LAPPEENRANTA
UNIVERSITY OF TECHNOLOGY

Faculty of Technology
Department of Mathematics and Physics
Laboratory of Applied Mathematics

Analysis of Patterns in Electricity Spot Market Time Series

The topic of this Master's thesis was approved
by the departmental council of the Department of Mathematics and Physics.

The thesis was supervised by Prof. Ph.D. Heikki Haario and Ph.D. Tuomo Kauranne.
The examiners of the thesis were Prof. Ph.D. Heikki Haario and Ph.D. Tuomo Kauranne.

Lappeenranta, April 13^{th}, 2010

Ngoga Kirabo Bob
Kaartinkatu 8 D 29
53850 Lappeenranta, Finland
Tel. +358 40 46 54 664
ngoga.kirabo.bob [at] lut.fi

Abstract

Lappeenranta University of Technology
Department of Mathematics and Physics

Ngoga K. Bob

Analysis of Patterns in Electricity Spot Market Time Series

Thesis for the Degree of Master of Science in Technology

2010

64 pages, 60 figures, 4 tables, 1 appendix

Examiners:Prof. Ph.D. Heikki Haario and Ph.D. Tuomo Kauranne.

Keywords: electricity spot price, Qlucore Omics Explorer, time series decomposition, trend, seasonality, regression.

Due to its non-storability, electricity must be produced at the same time that it is consumed, as a result prices are determined on an hourly basis and thus analysis becomes more challenging. Moreover, the seasonal fluctuations in demand and supply lead to a seasonal behavior of electricity spot prices. The purpose of this thesis is to seek and remove all causal effects from electricity spot prices and remain with pure prices for modeling purposes. To achieve this we use Qlucore Omics Explorer (QOE) for the visualization and the exploration of the data set and Time Series Decomposition method to estimate and extract the deterministic components from the series. To obtain the target series we use regression based on the background variables (water reservoir and temperature). The result obtained is three price series (for Sweden, Norway and System prices) with no apparent pattern.

Acknowledgements

I am grateful to the Department of Mathematics and Physics of Lappeenranta University of Technology for the financial support during the entire duration of my studies. Thank you very much.

I am also grateful to the supervisor of the thesis, PhD. Tuomo Kauranne and the examiner Prof. PhD. Heikki Haario, for giving valuable comments and guidance; and Matylda Jabłońska for assistance and insightful reviews.

This work would not have been possible without the help of many parents, friends and colleagues. I refrain from listing all their names here for fear that this section might otherwise exceed the others in length.

Ndabashimiye mwese (Kiitoksia kaikille).

Lappeenranta, April 13^{th}, 2010.

Ngoga K. Bob.

Contents

List of Tables

List of Figures

1 Introduction

Electricity is an essential part of modern life. It must be produced at the same time that it is consumed; as a result, prices are usually determined on an hourly basis. These are among reasons why electricity spot prices are among the most challenging data set for time series analysis. There are two main goals of time series analysis: identifying the nature of the phenomenon represented by the sequence of observations, and forecasting (predicting future values of the time series variables). Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, we can interpret and integrate it with other data. Regardless of the depth of our understanding and the validity of our interpretation of the phenomenon, we can extrapolate the identified pattern to predict future events. Most time series patterns can be described in terms of two basic classes of components: trends and seasonality. Trends are generally linear or quadratic and seasonality is a trend that repeats itself systematically over time.

There are two main approaches used to analyze time series: in the time domain or in the frequency domain. Many techniques are available to analyze data within each domain. Analysis in the time domain is most often used for stochastic observations. One common technique in case of electricity prices [9], [14] is the Box Jenkins ARIMA method. It uses moving averages, detrending, and regression methods to detect and remove autocorrelations in the data. Analysis in the frequency domain is often used for periodic and cyclical observations. Common techniques are spectral analysis, harmonic analysis, and periodogram analysis. A specialized technique is Fast Fourier Transform (FFT) [8].

It is well known that electricity demand exhibits seasonal fluctuations [13],[15]. They mostly arise due to changing climate conditions, like temperature and the number of daylight hours. In some countries also the supply side shows seasonal variations. Hydro units, for example, are heavily dependent on precipitation and snow melting, which varies from season to season. These seasonal fluctuations in demand and supply translate into seasonal behavior of electricity prices, and spot prices in particular. The aim of this thesis is to remove all patterns in the price series in the Nord Pool data set and remain with pure series to use for modeling purposes. The approach used is the method of classical time series decomposition to estimate and extract the deterministic components from the price series, and using regression model based on background variables to get the pure (cleaned) price series. For the visualization and exploration of the data to find the patterns, we use Qlucore Omics Explorer (QOE).

The structure of this thesis is as follows: the next Section goes through the description of the Industrial problem: an overall description of the background on the industrial process and the target problem. Section 3 describes the mathematical modeling approach, that is, the history and the background theory on Principal Component Analysis (PCA), regression analysis and time series decomposition methods. Section 4 describes the mathematical model for the problem, that is the description of how the theory in Section 3 is applied on the problem in Section 2. This section will describe the use of Qlucore Omics Explorer (QOE), and the regression analysis based on background variables and the time series decomposition. The results are presented in Section 5 and, finally, Section 6 concludes.

2 Description of the Industrial Problem

2.1 Background on the Industrial Process

2.1.1 Overview of the Electricity Trading on Market

Electricity is the electromagnetic field energy sent out by batteries and generators. In economic terms, electricity (both power and energy) is a commodity capable of being bought, sold and traded. It has many sources and it plays a big role in our daily life. It must be produced at the same time that it is consumed; as a result, prices are usually determined on an hourly or half-hourly basis.

An electricity market is a system for effecting purchases through bids to buy, sales, through offers to sell, and short-term trades, generally in the form of financial or obligation swaps. Bids and offers use supply and demand principles to set the price.

Trading is the buying and selling of electricity in an electricity market. The buying and selling occurs 24 hours a day, seven days a week, for each discrete hour of the day. Generators sell their generation to the market; retailers buy this generation and sell it to customers as electricity. That is, a trader determines whether to buy or sell electricity based on the spot price and the generation assets that they have available.

2.1.2 The Nordic Market

The Nordic commodity market for electricity is known as **Nord Pool**[19]. It was established in 1992 and at this time it was a Norwegian market, but in the years to follow Sweden (1996), Finland (1998) and Denmark (2000) joined in. Nord Pool is the oldest and one of the most mature power exchanges in the world. It was the world's first international power exchange. In this market, players from outside the Nordic region are allowed to participate on equal terms with local exchange members. To participate in the spot (physical) market, called **Elspot**, a grid connection enabling power to be delivered to or taken from the main grid is required. Additionally, a continuous hour-ahead **Elbas** market is also operated in Finland, Sweden and Eastern Denmark. In the financial **Eltermin** market power derivatives, like forwards (up to threes years ahead), futures options and contracts for differences (CfD; for price area differentials, using the system day-ahead price as the reference price) are being traded.

2.1.3 Price setting at Nord Pool

At Nord Pool the spot price is a result of two-sided uniform price for hourly time intervals. It is determined from the various bids presented to the market administrator up to the time when the auction is closed. The market for trading power for physical delivery is called Elspot which is a day-ahead market. What is traded are one-hour-long physical power contracts, and the minimum contract is 0.1 MWh. At noon (12 p.m.) each day, the market participants submit to the market administrator (Nord Pool) their (bid and ask) offers for the next 24 hours starting at 1 a.m. the next day. This information is provided electronically via the Internet (**Elweb**) with a resolution of one hour, i.e. one for each hour of the next day. Such information should contain both price and volume of the bids. There are three possible ways of bidding at Elspot.

- **Hourly bidding**: consisting of pairs of price and volume for each hour.
- **Block bidding**: here the bidding price and volume are fixed for a number of consecutive hours.
- **Flexible hourly bidding**: it is a fixed price and volume sales bid where the hour of the sale is flexible and determined by the highest (next day) spot price that is above the price indicated by the bid.

The market participants are free (for hourly bidding) to provide a whole sell and/or buy stack for each hour. For instance, a power generator could be more interested in selling larger quantities of electricity if the price is high than if it is low. To optimize their profit, power generators buy electricity during low price periods, and thereby saving own production potential for periods when the price is higher.

By 12 p.m. Nord Pool closes the bidding for the next day and for each hour proceeds to make cumulative supply and demand curves. Since there must be a balance between production and consumption, the system spot price for that particular hour is determined as the price where the supply and demand curves cross. Hence, the name of this operation is **market cross** or **equilibrium point**. Trading based on this method is called **equilibrium trading**, **auction trading** or **simultaneous price setting**. If the data does not define an equilibrium point, no transactions will take place for that hour [15].

After having determined the system price for a given hour of the next day's 24 hour period, Nord Pool continues by analyzing for potential bottlenecks (grid congestions) in the power transmission grid that might result from this system price. If no bottlenecks are found, the system price will represent the spot price for the whole Nord Pool area. However, if potential grid congestions may result from the bidding, so called area spot price (zonal prices) that are different from the system price will have to be computed. The idea behind the introduction of area (zonal) prices is to adjust electricity prices within a geographical area in order to favor local trading to such a degree that the limited capacity of the transmission grid is not exceeded. How the area prices are determined within Nord Pool differs between, say, Finland and Sweden. The system price is then the price determined by the equilibrium point independent of potential grid congestions. The area (zonal) prices will only differ from this price for those hours when transmission capacity in the central grid is limited. The system price is therefore typically less volatile than the area prices.

2.2 The Target Problem

One definition of a time series is that of a collection of quantitative observations that are evenly spaced in time and measured successively. The main goals of Time Series Analysis are: Description, Explanation, Forecasting, Intervention Analysis and Quality Control. Time series are analyzed in order to understand the underlying structure and function that produce the observations. Understanding the mechanisms of a time series allows a mathematical model to be developed that explains the data in such a way that prediction, monitoring, or control can be performed. Examples include prediction/forecasting, which is widely used in economics and business. Monitoring of ambient conditions, or of an input or an output, is common in science and industry. Quality control is used in computer science, communications, and industry.

In this work we are focused on the first goal, Descriptive, which consist on the Identification of patterns in the data set. The most common patterns are the trend pattern, the seasonal pattern and the cyclic pattern. The aim here is to remove all patterns in the original Series and remain with a pure Series to use for other goals of Time Series Analysis.

3 Description of the Mathematical Modeling Approach

3.1 Principal Component Analysis (PCA)

3.1.1 Overview

Principal Component Analysis is a useful statistical technique that has found application in many fields, and it is a common technique for finding patterns in data of high dimension. Depending on the field of application, it is also named **the discrete Karhunen-Loève Transform (KLT)**, **the Hotelling transform** or **Proper Orthogonal Decomposition (POD)**. It was invented in 1901 by Karl Pearson. Now it is mostly used as a tool in exploratory data analysis and for making predictive models. The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first *few* retain most of the variation present in *all* of the original variables. Mathematically, given a p-dimensional random variable $x = (x_1, x_2, \ldots, x_p)'$ with covariance matrix Σ_x, a PCA is concerned with using a *few* linear combinations of x_i to explain the structure of Σ_x.

3.1.2 Classical Principal Component Analysis

Classical Principal Component Analysis (PCA) is concerned with explaining the variance-covariance structure among p variables, $x = (x_1, x_2, \ldots \ldots, x_p)'$, through a few linear combinations of the components of x. Suppose we wish to find a linear combination

$$y = c'x = c_1 x_1 + \ldots + c_p x_p \tag{1}$$

of the components of x such that $Var(y)$ is as large as possible. Because $Var(y)$ can be increased by simply multiplying c by a constant, it is common to restrict c to be of unit length; that is, $c'c = 1$. Noting that $Var(y) = c'\Sigma_{xx}c$, where Σ_{xx} is the $p \times p$ variance-covariance matrix of x, another way of stating the problem is to find c such that

$$max_{c \neq 0} \frac{c'\Sigma_{xx}c}{c'c}. \tag{2}$$

Denote the eigenvalue-eigenvector pairs of Σ_{xx} by $\{(\lambda_1, e_1), \ldots (\lambda_p, e_p)\}$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, and the eigenvectors are of unit length. The solution to (2) is to choose $c = e_1$, in which case the linear combination $y_1 = e_1'x$ has maximum variance,

$Var(y_1) = \lambda_1$. In other words,

$$max_{c\neq 0}\frac{c'\Sigma_{xx}c}{c'c} = \frac{e_1'\Sigma_{xx}e_1}{e_1'e1} = \lambda_1. \tag{3}$$

The linear combination, $y_1 = e_1'x$, is called the first principal component. Because the eigenvalues of Σ_{xx} are not necessarily unique, the first principal component is not necessarily unique either. The second principal component is defined to be the linear combination $y_2 = c'x$ that maximizes $Var(y_2)$ subject to $c'c = 1$ and such that $Cov(y_1, y_2) = 0$. The solution is to choose $c = e_2$, in which case $Var(y_2) = \lambda_2$. In general, the k^{th} principal component, for $k = 1, 2, \ldots, p$, is the linear combination $y_k = c'x$ that maximizes $Var(y_k)$ subject to $c'c = 1$ and such that $Cov(y_k, y_j) = 0$, for $j = 1, 2, \ldots, k-1$. The solution is to choose $c = e_k$, in which case $Var(y_k) = \lambda_k$.

One measure of the importance of a principal component is to asses the proportion of the total variance attributed to that principal component. The total variance of x is defined to be the sum of the variances of the individual components, that is, $Var(x_1) + \ldots + Var(x_p) = \sigma_{11} + \ldots + \sigma_{pp}$, where σ_{jj} is the j^{th} diagonal element of Σ_{xx}. This sum is also denoted as $tr(\Sigma_{xx})$, or the trace of Σ_{xx}. Because $tr(\Sigma_{xx}) = \lambda_1 + \ldots + \lambda_p$, the proportion of the total variance attributed to the k^{th} principal component is given simply by $\frac{Var(y_k)}{tr(\Sigma_{xx})} = \frac{\lambda_k}{\sum_{j=1}^{p}}\lambda_j$ [12].

