LUND UNIVERSITY
School of Economics and Management

Push it to the limit

Testing the usefulness of Extreme Value Theory in electricity markets

Authors Henrik Fulgentiusson

Course: NEKP01

Tutors: Birger Nilsson and Rikard Green

Date: 2012-10-26

Abstract

We set out to investigate whether the methodologies used in extreme value analysis are applicable in estimating Value at Risk (VaR) for the spot price returns of the European Energy Exchange (EEX). An initial inspection of hourly data reveals a volatile behaviour where returns of extreme proportions occur frequently. Applying the two traditional extreme value theory methods to the data, *Block Maxima* and *Peaks Over Threshold,* we find that while the latter perform better, it is highly sensitive to the chosen threshold and thereby the confidence level.

Suspecting that spikes in the original data series might have distorted the estimated parameters and consequently the measures of VaR, we perform a similar analysis on daily data obtained through aggregating the hourly observations. Now we find a clear weekly dependence and an AR-GARCH-filtering approach suggested by McNeil and Frey (2000) is employed. Unfortunately, the risk measures based on daily data did not improve the overall picture as much as we had hoped, as they clearly still deviate from their theoretical values and could be rejected for most confidence levels.

Backtesting the daily data did improve the results and we found that a conditional AR-GARCH-model outperformed an unconditional one, however only slightly. Our findings would thus suggest that the two classical extreme value methodologies can be used to model the extreme tails of the return distribution, but that they are not as accurate as found in other electricity markets. In order to increase the accuracy, one would need to constantly update the model parameters. Furthermore, we believe that more advanced modelling, taking spikes and mean reversion of the data into account, could lead to improvements.

Keywords: Electricity, Extreme Value Theory, Peaks Over Threshold, Block Maxima, Value at Risk.

Acknowledgements

I would like to thank my tutors Associate Professor Birger Nilsson and PhD Rikard Green for providing the data and for their valuable input during the process of writing this thesis. I would also like to express my gratitude to my family for their support and encouragement.

Contents

1. INTRODUCTION 5
1.1. BACKGROUND 5
1.2. PURPOSE AND CONTRIBUTION 6
1.3. OUTLINE 6

2. THEORETICAL FOUNDATION 7
2.1. VALUE AT RISK 7
2.2. MAXIMUM LIKELIHOOD ESTIMATION 8
2.3. EXTREME VALUE THEORY 8
2.4. BLOCK MAXIMA 9
2.4.1.Estimating GEV parameters 10
2.4.2. Value at Risk estimation for GEV 11
2.5. PEAKS OVER THRESHOLD 11
2.5.1 Threshold selection 12
2.5.2. Estimating parameters GPD 12
2.5.3. Value at Risk calculation for GPD 13

3. PREVIOUS RESEARCH 14
3.1. APPLICATIONS IN STOCK MARKETS 14
3.2. APPLICATIONS IN ELECTRICITY MARKETS 14

4. DATA ANALYSIS 16
4.1. DATA DESCRIPTION 16
4.1.1. Negative Prices 16
4.2. DATA MANIPULATIONS 17
4.3 PROPERTIES OF RETURNS 18
4.3.1. Hourly returns 18
4.3.2. Daily returns 20

5.METHODOLOGY 21
5.1. HOURLY DATA 21
5.1.1. Block Maxima 21
5.1.2. Peaks Over Threshold 21
5.2. DAILY DATA 23
5.2.1. AR-GARCH-EVT 23
5.2.2. Block Maxima 25
5.2.3. Peaks Over Threshold 26
5.3. BACKTESTING 27
5.4. KUPIEC TEST 28

6. RESULTS 29
6.1. HOURLY RETURNS 29
6.2. DAILY RETURNS 30
6.3. BACKTESTING 31

7. CONCLUSION 33
7.1. SUMMARY 33
7.2. CONCLUDING REMARKS 34
7.3. SUGGESTIONS FOR FUTURE RESEARCH 35

BIBLIOGRAPHY 36

1. Introduction

1.1. Background

The introduction of the Basel II framework in the mid 90's calls for financial institutions to set aside a minimum amount of regulatory capital to cover potential losses from their exposure to financial risk. In light of this framework, a new measure of risk, called *Value at Risk* (VaR), emerged and quickly gained huge popularity as it directly shows the amount a firm stands to lose at a particular time horizon and at a given significance level. Considering that Basel II penalize small and frequent violations of VaR this makes finding an accurate estimate highly important. Especially since it allows the firm to efficiently allocate capital by minimizing the amount they set aside for their regulatory capital reserve.

Hence, VaR is conceptually easy to understand and is advantageous from several perspectives, but acquiring good estimates depends crucially on the underlying distribution of returns. Early studies assumed normality of returns, which makes VaR easy to derive using historical data. But as history has shown, financial returns do not behave normally and the tails of the distribution have been found to be much longer. Using the normal distribution as the basis for VaR estimates can thus cause investors and risk managers alike to seriously underestimate future risk in times of turmoil.

Another problem with using VaR as a risk measure is that it only shows what we stand to lose at a given probability. Thus, it is silent about the magnitude of losses should a rare- or extreme event occur. One way of mitigating this problem is by applying what is called *Extreme Value Theory* (EVT). The advantage of using extreme value models is that they direct their focus to finding a suitable distribution for the large- and extreme losses, which are what is relevant for obtaining VaR. Accordingly, studies implementing these kinds of methods, found that they were able to improve their forecasts of risk considerably.

The worldwide liberalization of electricity markets in the early 90's has led to a large increase in trade of electricity, both in the spot prices and the derivatives using these prices as the underlying asset. The electricity markets differ considerably from regular financial markets as prices often display mean-reversion, price spikes and time-varying volatility (Huisman, 2013). Whereas stock and similar financial assets appear to be fairly stable over time, we have seen that electricity prices can more than double from one hour to the next, and in some markets even exhibit negative prices. Trade conducted in highly volatile environments such as these thus further stresses the need for proper risk management, and the extreme behaviour of energy prices could potentially make extreme value analysis a candidate for these types of markets.

1.2. Purpose and Contribution

In this study we set out to apply extreme value theory on the spot price returns of the *European Energy Exchange* (EEX). We carry out our investigation by employing the two classic extreme value models: *Block Maxima* and *Peaks over Threshold* to see whether they are applicable in modelling the tails of the return distribution. We then test whether these methods could yield good estimates of Value at Risk and if a more dynamic model specification could further improve the accuracy.

Studies applying the EVT methodology to electricity data are scarce, and to our best knowledge there are not one studying the EEX spot market specifically. Our choice of data is also of particular interest as EEX is one of the largest electricity markets in Europe and that it exhibits extreme price swings where even negative prices occur frequently. The few studies that have investigated electricity markets using a similar methodology to what is used in the forthcoming paper have all investigated the right tail of the return distribution (e.g. Chan and Gray, 2006), which is why we decided to do the same.

Given that these studies have found that extreme value models yields improved estimates of VaR compared to standard parametric- and non-parametric approaches, we hope to further shed some light on the situation by using a similar methodology applied on another market.

1.3. Outline

The remainder of the paper is organized as follows. Section 2 provides the reader with the theoretical foundation on which extreme value analysis and Value at Risk are based. Section 3 presents the findings from previous research applied on electricity spot markets as well as from other markets using similar methodologies. Section 4 describes the data used in the study and also the manipulations that were undertaken. We also perform basic statistical tests to investigate the properties of the raw returns. In Section 5 we present the methodology that is used for carrying out the tests. The results from theses tests are thereafter presented and analysed in Section 6. Section 7 concludes the paper and suggestions for future research are provided.

2. Theoretical Foundation

2.1. Value at Risk

This study aims at describing the appropriateness of extreme value models in determining the risk of electricity prices. While there are many measures of risk, we have chosen to test this hypothesis in terms of *Value at Risk (VaR).* This is one of the most prominent risk measures and it shows the maximum amount one stands to lose on a portfolio over a predefined time period, not exceeded at a high probability. Mathematically, it can be defined as follows:

Definition: Value at Risk

For a confidence level $\alpha \in (0,1)$, VaR can be defined as the smallest future loss of a portfolio l, for which the probability of a loss (L) exceeding l is $(1-\alpha)$ (McNeil and Frey, 2000). Or, more formally as

$$VaR_\alpha = \inf\{l \in \mathbb{R}: P(L > l) \leq 1-\alpha\} = \inf\{l \in \mathbb{R} : F_L(l) \geq \alpha\},$$

which can be rewritten as:

$$VaR_\alpha = q_\alpha(F_L) = F_L^{-1}(\alpha)$$

From the formula above, we see that VaR can be viewed as the quantile of a profit- and loss-distribution (P/L). One advantage of VaR is that it can be applied on all asset classes regardless of the underlying P/L-distribution. Since it also gives us an estimate of risk in actual in figures it has come to grown highly popular measure of risk due to its simplicity (Dowd, 2005, p. 9-13).

Expressing VaR in terms of quantiles does however require some knowledge of the underlying P/L-distribution if one wants accurate estimates. Given the reasons mentioned in the introduction, accurate estimates of VaR allows the firm to efficiently allocate their capital while at the same time minimizing the risk of being penalized according to the Basel II framework. VaR estimation as proposed by J.P Morgan's RiskMetrics™ assumes multivariate normality, which is problematic as it is has no support from empirical investigations (McNeil and Frey, 2000). For instance, financial asset returns and exchange rates often display a positive degree of excess kurtosis, i.e. longer tails than suggested by the normal distribution. An immediate consequence of this is that the probability of observing large losses increases in comparison with the normal distribution (Duffie and Pan, 1997).

