ROYAL INSTITUTE OF TECHNOLOGY

MASTER OF SCIENCE THESIS PROJECT

A PIT – Based approach to Validation of Electricity Spot Price Models

Author: Hampus Engsner
hengsner@kth.se

Supervisor:
Sergey Zykov
sergey.zykov@vattenfall.de

Examiner:
Timo Koski
tjtkoski@kth.se

August 19, 2015

Acknowledgements

I would like to thank the Modelling and Methodology unit at Vattenfall on which behalf I conducted this research. In this group I especially thank my supervisor Sergey Zykov, whose many suggestions, thoughts and pointers during this project have been invaluable. Also I would really like to thank my examiner Timo Koski at the KTH mathematics department, who helped me make this thesis academically presentable and who also gave me many important suggestions and pieces of advice.

Hampus Engsner, August 19, 2015

Abstract

The modeling of electricity spot prices is still in its early stages, with various different competing models being proposed by different researchers. This makes model evaluation and comparison research an important area, for practitioners and researchers alike. However, there is a distinct lack in the literature of consensus regarding model evaluation tools to assess model validity, with different researchers using different methods of varying suitability as validation methods. In this thesis the current landscape of electricity spot price models and how they are currently evaluated is mapped out. Then, as the main contribution this research aims to make, a general and flexible framework for model validation is proposed, based on the Probability Integral Transform (PIT). The probability integral transform, which can be seen as a generalization of analyzing residuals in simple time series and regression models, transforms the realizations of a time series into independent and identically distributed U(0,1) variables using the conditional distributions of the time series. Testing model validity is with this method reduced to testing if the PIT values are independent and identically distributed U(0,1) variables. The thesis is concluded by testing spot price models of varying validity according to previous research using this framework against actual spot price data. These empirical tests suggest that PIT-based model testing does indeed point us toward the more suitable models, with especially unsuitable models being rejected by a large margin.

Sammanfattning

Modelleringen av spotpriser på el är fortfarande i ett tidigt stadium, med många olika modeller som förespråkas av olika forskare. Detta innebär att forskning som fokuserar på modellutvärdering och jämförelse är viktig både för berörda parter i näringslivet och forskare inom detta område. Det finns dock en klar brist på konsensusmetoder att utvärdera modellers validitet, då olika forskare förespråkar olika metoder av varierande lämplighet som valideringsverktyg. I den här uppsatsen kartläggs det nuvarande landskapet av spotprismodeller och de metoder som används för att utvärdera dem. Sedan, som det huvudsakliga forskningsbidraget av detta arbete, presenteras ett generellt och flexibelt valideringsramverk som baseras på vad som kallas ”Probability Integral Transform” (PIT). PIT, vilken kan ses som en generalisering av att undersöka residualer i enkla tidsserie- och regressionsmodeller, transformerar utfallet av en tidsserie till oberoende och identiskt fördelade U(0,1) variabler med hjälp av tidsseriens betingade fördelningar. Att testa modellens validitet reduceras med denna metod till att testa om PIT – värdena är oberoende och identiskt fördelade U(0,1) variabler. Uppsatsen avslutas med tester av spotprismodeller av varierande validitet enligt litteraturen med hjälp av detta ramverk mot faktiskt spotprisdata. De empiriska testerna antyder att PIT – baserad modellvalidering faktiskt stämmer överrens med modellers validitet baserat på nuvarande konsensus, där särskilt opassande modeller förkastas med stora marginaler.