3.2 Regression Analysis

The earliest form of regression was the *method of least squares*, which was published by *Legendre* in 1805, and by *Gauss* in 1809 [20]. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun. Gauss published a further development of the theory of least square in 1821, including a version of the Gauss-Markov theorem. The term ”regression” was coined by *Francis Galton*, a cousin of *Charles Darwin*, in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average. For Galton, regression had only this biological meaning, but his work was later extended by *Udny Yule* and *Karl Pearson* to a more general statistical context. In the work of Yule and Pearson [11], the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by *R.A. Fisher* in his works of 1922 and 1925. Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is close to Gauss' formulation of 1821. Regression methods continue to be an area of active research. In recent decades, new methods have been developed for *robust regression*, regression involving correlated responses such as time series and growth curves, regression in which the predictor or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, *nonparametric regression, Bayesian* methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression.

3.2.1 Time Series Regression Model

Consider *the linear time series regression model*

$$Y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + \epsilon_t = x_t'\beta + \epsilon_t, t = 1, \ldots, T \tag{4}$$

where $x_t = (1, x_{1t}, \ldots, x_{kt})'$ is a $(k+1) \times 1$ vector of explanatory variables, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$ is a $(k+1) \times 1$ vector of coefficients, and ϵ_t is a random error term. In matrix form the model is expressed as

$$Y = X\beta + \epsilon \tag{5}$$

where Y and ϵ are $(T \times 1)$ vectors and X is a $(T \times (k+1))$ matrix. The standard assumptions of the time series regression model are [17]:

1. the linear model (Equation (4)) is correctly specified
2. y_t, x_t is jointly stationary and ergodic
3. the regressors x_t are predetermined: $E[X_{is}\epsilon_t] = 0$ for all $s \leq t$ and $i = 1, \ldots, k$.
4. $E[x_x x_t'] = \Sigma_{xx}$ is of full rank $k+1$
5. $x_t\epsilon_t$ is an uncorrelated process with finite $(k+1) \times (k+1)$ covariance matrix $E[\epsilon_t^2 x_t x_t'] = S = \sigma^2\Sigma_{xx}$.

The second assumption rules out trending regressors, the third rules out endogenous regressors but allows lagged dependent variables, the fourth avoids redundant regressors or exact multicollinearity, and the fifth implies the error term is a serially uncorrelated process with constant unconditional variance σ^2. In the time series regression model, the regressors x_t are random and the error term ϵ_t is not assumed to be normally distributed.

3.2.2 Least Squares Estimation

Ordinary least squares (OLS) estimation is based on minimizing the sum of squared residuals

$$SSR(\beta) = \sum_{t=1}^{T} (Y_t - x_t'\beta)^2 = \sum_{t=1}^{T} \epsilon_t^2 \tag{6}$$

and produces the fitted model

$$Y_t = x_t'\widehat{\beta} + \widehat{\epsilon}_t, t = 1, \ldots, T \tag{7}$$

where $\widehat{\beta} = (X'X)^{-1}X'Y$and $\widehat{\epsilon_t} = Y_t - \widehat{Y_t} = Y_t - x_t'\widehat{\beta}$. The error variance is estimated as $\widehat{\sigma^2} = \frac{\widehat{\epsilon}'\widehat{\epsilon}}{(T-k-1)}$. Under the assumptions described above, the OLS estimates $\widehat{\beta}$ are consistent and asymptotically normally distributed.

Goodness of Fit

Goodness of fit is summarized by the R^2 of the regression model

$$R^2 = 1 - \frac{\widehat{\epsilon}'\widehat{\epsilon}}{(Y - \overline{Y})'(Y - \overline{Y})} \tag{8}$$

where $\overline{Y}$ is the sample mean of Y_t and 1 is a $(T \times 1)$ vector of ones. R^2 measures the percentage of the variability of Y_t that is explained by the regressors x_t.

3.3 Time Series Decomposition

The analysis of Time Series is based on the assumption that successive values in the data set represent consecutive measurements taken at equally spaced time intervals. There are two main goals of Time Series analysis: Identifying the nature of the phenomenon represented by the sequence of observations, and forecasting (predicting future values of the time series variables. Both of these goals require that the pattern of observed time series data is identified and, in a sense, formally described. Once the pattern is established, we can interpret and integrate it with our data (i.e. use it in our theory of the investigated phenomenon).

3.3.1 Identifying Patterns in Times Series Data

As in most other analyses, in Time Series Analysis it is assumed that the data consist of a systematic pattern (usually a set of identifiable components) and a random noise which usually makes the pattern difficult to identify. Most time series analysis techniques involve some form of filtering out noise in order to make the pattern more salient.

Most time series patterns can be described in terms of two basic classes of components: **trend** and **seasonality**. The former represents a general systematic linear or (most often) an nonlinear component that changes over time and does not repeat or at least does not repeat within the time range captured by the data. The latter may have a formally similar nature, however, it repeats itself in systematic intervals over time. Those two general classes of time series components may coexist in real-life data. For example, electricity prices in Finland can grow over years but they still follow consistent seasonal patterns as presented in Figure 1.

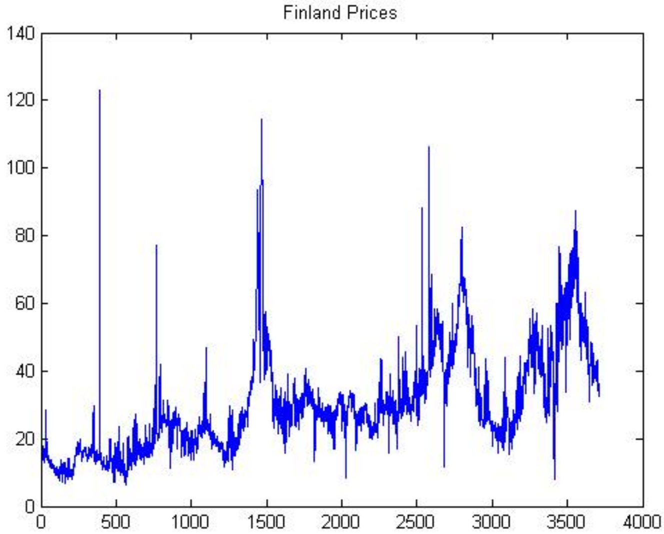

Figure 1: An example of Patterns in Time Series.

From this plot we can see a clear, almost linear trend, indicating that electricity prices in Finland follow a steady growth over years. At the same time, we can see an almost identical pattern each lets say winter season (e.g., prices are higher in winter time).

3.3.2 Estimation and Elimination of Trend and Seasonal Components

A general approach to a Time Series Modeling is constituted of the following steps[3]:

- Plot the series and examine the main features of the graph, checking in particular whether there is a trend, a seasonal component, any apparent sharp changes in behavior and any outlying observations.

- Remove the trend and seasonal components to get stationary residuals.
- Choose a model to fit the residuals.
- Forecast.

After the first step, inspection of the graph may suggest the possibility of representing the data as a realization of the processes [1],[10],[16]:

$$X_t = m_t s_t Y_t \quad \text{(Multiplicative Decomposition Model)} \tag{9}$$

or

$$X_t = m_t + s_t + Y_t \quad \text{(Additive Decomposition Model)} \tag{10}$$

where m_t is the **trend component**, s_t is the **seasonal component** and Y_t is a **random noise component**. The aim is to estimate and extract the deterministic components m_t and s_t in the hope that the residual or noise Y_t will turn out to be a stationary time series. We can use the theory of such processes to find a satisfactory probabilistic model for the process Y_t, to analyze its properties, and to use it in conjunction with m_t and s_t for purposes of prediction and simulation of X_t.

Definitions

Let X_t be a time series with $E(X_t^2)\infty$. The **mean function** of X_t is $\mu X(t) = E(X_t)$.

The **covariance function** of X_t is $\gamma X(r, s) = Cov(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$ for all integers r and s.

X_t is (**weakly stationary**) if:

1. $\mu_{X(t)}$ is independent of t, and
2. $\gamma_{X(t+h,t)}$ is independent of t for each h.

This means that both mean of X_t and covariance between X_t and X_{t+h} are time invariant, where h is an arbitrary integer. In practice, suppose that we have observed N data points, $\{X_t/t = 1, \ldots, N\}$. The weak stationarity implies that the time plot of the data would show that the N values fluctuate with constant variation around a fixed level. In applications, weak stationarity enables one to make inferences concerning future observations (e.g., prediction).

3.3.3 Estimation and Elimination of Trend in the absence of Seasonality

In the absence of a seasonal component the model in (10) becomes the following.

$$X_t = m_t + Y_t,\ t = 1, \dots, n,\ \text{where } EY_t = 0. \tag{11}$$

If $EY_t \neq 0$, then we replace m_t and Y_t in (11) with $m_t + EY_t$ and $Y_t - EY_t$ respectively.

Method 1: Trend Estimation

Smoothing with a finite moving average filter

Let q be a nonnegative integer and consider the two-sided moving average $W_t = (2q + 1)^{-1} \sum_{j=-q}^{q} X_{t-j}$ of the process X_t defined by (11). Then for $q + 1 \leq t \leq n - q$,

$$W_t = (2q+1)^{-1} \sum_{j=-q}^{q} m_{t-j} + (2q+1)^{-1} \sum_{j=-q}^{q} Y_{t-j} \approx\ t$$

assuming that m_t is approximately linear over the interval $[t - q, t + q]$ and that the average of the error terms over this interval is close to zero [4]. The moving average thus provides us with the estimates

$$\hat{m}_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j}, \quad q + 1 \leq t \leq n - q. \tag{12}$$

Polynomial Fitting

A useful technique for estimating m_t is the **method of least squares**. In the least squares procedure we attempt to fit a parametric family of functions, e.g., $m_t = a_1 + a_1 t + a_2 t^2$, to the data $x_1, \dots, x_n$ by choosing the parameters, in this illustration a_0, a_1 and a_2, to minimize $\sum_{t=1}^{n} (Y_t - m_t)^2$. This technique can also be used to estimate higher-order polynomial trends in the same way.

Method 2: Trend Elimination by Differencing

Instead of attempting to remove the noise by smoothing as in Method 1, we now attempt to eliminate the trend by differencing. We define the lag-1 difference operator ∇ by

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t,$$

where B is the backward operator, $BX_t = X_{t-1}$. Powers of the operators B and ∇ are defined in the obvious way, i.e., $B^j(X_t) = X_{t-j}$ and

$$\nabla^j(X_t) = \nabla(\nabla^{j-1}(X_t)), j \geq 1,$$

with $\nabla^0 = X_t$. Polynomials in B and ∇ are manipulated in the same way as polynomial functions of real variables. If the operator ∇ is applied to a linear trend function $m_t = c_0 + c_1 t$, then we obtain the constant function

$$\nabla m_t = m_t - m_{t-1} = c_0 + c_1 t - (c_0 + c_1(t-1)) = c_1.$$

In the same way any polynomial trend of degree k can be reduced to a constant by application of the operator ∇^k. For example, if $X_t = m_t + Y_t$ where $m_t = \sum_{j=0}^{k} c_j t^j$ and Y_t is stationary with mean zero, application of ∇^k gives

$$\nabla^k X_t = k! c_k + \nabla^k Y_t,$$

a stationary process with mean $k! c_k$. These considerations suggest the possibility, given any sequence $\{X_t\}$ of data, of applying the operator ∇ repeatedly until we find a sequence $\nabla_k X_t$ that can plausibly be modeled as a realization of a stationary process. It is often found in practice that the order k of differencing required is quite small, frequently one or two [4],[2]. This relies on the fact that many functions can be well approximated, on an interval of finite length, by a polynomial of reasonably low degree.

3.3.4 Estimation and Elimination of Both Trend and Seasonality

Method 1: Estimation of Trend and Seasonal Components

Suppose we have observations $x_1, \ldots, x_n$. The trend is first estimated by applying a moving average filter specially chosen to eliminate the seasonal component and to dampen the noise. If the period d is even, say $d = 2q$, then we use

$$\hat{m_t} = (0.5x_{t-q} + x_{t-q+1} + \ldots + x_{t+q-1} + 0.5x_{t+q})/d, q < t \leq n - q.$$

If the period is odd, say $d = 2q + 1$, then we use the simple moving average. The second step is to eliminate the seasonal component. For each $k = 1, \ldots, d$, we compute the average w_k of the deviations $(x_{k+jd} - \hat{m_{k+jd}})$, $q < k + jd \leq n - q$. Since these average deviations do not necessarily sum to zero, we estimate the seasonal component s_k as $\hat{s_k} = w_k - d^{-1} \sum_{i=1}^{d} w_i$, $k = 1, \ldots, d$, and $\hat{s_k} = \hat{s_{k-d}}$, $k > d$. The *deseasonalized* data is then defined to be the original series with the estimated seasonal component removed, i.e., $d_t = x_t - \hat{s_t}$,$t = 1, \ldots, n$. Finally, we reestimate the trend from the deseasonalized data d_t using one of the methods already described. We can for example fit a least squares polynomial trend $\hat{m_t}$ to the deseasonalized series. In terms of this reestimated trend and the estimated seasonal component, the estimated noise series is then given by $\hat{Y_t} = x_t - \hat{m_t} - \hat{s_t}$, $t = 1, \ldots, n$. The reestimation of the trend is done in order to have a parametric form for the trend that can be extrapolated for the purposes of prediction and simulation.

Method 2: Elimination of Trend and Seasonal Components by Differencing

The technique of differencing that we applied earlier to nonseasonal data can be adapted to deal with seasonality of period d by introducing the lag-d differencing operator ∇_d defined by

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t.$$

Applying the operator ∇_d to the model in Equation (10), where s_t has period d, we obtain

$$\nabla_d X_t = m_t - m_{t-d} + Y_t - Y_{t-d},$$

which gives a decomposition of the difference $\nabla_d X_t$ into a trend component $(m_t - m_{t-d})$ and a noise term $(Y_t - Y_{t-d})$. The trend can then be eliminated using the methods already described, in particular, by applying a power of the operator ∇.

3.3.5 Testing the Estimated Noise sequence

The objective of the data transformations described above is to produce a series with no apparent deviations from stationarity and, in particular, with no apparent trend and seasonality. Assuming that this has been done, the next step is to model the estimated noise sequence (i.e., the **residuals** obtained either by differencing the data or by estimating and subtracting the trend and seasonal components). If there is no dependence among these residuals, then we can regard them as observations of independent random variables, and there is no further modeling to be done except to estimate their mean and variance. However, if there is significant dependence among the residuals, then we need to look for more complex stationary time series model for the noise that accounts for the dependence. This will be our advantage, since dependence means in particular that past observations of the noise sequence can assist in predicting future values.

There are simple tests for checking the hypothesis that the residuals are observed values of independent and identically distributed (iid) variables. If they are, then our work is done. If not, then we must use the theory of stationary processes to find a more appropriate model.