A drawback of VaR is that it does not give us any information about the size of losses occurring with a probability smaller than $(1-\alpha)$. There is also an issue regarding its coherency, as VaR is not subadditive

and hence not coherent. The interpretation of the latter suggests that diversification does not reduce risk, which goes against findings found in portfolio choice theory. [1]

2.2. Maximum Likelihood Estimation

In the coming sections, we use *Maximum Likelihood Estimations* (MLE) to estimate the parameters in the different models. The parameter estimates retrieved from MLE are those assigned the highest probability given the data. We begin by defining the likelihood function:

Definition: Likelihood function

Assuming that $x_1, \dots x_n$ are independent realisations of a random variable with probability density function $f(x_i; \theta)$, the log likelihood function is given by:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

By taking the logarithm of the above expression we retrieve the log-likelihood function, which is then maximized:

$$l(\theta) = \ln L(\theta) = \sum_{1=1}^{n} \ln \left(f(x_i; \theta)\right)$$

2.3. Extreme Value Theory

In this section, we present the theory underlying extreme value theory and how it can be used to carry out the coming tests. The following two sections will be devoted to the two classical extreme value models, *Block Maxima (BM)* and *Peaks Over Threshold (POT)*. After providing the reader with the background for these models, we will show how to use them in order to estimate VaR. The coming exposition is fairly technical and requires some knowledge of probability theory and statistics, even though we will try to present it as simplistic as possible. For readability, some steps have purposely been left out when deriving the results, but the interested reader will be referred to the original sources. The notation used throughout this section corresponds to that used by Coles (2001).

A simple way of describing extreme value analysis is to think of it in a similar way as the *Central Limit Theorem* (CLT). But instead of the mean, one directs attention to the behaviour of extremes. In order to describe these methods we must first pose a definition, which is used in both of these models:

[1] Other properties of VaR and the discussion of alternative risk measures can be found in among others Dowd (2005, Chapter 1-2), Embrechts (2003, p. 4-6) and Jorion (1997).

Definition: Maxima

If we let $X_1, \dots, X_n$ be a sequence of independent and identically distributed (iid) random variables with distribution function $F(x) = Prob(X \leq x)$, we can express the maxima of this sequence as

$$M_n = \max(X_1, \dots X_n)$$

where M_n has a distribution function $F_{M_n}(z) = \mathbb{P}(\mathrm{M_n} \leq z) = \{F_{X_1}(z)\}^n$

2.4. Block Maxima

Theoretically, it is possible do derive the distribution of M_n exactly, but as we do not know what distribution *F* really has this is of little help practically. Estimating *F* from observed data is possible, but even small biases in the estimation could give us large estimation errors of F^n (Coles, 2001, p. 46).

Using the insights found in the CLT, we could estimate the distribution of M_n by investigating its asymptotic properties. By considering the limiting distributions of:

$$\frac{M_n - b_n}{a_n}$$

where a_n and b_n are sequences of normalising constants where $a_n > 0$. These can be thought of constants similar to those used in CLT for deriving a population mean from a sample. In a similar way, the *Extremal Types Theorem* can be used to show the range of possible limit distributions for extreme values.

Definition: Extremal Types Theorem

If we have a sequence of constants $a_n > 0$ and b_n and where G is a non-degenerate[2] distribution function,

$$\Pr\{(M_n - b_n)/(a_n \leq z\} \to G(z) \text{ when } n \to \infty,$$

G is said to belong to of the following three *extreme value distributions*:

$$\mathbb{I}: Gumbel: G(z) = \exp\left\{-exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \qquad -\infty < z < \infty$$

$$\mathbb{II}: Fréchet: G(z) = \begin{cases} 0, & z \leq b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b \end{cases}$$

$$\mathbb{III}: Weibull: G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b \\ 1, & z \geq b \end{cases}$$

for $a > 0$, b, and for II and III, $\alpha > 0$. These families all have location- and scale parameters, a and b, while families II and III also has a shape parameter, α. We can then rewrite the above expressions to obtain a distribution function that subsumes all three families, called the *Generalized Extreme Value-distribution (GEV):*

[2] A non-degenerate distribution function is not restricted to take on a single value.

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

which is defined for $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$.

In this specification, it holds that $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. Despite the notation, these parameters are not to be confused with the mean and standard deviation used for the normal distribution (Nilsson, 2012). Instead, they should be interpreted as the location-, scale- and shape parameters of the distribution respectively. We also see that it is the sign of the shape parameter that determines what family the distribution belongs to. In addition, this formulation shows that for a fixed μ, it is continuous in the shape parameter, which makes it easier to work with in statistical modelling (McNeil et. al, 2005). Out of these families, we are normally interested in the Gumbel- and Fréchet-distribution, which both have infinite endpoints. (Ibid.)

2.4.1.Estimating GEV parameters

As a result of the proofs just outlined, we can now estimate the parameters of the GEV. While there are different ways in which we can accomplish this, we choose the Maximum Likelihood Estimator.

Assuming that $Z_1, \dots, Z_m$ are independent variables following the GEV distribution, the log-likelihood functions for estimating the parameters are given by:

(1) For $\xi \neq 0$:

$$l_G(\mu, \sigma, \xi) = -mln(\sigma) - (1 + \frac{1}{\xi})\sum_{i=1}^{m} ln\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi},$$

provided that all maxima satisfies $1 + \xi\left(\frac{z_i-\mu}{\sigma}\right) > 0$ for $i = 1, \dots, m$.

(2) For $\xi = 0$:

$$l_G(\mu, \sigma) = -mln(\sigma) - \sum_{i=1}^{m}\left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\}.$$

The advantage of using MLE to estimate the parameters is that provided $\xi > -1/2$, the estimates are consistent and asymptotically normal. Meanwhile, the parameters do not have closed form solutions, meaning that an appropriate numerical solution method is required. In addition, small samples and software issues could also give us biased estimates. (Dowd, 2005, p. 195-196).[3]

[3] For an elaborate discussion of other estimators and their properties, please refer to Dowd (2005), Embrechts et. al (1997) and McNeil et. al (2005).

2.4.2. Value at Risk estimation for GEV

After sorting the data into *m* blocks of length *n*, which gives us a sequence of maxima $M_{n,1}, \dots, M_{m,n}$, we maximize the log-likelihood function as specified above. We can then proceed by extracting the estimated parameters $\mu, \sigma \; and \; \xi$, which in turn allows us to calculate VaR estimates for the Fréchet- and Gumbel distribution respectively. At a given confidence level *p*, these are given by:

$$VaR_\alpha = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}[1 - \{-nln(1-\alpha)\}^{-\hat{\xi}}], for\ \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma}\{-nln(1-\alpha)\} \qquad\quad , for\ \hat{\xi} = 0 \end{cases}$$

2.5. Peaks over Threshold

While the Block Maxima framework has several advantages from a practical perspective, arranging the data into blocks could potentially mean that we waste information in our analysis. This is particularly true for financial data as these are prone to clustering, and the likelihood of finding two or more extreme observations in each block increases. Focusing only on the maxima in each block could therefore lead us to draw wrongful conclusions. When dealing with extremes, we are by definition left with a smaller set of data on which to perform our analysis, which shows the importance of making use of all information available. The parameters are also sensitive to the choice of block length, and a bias-variance trade-off emerges. An alternative method is the *Peaks Over Threshold* (POT), in which we direct our focus to observations that are exceeding a predefined threshold.

If we let *u* denote the threshold and *y* observations exceeding *u* for an arbitrary term X_i in *X*, the following conditional probability can be used to describe the stochastic behaviour of extreme events:

$$\Pr\{X > u + y | \, X > u\} = \frac{1 - F(u+y)}{1 - F(u)}, \;\; y > 0$$

As was the case for the generalized extreme value distribution, the fact that *F* is unknown means that we have to find an approximation to this distribution. Luckily, Gnedenko-Pickands-Balkema-deHaan was able to show that as *u* gets large, the distribution function *F* converges to a *Generalized Pareto Distribution (GPD)* (Dowd, 2005, Chapter 7).

Definition: Generalized Pareto Distribution

Suppose that we have a sequence of independent and identically distributed random variables $X_1, X2, \dots$ with distribution function *F*. We then define the maxima as:

$$M_n = \max\{X_1, \dots, X_n\}$$

If we assume that the distribution function *F* satisfies the GEV and denoting an arbitrary term in M_n, X, we have that for large *n*

$$\Pr\{M_n \leq z\} \approx G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

for $\mu, \sigma > 0$ and ξ. If u is sufficiently high, it follows that the distribution of $(X - u)$, given that $X > u$, is approximately

$$H(y) = 1 - (1 + \xi y/\hat{\sigma})^{-1/\xi}$$

defined on $\{y: y > 0 \; and \; (1 + \xi y/\tilde{\sigma}) \; > 0\}$, where $\tilde{\sigma} = \sigma + \xi(u - \mu)$. This type of distribution is said to belong to the *generalized Pareto family.* It tells us that if block maxima have the distribution function *G*, then there exists an approximate distribution for the observations exceeding the threshold in the generalized Pareto family. (Coles, 2001, p. 75)

2.5.1 Threshold selection

The problems with the block maxima method can to some extent be remedied by using POT. However, choosing a suitable threshold over which to model the exceedances is a cause of concern as it also induces a bias-variance trade-off. In order for the asymptotic theory of the GPD-distribution to be valid, we must set a high enough threshold. At the same time, setting a high threshold obviously reduces the number of observations on which to estimate the parameters.

Consulting the literature for guidance does not give a clear answer as to how to find the best threshold, even though a few methods are suggested. The main insight from these methods lies in finding stability for the tail- and shape parameter. McNeil and Frey (2000) used a Monte-Carlo simulation in order find stability of the parameter, whereas others employ a graphical inspection of the exceedances (see for instance Coles, 2001, Chapter 4 and Gençay, Selçuk and Ulugülyagci, 2003). In the coming tests, we decided to employ the latter approach, and will not discuss this issue further here. We will however return to threshold selection in Section 5, were we present the methodology of this study.

2.5.2. Estimating parameters GPD

After choosing a threshold, we can proceed to estimate the parameters $\hat{\sigma}$ and ξ, which are needed to estimate VaR. This is once again carried out with the help of maximum likelihood estimation. The log likelihood functions are then given by:

1. For $\xi \neq 0$:

$$l(\sigma, \xi) = -kln(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \ln(1 + \xi y_i/\sigma),$$

whenever $(1 + \sigma^{-1}\xi y_i) > 0$ for $i = 1, \dots, k$, and $l(\sigma, \xi) = -\infty$ otherwise.