Table of Contents

1 Introduction 6

2 Background 8

3 Literature Review 11

3.1 Models of the electricity spot price in the literature 11

3.2 Model validation methods currently observable in the literature 17

4 Mathematical Background 22

4.1 The Probability Integral Transform (PIT) 22

4.2 Testing for distribution and independence 25

4.3 Ranking validated density forecasts using sharpness 29

4.4 Evaluating interval forecast performance of models 30

4.5 Motivation of the use of PIT-based validation methods 31

5 Methodology 36

5.1 Data description 36

5.2 Deterministic Component 39

5.3 Diffusion Models 40

5.4 Methodology for model validation 44

5.5 Parameter estimation 45

6 Results 49

6.1 Parameter estimation results 49

6.2 Visual results of PIT transforms 54

6.3 Test statistics 55

7 Discussion 56

8 Conclusion 59

9 References 60

10 Appendix A1: Tables 63

1 Introduction

The last decades have seen a liberalization of electricity markets around the world, leading to electricity prices being determined by supply and demand. A multitude of contracts which all depend on the electricity spot price are traded on the markets or over-the-counter, making the spot price an entity of great importance for risk management and valuation for market participants. However, due to the non-storability of electricity as a commodity, its inelasticity and its dependence of weather conditions and consumer patterns, the electricity spot price has several unique characteristics. The three most pronounced characteristics are seasonality of mean price levels on various timescales, a mean-reverting fluctuation around said mean levels and the existence of large, but short lived, changes in price called "price spikes". These characteristics have lead to many different modelling approaches in academia, with different researchers championing different modelling methods. As (Meyer et al 2015) put it: "*Electricity price modeling is complex and still in its infancy*".

The multitude of choices for the practitioner makes the process of model selection and validation an important related subject. However, while several papers handle cases of model selection, statistically rigorous validation or comparison procedures for models have been rare and different metrics and levels of statistical rigor are used by different authors. In this thesis, the concept of validation is interpreted as a method to determine a model's adequacy, preferably in absolute terms rather than in a relative sense to some other model, but also possibly by showing model superiority to some benchmark model. As literature specifically on the subject of model validity is scant, the reasoning of this thesis relies heavily on the statistical adequacy model selection criterion outlined by (Spanos 2010), originally invented by Fischer:

"the task [of model selection] *is understood as one of selecting a statistical model that renders the data a typical realization thereof, or equivalently, the postulated statistical model `accounts for the regularities in the data'"*

In this thesis the term model validity and model validation will primarily refer to statistical adequacy and methods of measuring it. This definition also calls for model validation methods that do not depend on other models to benchmark against, but are freestanding.

Based on the above observation, Vattenfall's "Models and Methodology" unit has requested a comprehensive overview of validation procedures that can be found in the literature and, importantly, requested suggestions for rigorous methods of model validation. Based on a thorough investigation of the existing electricity spot price literature, a general validation method is suggested which is based on the so-called Probability Integral Transform, an out-of-sample approach that transforms a time-series with known stochastic behavior into independent and identically distributed uniform random variables. The hypothesis of a correctly specified model thus amounts to testing transformed data for independence and distribution. Furthermore, this validation method is used on some of the most common spot-price models in the literature. The purpose of this testing is *not* to absolutely determine which of the various relevant spot price models is better suited for modelling the electricity spot price, but rather to investigate what triggers acceptance and rejection depending on approach to check the transformed values and to showcase the suggested method of validation in action.

The research contribution this thesis aims to make is to propagate this method of validation for use in future research papers as a basic test of general model adequacy, so that models in different papers may be compared more readily. Furthermore, the subject of model validation, rather than the models themselves, has not been the sole focus of any academic paper so far observed in the literature and can thus be said to constitute somewhat of a gap in the research.

The outline of this thesis is as follows: First, in section 2, a background of the electricity markets will be given and the electricity spot price and its qualities will be outlined. Then, the literature review in section 3 will give a description of the most common models used for the electricity spot price as well as what validation procedures are used in the literature. This mapping of validation procedures should be seen as one of the key results of the thesis. In the mathematical background in section 4 the Probability Integral Transform (PIT) and associated testing procedures will be mathematically described and the choice of PIT-based validation will be mathematically motivated. As a secondary result, some limit theorems will be presented regarding the calculation of order-invariant statistics with regards to certain classes of mean-reverting time series, to shed light on some of the validation methods observed in the literature. Section 5 and 6 are devoted to testing some common models using PIT-based validation, importantly testing both known good models and known bad models of the electricity spot price. This testing can be seen as the second main result of the thesis. Sections 7 and 8 are the Discussion and Conclusion chapters of the thesis.

2 Background

2.1.1 The Electricity markets

During the last decades, the electricity market in many countries has been increasingly liberalized, transforming the markets from heavily regulated and government controlled entities to deregulated and competitive ones (Bierbrauer et al 2007). Furthermore, most of these electricity markets are local, meaning that each market trades electricity in a certain zone, which may be a whole country. Most of the recently developed European markets are Power Exchanges, whose primary purpose is to match supply with demand and announce a market clearing price, known as the *spot price*. This is generally done via auctions the day before, where the bids concern the prices for different hours of the next day. However, it should be noted that exact bidding processes vary across different markets. There also exist *balancing* or *real-time* markets for delivery within short time horizons (Weron 2006).