The Sample Autocorrelation function (ACF).

For large n, the sample autocorrelations of an iid sequence $Y_1, \ldots, Y_n$ with finite variance are approximately iid with distribution $N(0, 1/n)$ [5],[7]. Hence, if $y_1, \ldots, y_n$ is a realization of such an iid sequence, about 95% of the sample autocorrelations should fall between the bounds $\pm 1.96/\sqrt{n}$. If we compute the sample autocorrelations up to lag 40 and find that more than two or three values fall outside the bounds, or that one value falls far outside the bounds, we therefore reject the iid hypothesis.

4 Description of the Mathematical Model for the Problem

The purpose of this section is to describe how we will apply the theory in Section 3 to the problem defined in Section 2. That is, we will describe the use of Qlucore and the application of Regression analysis and Time Series decomposition on the Nord Pool data set.

4.1 Qlucore Omics Explorer (QOE) Analysis

4.1.1 Overview

Qlucore Omics Explorer (QOE) is a tool for *exploration* and *visualization* of high dimensional data set. It is a powerful interactive visualization environment which helps to uncover hidden structures and find patterns in data sets, and it can be used to analyze different types of data sets such as: gene expression, protein expression, image analysis data and any multivariate data of sizes up to 1000 samples and 500000 variables or 1000 variables and 500000 samples.

Qlucore is a company from Lund in Sweden, that provides bioinformatics software for the life science and biotech industries. It was founded in early 2007 and started as a collaborative research at Lund University, with researchers at the Department of Mathematics and Clinical Genetics. The main problem that the early Qlucore project faced was the vast amount of high dimensional data generated with microarray gene expression analysis. It was recognized that an interactive scientific software tool was needed to conceptualize the ideas evolving from the research collaboration.

The basic concept behind the software, which is based on principal component analysis, is to provide a tool that takes full advantage of the most powerful pattern recognizer that exists, the human brain. The result is a core software engine that visualizes the data in 3D and aids the user in identifying hidden structures and patterns. The first released product of Qlucore was the *Qlucore Gene Expression Explorer 1.0*, and in June 2009 *Qlucore Omics Explorer (QOE)* was released.

4.1.2 PCA plot and Dynamic PCA

An important basic operation used in QOE environment, in addition to scatter plots, heat maps and data tables, is the **dynamic principal component analysis** (dynamic PCA) that presents and makes it possible to visualize high dimensional data in lower dimension. In QOE, all data is then presented in three dimensions and this three-dimensional representation is plotted on the computer screen. When working with QOE, the user has the possibility to interactively and in real time modify the different PCA plots directly on the computer screen by using several available mathematical and statistical algorithms and at the same time work with all annotations and other links in a fully integrated way. The dynamic PCA functionality is a unique feature of QOE.

The basic meaning of the PCA plot of any multidimensional data in QOE is that data points that are similar are also presented close together in the general plots. The PCA operation is characterized by the feature that it preserves as much of the originally available information as possible in the generated three-dimensional plots. The information content is then measured by the statistical variance in the data when applying PCA.

The PCA operation does not make any assumptions regarding the data set. If there are visible structures and patterns it is then because that structure is present. Some statistical methods provided in QOE (such as ANOVA) may create patterns even from random data. These patterns are then, with very high probability, not statistically stable and the statistical significance of the structure discovered must then be looked at [18]. QOE comes with several available tools for controlling statistical significance. They include *cross validation*, *randomization* or *permutation tests*. QOE also provides p-values and q-values for the chosen statistical methods, making it easy to dynamically check the statistical significance of the structures discovered.

The PCA operation is used to reduce dimension and, hence, there is in general a loss of information in the three-dimensional presentations. The PCA operation is, nevertheless, a stable and in a certain sense optimal method for dimensionality reduction and by using the flexibility of the dynamic PCA functionality in QOE the risk of missing important structures is minimized. The use of graphs and nonlinear methods such as ISOMAP proveded in QOE is also a way to minimize the risk of missing vital information concerning the data set.

4.1.3 Nord Pool Data Set in QOE

Original Data Set.

The Nord Pool data set we are using in this work covers the period from January 1999 to February 2009, it is constituted of over ten years of electricity daily observations (Prices, Consumptions and Productions) for the Nordic countries: Finland, Sweden, Denmark East, Denmark West and Norway. In addition, we have data for two background variables which are the Water reservoir and the Temperature for Sweden and Norway. In the following we are visualizing the data set in QOE.

When using QOE the original version of the data set with dimension 3712 × 3, that is 3712 daily observations are considered as Variables (genes) and 3 columns of prices, consumptions and productions are considered as Samples. In QOE the variables have two annotations: the *Types* (prices, consumptions and productions) and *Regions* (Finland, Sweden, Denmark East, Denmark West and Norway).

Figure 2 represents the plot of the Nord Pool data set in QOE. We have in red the prices, in green the consumption and in blue the production. When rotating the figure it was hard to discern any structure or pattern in the plot. The reason for this is that all genes (variables) take part in the analysis. Most of them have possibly very little to do with the different variations that we are interested in, but all of them contribute to the noise in the data by small random fluctuations. To remedy to this in QOE we apply the *Filter by variance* slider up to a value x, i.e. only genes having a variance greater or equal to x of the variance of the gene having the largest variance over the samples now take part in the analysis. In other words, we are selecting the genes that contribute most to the variation over the data set and discarding the genes that only exhibit small (possibly random) fluctuations.

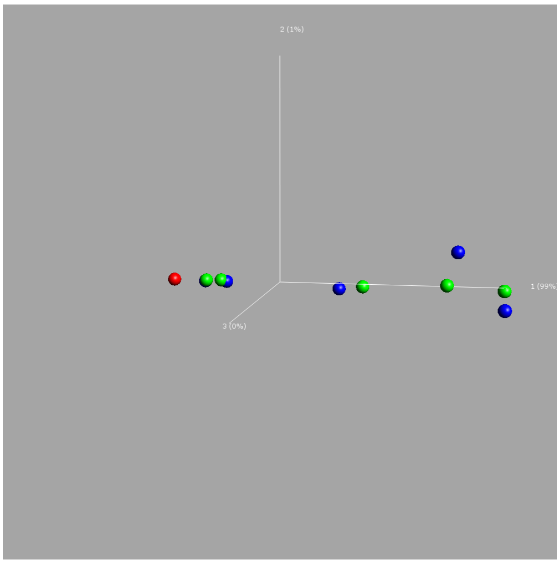

Figure 2: Original Nord Plool Data set plot in QOE.

Figure 3 shows the Nord Pool data set filtered at 0.9 and colored by types of samples, that is prices in red, productions and consumptions in green and blue respectively. We can see that prices altogether form one very integrated group, whereas productions and consumptions mix among each other. This is mostly due to different production/consumption levels respective to countries' size and population. The groups closest to price group represent both Denmark areas, the group most on the right stands for Sweden and Norway, on the remaining one for Finland.

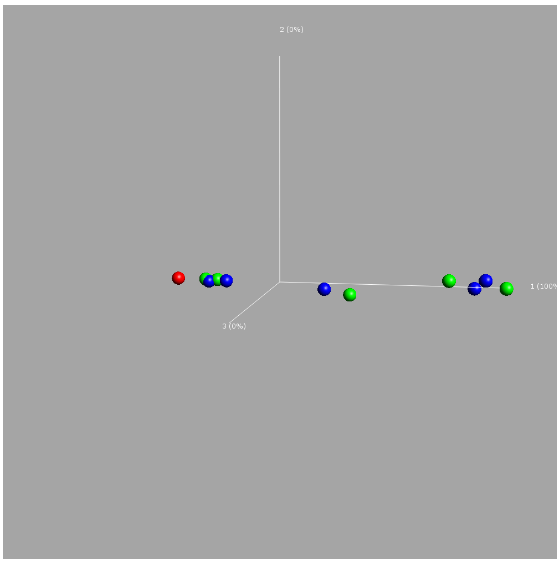

Figure 3: Original Nord Plool Data set Filtered and Colored by Type.

This is also clearly seen in Figure 4 which shows the Nord Pool data set filtered at 0.8 and colored by regions, that is Finland in green, Denmark West in white, Denmark East in yellow, Norway in blue and Sweden in orange.

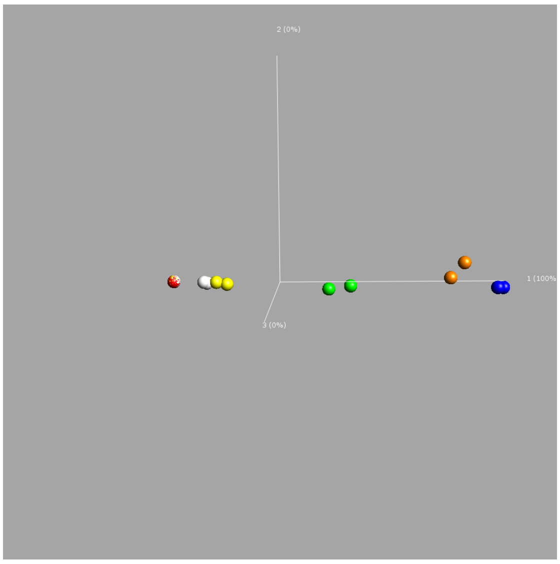

Figure 4: Original Nord Plool Data set Filtered and Colored by Region.

Preprocessed Data Set.

In Figure 5 we have the plot of the Nord Pool data set transposed so that we can have the prices, productions and consumptions as variables and as samples the daily observations. We can see that 54% of the data are on the first principal component, 29% on the second and 5% on the third.

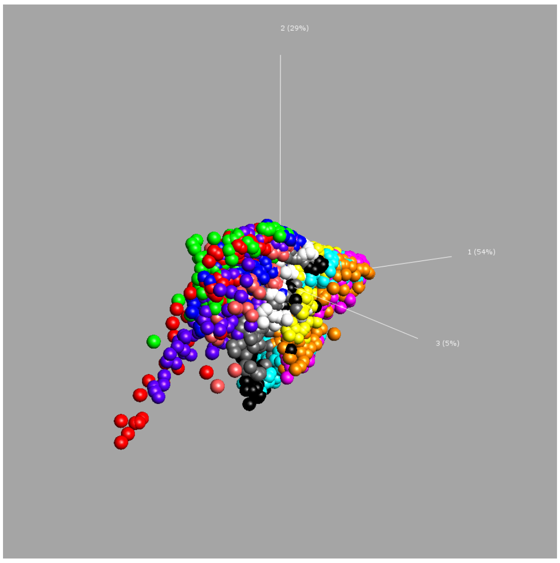

Figure 5: Nord Pool data Transposed

In Figure 6 we have the filtered data set (at 0.24) and colored by month and we can not see clearly any separate groups. However, the colors' layout does not form a purely random mix. Particular months do remain distinguishable, like most cold months on the left (December in violet, January in red and February in green) and on the right we have the most warm months (June in orange, July in pink and August in cyan) and in the middle we have April in white, September in black and May in yellow.

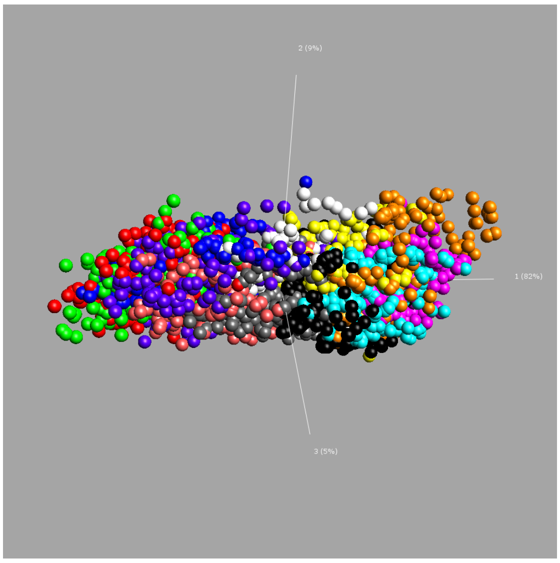

Figure 6: Nord Pool data Transposed Filtered and Colored by Month.

In Figure 7 we have the filtered data set (at 0.24) and colored by weekdays and the grouping seems less clear than in case of monthly coloring.

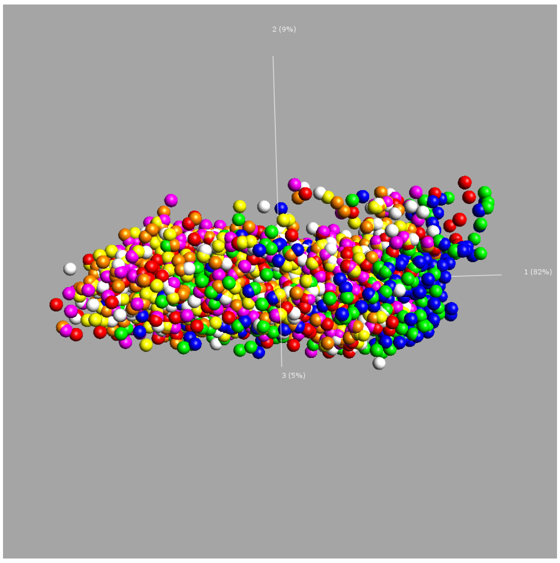

Figure 7: Nord Pool data Transposed Filtered and Colored by Weekday.

4.2 Regression based on Background variables and Time Series Decomposition

In the following section we describe the application of the regression model based on background variables (temperature and water reservoir) and consumption for both Sweden and Norway.

4.2.1 Sweden Case

Applying the regression model in Equation (4), with Y the Swedish price and X the matrix constituted by the background variables and the consumption; we get the result presented in Figure 8.

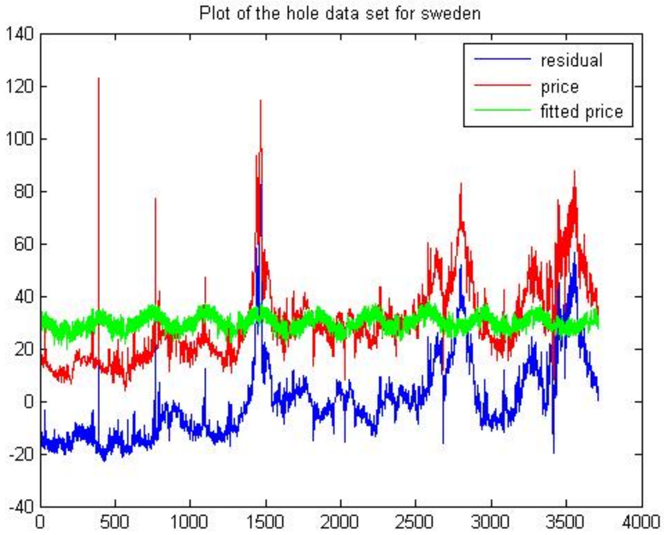

Figure 8: Plot of price fit by regression of the whole data set.

