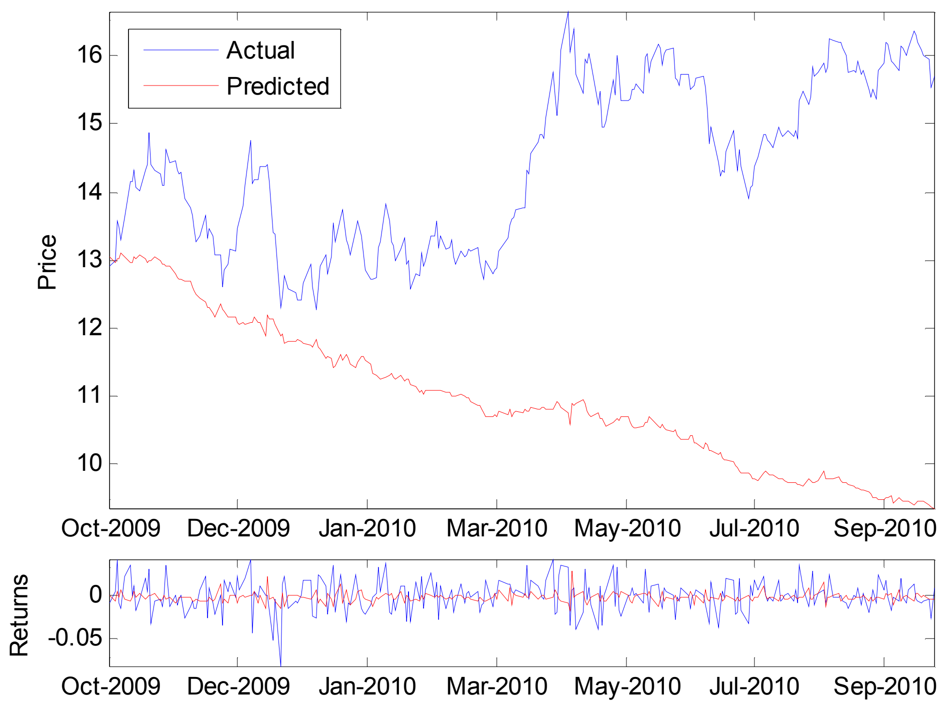

Figure 5: The actual vs. predicted lag-1 day price and returns for out-of-sample EUAs using the backward elimination PCA.

Repeating the test for data lagged by one business day yielded the results shown in table 14. Goodness of fit parameters are much better for the out-of-sample data. The model manages to capture the volatility best during the first quarter of the prediction horizon as shown in figure 6. This model outperforms the lag-1 base-line model as all goodness of fit parameters show improvement. However, none of the betas, except WTI crude, are statistically significant at the 95% confidence interval.

Table 14: Out-of-sample regression results for lag-1 day test data for the forward selection PCA.

Variable	β	t-stat	p-value	HC4	p_{HC4}-value
Constant	-0.0014	-0.7830	0.4343	-0.9979	0.3193
DS	0.0113	0.0720	0.9427	0.0844	0.9328
Coal	-0.0028	-0.0313	0.9751	-0.0310	0.9753
CDS	0.0753	0.3232	0.7468	0.3915	0.6957
CSS	-0.0289	-0.6587	0.5106	-0.8629	0.3890
WTI crude	0.4946	4.7854	0.0000	4.9233	0.0000
Goodness	MSE	R^2	R^2_{adj}	AIC	BIC
of fit	5.2065 E-04	15.60%	14.01%	-2,043	-2,021

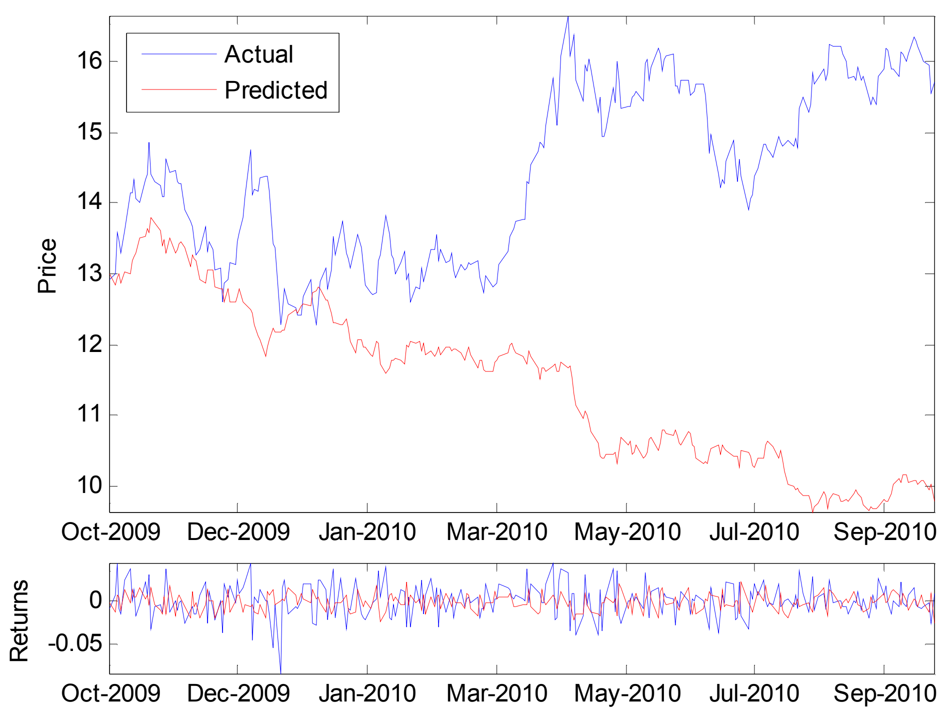

Figure 6: The actual vs. predicted lag-1 day price and returns for out-of-sample EUAs using the forward selection PCA.