The creation of the liberalized electricity markets has led to the trading of a variety of contracts on electricity, for instance futures contracts and options, which may be either sold "Over-the-counter" (i.e. bilaterally) or on the market. However, while for instance futures and forwards can be seen as long-term contracts, very short term *spot contracts* are also sold (Weron, 2006).

2.1.2 Characteristics of the electricity spot price

Since many contracts available on the electricity market depend on the electricity spot price, the understanding of the electricity spot price, and electricity as a commodity, is vital for pricing derivatives and calculating risk.

The characteristics of electricity as a commodity, to begin with, are rather unique. First of all, electricity is a non-storable commodity and the stability of power systems requires a match between inputs and outputs. Secondly, in relation to non-storability, electricity exists only as a flow and is thus buyable only in terms of a certain effect (in watts) for a certain time (usually hours). Electricity prices also depend on weather conditions and on the real-time activities of its consumers (Weron 2014). Finally, the continuous flow of electricity is essential to many businesses, industries and private consumers, making electricity as a commodity inelastic in the short term (Weron 2006, Geman and Roncoroni 2006).

The unique features of electricity as a commodity, often mentioned in papers on electricity spot prices, lead to three main characteristics which almost no article on electricity spot prices will fail to mention: *mean reversion*, *seasonality* and *price spikes.*

Mean reversion is a characteristic observed in many commodity prices, notably described in a mathematical sense by Schwartz (1997). Mean reversion means that spot prices fluctuate around some mean level that represents marginal cost (Geman and Roncoroni 2006). This mean level can possibly be time-varying and can be explained by fluctuation of demand leading to increased marginal costs (Bierbrauer et al 2007). Notably, the mean reversion in electricity prices is rather strong (Bierbrauer et al 2007), with some authors finding that prices return to mean level within days (see e.g. Cartea and Figueroa 2005).

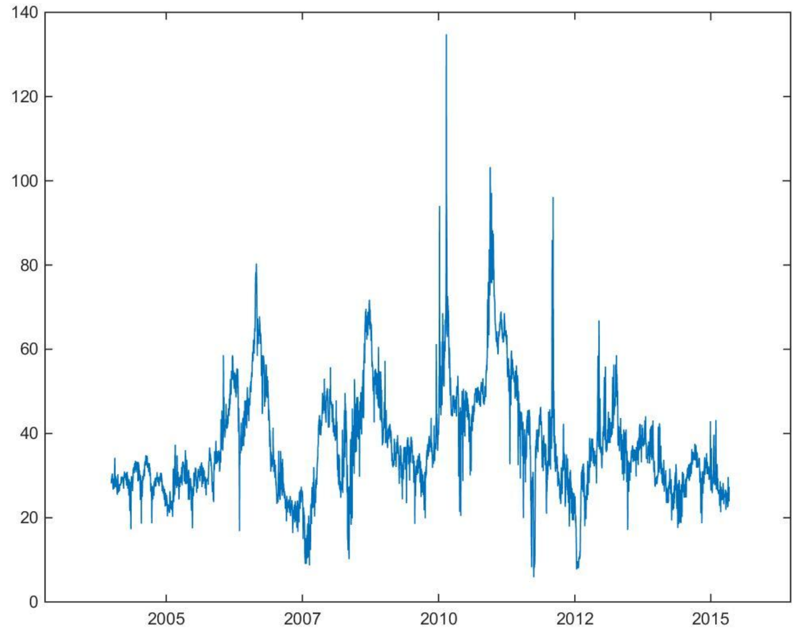

Figure 1: daily baseload price of the System price of the Nordpool market. The prices are quoted in EUR.

This mean level of electricity prices also displays a more pronounced *seasonality* than any other commodity (Bierbrauer et al 2007). Since the electricity price depends heavily on weather conditions and the real-time activities, the electricity price experience daily seasonality (with some hours being "peak" hours with increased demand), weekly seasonality (weekdays and weekends have different demands) and annual seasonality (weather conditions change throughout the year and may affect power consumption) (Weron 2014).