The plot shows the original price in red, the fitted price in green and the residuals in blue. Due to high volatility in the original price, the fitted price does not follow the original price. Therefore, instead of doing one global regression model (on the whole series at once) we try to remedy that by running a moving regression with different time windows: two years, one year and half a year. In particular, to estimate the price on the next day we use the observations from the past only in range of two, one or half a year. In this manner we get new regression estimates for each following day, as the historical data used for regression changes one by one.

As a result we obtain results for different moving regression horizons as presented in Figures 9, 10 and 11 respectively.

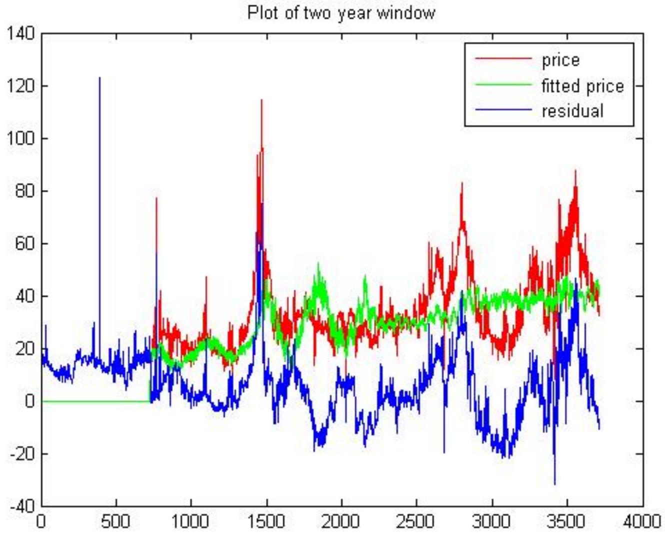

Figure 9: Plot of the moving regression fit with 2-year regression window (Sweden).

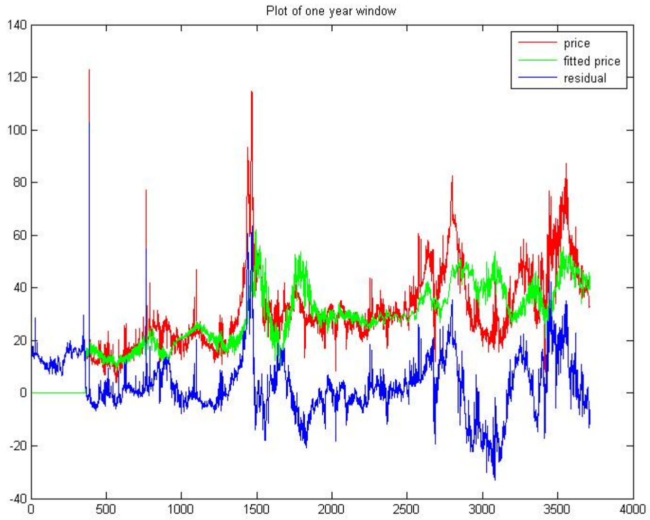

Figure 10: Plot of the moving regression fit with 1-year regression window (Sweden).

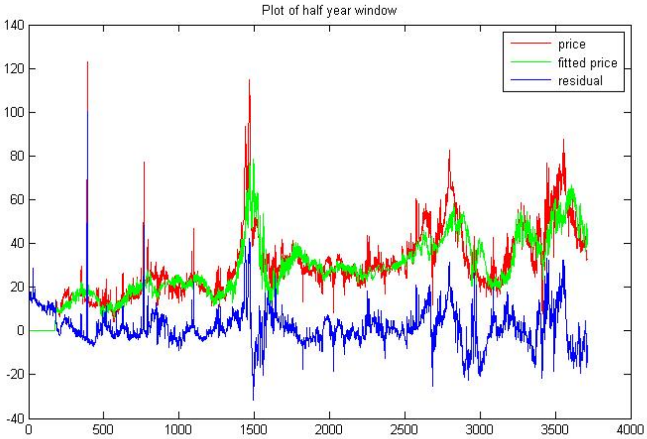

Figure 11: Plot of the moving regression fit with half a year regression window (Sweden).

In Figure 11, where we used a half year horizon we clearly see the fitted price trying to follow the original price series considerably well. The half year horizon will be used in the rest of the work for good visualization. The values of basic statistics are collected in Table 1. We can see that the residuals series have a high and positive Kurtosis, which indicates that more of the variance is the result of infrequent extreme deviations. As this value is very far from three (three is the Kurtosis value of a normal distribution) we conclude that we need to find and remove the causes of these infrequent deviations before applying again the regression model.

Table 1: Basic statistics of regression fit for Sweden prices.

Statistics	Original Price	Regression fit	Residuals	Returns
Mean	30.3644	30.3644	1.7320	-0.0001
Std	15.4189	3.1198	8.2368	0.1090
Skewness	1.2774	-0.1408	1.0136	1.0997
Kurtosis	5.4572	2.5035	10.6587	31.9881

Autocorrelation (ACF) and Partial Autocorrelation (PACF) Functions.

Visual inspection of the plots in Figure 12 and 13 for both price and residuals series shows that not most (95%) of the sample autocorrelation falls inside the bounds ($\pm 1.96 \div \sqrt{N}$). This implies that the residual series is not iid. We can also see that we have a positive and very slowly decreasing sample ACF for the price series, this implies the presence of a trend pattern. From the PACF plot of the price series we see a strong 7-day dependence which can be explained as a weekly seasonal pattern. All these findings lead to the detrending and deseasonalizing steps which are presented in the following section.

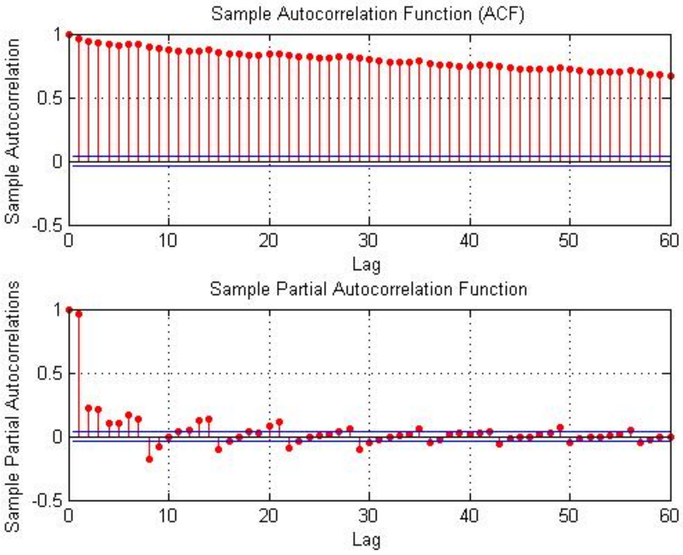

Figure 12: Plot of the ACF and PACF of the price series (Sweden).

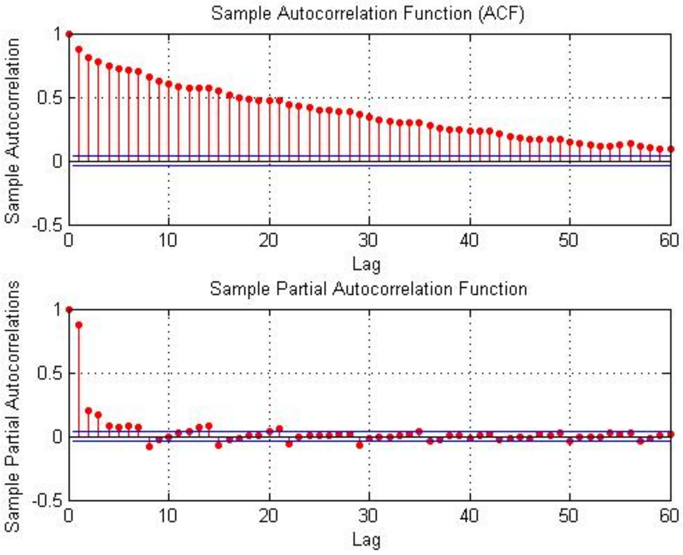

Figure 13: Plot of the ACF and PACF of the residuals series (Sweden).

4.2.2 Norway Case

Figure 14 shows the result of regression model for the Norway price series on a half year horizon, and Figures 15 and 16 the ACF and PACF of both price series and the residual series.

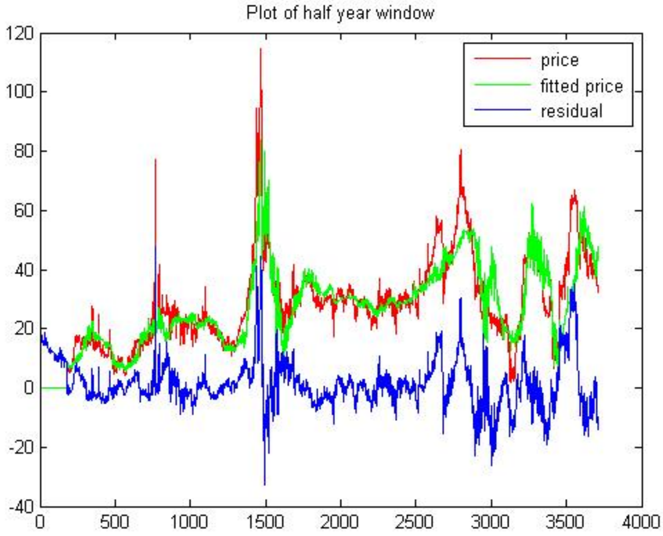

Figure 14: Plot of the moving regression fit with 182-day regression window (Norway).

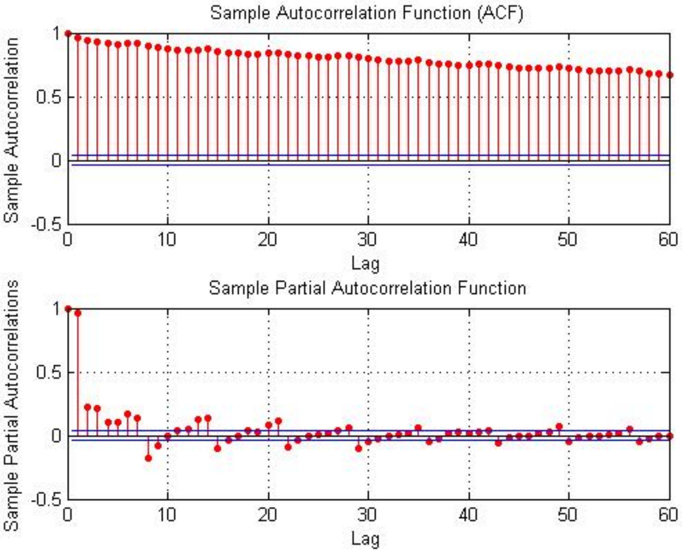

Figure 15: Plot of the ACF and PACF of the price series (Norway).

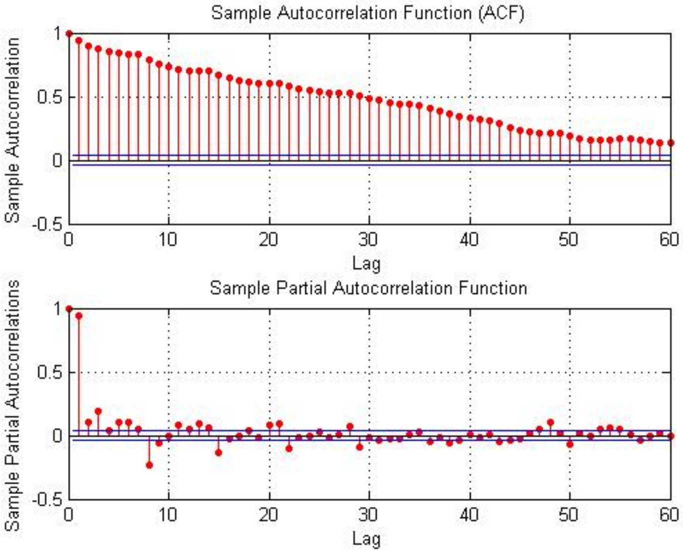

Figure 16: Plot of the ACF and PACF of the residuals series (Norway).

Table 2 shows the basic statistics and same conclusions as from Sweden case can be drawn from these results.

Table 2: Basic statistics of regression fit for Norway prices.

Statistics	Original Price	Regression fit	Residuals	Returns
Mean	28.6304	28.6304	1.5851	-0.0002
Std	14.9254	3.2132	8.2658	0.0764
Skewness	1.1934	0.1148	1.0102	1.3227
Kurtosis	5.6909	2.2593	6.2505	35.4161

5 Results

5.1 Results from Qlucore Omics Explorer (QOE)

From the findings in Section 4.1.3 we were not able to see clear patterns in Nord Pool data set plotted in QOE. In this section we are analyzing one of the most pronounced features of electricity markets which are the abrupt and generally unanticipated extreme changes in spot prices known as *jumps* or *spikes*.

Analysis of Spikes in QOE

Here we are presenting the Spikes analysis from Finland data set and the other results from other Nordic countries are collected in Appendix 1.

In Figure 17 we have the spikes representation and colored by weekdays in red and weekends in green. More spikes are seen in weekdays and when considering day by day, Wednesday is the most spiky day in Finland.

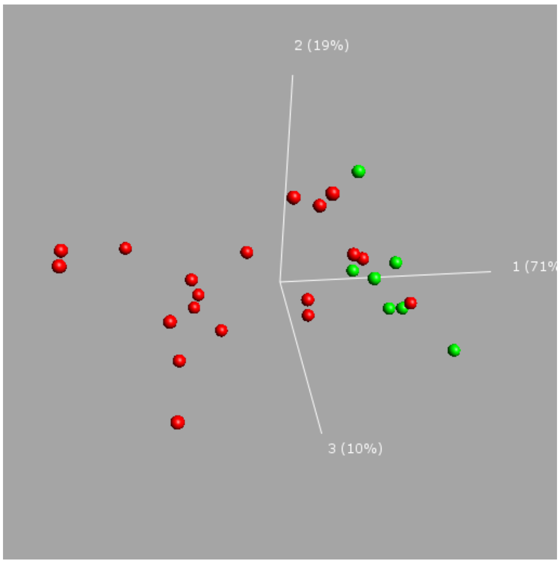

Figure 17: Plot of spikes in Finland.