2. For $\xi = 0$:

$$l(\sigma) = -kln(\sigma) - \sigma^{-1} \sum_{i=1}^{k} y_i$$

2.5.3. Value at Risk calculation for GPD

After obtaining the estimates of the GPD parameters $\hat{\xi}$ and $\hat{\sigma}$ using the Maximum Likelihood Estimation, we can proceed to calculate Value at Risk for the POT-method by implementing the following formula:

For $\hat{\xi} > 0$:

$$VaR_\alpha = u + \frac{\hat{\sigma}}{\hat{\xi}}\left\{\left[\frac{n}{N_u}(1-\alpha)\right]^{-\hat{\xi}} - 1\right\},$$

For $\hat{\xi} = 0$:

$$VaR_\alpha = u - \hat{\sigma}\ln\left(\frac{n}{N_u}(1-\alpha)\right)$$

where u denotes the threshold, $\hat{\sigma}$ and $\hat{\xi}$ are the estimated parameters, n is the sample length, N_u is the number of exceedances and α is the quantile. Since we are dealing with financial data, we can expect to mainly be dealing with the former case, as this kind of data often displays longer tails than the normal distribution.

3. Previous Research

Extreme Value analysis has grown to become a popular tool in areas where risks are involved. Hydrologists wanting to determine future sea levels, and professional actuaries trying to estimate extreme insurance claims are just some of the fields in which EVT has been used. Finance is no exception, and in the light of the Basel II framework the need of properly accumulating capital for VaR has led many firms and portfolio managers to employ extreme value analysis in order to optimize their capital reserve. Due to the inherent nature of financial assets, such as volatility clustering and non-normality, estimating VaR based on the normal distribution have proved to yield inaccurate results (see Dowd, 2005 and Daníelsson, 2011). By instead acknowledging these stylized facts, many studies have instead seen the potential of EVT as it accounts for the large losses occurring in the tail of the assets distribution.

3.1. Applications in stock markets

Most research that aimed at estimating VaR with the help of EVT has mainly focused on for various stock indices such as S&P500 (Xiu-min & Fa-chao (2006)), the ISE-100 (Gençay, Selçuk and Ulugülyagci (2002)) and the DJIA (Byström (2004)) or exchange rates (Bali, 2003). Generally, while the studies have used different time horizons and data, the results all point in the same direction. When applying extreme value analysis, it is found that returns obey an Fréchet-distribution, which is to be expected considering the nature of asset returns. EVT is also found to yield better VaR forecasts than for example Historical Simulation (HS) and other methods where one is assuming normality of returns.

The fact that returns are not normally distributed is problematic considering that EVT requires that the observations are iid. Aware of this, McNeil and Frey (2000) developed a strategy in which they first filtered the returns using an AR-GARCH model.[4] This captured the conditional heteroskedasticity of the data and produced residuals that were approximately iid. In the second stage, they used the POT-method on the standardized residuals, which proved to increase the accuracy of their estimates compared to when applying EVT directly to the return series. This methodology has since then been employed by many, there among Byström (2004), Bali and Neftci (2003) and Fernandez (2005).

3.2. Applications in electricity markets

As such, there have been numerous studies conducted where EVT has been applied to financial data, and of all these the majority have been aimed at stock markets. The evidence found in these studies has then led a few to investigate whether similar improvements in risk management could be made in a more volatile financial environments such as the electricity market.

One of the first studies to test the appropriateness of EVT on electricity markets was made by Andrews and Thomas (2002). They used a combination of historical simulation and a POT-model that was able to produced better tail estimates than the normal distribution. Around the same time, de Rozario (2002)

[4] A similar method was employed by and Diebold et. al (1999).

studied VaR for the Victorian electricity market using half-hourly electricity returns. He also found that a POT-model outperformed the normally distributed model, which performed poorly even at the 95% level. However, the threshold based model performed well at the 95%- and 99% level, while its accuracy diminished drastically further out in the tail. This, he argued, was due to clustering among the extreme observations in the data, something that his model could not account for.

Studying hourly returns on the Nord Pool spot market, Byström (2005) found that the generalized Pareto distribution produced a good fit, displaying an Fréchet-distribution. In his study, he used an approach similar to that of McNeil and Frey (2000) where the returns were filtered with an AR-GARCH-model governed by a t-distribution to account for the long tails. The AR-terms in his model captures the daily- and weekly dependence found in his data while the GARCH-terms captured volatility clustering. After applying the POT-method on the standardized residuals, he found that this method yielded better estimates as well as forecasts of extreme quantiles than the Block Maxima and regular AR-GARCH model.

Chan and Grey (2006) follow a similar methodology to the one used by Byström (2005) when they investigated five international electricity markets with different characteristics. Their methods differed in that they combined the autoregressive model with Nelson's EGARCH-model when filtering the data instead of the regular GARCH-model. EGARCH-models have the advantage that they can capture potential leverage effects in the conditional volatility (Verbeek, 2004). They motivate this GARCH-specification on a finding made by Knittel and Roberts (2001), who argued that positive demand shocks have a larger impact on electricity prices than negative demand shocks due to the marginal costs of electricity generation being convex.

Similar to the other studies, they found that their extreme value model gave better out of sample estimates than regular methods of estimating VaR, there among parametric and non-parametric approaches, particularly in the more volatile markets exhibiting higher levels of skewness and excess kurtosis. Noticeably, a simple Historical Simulation model outperformed more sophisticated models in several markets. However, they quickly pointed out that this most likely was due to these markets being much less skewed than others, and that the HS-model failed tests of conditional coverage. Meanwhile, their filtered POT-model performed best in the more volatile markets as well as providing satisfactory conditional coverage in all markets.

4. Data Analysis

4.1. Data description

The data used throughout this study consists of hourly spot prices of the European Energy Exchange (EEX) from 2005-01-01 to 2011-05-01. EEX is a spot market of the day-ahead type, meaning that each day at 12am, electricity prices for every hour the *coming day* is determined at an auction. Hence, we should point out that we are not dealing with regular spot prices per se. For this reason, our study should merely be viewed as a tool of finding the distribution of price changes. Even so, this study could still be of interest given the trade of derivatives that are priced based on the spot prices, as pointed out by Byström (2005). Similarly, we will treat the EEX spot prices *as if* they were regular spot prices. Figure 1 below displays the evolution of spot prices during the sample length:

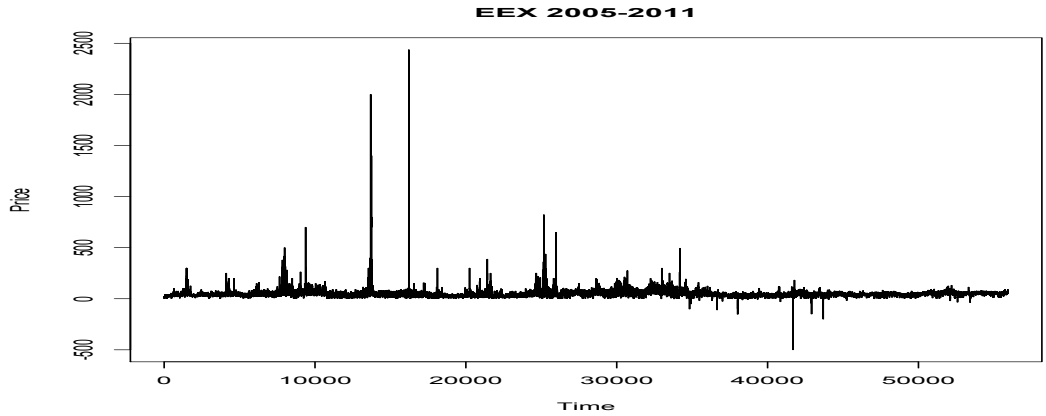


Figure 1. Electricity Prices (EUR/MWh), January 1, 2005 to May 20th, 2011.

As can be seen above, the data has several spikes and is very volatile, which are common findings when dealing with electricity data (see for instance Huisman, 2013 and Schneider, 2010). After a closer inspection of the data we also notice that there are several instances where prices are equal to zero or even negative Below, we will briefly discuss the reasons for this and the implications it brings for our study.

4.1.1. Negative Prices

Even though negative pries have been around for quite some time in international electricity markets, they are a fairly new phenomenon in the European region. EEX did not introduce this feature until the autumn of 2008 (Schneider, 2010), and since then, other large markets such as NordPool have followed.[5]

The reason for the negative prices is a consequence of a very low demand combined with a high supply, where the cost of providing the electricity at a discount is more profitable than shutting the plant down. This would not pose a problem if it were not for the fact that electricity could not be stored. In the German market, it is often large nuclear plants that provide the electricity and shutting the plants down at times of low demand would simply incur start-up costs far exceeding the cost of paying to get rid of the

[5] http://www.nordpoolspot.com/About-us/History/

electricity. This phenomenon mostly appears during so-called off-peak hours such as night time, weekends and public holiday when demand is much lower than during peak hours (8am-8pm). Dettmer and Jacob (2009) and Schneider (2010) also propose that slumps in industrial activity caused by the recent economic crisis have contributed to negative prices.

4.2. Data manipulations

As we are investigating the returns instead of prices, negative prices causes some computational problems: Firstly, the fact that negative prices, 107 observations in our data, are present has the immediate consequence that one cannot use logarithmic returns, as is normally the case when dealing with financial data. Secondly, neither logarithmic- nor simple returns can be used should the prices be equal to zero, which is the case for 62 observations. While it would be possible to simply exclude these observations from our data set, this again causes problems if one wants to account for seasonal behaviour in the data. There are few guidelines to be found in the literature and different authors propose different methods. Among these are setting a lower bound on the prices and thereafter rescaling them (Sewalt and de Jong, 2003), or using an affine transformation (Mayer et. al, 2011).