A final very characteristic feature of the electricity spot price is the occurrence of what is often called *price spikes* in the literature (Weron 2006, Benth et al 2008, Weron 2014, Geman and Roncoroni 2006). Price spikes are extreme moves in the spot price that quickly reverts back to a normal level. The reasons for price spikes include the inelasticity of the electricity price, the increasing marginal cost of producing electricity and the bidding strategies of producers (Benth et al 2008). A famous example of such a spike is the June 1998 Cinergy Price spike in mid-western USA, which momentarily saw the price rise from the typical level of 30 USD/MWh to a peak of 7500 USD/MWh with a daily average of 183.33 USD/MWh. This price spike resulted in the default of power obligations of at least two companies, Federal Energy Sales and Power Company of America, PCA, which had to file for bankruptcy (Weron 2006). A feature of price spikes in some markets is that the intensity of price spikes, be it the daily or annual, is observed to be time-inhomogeneous. Periods such as peak hours or the winter season in Scandinavia or the summer in Western USA are especially prone to price spikes (Weron 2006).

A fact not always mentioned about the electricity spot price is that it is, due to transportation cost, regional. For instance, the 1998 Cinergy spike did not affect nearby markets. This is another factor that makes electricity as a commodity different from other financial and most commodity markets (Weron 2006).

Finally, what is not always noted is that, due to the non-storability of electricity, the spot price behavior versus that of derivatives on it is not consistent with the usual no-arbitrage pricing formulae, as the dynamic hedging and cash-and-carry arguments in arbitrage pricing do not work in this context (Cartea et al 2009). This means that the electricity market is *not* complete (Benth et al 2012).

3 Literature Review

The literature review is divided into three parts. First, the landscape of electricity spot price models and their uses will be mapped out. Secondly, the methods of validation of these models that are currently observable in the literature will be reviewed. Third, the proposed main method of validation proposed in this thesis, the LR test on PIT values, will be presented and its grounding in the literature established.

3.1 Models of the electricity spot price in the literature

3.1.1 Two classes of models in the literature and their uses

Weron (Weron 2006, Weron 2014), amongst others, identifies two types of probabilistic models, with different uses: Quantitative (or reduced form models), which refers to a continuous time stochastic process approach to modeling the electricity spot price. The use for these kinds of models is, according to Weron, to "characterize the statistical properties of electricity prices over time, with the ultimate objective of derivatives evaluation and risk management" (Weron 2006). Their primary use is thus not point forecasting (Weron 2014). (Cartea et al 2009) also write that the main purpose of reduced-form models is to capture the main characteristics of electricity prices.

The second type of models is Statistical models, which refers to time-series models of type ARMA, ARIMA, GARCH, etc. The primary use of these, according to Weron, is point forecasting (Weron 2006). Since these models are often compared based on different properties than the reduced-form models above, and are not generally used for valuation, they will not be the focus of this thesis.

The rest of the literature does indeed confirm this partition of models: Reduced-form models in for instance (Benth et al 2011, Geman and Roncoroni 2006) are primarily used for futures valuations, while time series models in for instance (Weron 2006, Escribano et al 2011) are primarily used for point forecasting.

All further discussion in this literature review will thus only be concerned with models used for risk and valuation.

3.1.2 Seasonal and Stochastic components

The modeling of the electricity spot price observed in the literature consists of two initial steps. The first step is to de-seasonalize the data to identify the seasonal component, which is often done using dummy variables and/or sinusoidal functions. The second, much more complicated, step is to model the stochastic deviation from the seasonal component .This approach is, amongst others, used by (Benth et al 2011, Geman and Roncoroni 2006, Cartea et al 2009, Janczura et al 2010, Higgs and Worthington 2008, Escribano et al 2011). However, whether to divide prices or log-prices in a stochastic and seasonal component differ between authors. Janczura et al (2013), for instance, identifies adding a seasonal component to the prices as an industry standard, while other authors (e.g. Benth et al 2012, Cartea et al 2009, Bierbrauer et al 2007) divide log-prices into a seasonal component *f*(t) and stochastic component X(t) illustrated in equation (1).

$$\ln(P(t)) = f\,(\mathrm{t}) + \mathrm{X(t)} \tag{1}$$

In (1), P(t) is the spot price process, *f*(t) is the deterministic part and X(t) is the stochastic part. Now, it is worth noting that while models in literature are generally identified by the stochastic component X(t), the model in its entirety is defined by the pair (*f*(t), X(t)). Thus a short review of the modelling of the deterministic part *f*(t) is in order, before the more extensive modelling of X(t) is handled. The method proposed by (Bierbrauer et al 2007) is a combination of a trend, a sinusoidal function and dummy variables:

$$f\,(t) = \alpha + \beta t + \mathbf{d}^{\mathrm{T}}\mathbf{D}_{\mathrm{day}} + \mathbf{m}^{\mathrm{T}}\mathbf{D}_{\mathrm{month}} + \gamma \sin\left((t+\tau)\frac{2\pi}{365}\right) \tag{2}$$