The spike intensity is non-homogeneous in time. The spikes are especially notorious during peak hours, i.e. around 09:00 and 18:00 on weekdays, and during high-consumption periods like in winter time for Finland, as it can be seen in Figure 18, where we have January spikes in red, February spikes in green and December spikes in violet. We can easily see that January is the most spiky month in winter.

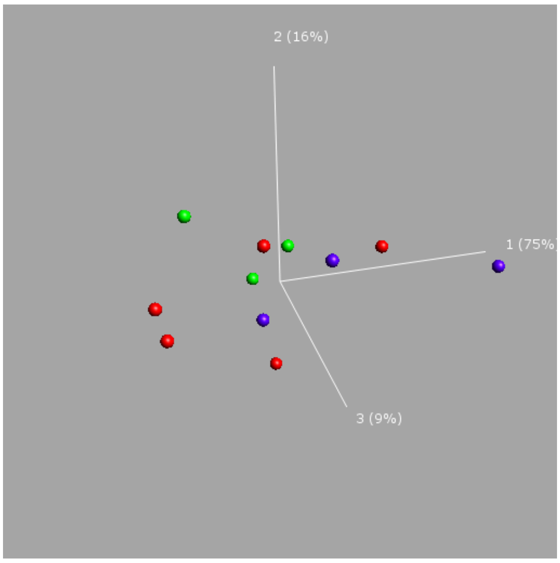

Figure 18: Plot of spikes in Winter.

The spiky nature of spot prices is the effect of non storability of electricity. Electricity to be delivered at a specific hour cannot be substituted for electricity available shortly after or before. As currently there is no efficient technology for storing vast amounts of electricity, it has to be consumed at the same time as it is produced. Hence, extreme load fluctuations, caused by severe weather conditions often in combination with generation outages or transmission failures, can lead to price spikes. The spikes are normally quite short-lived, and as soon as the weather phenomenon or outage is over, prices fall back to a normal level.

In Figure 19 we have spikes in summer colored by month. June in orange, July in pink and August in cyan. We can see that July is the most spiky month in Summer.

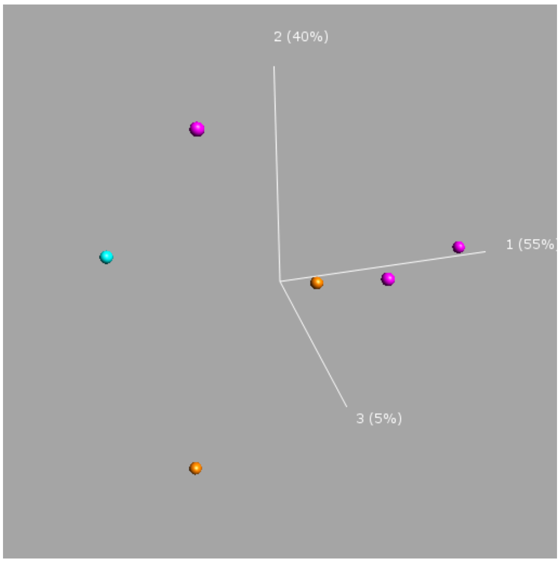

Figure 19: Plot of Finland spikes in Summer.

In Figure 20 we have spikes in spring with March in blue and May in yellow and in Figure 21 we have spikes in Fall where all the spikes appears in September.

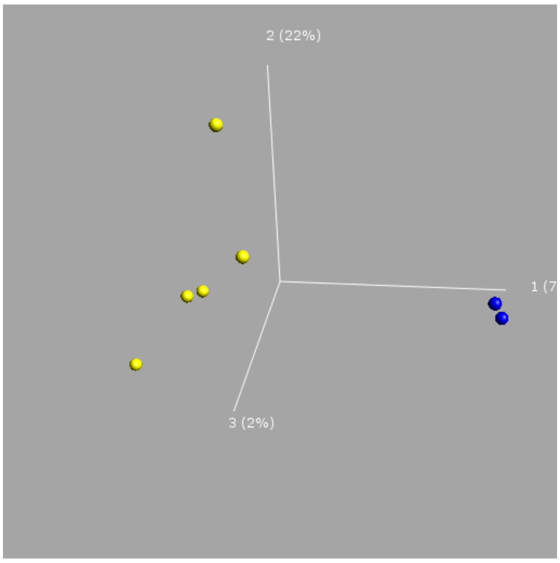

Figure 20: Plot of Finland spikes in Spring.

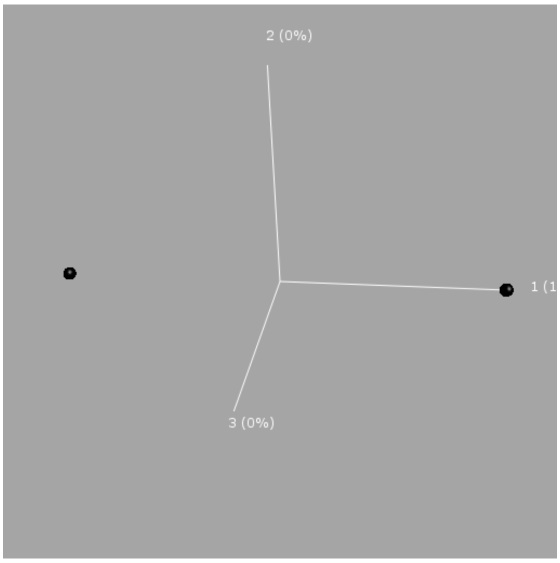

Figure 21: Plot of Finland spikes in Fall.

What we can see from the whole analysis is that also visual representation of PCA results on the Nord Pool data set does show some patterns in the prices and spikes distributions. We did see some grouping in the monthly representation and slightly less clear in case of weekday representation. Also, when presenting the weekday and weekend spikes, we can notice that in most cases these groups do not mix much. These features are related to two main periodicity types in spot prices, weekly and weather-related. Therefore, in the following section we present results from the price time series decomposition, where we deseasonalize and detrend the data, as to leave the indeterministic part for modelling purposes.

5.2 Results from Regression based on Background variables and Time Series Decomposition

5.2.1 Elimination of Trend in the Price Series

Applying Method 1 described in Section 3.3.3 (with Polynomial Fitting) on the price series we detrend the series and then use it in the regression model based on the background variables.

In Figure 22 we have the result from this analysis, and we can see in pink the original Sweden price series, in green the detrended series, in red the residuals series from the regression model and in blue the returns of the residuals series.

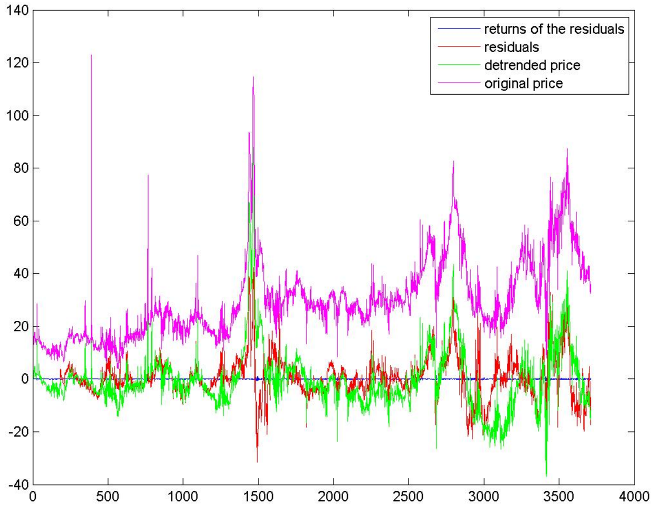

Figure 22: Plot of the regression model on the detrended Sweden price.

In Table 3 we have the basic statistics of the previous series and we can clearly see that still the residuals series is not iid.

Table 3: Basic Statistics of the detrended price (Sweden).

Statistics	Original Price	Regression fit	Residuals	Returns of the residuals
Mean	0	0	0.6444	-0.0001
Std	11.6506	2.5816	7.9165	0.1102
Skewness	2.1895	0.0147	1.2464	1.0596
Kurtosis	13.8403	2.5035	12.8790	31.0069

From Figure 23 and Table 4 the above findings can be seen in the case of Norway price series.

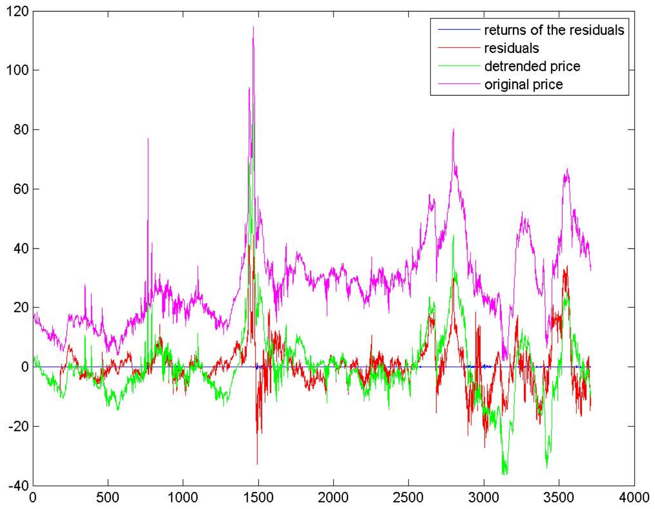

Figure 23: Plot of the regression model on the detrended Norway price.

Table 4: Basic Statistics of the detrended price (Norway).

Statistics	Original Price	Regression fit	Residuals	Returns of the residuals
Mean	-0.0000	-0.0000	0.5252	-0.0001
Std	12.4520	3.8185	7.9473	0.0805
Skewness	1.5851	0.2296	1.2279	1.4735
Kurtosis	11.2283	3.7073	7.4958	39.7755

5.2.2 Elimination of Trend and Seasonality in the Price Series

After eliminating and estimating the trend in the previous section, we use Method 1 from Section 3.3.4 to eliminate the seasonal component from the price series.

In Figure 24 we can see the detrended series in red, in blue the original series and in green the detrended and deseasonalized series in the case of Sweden price series and in Figure 25 in the case of Noway price series.

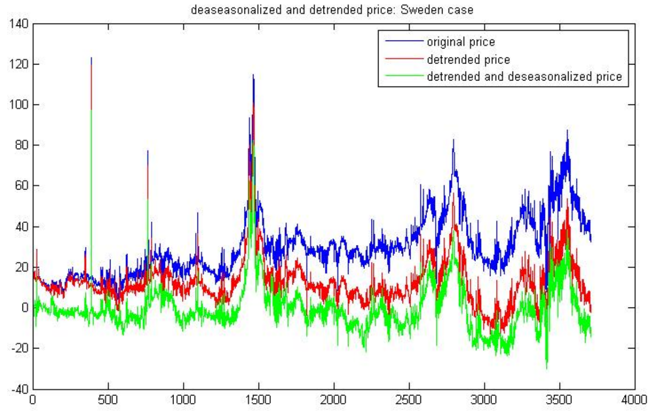

Figure 24: Plot of the detrended and deseasonalized Sweden price.

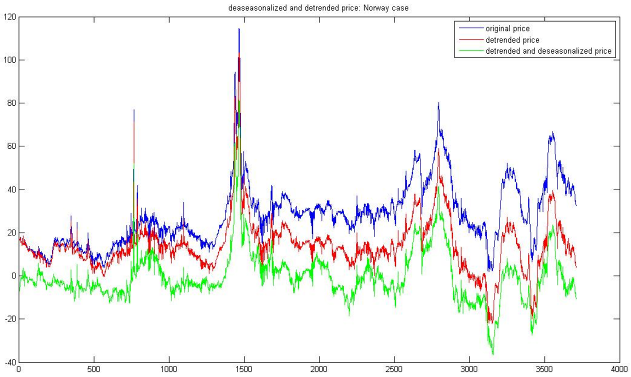

Figure 25: Plot of the detrended and deseasonalized Norway price.

The Cross Correlation Function (XCF)

The cross correlation is a standard method of estimating the degree to which two series are correlated [5],[6]. Consider two series $x(i)$ and $y(i)$ where $i = 0, 1, 2, \ldots, N-1$. The cross correlation r at delay d is defined as:

$$r = \frac{\sum_i [(x(i) - \bar{x})(y(i-d) - \bar{y})]}{\sqrt{\sum_i (x(i) - \bar{x})^2}\sqrt{\sum_i (y(i-d) - \bar{y})^2}} \tag{13}$$

where $\bar{x}$ and $\bar{y}$ are the means of the corresponding series. If the above is computed for all delays $d = 0, 1, 2, \ldots, N-1$ then it results in a cross correlation series of twice as long as the original series.

$$r(d) = \frac{\sum_i[(x(i)-\bar{x})(y(i-d)-\bar{y})]}{\sqrt{\sum_i(x(i)-\bar{x})^2}\sqrt{\sum_i(y(i-d)-\bar{y})^2}} \tag{14}$$

There is the issue of what to do when the index in the series is less than 0 or greater than or equal to the number of points ($i-d<0$ or $i-d \geq N$). The most common approaches are to either ignore these points or assuming the series x and y are zero for $i<0$ and $i \geq N$. The denominator in the expression above serves to normalize the correlation coefficients such that $-1 \leq r(d) \leq 1$, the bounds indicating maximum correlation and 0 indicating no correlation. A high negative correlation indicates a high correlation but of the inverse of one of the series.

In Figures 26 and 27 we have the plots of the cross correlation functions (XCF) of the seasonal component of the price series in Sweden and the background variables, that is water reservoir and temperature respectively. In Figures 28 and 29 we have the analogical outcome for Norway.