We employ a rather simple method to account for the 'zero-prices' by setting them equal to 0.01. This allows us to obtain returns for the full period while still having a series whose prices are close to their true value (i.e. =0). Studies employing extreme value analysis normally work with logarithmic returns, as they serve as an approximation for stationatity (McNeil et. al, 2005). However, due to the extreme behaviour of the data, large price changes would cause some returns to go far below -100% which is not very realistic. On the other hand, simple returns give us a lower bound of -100%, but since this appears to be the common choice when investigating electricity markets, this method is employed here as well. This obviously causes the return distribution to be skewed for large returns as mentioned by Byström (2005), which would be problematic if one is interested in losses rather than gains. However, as we are merely interested in the right tail of the distribution, i.e. large gains, this lower bound has no impact on the results.

The simple returns are calculated with the following formula:

$$\tilde{r_t} = \frac{s_t - s_{t-1}}{s_{t-1}}$$

where s_t denotes the electricity spot price at time t.

Another issue with the data set is that changes from winter to summertime leaves us with one hour less in the last Sunday of March each year. Here the missing observations are added, defined as the average price of the two closest observations.

4.3 Properties of returns

4.3.1. Hourly returns

In this section, we investigate the properties of hourly returns. Finding accurate estimates of Value at Risk depends to a large extent on what kind distribution the returns are following, and different methods apply depending on the distribution. As such, we should use statistical tools to test if the data follows a normal distribution and whether or not it is stationary. But before doing so, we perform a graphical analysis of the data and the following figure depicts the hourly price changes over our sample:

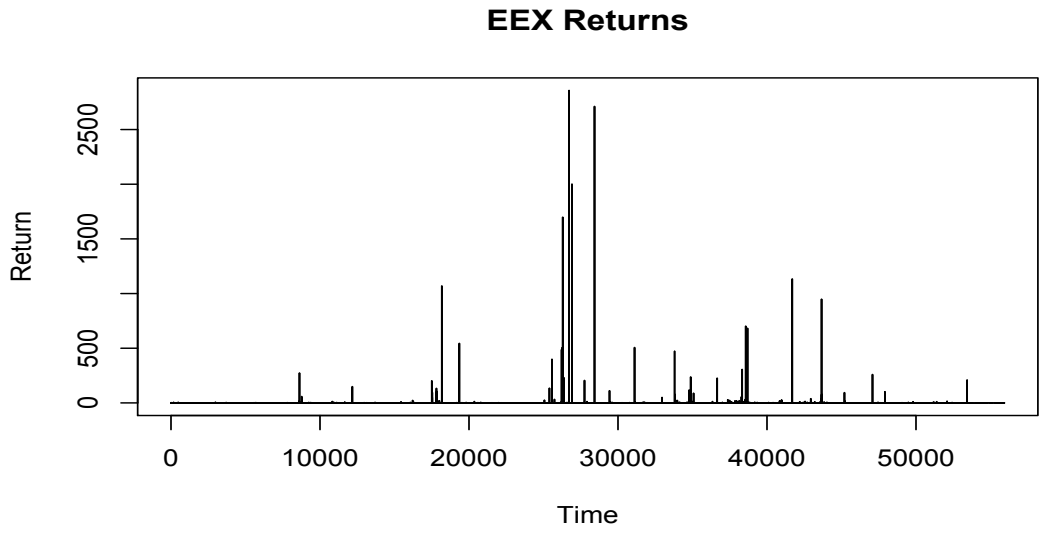


Figure 2. Electricity price changes (%), January 1, 2005 to May 20th, 2011

A quick glance at the return series suggests that the data displays a higher degree of volatility and that it is more extremes than was found in past studies. We can see several spikes in the series, where returns from one hour to the next goes far beyond several hundred per cent. To further explore the properties of our data we present some descriptive statistics presented in Table 1:

Table 1. Descriptive statistics, hourly price changes.

Mean	Std deviation	Skewness	Kurtosis	Q(6)	Q(24)	$Q^2(6)$	$Q^2(24)$	JB	ADF
0.41	22.71	48.6	6747.4	0.272	6.324	0.003	0.017	0.000	<0.001

This table presents the mean, standard deviation, skewness, excess kurtosis, Q-statistics at lag 6 and 24 for the return- and squared return series respectively. The Jarque-Bera- and Augmented-Dickey-Fuller tests for normality and a unit root, and we can reject the null-hypothesis for both, as indicated by their p-value.

Table 1 highlights some of the stylized facts of electricity markets. The mean return is equal to 41% while the standard deviation reaches a staggering 2271%. The non-normality of the return series is even more obvious when looking at the skewness and excess kurtosis, both suggesting a very highly peaked distribution together with a very long right tail, the latter further increased by the lower bound of -100%. Evidence of non-normality can be found by looking at Jarque-Bera statistics, where the p-value tells us that we can strongly reject the normality assumption.

The Ljung-Box Q-statistics of the returns and squared returns are included as they serve as indicators of temporal dependence and heteroskedasticity, respectively (Daníelsson, 2011). However, it appears as if any potential weekly-, and/or daily dependencies in the prices seem to have been captured by our return transformation. These findings stand in vast contrast to other studies, where these statistics were highly significant (see Byström, 2005, Chan and Gray, 2006 and de Rozario, 2002). One possible explanation for these findings might be that the price spikes tend to distort the overall picture. This view is supported by that the median (not shown above) is equal to -13%, suggesting that the extreme observations in the data drive the mean above zero (the maximum reaching 2858% and 6 observations exceeding 1000%!). Another explanation might be that the sample is to large for us to find any form of seasonality, and from Figure 1 it is clear that the price swings have decreased to some extent over the last few years.

Our findings are further illustrated in the following figure, which displays a graph of the autocorrelation function, a QQ-plot plot against the normal distribution and a histogram, which are all graphical methods used to determine the properties of the data. For normality, the QQ-plot is supposed to plot as a diagonal line, which it clearly does not. We can also deduce that no lags are significant in the autocorrelation function where the blue lines are confidence bands for the 95% level [6]. Finally, from the histogram it is clear that the data has a very long right tail, which rejects normality.

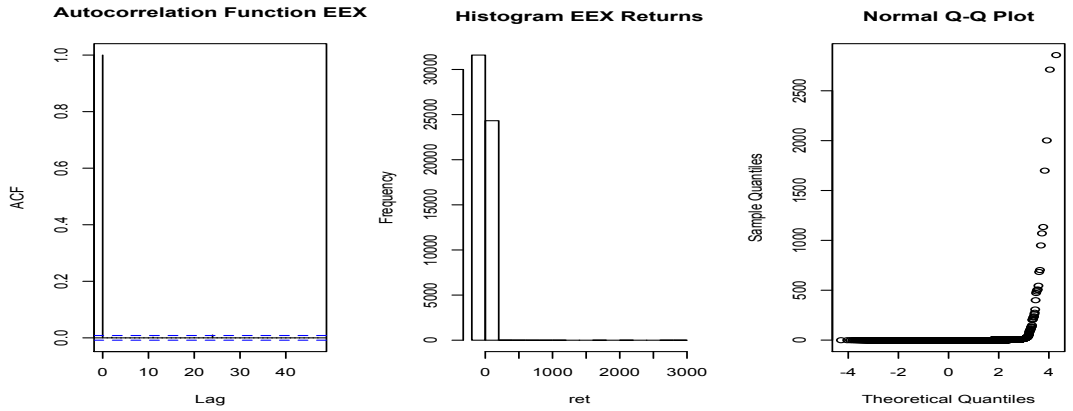


Figure 3. Autocorrelation function, Histogram and QQ-plot for hourly returns.

It is troubling that we do not find a clear dependence structure in the series. For in order to apply extreme value analysis, we are required to have iid observations, and considering that the data obviously is non-normal, this begs the question of how to remove this non-normality. The AR-GARCH procedure suggested by McNeil and Frey is only helpful if we can find any form of structural dependence, which our data clearly is lacking[7].

[6] For a longer explanation of the autocorrelation function please refer to Verbeek (2004) Chapter 8.
[7] During the process of writing this paper, attempts were made to filter the series regardless of the lack of autocorrelation. However, these attempts were futile as they instead further increased the kurtosis of the data while the skewness remained unchanged.

Research attempting to forecast electricity prices have tried different ways of accounting for this peculiar behaviour. These all employ fairly advanced methods such as regime-switching- and jump-diffusion models allowing for mean reversion (see Deng, 2005 and Mayer, 2011). Since this is beyond the scope of this thesis we choose a different route and decided to employ the extreme value methods on the raw return series.

4.3.2. Daily returns

For the reasons mentioned above, we decided to investigate whether or not daily data would exhibit other properties, as we suspect that the price spikes would not have as large of an impact. The following table reports the descriptive statistics for daily data, obtained by using daily averages of the original series:

Table 2. Descriptive statistics, daily returns,

Median	Mean	Std deviation	Skewness	Kurtosis	Q(7)	$Q^2(7)$	JB	ADF
-0.0256	0.042	0.352	2.8	18.3	862.4	60.1	0.000	<0.001

This table presents the mean, standard deviation, skewness, excess kurtosis, Q-statistics at lag 7 for the return- and squared return series respectively. The Jarque-Bera- and Augmented-Dickey-Fuller tests for normality and a unit root, and we can reject the null-hypothesis for both, as indicated by their p-value.

Just as suspected, the daily data displays results more in line with earlier studies. To illustrate this, we present the same types of graphs as before applied to the daily data for the raw return series. From the ACF we see that the returns appear to follow a mean-reverting process, as lags 1 until 6 are negative, whereas the 7th lag is clearly positive, all being significant. The histogram indicates that the data is much less volatile, but that it still has a very long right tail. Finally, the non-normality is evident from looking at the QQ-plot, where a normal distribution ought to be plotted as a straight line, not the kinked form that we see here. This would then mean that an AR-GARCH filtering proposed by McNeil and Frey (2000) could help us to remove the seasonal dependence and obtain observations at least closer to being iid.