Here the vectors $\mathrm{D}_{\mathrm{day}}$ and $\mathrm{D}_{\mathrm{month}}$ are vectors of indicator functions for different days and months (to avoid issues with colinearity, these should include all but one day and month), and the vectors d and m are parameter vectors. The authors, similar to most other authors on this subject, use non-linear least-squares regression to estimate *f*(t) to data.

Some combination of constant, trend, sinusoidal and dummy variables is used by most authors. For instance, (Benth et al 2012) models *f*(t) as a sum of a constant component, a trend component and two sinusoidal components, the first with one-year periodicity and the second with six month periodicity. (Cartea and Figueroa 2005), as another example, fit a Fourier series of order 5 to monthly average data.

3.1.3 Jump-Diffusion models

The most common reduced-form model observed in the literature so far is some variation of the Jump-Diffusion model, described in various forms by amongst others (Benth et al 2011, Weron 2006, Cartea and Figueroa 2005, Geman and Roncoroni 2006, Weron et al 2004). One of the simplest Jump diffusion models is described by the dynamics (Cartea and Figueroa 2005)

$$\mathrm{d}Y_t = -\alpha Y_t \mathrm{d}t + \sigma(t)\mathrm{d}W_t + \ln(J)\mathrm{d}q_t \tag{3}$$

Here Y_t is the logarithm of the de-seasonalized spot-price and W_t is a standard Brownian Motion. With a slight abuse of notation, $\ln(J)dq_t$ represent a compound Poisson process, i.e. jumps of size ln(J) (where J is an IID random variable for each jump) occurring at exponentially distributed intervals. Note, first off all, that without the last term this is an Ornstein-Uhlenbeck process. In fact, most of the reduced-form models encountered in the literature consist of an Ornstein-Uhlenbeck process, with some added dynamics to account for price spikes and varying volatility. These models are often abbreviated as MRJD (Mean Reverting Jump Diffusion) models.

The dynamics of (3) as described by (Cartea and Figueroa 2005) are fairly simple: $\sigma(t)$ is assumed to be rolling historical volatility, dq_t is a time-homogeneous Poisson process with the intensity l, and J is log normally distributed such that $E[J] = 1$.

However, any number of dynamics could be assigned to the model. For instance, since price spikes in some markets tend to be seasonal; (Geman and Roncoroni 2006) suggest introducing a time varying jump intensity, making the Poisson process in (3) time-inhomogeneous. The authors modifies the models in many other ways: They introduce time-varying volatility (in their case rolling historical volatility), giving the jumps a truncated exponential distribution and giving jumps signs according to whether or not the price is above or below some threshold (in this case jumps are given positive sign if the price is above the threshold). Due to this last change, (Benth et al 2012) refer to this model as the "threshold model". Note that this does not automatically mean that positive jumps are followed by negative jumps, thus this signing of jumps does not automatically generate the familiar spike-shapes of electricity spot prices. (Weron et al 2003), in contrast to the model in (Geman and Roncoroni 2006) dictates that a jump must be followed by a negative jump in order to achieve the familiar spike-shape of price spikes.

Similarly, in many of the above and following models a large variety of different jump distributions have been tried by for instance (Benth et al 2012, Geman and Roncoroni 2006). Alternative spike-distributions include truncated exponential, Pareto and Gamma distributions.

In total, variations on the Ornstein-Uhlenbeck Jump Diffusion model account a large part of reduced-form models in the literature, and could possibly be said to be the industry standard, based on the literature reviewed so far.

3.1.4 Non-Gaussian diffusion models

A slightly different approach propagated by Benth et al (2008) and Benth et al (2007) is an arithmetic model:

$$S(t) = \mu(t) + X(t) \tag{4}$$

$$X(t) = \sum_{i=1}^{n} w_i Y_i(t) \tag{5}$$

Where μ(t) is a deterministic and periodic function (representing the deterministic lower bound, not the mean, of the price process) and the non-Gaussian Ornstein-Uhlenbeck processes $Y_i(t)$ are governed by the dynamics:

$$\mathrm{d}Y_i(t) = -\lambda_i Y_i(t)\mathrm{d}t + \sigma_i(t)\mathrm{d}L_i(t) \tag{6}$$

The processes $L_i(t)$ i = 1, ...,n are assumed to be independent increasing càdlàg pure jump processes. The idea of this model is that large jumps can be captured by one of the benefits of this model is that the price spikes can be represented by one of the summands in (5), which may have a very high mean reversion (creating the typical spike-shape), while more common price moves can be represented by the other processes with less mean-reversion.