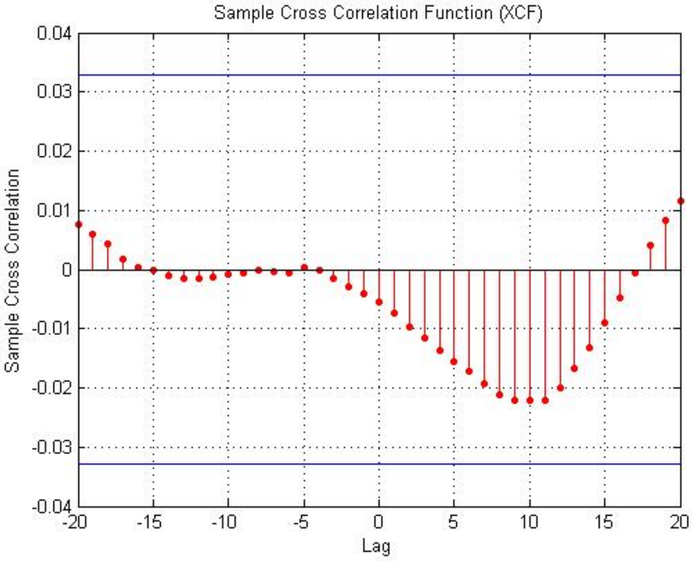

Figure 26: Plot of the cross correlation function of the seasonal component and the water reservoir variable (Sweden).

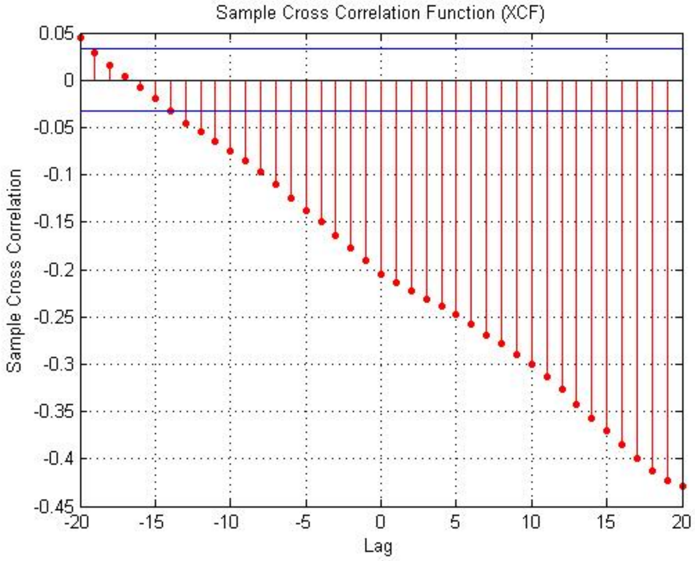

Figure 27: Plot of the cross correlation function of the seasonal component and the temperature variable (Sweden).

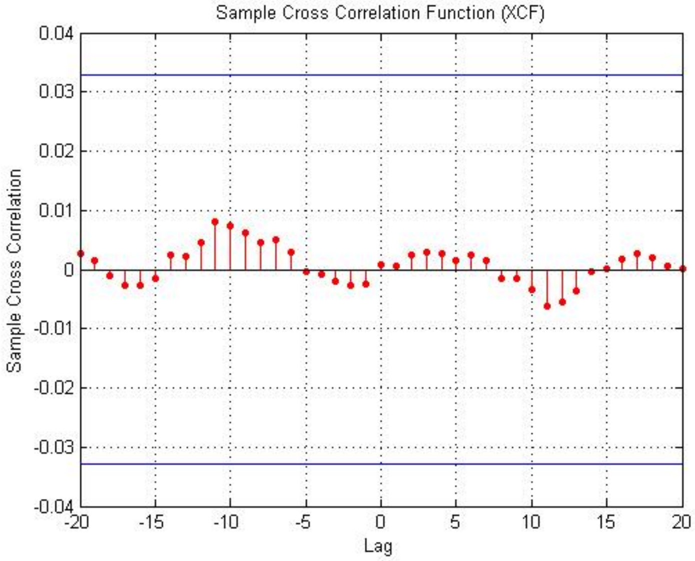

Figure 28: Plot of the cross correlation function of the seasonal component and the water reservoir variable (Norway).

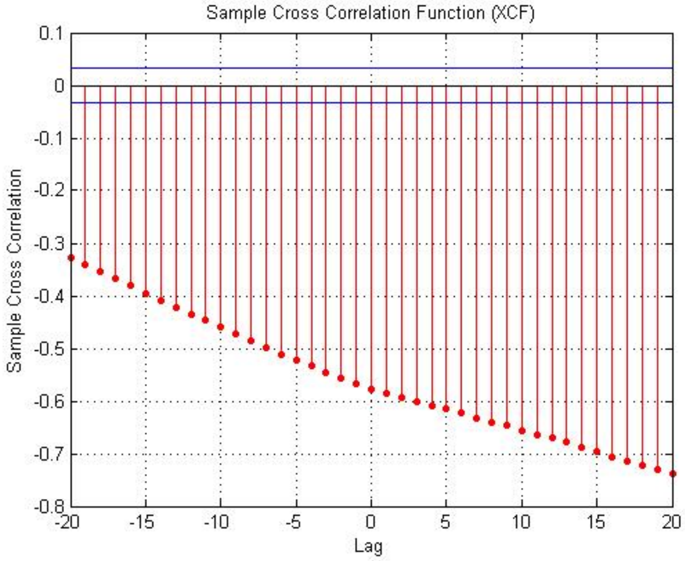

Figure 29: Plot of the cross correlation function of the seasonal component and the temperature variable (Norway).

From Figure 26 we can see that price on a given day is mostly correlated with water reservoir in Sweden 10 days ahead. This comes from the fact that generators use hydrological storage forecasts for generation planning. Therefore, for the regression model we will use the price series lagged by 10 days with respect to water reservoir storage in case of Sweden and by 11 days in case of Norway (see Figure 28).

ACF and PACF

In the following, in Figure 30 we have the ACF and PACF plots of the original as well as the detrended and deseasonalized price series for Sweden and in Figure 31 for Norway.

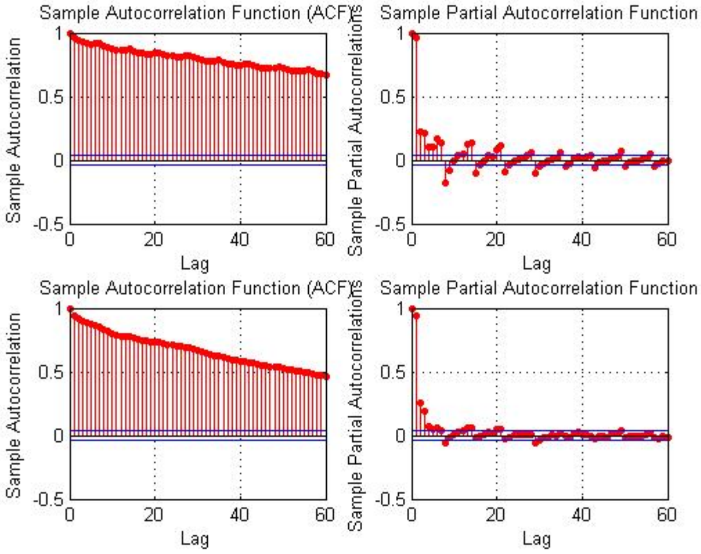

Figure 30: Plot of the ACF and the PACF of the original price and the detrended and deseasonalized price series(Sweden).

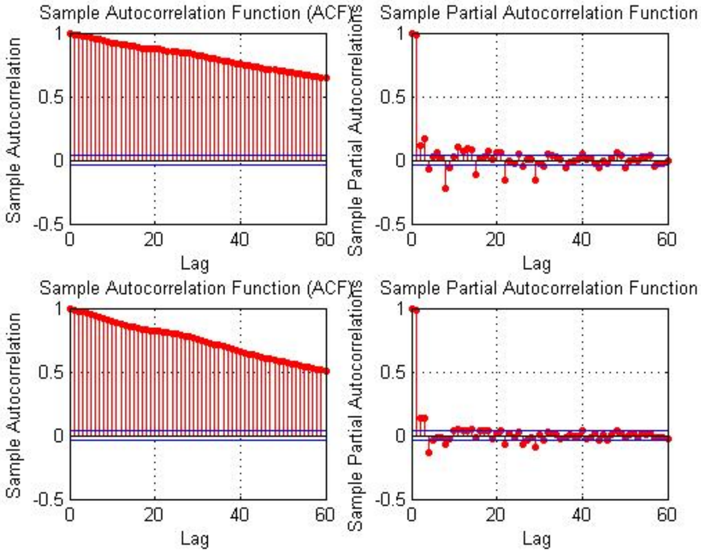

Figure 31: Plot of the ACF and the PACF of the original price and the detrended and deseasonalized price series(Norway).

We can see that we do not any longer have the trend or seasonal patterns.

5.2.3 Pure market Price Series

In the following section we show the results from the regression model of the detrended and deseasonalized price series using the detrended and deseasonalized background variables. The water reservoir series are shifted as mentioned previously.

Sweden Case

In Figure 32 we can see the plots of the price series (the detrended and deseasonalized price) in red, the fitted price by regression model in green and the residuals in blue. In Figure 33 we have the ACF and PACF plots of the cleaned series (the residuals series from the regression model) and the ACF and PACF plots of the original price series.

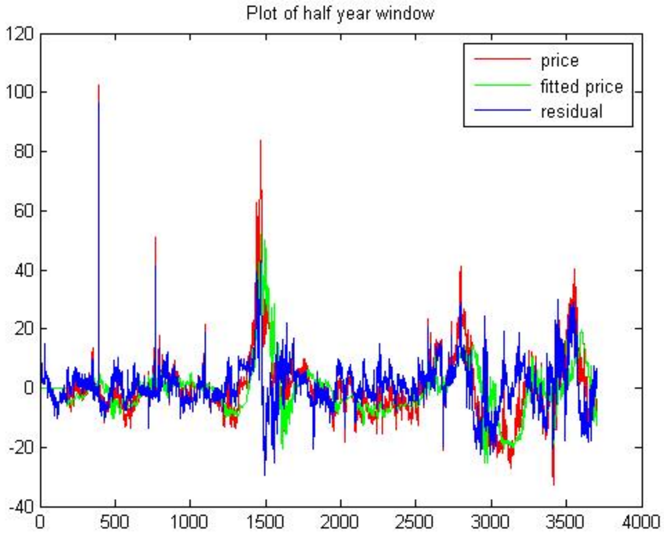

Figure 32: Plot of the regression model for a half year window (Sweden).

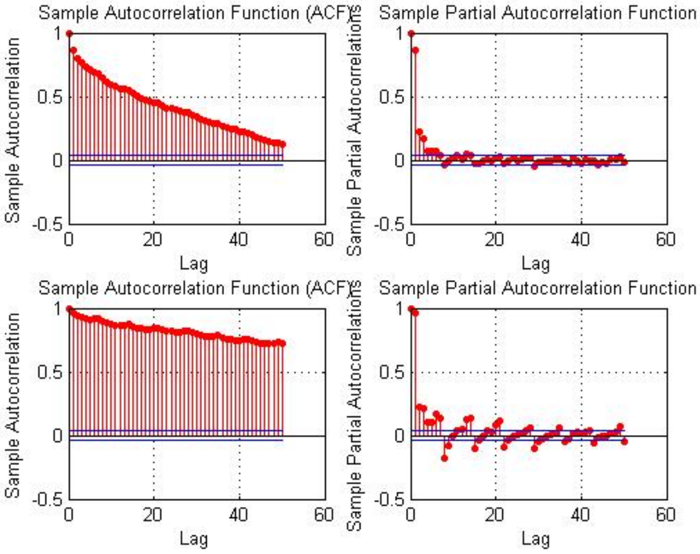

Figure 33: Plot of the ACF and PACF for the cleaned price and the original price (Sweden).

In Figure 34 we have our target series which is the residuals series from the regression model of the detrended and deseasonalized price series using the detrended and deseasonalized background variables.

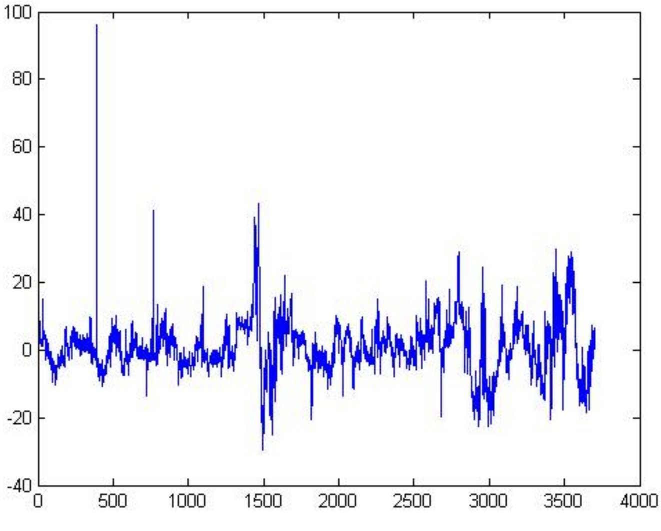

Figure 34: Plot of the cleaned price (Sweden).

Norway Case

The previous interpretation in Sweden case of Figures 32, 33 and 34 can be done for Norway case as presented in Figures 35, 36 and 37.

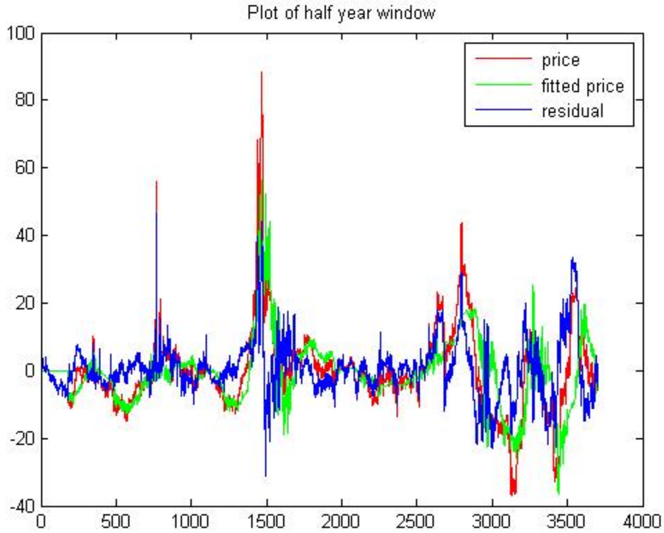

Figure 35: Plot of the regression model for a half year window (Norway).