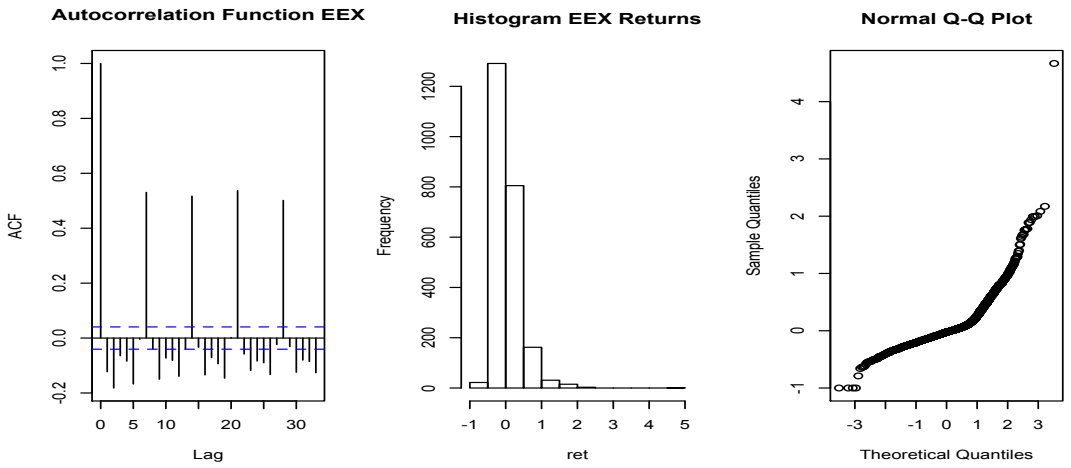


Figure 4. Autocorrelation function, Histogram and QQ-plot for daily returns.

5.Methodology

The following section is devoted to explaining the steps undertaken in order to calculate Value at Risk for the extreme value models. A statistical test to determine the adequacy of the models is also described.

5.1. Hourly data

We begin by employing the EVT-methods on hourly data. Since we could not find any clear-cut dependence of the time series, we decided to apply the extreme value methodologies without pre-filtering the returns with an AR-GARCH-model.

5.1.1. Block Maxima

Before we apply the block maxima method we must first choose the block size, n. In Section 2.3, we highlighted the importance of picking the right block size because of the trade-off that emerges. We must choose a sufficiently high n in order for the distribution of maxima to converge to a GEV distribution. At the same time, we must obtain enough observations to efficiently estimate the parameters. We chose to specify the blocks on a weekly basis, i.e. n=168 which yields a total of 334 maxima, (m=334) for the hourly returns series, which we deem to be sufficiently large to avoid clustering. Our basis for choosing this block size is that we want to make sure that the series of maxima are independent of each other, for reasons mentioned above. Furthermore, the only other study applying the BM-method on electricity returns (e.g. Byström, 2005) used the same block length in his study.

The estimated parameters and their standard errors are presented below:

Table 3. GEV parameters hourly returns

	N	$\hat{\mu}$	$\hat{\xi}$	$\hat{\sigma}$
Estimate	168	1.074 (0.072)	1.409 (0.073)	1.239 (0.127)

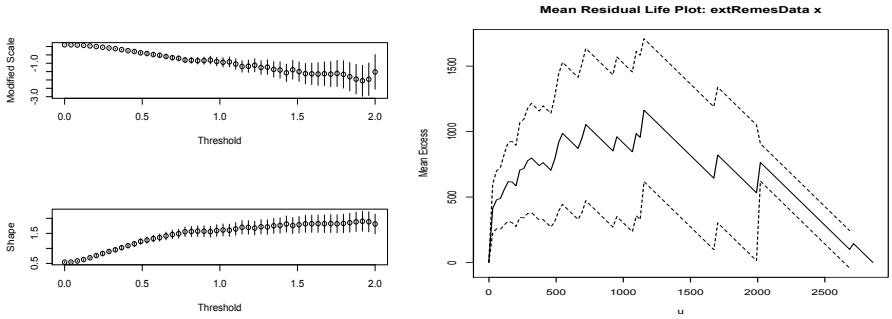
This table presents the estimated GEV-parameters obtained from maximizing the Log-likelihood function, using hourly data. Standard errors are shown in parenthesis.

The estimate of the $\hat{\xi}$ give further evidence of the data being heavily right tailed, and also that that the distribution function have an infinite mean and variance (Coles (2001, p. 79). As was shown by Neslehova et. al (2006), the estimated VaR grows exponentially with $\hat{\xi}$, thus making the measurements very unstable. This, they find, is especially true for $\hat{\xi} > 1$. One explanation for retrieving parameters this high is data contamination, meaning that some observations in the data sample follow a different distribution than others. This appears to be a plausible explanation after once again inspecting Figure 2. The actual impact on our VaR-estimates will be further discussed in Section 6., where we present our findings.

5.1.2. Peaks Over Threshold

Before estimating the parameters, we must choose a reasonable threshold over which to model the exceedances. As already mentioned, choosing the correct threshold is the weak point of the POT-method. For reasons similar to the BM-approach, it implies a choice between bias and variance (Coles, 2001, p. 78). For the asymptotic theory to be valid we must choose a threshold that is high enough, but this naturally leaves us with fewer observations from which we to estimate our parameters. It does not help that the

suggestions given in the literature cannot reach general agreement over which method to use. A few methods are available, and we will use graphical tools in which we seek linearity and/or stability in the parameters.[8] Below, we shall depict two such methods. Should the data belong to one of the GPD families, the *mean-excess plot* tells us to search for an increased linearity above a threshold u[9]. If this is so, we should then choose a threshold about where the linearity starts. In our study, we expect to find a positive slope since this would suggest an Fréchet-distribution (McNeil et. al 2005, p.280). The purpose of the second graphical tool is to fit the model using a range of thresholds. We should then look for a threshold where the scale-, and shape parameter are stable (Coles, 2002, p.83).



As can be seen from the mean excess plot on the right hand side, it is hard to deduce any useful information. Perhaps one could argue that after an initial steep slope, there is some linearity, and the line is thereafter kinked. However, there are several points in the graph that could be deemed linear. Also worrying is that the two graphs displaying the scale- and shape parameter do not appear stable at all until the threshold reaches approximately 1.4. Here they are still not completely stable which makes it difficult to choose. While one could choose a threshold based on a 10%-rule proposed by DuMouchel (1983), this is not a theoretically sound procedure (Scarrott and Macdonald, 2012). Another method was proposed by Ferreira et al. (2003) who they let the square root of the sample determine the threshold. But, as Scarrott and Macdonald (2012) point out, the validity of the latter can be put in question. Therefore, we decided to try two different methods. First, choose to pre-set the number of exceedances to equal 3000, which amounts to roughly 5% of the sample (a similar approach was used by Byström, 2005). We also estimate the model using a threshold that equals 1.4 for which we obtain 427 observations exceeding the threshold. The resulting estimates are demonstrated below:

Table 4. GPD parameters hourly returns (full sample)

	N=3000	N=417
$\hat{\xi}$	1.035 (0.035)	1.763 (0.132)
$\hat{\sigma}$	0.199 (0.007)	1.063 (0.118)
u	0.379	1.400

This table presents the estimated GPD-parameters obtained from maximizing the Log-likelihood function, using hourly data. Standard errors are shown in parenthesis.

[8] A thorough review of the various methods available for threshold selection can be found in Resnick (2007, Chapter 4) and Embrechts et. al, (1997, Chapter 6).
[9] For an elaborate discussion of the mean excess plots, please refer to Ghosh and Resnick (2010).

As was the case for the Block Maxima method, we obtain very large estimates of $\hat{\xi}$, and we are once more dealing with an infinite-mean process. The sketchy look of the mean-residual life plot is most likely a consequence of this, as it is derived from the conditional expected value of the GPD. It can be shown that this is determined by the following formula (Coles, 2001, p. 79):

$$E(X - u|X > u) = \frac{\sigma_u}{1 - \xi}$$

Only by drastically reducing the threshold were we able to get parameters below 1, which makes no intuitive sense if we wish to model the extremes exclusively. With a scale parameter far below the tail parameter, a glance at the VaR formula would indicate that our model would overestimate the risk, at least for lower confidence levels and vice versa.

5.2. Daily data

5.2.1. AR-GARCH-EVT

Both the Block-Maxima and POT approach described above have proven to be well suited for forecasting VaR and ES over longer horizons. One problem, however, is that these two both treat the random variable unconditionally. As pointed out by among others Dowd (2005) and McNeil and Frey (2000), the choice of data might instead require us to treat it conditionally. Due to the inherent nature of electricity returns mentioned in the data description, treating the data conditionally allows us to both look at shorter horizons as well as model the dynamics of the data set at hand. Historically, studies of financial markets have shown that the volatility of returns is not constant over time and that they tend to cluster (i.e. periods with high (low) volatility tends to be followed by more periods of high (low) volatility (Verbeek, 2004, p. 311). Hence, failing to properly acknowledging this might lead to wrong decisions when market conditions changes. One way of solving this problem was proposed by McNeil and Frey (2000). Combining findings from econometrics and extreme value analysis, they were able to improve VaR forecasts by taking current market conditions into account, thereby making their model more dynamic (Nilsson, 2012).

Before we begin to estimate the shape and tail parameters, we want to remove the autocorrelation and heteroskedasticity found in the data. This is important since the generalized extreme value theorem is based upon the iid assumption. McNeil and Frey (2000) pointed out that stochastic volatility and fat-tailed P/L-distributions is often found in real data, together complicating the procedure of obtaining accurate measurements of VaR (Daníelsson and de Vries, 1997b). Large tails can obviously be modelled using the classical EVT models, but their largest problem is accounting for stochastic volatility, which has the immediate effect that the returns are not independent.

In contrast, regular GARCH-models help in removing the stochastic volatility, but they assume normality (McNeil and Frey, 2000). As mentioned repeatedly, neither this is the case for real data where the residuals from GARCH-estimations have fat tails.