Common distributional choices for the stationary distribution of L are the Gamma or Normal Inverse Gaussian (NIG) distributions.

Noteworthy, and also useful as an example, is that the Gamma stationary distribution simply corresponds to a time-homogeneous compound Poisson process with exponential jumps. If the Levy process L is specified as having the stationary distribution $L(1) \sim \Gamma(\nu, \alpha)$, this corresponds to jumps with an Exp(α) distribution arriving according to a Poisson process with intensity ν.

A positive aspect of this kind of model is that it allows for a comparatively simple calculation of futures prices given that it is arithmetic.

The theory on these kinds of processes is detailed in (Benth et al 2008) and we won't discuss it further here.

3.1.5 Visible Regime-Switching models

A different approach from the above involves modeling the spot price using a regime-switching model, such as in (Weron 2006, Rambharat et al 2005, Weron et al 2004, Janczura et al 2010, Cartea et al 2009). Noteworthy is that these kinds of models are taken up in Weron (2006) both under "Quantitative" models and "Statistical" models, implying that the classification of different spot price models, as might be imagined, is a bit fuzzy.

Weron (2006) roughly classifies regime-switching models in two categories: Ones where the regime is readily observable and one deterministically can determine historic and current regimes and models where the regime is some unobserved hidden variable, whose possible historic values can only be inferred. Note that these models are discrete-time models, but their use seems to be mostly valuation as in for instance (Janczura et al 2010).

A simple class of observable-regime models that Weron (2006) introduces is called a Threshold AutoRegressive (TAR) model. The model has the following dynamics:

$$\begin{cases} \phi_1(B)P_t = \varepsilon_t,\ v_t \geq T \\ \phi_2(B)P_t = \varepsilon_t,\ v_t < T \end{cases} \tag{7}$$

Where t is a threshold and $\phi_i(B) = 1 - \phi_{i,1}B - \ldots\ - \phi_{i,p}B^p$ where B is the backward shift operator. The threshold variable v_t can be, for instance the lagged price P_{t-d} or some function thereof. As Weron (2006) points out, this type of model is quite rarely applied to the electricity spot price in the literature. However, for instance (Rambharat et al 2005) compare a TAR model with regime dependent distribution and mean reversion to a standard mean-reverting jump-diffusion model. Also, (Cartea et al 2009) employ an observable regime-switching model decided by the following variant of the jump-diffusion model:

$$\mathrm{d}y_t = -\beta(t)y(t)\mathrm{d}t + \rho(t)\ln(J)\mathrm{d}N(t) + (1-\rho(t))\mathrm{d}Z(t) \tag{8}$$

Here $\rho(t)$ is the regime parameter which takes the binary value of 0 or one, N(t) is a Poisson process and Z(t) is a Lévy process. In (Cartea et al 2009), $\varrho(t)$ is determined as the quotient between a demand forecast and a generation capacity forecast, figures which are readily available for the practitioner.

3.1.6 Hidden Markov models

More common than the above approach, however, is the hidden Markov or Markov regime-switching models (commonly denoted HMM or MRS models) (Weron 2006, Janczura et al 2010, Weron et al 2003). An MRS model works as In the following way: Let R_t be a n-state time-homogeneous Markov chain, i.e. R_t is a discrete-time random process which takes values in {1, ...,n} and with the property (9):

$$\mathrm{P}(R_t = j \mid R_{t-1} = i,\ R_{t-2} = k,\ \ldots) = \mathrm{P}(R_t = j \mid R_{t-1} = i) = \mathrm{P}(R_2 = j \mid R_1 = i) \tag{9}$$

This yields that the distribution of R_t in vector form is equal to $Q^T e_j$ if $R_{t-1}=j$, where e_j is the j: the unit vector in R^n. $Q = (Q_{ij})_{i,j} = (P(R_2 = j|R_1= i))_{i,j}$ is called the transition matrix of the Markov chain R_t (Janczura et al 2010).