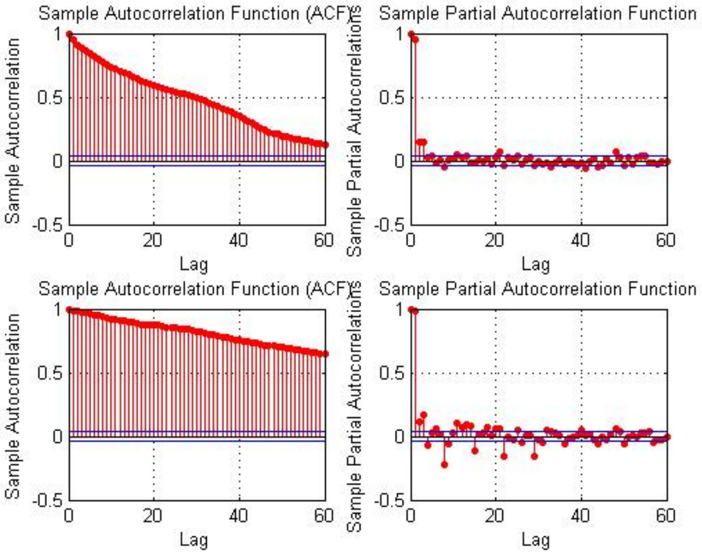

Figure 36: Plot of the ACF and PACF for the cleaned price and the original price (Norway).

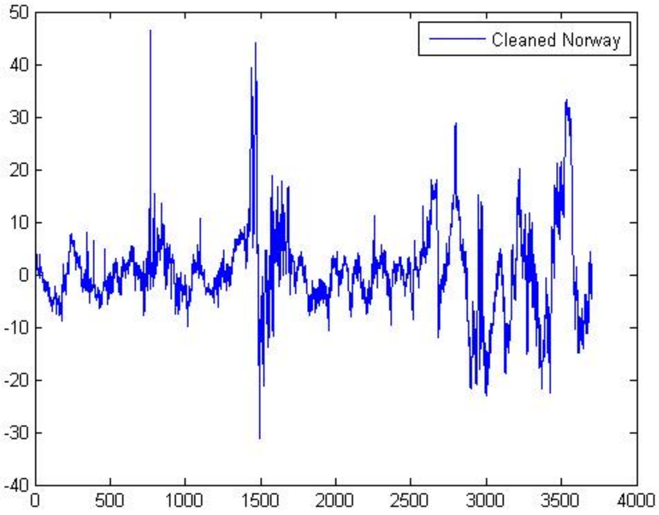

Figure 37: Plot of the cleaned price (Norway).

System Case

For the following results we have proceeded as in the case of Norway, explaining the detrended and deseasonalized System price series using the detrended and deseasonalized Norway background variables. The motivation for that is that Norway has the highest water reservoir levels as well as the highest hydro production and, therefore, we consider those most influential on the System price. Moreover, Norway price itself is in behavior closed to the System price out of all area prices.

Figures 38, 39 and 40 present the outcome for System price, analogical to the previous Sweden and Norway analysis.

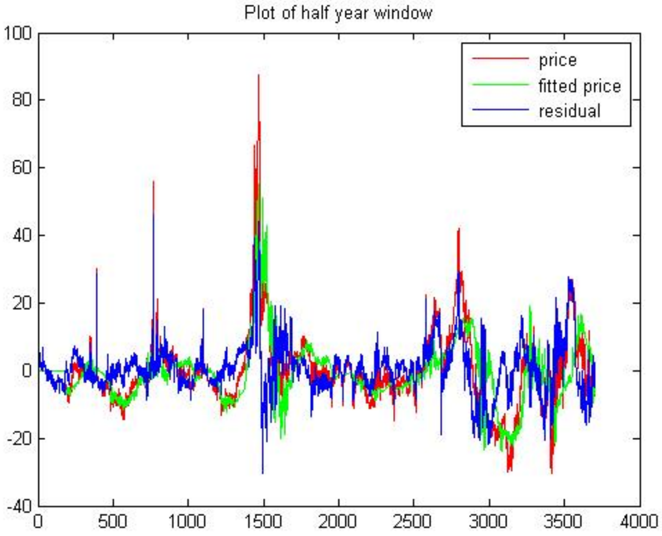

Figure 38: Plot of the regression model for a half year window (System).

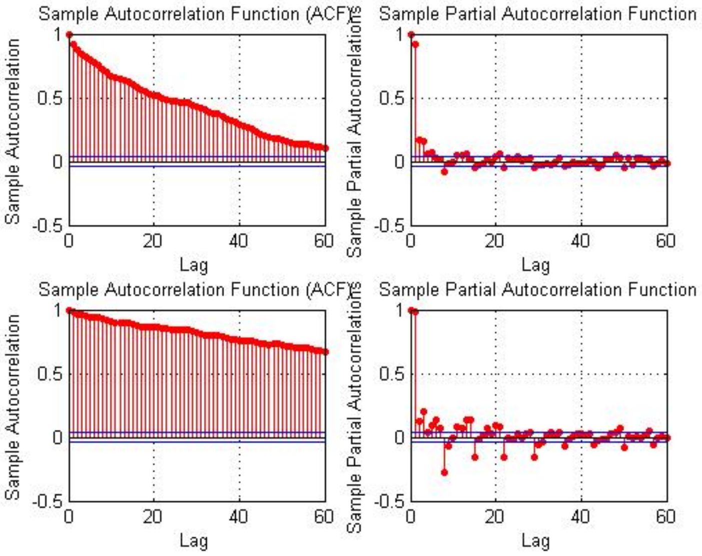

Figure 39: Plot of the ACF and PACF for the cleaned price and the original price (System).

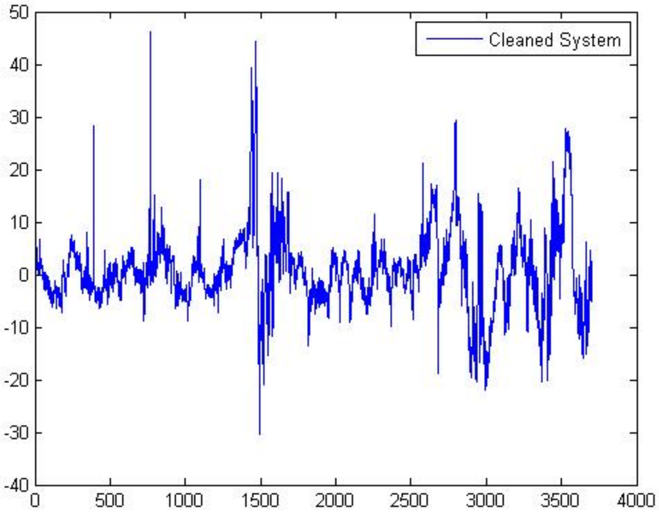

Figure 40: Plot of the cleaned price (System).

6 Conclusions

In this study an extensive analysis of the Nord Pool data set was performed. The data set covers the period from January 1999 to February 2009, it is constituted of over ten years of electricity daily observations (Spot Prices, Consumptions and Productions) for the Nordic countries: Finland, Sweden, Denmark East, Denmark West and Norway. In addition, we have data for two background variables which are the water reservoir and the temperature for Sweden and Norway.

The analysis started with an exploration and a visualization of the data set using Qlucore Omics Explorer (QOE). The aim was to find patterns in the data set. We saw some clear patterns in monthly representation and slightly less clear in weekday representation. From Qlucore we saw that Denmark East is the most spiky country and that spikes are more frequent in weekdays than in weekends. The most frequent spiky days are Wednesday and Sunday and the most spiky months are January in Winter, June in Summer, May in Spring and September in Fall.

All the features mentioned above are related to two main periodicity types in spot prices: weekly and weather related. Therefore, the aim of the next step was to work on the price time series decomposition, to deseasonalize and detrend the series as to leave the indeterministic part for modeling purposes. This operation was performed in two steps. First, the prices were detrended and deseasonalized with use of classical additive decomposition methodology, with trend assumed to be linear and two types of periodicities: weekly and annual (365 days). The resulting series had visually less obvious seasonalities, though still holding some patterns.

Some particular physical factors, that is, hydrological storage levels and temperatures, are known for having significant influence on electricity spot price behavior. Therefore, the second step after classical approach was to use the obtained detrended and deseasonalized price series in a regression model as the dependent variable. Before estimating the desired model the explanatory variables were initially detrended and deseasonalized as well, to have them treated analogically to the prices. Also, to get the best regression fit we had to make sure that the independents were properly aligned with the dependent variable in time. For that purpose the crosscorrelations between the time series were studied. As the result, we found out that prices should be lagged with respect to water reservoir levels by 10-11 days, which was connected with the hydro generators' 1-2 week ahead planning.

When having the dependent and explanatory variables properly aligned, we estimated the least-squares-optimal regression model. However, the fit was not done globally on the whole data set at once, but in a moving regression fashion, where every day a half-a-year history was used to project the resulting price for the given moment. Finally, we constructed the resulting residual series which is claimed to be the pure market series representing electricity trading characteristics.

The results still leave some space for discussion on the explanatory variables used in the regression model. One could argue that there could be some more, for instance, economical information used, like prices of fossil fuels (very influential on thermal power generation). However, the outcome of this study is considered useful, as we were able to eliminate the obvious weekly and annual periodicities, as well as the weather influence on the prices.

References

[1] Bottazzi, G., Sapio, S. and Secchi, A. (2005). *Some statistical investigations of the nature and dynamics of electricity prices.* Pysica A 355, 54-61.

[2] Botterud, A., Bhattacharyya, A.K. and Ilic, M. (2002). *Futures and spot prices. An analysis of the Scandinavian electricity market.* Proceedings of North American Power Symposium 2002, Tempe, Arizona.

[3] Box, G.E.P. and Jenkins, G.M. (1976).*Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco.

[4] Brockwell, P.J. and Davis, R.A. (1996). *Introduction to Time Series and Forecasting.* Springer-Verlag, New york.

[5] Brockwell, P.J. and Davis, R.A. (1991). *Time Series:Theory and Methods.* Second Edition, Springer-Verlag, New york.

[6] Brooks, C. (2002). *Introductory econometrics for finance.* Cambridge University Press, United Kingdom.

[7] Broszkiewicz-Suwaj, E., Makagon, A., Weron, R. and Wylomanska, A. (2004). *On detecting and modeling periodic correlation in financial data.* Physica A336, 196-205.

[8] Conejo, A.J., Plazas, M.A., Espinola, R.M. (2005). *Day ahead electricity price forecasting using the wavelet transform and ARIMA models.* IEEE Transactions on Power Systems, vol.20, No.2, pp. 1035-1042.

[9] Contreras, J. et al.; (2003). *ARIMA models to predict next day electricity prices.* IEEE Transactions on Power Systems, vol.18, No.3, pp. 1014-1020.

[10] Ghysels, E. and Osborn, D.R. (2001).*The econometric analysis of seasonal Time series.* Cambridge University Press.

[11] Pearson, K., Yule, G.U., Blanchard, N., and Lee, A. (1903). *The Law of Ancestral Heredity.* Biometrika, vol.2,No.2 (Feb.,1903),pp.211-236.

[12] Schumway, R.H., and Staffer, D.S. (2006). *Time Series Analysis and its Application. With R Examples* Second Edition, Springer.

[13] Weron, R., Simonsen, I. and Wilman, P. (2004).*Modeling highly volatile and seasonal markets: Evidence from the Nord Pool electricity market.*In: The application of Econophysics, H.Takayasu (ed.). Springer, Tokyo, pp.182-191

[14] Weron, R. and Misiorek, A. (2005). *Forecasting spot electricity prices with time series models.* Proceedings of the European Electricity Market EEM-05 conference, Lodz, 133-141.

[15] Weron, R. (2006). *Modeling and forecasting electricity loads and prices. A statistical Approach.* John Wiley & Sons,Ltd.

[16] Zarnowitz V., A. Ozyildirim (2006). *Time Series decomposition and measurement of business cycles, trends and growth cycles.* Journal of Monetary Economics, Elsevier.

[17] Zivot E. and Jiahui Wang (2002). *Modeling Financial Time Series with S-Plus.*