For these reasons, we follow the proposed two-stage method given by McNeil and Frey (2000) in search of observations that are closer to being iid. Following previous studies, we filter return series assuming both the normal- and Student's t-distribution, where the latter is supposed to acknowledge that the data exhibits longer tails than suggested by the former.

However, before we proceed we must specify a model to capture these characteristics. Our choice of model is a mixture of models used in previous studies, where we choose to specify an AR(7)-GARCH(1,1)-model. Here, the autoregressive terms are supposed to capture the weekly dependence in our data while the GARCH-terms should help in removing the heteroskedasticity, as was proposed by Bollerslev (1986). The motivation for including seven lags in the mean equation is due to the dependence found above, where we observed negative autocorrelation for lags 1-6 followed by a positive spike at lag 7. Modelling the heteroskedasticity with an GARCH(1,1) process is mainly due to it being parsimonious and it is usually found that it outperforms more advanced models despite its simplicity . The full model is specified as follows:

$$r_t = \phi_o + \sum_{j=1}^{7} \phi_j r_{t-j} + \varepsilon_t$$

$$\sigma_t^2 = \beta_0 + \beta_1 \varepsilon_{t-1}^2 + \beta_2 \sigma_{t-1}^2$$

where σ_t^2 denotes the conditional variance of ε_t, $\varepsilon_t = \sigma_t \eta_t$and where η_t follows either the normal- or the t-distribution. After estimating the model, we run the extreme value analysis in the same way as before, but now we do it on the standardized residuals instead of the original return series. We fit the model to the residuals by maximizing the likelihood function, and obtain estimates of the parameters that will later be used to estimate the quantile. A conditional VaR-estimate is then calculated by scaling the quantile with the estimated parameters from the AR-GARCH specification. If we let q_α denote the quantile obtained from the unconditional VaR calculation, the conditional estimate can be calculated as follows:

$$VaR_\alpha = \phi_o + \sum_{j=1}^{7} \phi_j r_{t-j} + \sigma_t q_\alpha$$

The following table presents the coefficients of the AR-GARCH model along with some descriptive statistics of the standardized residuals. For comparability, we include the relevant statistics for the raw data once again:

Table 5. AR-GARCH parameters (full sample, daily returns)

	Normal		Student's t	
ϕ_0	0.033076 ***	(0.001807)	0.018482 ***	(0.002393)
ϕ_1	-0.352197 ***	(0.020249)	-0.195659 ***	(0.019739)
ϕ_2	-0.350205 ***	(0.021516)	-0.204532 ***	(0.019767)
ϕ_3	-0.294402 ***	(0.021322)	-0.189159 ***	(0.018478)
ϕ_4	-0.275097 ***	(0.021982)	-0.184752 ***	(0.017491)
ϕ_5	-0.229415 ***	(0.021969)	-0.199992 ***	(0.016666)
ϕ_6	-0.171463 ***	(0.019217)	-0.159585 ***	(0.016094)
ϕ_7	0.429850 ***	(0.018889)	0.452452 ***	(0.019399)
β_0	0.004572 ***	(0.000753)	0.012565 ***	(0.003981)
β_1	0.333086 ***	(0.029040)	0.267689 ***	(0.073574)
β_2	0.665914 ***	(0.024170)	0.731311 ***	(0.047147)
v			2.493793 ***	

Standardized Residuals	Median	Mean	Std.dev	Skewness	Kurtosis	Q(7)	$Q^2(7)$
Normal	-0.448	-0.232	1.009	0.721	5.37	58.06	54.80
Student's t	0.019	0.098	0.802	1.805	12.17	99.4	27.46
Raw returns	-0.0256	0.042	0.352	2.8	18.3	862.4	60.10

This table presents the estimates obtained from applying the AR-GARCH filter using both normal- and t-distributed innovations as well as descriptive statistics for the standardized residuals.

The estimated AR-GARCH parameters are all significant at the 99% level, suggesting that our model specification is well suited for the data. Comparing the residuals to the original data strengthens this view, as we can see that both models help in reducing the autocorrelation and heteroskedasticity, though not fully. Using the t-distribution removes the latter to a larger extent, whereas the normal distribution reduces more of the autocorrelation. Also worth mentioning is that the estimated GARCH-terms are very close to 1, and we can thus not rule out an infinite unconditional variance. In other words, it suggests a high persistence in the volatility (Verbeek, 2004, p. 313).

5.2.2. Block Maxima

Using daily data, the number of observations from which we estimate the parameters puts even more weight on choosing block length. We must therefore be cautious when interpreting our results due to the small sample biases our reduced data set might incur. (see for instance Dowd, 2005), Embechts et. al, 1997 and Coles, 2001). While many studies employ quarterly maxima (see for instance Byström, 2004), our short sample length would give us too little data to estimate the parameters. Instead we sort the data into blocks of 14, giving us a total of 165 maxima. The estimated parameters for the two models are presented below:

Table 6. GEV parameters daily returns

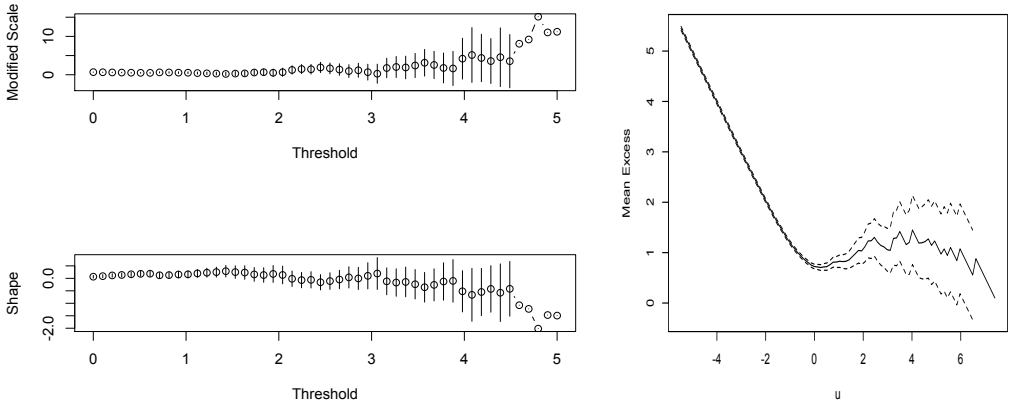
	N	$\hat{\mu}$	$\hat{\xi}$	$\hat{\sigma}$
Normal distr.	165	1.316 (0.065)	0.128 (0.073)	0.756 (0.049)
t-distr.	165	1.265 (0.060)	0.179 (0.064)	0.679 (0.047)

This table presents the estimated GEV-parameters obtained from maximizing the Log-likelihood function using daily data. Standard errors are shown in parenthesis.

Here the results appear more plausible than they did for daily data. The distribution is no longer displaying an infinite mean and the estimated parameters are positive, indicating that they too are following an Fréchet-distribution.

5.2.3. Peaks Over Threshold

Repeating the procedure from 5.1.2, we begin by using the graphical tools in order to determine a suitable threshold. In order to save space, we only depict the plots obtained for the residuals following the normal distribution. These are shown below:



For daily data, the parameters behave much more stable up until just below 2. For the mean residual life plot, we can distinguish a similar pattern, where it becomes increasingly linear around 1.4-1.7. Based on these graphical methods, we decide to put the threshold at 1.5 for the residuals specified by the normal distribution and to 1.4 for the t-distribution. This results in 131 and 127 exceedances for the normal-, and t-distributed residuals respectively. The resulting parameter estimates are shown below:

Table 7. GPD parameters daily returns

	Normal distr.	Student's t-distr.
$\hat{\xi}$	0.267 (0.126)	0.158 (0.106)
$\hat{\sigma}$	0.664 (0.101)	0.748 (0.103)
u	1.500	1.400

This table presents the estimated GPD-parameters obtained from maximizing the Log-likelihood function, using daily data. Standard errors are shown in parenthesis.

Now, we retrieve estimates that are much more in line with those obtained in other studies. The estimated shape parameters are both positive where the one obtained from t-distributed errors are a bit lower. We can thereby conclude that the filtered daily data follows an Fréchet-distribution. The influence of spikes in the

data appears to have been reduced considerably and the filtering also reduced the non-normality. We would thereby expect to obtain more accurate VaR-estimates when using daily data for both models.

5.3. Backtesting

Thus far, our tests have all been performed based on in-sample data, meaning that we estimate the parameters and calculate VaR based on the full data set. From a risk management perspective, this is of limited use as one strives to continuously acknowledge recent market information. For instance, the use of electricity varies much depending on season and what time of day it is and the season, something that manifests itself in the price. To truly appreciate the extent to which EVT is accurate in modelling the tails of the P/L-distribution, we want to perform out-of-sample analysis. One popular method is to capture the recent market movements by applying the conditional extreme value analysis as described above. The only difference compared to the earlier conditional methodology is that one use one-day forecast of the mean equation and the volatility, i.e. $\hat{\mu}_{t+1}$ and $\hat{\sigma}_{t+1}$.

While there are several methods that can be included in order to compare the accuracy, we restrict ourselves to a conditional- and pre-filtered POT-model using t-innovations and an unfiltered unconditional model for daily data exclusively. The reason for choosing these two is that we wish to test whether the poor performance from before was due to the parameter estimates being out of date and whether the conditional framework actually can improve the results as suggested in the literature. As was argued by Byström (2005), a t-distributed model is more likely to capture the long tails in the distribution, which is why it is also used here. Even though we included the Block Maxima approach in the previous section, our sample is reduced even further. Problems obtaining proper parameter estimates are likely to emerge and, as pointed out by McNeil and Frey (2000), GDP-models are superior when left with smaller samples. What we do is essentially to continuously estimate our parameters in order to obtain VaR estimates of the coming day, given the information available up until that point in time.

We must therefore specify an estimation window in which we estimate the parameters, and thereafter calculate VaR for each day. Following the procedure in McNeil and Frey (2000), we set this estimation window to equal 1000 observations, after which we filter the returns with the use of our AR-GARCH model for the conditional model. After standardizing the residuals obtained from the filtering, we estimate the EVT parameters and calculate VaR for the first day out of sample. The other model is left untouched and is applied directly on the returns, using the same estimation window. By continuously moving the estimation window one day forward, we repeat the steps just described and extract estimates for VaR for the remaining period. For simplicity we set the number of exceedances to equal 100, something that was also employed by McNeil and Frey (2000) in their study. Performing these steps results in a total of 1330 out-of -sample VaR estimates that are compared to the expected number of violations.

5.4. Kupiec test

In order to statistically determine the adequacy of the above models, we chose to employ Kupiec's test. This is a binomial test with which we can test whether the number of VaR-violations generated by our models is consistent with the observed frequency (Dowd, 2005, p. 324). Should we find that the deviation of the actual- and the predicted number of violations are too large, we can reject the model used to estimate VaR. If our models are consistent with the data, the losses follow a binomial distribution (Ibid, p. 325). With n observations, x number of violations and a predicted tail frequency p, the probability of obtaining x violations is:

$$\Pr(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

The cumulative probabilities can be used to calculate the probability of observing less than the actual number of observations as (Nilsson, 2012):

$$\Pr(X \leq x) = \sum_{i=0}^{x} \binom{n}{k} p^i (1-p)^{N-i}$$

From this formula it is possible to construct a confidence interval for the number of violations, giving us a two-sided test. However, instead of constructing a confidence interval for the number of violations from the model, we can do the same for the expected number of violations, where we accept the model if the violations are within this range. This is the procedure carried out in this study, and it is easily carried out through finding the inverse of the cumulative binomial distribution for the upper and lower limit respectively. In the forthcoming section, this confidence interval will be denoted *Range*.

We should point out that while the Kupiec test is a simple way of testing the models under study, it does suffer from a few drawbacks. One problem is that it lacks power unless the sample size is large. With a smaller sample size, it could potentially throw away valuable information. This is because it only focuses on the number of violations, and not on the pattern of the violations and the forecasts of VaR. (Dowd, 2005, p.327). As such, this could obviously pose a problem for the two latter tests, when the sample size is considerably reduced. [10]

[10] For a more elaborate discussion of the Kupiec test and its alternatives, please refer to Dowd (2005, Chapter 15)

6. Results

In this section, we use the estimated parameters from the Block Maxima- and the Peaks Over Threshold methods to calculate Value at Risk by implementing the formulas provided in section 2.4.2, and 2.5.3 respectively.

6.1. Hourly returns

After carrying out the calculations presented in the previous section, we now present our results from the VaR-analysis using the hourly data. The expected number of violations are $n(1-p)$ where n is the total number of observations in the sample, and p is the confidence level. The results are presented in the following table:

Table 8. VaR violations for hourly returns (full sample)

Probability	Expected	Range	GEV	GPD (u=1.03)	GPD(u=1.4)
95.00%	2797	2697-2899	5755	2846	856
99.00%	559	500-621	1351	472	532
99.50%	280	234-328	455	239	275
99.90%	56	33-82	81	77	55
99.95%	28	12-48	47	52	37
99.99%	6	0-17	16	28	4

This table presents the VaR-violations from the GEV-, and GPD model for hourly returns on the full sample. The number of violations is supposed to be as close to the expected value as possible. Range denotes the confidence interval for the expected number of violations.

We can see that the Block Maxima-model clearly underestimates the risk at the 95% confidence level as it yields more than double the allowed number of violations derived from the Kupiec test. The model continues to underestimate risk for higher confidence levels as well, even though we cannot reject the model as the violations are just within the range from 99.90% and upward.

For the Peaks Over Threshold-model using a threshold based on 3000 exceedances, we see a pretty good fit up at 95%, after which it is either rejected or very close to be rejected for all levels. After initially overestimating the risk, our other POT-model display the reverse pattern, only getting close to the expected number of violations at extreme confidence levels. It is hard to draw any conclusions as the violations from this model appears to lie within the range for higher confidence levels whereas they are far too low at 95%. It would thus seem that neither of the models are particularly useful when applied to hourly data when looking at the full range of confidence levels. Obviously, the higher threshold provides a better fit for the extreme quantiles, which might be of more interest to practitioners.

As we pointed out earlier in the study, our choice of block size and threshold level could potentially be part of the explanation for the varying results. We believe that the extreme returns found in the data could mean that we are dealing with some form of data contamination, as was argued by Neslehova et. al (2006). The graphical tools normally used to determine a proper threshold did not help us at all as we struggled to

find stability of the parameters and could not draw any conclusion with regards to the linearity of the mean residual life plot. Setting the threshold to 1.4 thus includes some guesswork, and these results should be interpreted with care.

Compared to earlier studies, our parameter estimates are much higher than what was found in other electricity spot markets. The resulting model fit obviously differs, as we could not filter the data before applying the extreme value methodology[11]. A likely explanation is that our observations are not iid, as required by both models to be accurate and the plots presented in Section 4 further proved this. Attempting to reduce the non-normality in the data proved to be hard, as we could not find any form of dependence or stochastic volatility. It is possible that the very long sample length in addition to several (very) extreme returns might be the reason for this. Apart from using data from different markets, the other studies had fewer observations in their sample than ours with fewer spikes where none were of this magnitude. Nevertheless, we believe that more advanced time series methods where due consideration is given to modelling the spikes might improve the overall results. For these reasons, it is hard to draw any final conclusions other than that the best model depends highly on the chosen confidence. Instead, we will now direct out attention to the daily returns.

6.2. Daily returns

With a reduced data set, we decided to perform our tests for the daily data using confidence levels 95%-99.5%. After the initial filtering procedure described above, we perform the extreme value parameter estimation on the standardized residuals to estimate the quantiles. Then, we proceed by scaling these quantiles with the AR-GARCH parameters to obtain conditional VaR estimates using both normal-, and t-distributed innovations. The results are presented below:

Table 9. VaR violations for daily returns

Probability	Expected	Range	GEV-norm	GEV-t	GPD-norm	GPD-t
95.00%	116	96-138	47	80	290	92
99.00%	23	12-37	1	1	13	1
99.50%	12	3-22	1	1	1	1

This table presents the VaR-violations from the GEV-, and GPD model for daily returns on the full sample. The number of violations should be as close to the expected value as possible. These are rounded to the closest integer. Range denotes the confidence interval for the expected number of violations.

As can be seen above, the conditional GEV-models perform rather poorly throughout the test and yield the same results except at 95% confidence level where t-distributed innovations provide a better fit. They seriously underestimate the risk further out in the tail and both can be rejected for all confidence levels. As such the Block Maxima-method is actually worse when applied on daily data than it was for hourly data, which highlights its inaccuracy in smaller samples.

[11] Byström (2005) and Chan and Gray (2006) both find $\hat{\xi}$ to be in the range of 0.3-0.5 for their *filtered* GPD-models.

By instead looking at the POT-models we obtain a better fit at 95% using t-innovations, but apart from that the results are not satisfactory. It continues to overestimate the risk for the remainder of the test and can be rejected at all confidence levels. More surprising is that POT-model based on the normal distribution is the only model that cannot be rejected for all confidence levels, as it is within the range at 99%. Based on theory and previous studies, we expected that models using t-innovations should provide a better fit further out in the tail but we see the reverse pattern.

Interestingly, it would thus appear as if the filtered- and conditional GPD models on daily data do not improve the results as much as one would expect from earlier studies. It could of course be that the reduced sample did not allow us to extract enough data to give us accurate parameter estimates, but this does not change the fact that the models do a poor job in modelling the tail. It is also possible that this is due to clustering similar to what was found by de Rozario (2002), which he found gave a worse fit further out in the tail.

6.3. Backtesting

Considering the poor accuracy provided by the previous models, we would expect at least some improvement from our backtesting models. The results obtained from performing this analysis is presented below:

Table 10. VaR violations for Backtest

Probability	Expected	Range	GPD uncond	AR-GARCH-GPD-t
95.00%	67	51-81	40	49
99.00%	13	5-23	11	11
99.50%	7	1-15	4	5

This table presents the VaR-violations from the GEV-, and GPD model for daily returns on the full sample. The number of violations should be as close to the expected value as possible. These are rounded to the closest integer. Range denotes the confidence interval for the expected number of violations.

It is clear that continuously updating the parameters yields much better VaR estimates for both models than our earlier static estimations did. Much to our surprise, there does not seem to be much difference between an unconditional model and one where we pre-filtered the returns from autocorrelation and heteroskedasticity. We see a somewhat different pattern to what we have seen in earlier tests as they both initially overestimate the risk and continue to do so. The filtered model is close to satisfying the statistical test for the lowest confidence level, and cannot be rejected for higher levels. The same can be said for the unconditional model applied directly on the return series. While the filtered model does perform *a little* better, one can question whether all the extra work is necessary since the biggest improvement is at 95%.

Choosing the threshold beforehand will inevitably case some bias in the parameter estimation, but considering that it is an often-employed procedure and that both models use the same number of threshold exceedances we figured that it ought to improve the fit a bit more for the filtered model. The fact that the models yields more accurate predictions at all confidence levels compared to what we found previously, suggests that in order for extreme value analysis to be useful in electricity markets, one would need to

update ones extreme value parameter estimates continuously, which makes sense. Other possible explanations for the improvements could be the reduced number of observations in addition to the market becoming less turbulent towards the end of the sample.

7. Conclusion

7.1. Summary

Initially, our goal was to see whether the two classic extreme value models, the Block Maxima and Peaks Over Threshold, could be used to describe the tail of the profit- and loss distribution of hourly EEX spot prices. However, we soon found out that the extreme behaviour of the data made it impossible to find any form of dependence even though we were clearly dealing with non-normality. Extreme value analysis requires iid observations in order to work properly, but since we could not find an easy way to normalize the data, we had to perform our analysis directly on the return series. This led to problems when trying to find a suitable threshold level for the POT model, as the graphical tools suggested in the literature did not tell us anything. It is possible that the data suffers from contamination, where parts of the data follow a different distribution, causing the parameters to behave very unstable and not displaying any linearity. We then had to make an arbitrary choice and decided to set the number of exceedances to equal 3000, which accounts for roughly 5% of the data. This gave us parameters that exceeded 1, suggesting that we were dealing with a process displaying infinite mean and variance. This was compared to a threshold of 1.4, where the parameters appeared a little more stable, which resulted in even higher parameter estimates.