[18] *www.qlucore.com/documentation.aspx*

[19] *www.nordpool.com*

[20] *wikipedia.org/wiki/Regression-analysis.*

APPENDIX I – Qlucore Results

Sweden Spikes

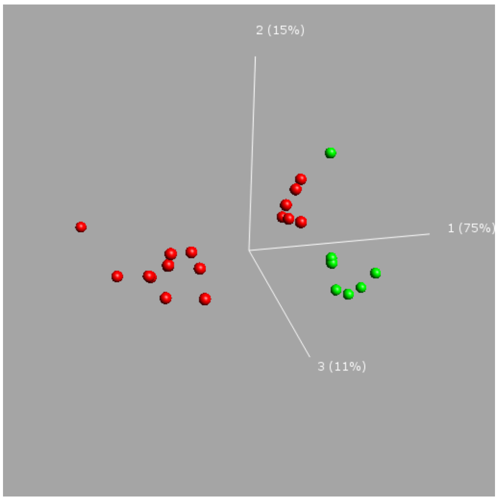

Figure 41: Sweden spikes

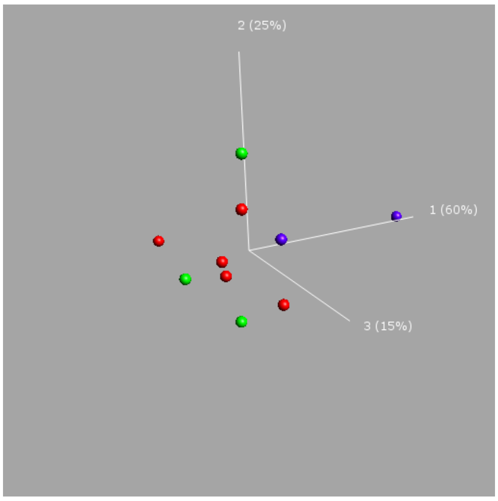

Figure 42: Sweden spikes in Winter

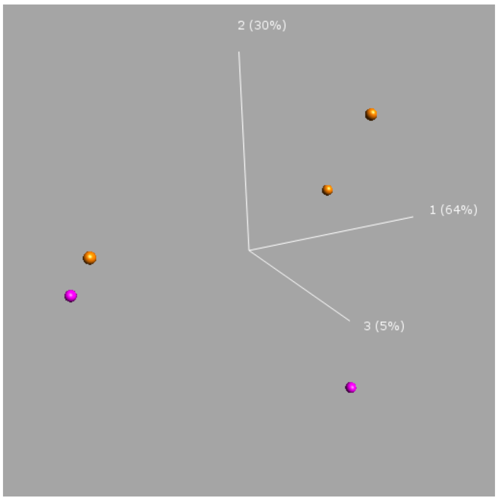

Figure 43: Sweden spikes in Winter

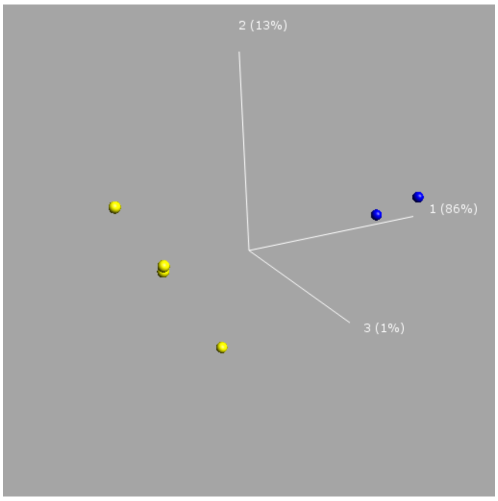

Figure 44: Sweden spikes in Summer

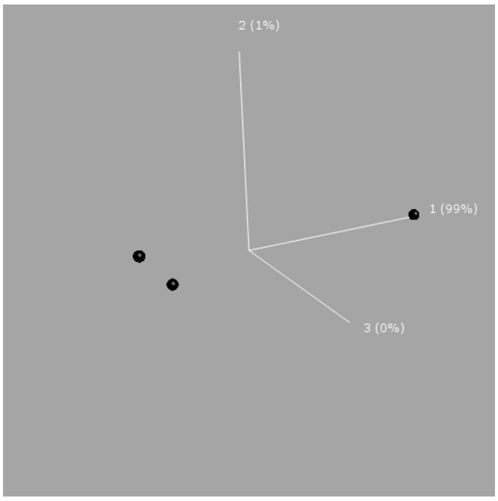

Figure 45: Sweden spikes in Spring

Norway Spikes

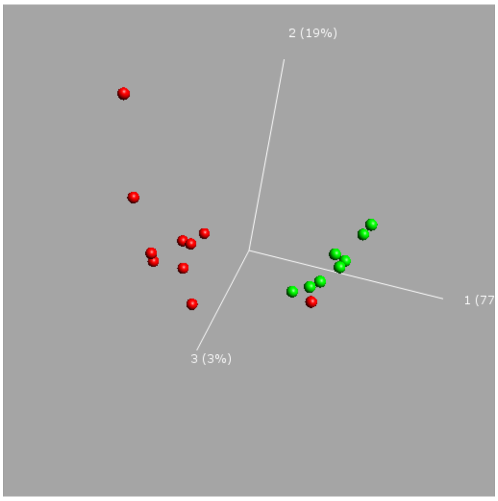

Figure 46: Norway spikes

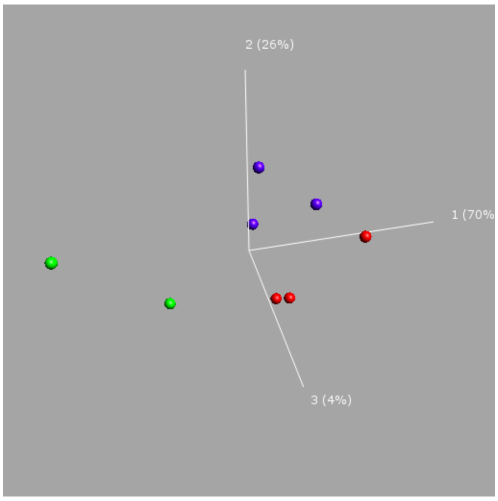

Figure 47: Norway spikes in Winter

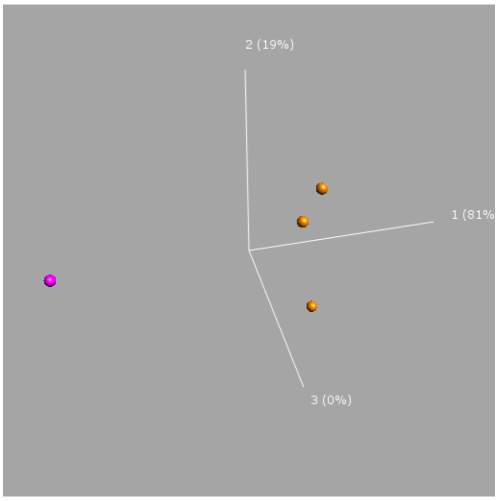

Figure 48: Norway spikes in Summer

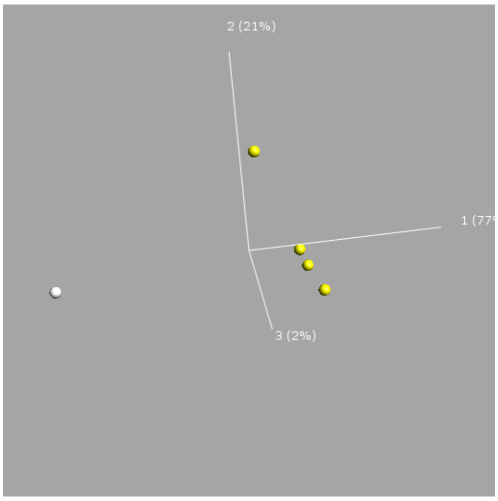

Figure 49: Norway spikes in Spring

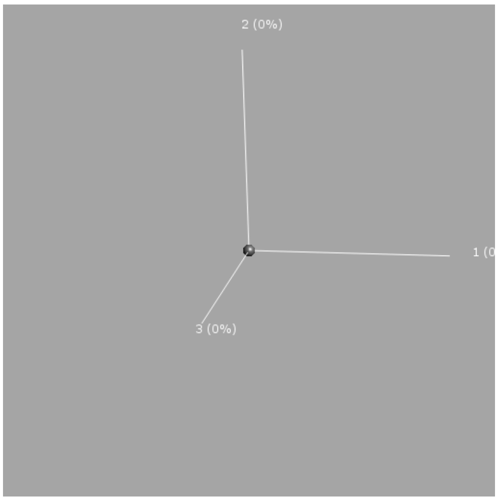

Figure 50: Norway spikes in Fall

DenmarkE Spikes

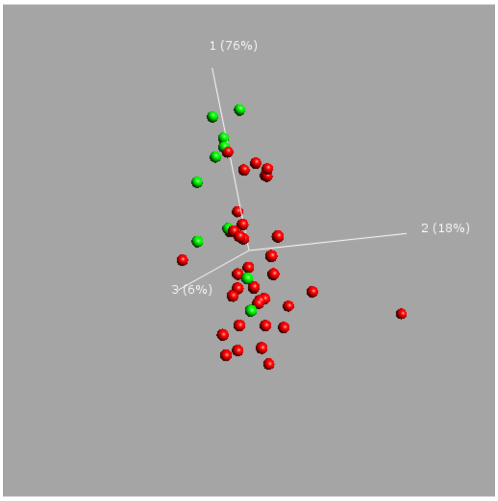

Figure 51: DenmarkE spikes

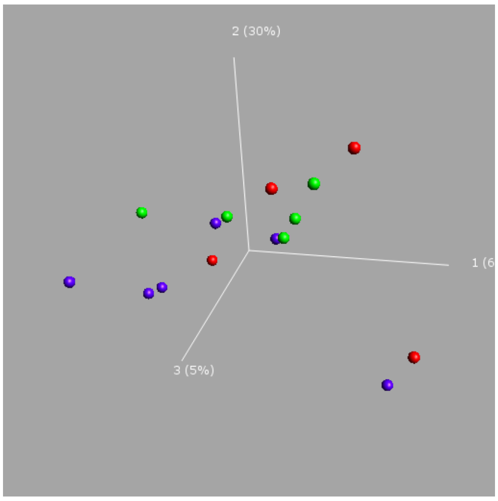

Figure 52: DenmarkE spikes in Winter

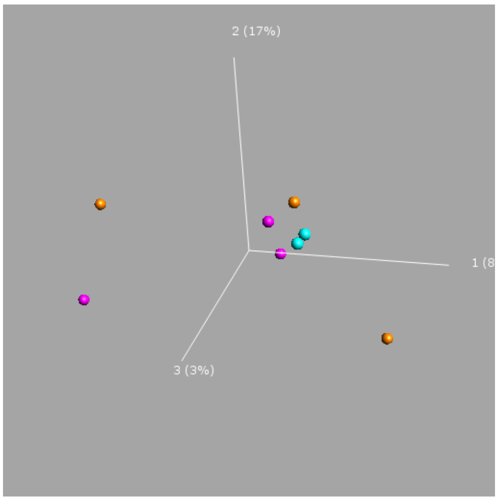

Figure 53: DenmarkE spikes in Summer

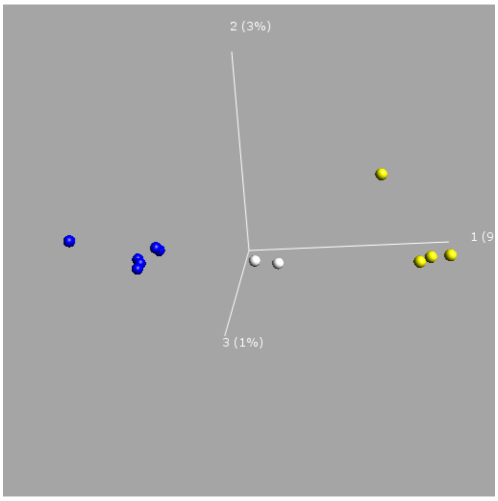

Figure 54: DenmarkE spikes in Spring

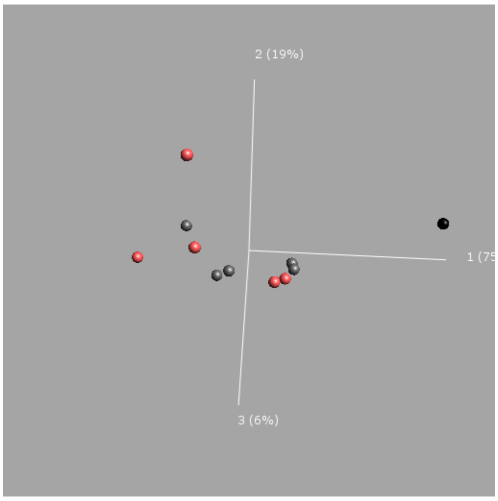

Figure 55: DenmarkE spikes in Fall

DenmarkW Spikes

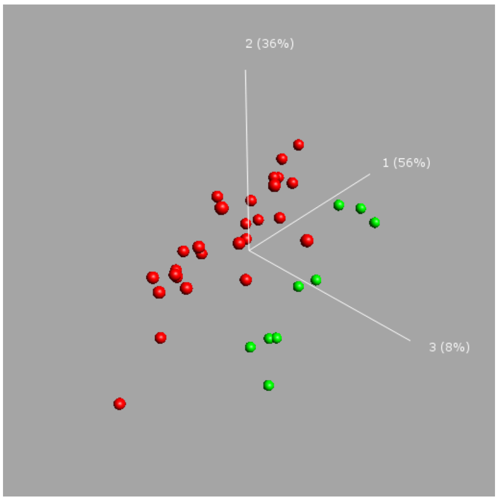

Figure 56: DenmarkW spikes

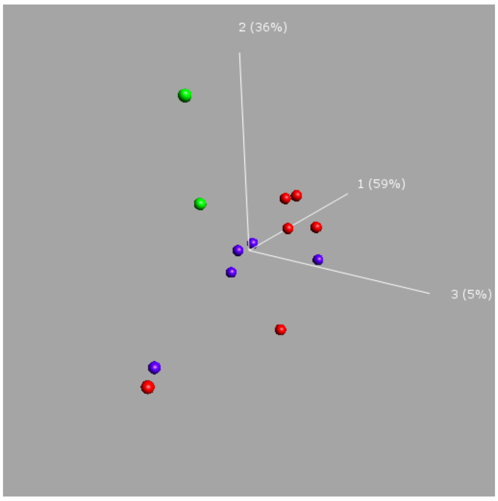

Figure 57: DenmarkW spikes in Winter

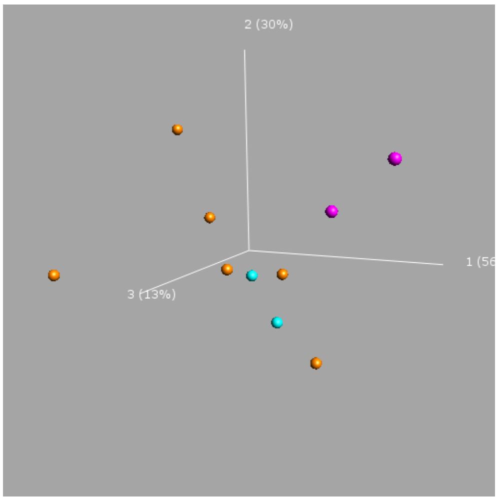

Figure 58: DenmarkW spikes in Summer

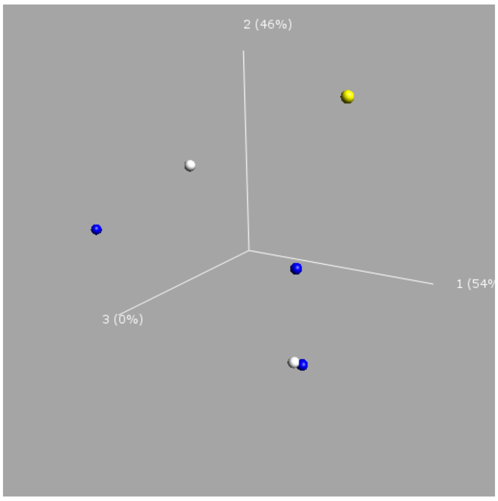

Figure 59: DenmarkW spikes in Spring

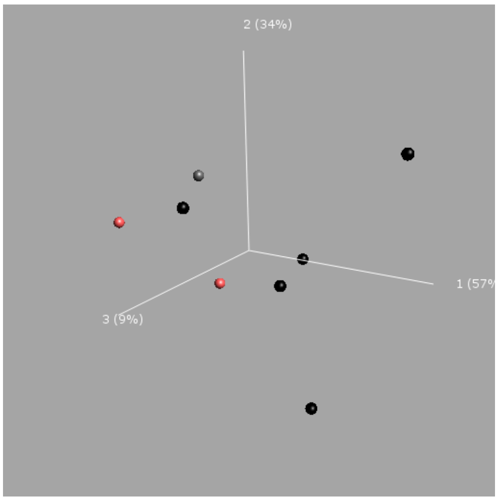

Figure 60: DenmarkW spikes in Fall