As suggested in the literature, this led to highly unstable estimates of Value at Risk where the lower threshold led to a slight underestimation of the risk at lower quantiles (95%) whereas it overestimated the risk further out in the tail. The opposite behaviour was found using the higher threshold, which proved to be quite successful at higher confidence level while seriously overestimating the risk at 95%. From a risk manager perspective, this is very unsatisfactory considering that an implementation of this type of methodology could give raise to inefficient use of capital depending on what quantile one is interested in. Similar to what has been found in earlier studies, the Block Maxima method proved to be even worse, underestimating VaR for all confidence levels, even though they could not be rejected at higher quantiles.

We thereby concluded that neither of these ways of estimating VaR could consistently provide us with accurate estimates with this kind of data, even though the models could not be rejected for all confidence levels. Apart from contamination, we believe that a larger fraction of extreme returns compared to other studies might be the reason for this.

For these reasons, we directed our attention to daily data instead. Here we were able to find the weekly dependence and heteroskedasticity often encountered in studies directed on electricity spot markets. We could then account for this by filtering the data using an AR-GARCH model in order to capture this dependence and obtain observations that were closer to iid than the raw data. Much to our surprise, we were not able to improve our estimates, as all models struggled to provide accurate estimates further out in the tail. Moreover, the conditional model following the t-distribution performed worse compared to the model using the normal distribution, at least further out in the tail. The former followed the same pattern

as many of our earlier findings. At first it underestimated the risk, followed by an overestimation at higher confidence levels. Surprisingly, the only model that could be accepted was the POT-model using normally distributed errors, which speaks against earlier findings. All other models could be rejected, suggesting a worse fit than for hourly data.

Again, it seems as if the data displays a range too wide to be properly modelled by extreme value models. The problem of data contamination might still pose a problem, since we still experience several returns that can be deemed extreme. Furthermore, using daily data reduced our sample drastically, and lack of observations could obviously be part of the reason for these findings. Another reason might be a wrongly picked threshold, even though the graphical methods could be used in this instance.

Finally, we performed an out of sample analysis to see whether we could further improve the accuracy of our models. Here we compared the number of VaR-violations produced by a conditional- and filtered POT-model, to those of an unfiltered and unconditional POT-model. Our choice of models to include in this test was mainly to see whether the pre-filtering process was beneficial. We found that both models were able to improve the accuracy compared to our earlier findings, suggesting that updated parameters are required for accuracy. While we were expecting to see a clear improvement with the conditional model, the accuracy only increased very little and the statistical test suggested a slight rejection at the 95% confidence levels for both models while they could not be rejected for higher levels. As such, it does appear as if the refined EVT methodologies could help in increasing the accuracy of the VaR estimates, but that they are still unsatisfactory from a risk management perspective.

7.2. Concluding remarks

Our findings thus suggests that while the two classical methods for dealing with financial risk are can be applied on the EEX spot market even though they were not as accurate as found in earlier studies. Considering that the data used throughout this study exhibited very different characteristics compared to theirs, this was to be expected. The accuracy of the models depend highly on what confidence level that one is choosing and which block size- and threshold that is implemented. If one is interested in the extreme quantiles only, the POT-models with higher threshold proved to be the most accurate for all sets of data. But with the hourly data, these are hard to specify using the available methodology, making these findings questionable. The models obviously benefited from updated parameters and particularly so the conditional ones.

Even if we could not reject the models at all confidence models in a statistical sense, one should be careful in interpreting the results considering the drawbacks of the Kupiec test used to test the validity, especially for the tests using smaller samples. By applying more advanced statistical tests, such as the Christoffersen frequency tests, we might have reached a different conclusion

7.3. Suggestions for future research

Extreme value theory applied to electricity returns is scarce, and our findings suggest that it could help to improve VaR estimates to some extent. However, further improvements could be accomplished by properly model the return series. In order to do so, one would have to resort to more advanced models where special care is given to the spikes in the data, something we found to seriously distort the overall picture. Studies directed at modelling electricity prices have done exactly this, and it is possible that a combination of these to methods could help in improving the findings. We feel that this is most obvious when studying hourly data. Perhaps it also is a good idea to test the extreme value theories directly applied on the data instead of the returns, given that many electricity markets are of the day-ahead type. Looking at EEX in isolation could also be of interest, as no studies have looked at this at all. For electricity markets in general, it might be a good idea to investigate spill over effects between related markets, such as NordPool and the EEX, and see whether there is correlation in spikes across markets. This might be obtained by applying Copula theory, something that is often performed to find dependence between assets, and regular stock markets.

Bibliography

Andrews, N., Thomas, M. (2002). At the end of the tail. ERPM 75-77.

Bali, T. G. (2003). An extreme value approach to estimating volatility and value at risk. Journal of Business, 76, p. 83-107.

Bali, T. G., Neftci, S. N. (2003). Disturbing extremal behaviour o spot price dynamics. Journal of Empirical Finance, 10, p. 455-477.

Bollerslev, T. (1986). Generalised Autoregressive Conditional Heteroskedasticity Journal of Econometrics, 51, 1986, p. 307–327.

Byström, H. N. E. (2004). Managing extreme risks in tranquil and volatile markets using conditional extreme value theory. International Review of Financial Analysis 13, p. 133-152.

Byström, H. (2005). Extreme value theory and extremely large electricity price changes. International Review of Economics and Finance, 14, p. 41 – 55.

Chan, K.F., Gray, P. (2006). Using Extreme Value Theory to Measure Value-at-Risk for Daily Electricity Spot Prices. International Journal of Forecasting, Vol.22, No 2. P. 283-300.

Coles, S. (2001). An Introduction to Statistical Modelling of Extreme Values. Springer, London.

Danielsson, J., de Vries, C., (1997b). Tail index and quantile estimation with very high frequency data. Journal of Empirical Finance 4, p. 241–257.

Daníelsson, J. (2011). Financial Risk Forecasting: The Theory and Practice of Forecasting Market Risk with Implementation in R and Matlab, (The Wiley Finance Series), John Wiley & Sons.

Deng, S.J and Jiang, W.J. (2005). "Levy Process Driven Mean-reverting Electricity Price Model: a Marginal Distribution Analysis," Decision Support Systems, Vol.40, Issues 3-4 , October, p.483-494.

de Rozario, R. (2002). Estimating value at risk for the electricity market using a technique from extreme value theory. Working paper. University of New South Wales.

Dettmer, F. and Jacob, M., (2009). Stunden, in den Strom kein Gut ist. EMW 5, p. 70-72.

Diebold, F., Schuermann, T., Stroughair, J. (1999). Pitfalls and opportunities in the use of extreme value theory in risk management, Advances in Computational Finance.

Dowd, K. (2005). Measuring Market Risk. John Wiley & Sons Ltd, New York.

Duffie, D., Pan, J. (1997). An overview of Value-at-Risk, The Journal of Derivatives, p. 4-49.

DuMouchel, W.H. (1983). Estimating the stable index α in order to measure tail thickness: A critique, Ann. Statist, 11, p. 1019–1031.

Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance. Springer-Verlag, Berlin.

Fernandez, V. (2005). Risk management under extreme events, International Review of Financial Analysis, 14, p. 113-148.

Ferreira, A. de Haan, L. & Peng, L. (2003). On optimising the estimation of high quantiles of a probability distribution, Statistics, 37, p. 401–434.

Gencay, R. Selcuk, F, Ulugulyagci, A. (2003). High volatility, thick tails and extreme value theory in value-at-risk estimation. Insurance, Mathematics and Economics, 33, p. 337-356.

Gosh, S., Resnick, S. (2010). A discussion on mean excess plots, Stochastic Processes and their Applications 120, pp. 1492-1517.

Huisman, R. & Kilic, M. (2013). A History of European Electricity Day-Ahead Prices. Applied Economics, 45, p. 2683-2693.

Jorion, P. (1997) Value-at-Risk: The New Benchmark for Managing Financial Risk. McGraw-Hill, New York.

Knittel, C., Roberts, M. (2001). An empirical examination of deregulated electricity prices. Working paper, University of California.

Mayer, K., Schmid, T., Weber, F. (2011). Modeling electricity spot prices – Combining Mean-Reversion, Spikes and Stochastic Volatility. CEFS working paper series, No. 2011-02.

McNeil, A. J., Frey, R. (2000) Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. Journal of Empirical Finance 7, p. 271-300.

McNeil, A. J., Frey, R. and Embrechts, P. (2005) Quantitative Risk Management: Concepts, Techniques and Tools, Princeton University Press.

Neslehova, J., Embrechts, P., and Chavez-Demoulin, V. (2006). Infinite mean models and the LDA for operational risk. The Journal of Operational Risk 1(1), p. 3–25.

Nilsson, B. (2012) Financial Valuation and Risk Management, Lecture Notes, Ekonomihögskolan vid Lunds Universitet.

Reiss, R.-D., Thomas, T. (2007) Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Birkhäuser Verlag, Basel; Boston; Berlin.

Resnick, S. (2007). Heavy-tail Phenomena: Probabilistic and Statistical Modeling. Springer, New York.

Sewalt, M. and De Jong, C. (2003). Negative prices in electricity markets", Commodities Now, June 2003, p. 74-77.

Scarrott, C., Macdonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification, REVSTAT – Statistical Journal, Vol. 10, No. 1 p. 33-60.

Verbeek, M. (2008). A Guide To Modern Econometrics. Wiley 3rd Edition.

Xiu-min, L., Fa-chao, L. (2006) Extreme value theory: an empirical analysis of equity risk for Shanghai stock market. *2006* International Conference on Service Systems and Service Management 2, p. 1073-1077.

Electronic and Other Sources:

European Energy Exchange (EEX): www.eex.com
NordPool: www.nordpool.com