Now, in a MRS model, we simply assume that the spot price process has n possibly distinct distributions depending on the value at time t of the non-observable regime variable R_t. In the observed literature so far, two- and three-regime models has been tried (Weron 2006). In general, a hidden Markov model can be described in equation form as follows (in the framework of for instance (Janczura et al 2010)

$$\mathrm{d}X_{t,b} = \mu_b(t, X_{t,b})\,\mathrm{d}t + \sigma_b(t, X_{t,b})\mathrm{d}W_t \tag{10}$$

The pair t, b indicates time and regime respectively. Possibly, as in (Bierbrauer et al 2007), one could restrict the effects of regime-switching to just the diffusion parameters. The idea behind a two-regime model is to have one "normal" regime in which the spot price behaves as it usually does and one "spike" regime where we see a large increase or decrease in the spot price. As mentioned above, there is a great deal of effort in getting the spike-shape of a price spike just right. A proposed version of a *three-regime* MRS model solves this by having one "normal", one "spike" and one "drop" regime. As the name imply, the "spike" regime consists of a drastic increase in the price process, and the transition matrix Q is sometimes specified so that this regime is immediately followed by the "drop" regime which assures a drastic downturn of the process from the spike value. Furthermore, Q is in this case also specified so that the "drop" regime is followed by the "normal" regime. (Weron 2006)

Since the user is free to specify basically any dynamics for the different regimes, this class of models is obviously very versatile. In the comparative article by (Janczura et al 2010), three different MRS models are tested and compared. In the paper the best suited model according to the tests is found to be a three-regime model with a heteroskedastic base regime and median-shifted log-normal "spike" and "drop" regimes. Noteworthy for this model is that the states in this model are persistent, i.e. there is, for all states, a fairly large probability of staying within the regime. Hence, this particular three-state MRS model does not necessarily has to have pre-specified "spike" and "drop" regime probabilities as described above.

3.2 Model validation methods currently observable in the literature

In this chapter various validation procedures, or more commonly goodness-of-fit statistics, that can be found literature are presented. The distinction between these terms in this thesis is that a validation procedure contains some absolute accept-reject criterion and aims to ascertain statistical adequacy, i.e. the user is supposed to know if a model is unsuitable just based on the model performance in and of itself. By this distinction, goodness-of-fit statistics are measures of how well the estimated model fit the estimation data and does usually not contain an accept-reject criterion. Rather, in model selection procedures, the model with the best fit to the data is chosen above the others. This does not necessarily mean goodness-of-fit statistics cannot be converted to a validation procedure, however. Rather, by comparing a proposed model to some benchmark industry standard model, a goodness- of fit test turn into a validation procedure. However, in comparison papers this is rarely done, i.e. none of the models compared is explicitly considered a “benchmark”. Furthermore, since goodness-of-fit tests test models against the same data used for estimation, it can be argued that this setting puts the modeler in an unrealistic situation and hence is generally inappropriate for model validation.

Another distinction that can be made is in-sample and out of sample tests. In-sample tests are tests in which the same data that is used for calibration is also used to produce the test statistic. Goodness-of-fit tests are by the definition made above in-sample tests.

In out-of sample tests, however, any realized data point may only enter into the test using information that is readily available *before* the data point is available. This hypothetically puts the modeler in a world where some portion of the data is unknown at the time of parameter estimation. An example of this is if we calibrate a model for one time-period, but test the model on a later time period. Another approach is to actually re-calibrate the model several times over the testing period.

3.2.1 In sample ocular inspection

Most authors use, as a complimentary qualitative evaluation tool, ocular inspection of a simulated curve compared to the dat. This is generally done in an in-sample manner, simulating the model with parameters estimated from the same data set which it is then compared to. One can then argue about visible differences or similarities between the curves (see for instance Geman and Roncoroni 2006, Cartea and Figuerora 2005, Cartea et al 2009, Benth et al 2012). A typical example on what can be achieved by this simple method is that one can make observations on mean reversion speed, apparent volatility of the price during non-spike periods or simply investigate if mean reversion and price spikes seem to be accurately represented (e.g. Benth et al 2012, Cartea and Figuerora 2005). Especially, if a model contains some glaring shortcomings, this could be spotted in this manner (too high/low mean reversion, say).

3.2.2 In sample comparison of descriptive statistics

A common in-sample approach to goodness-of-fit is to compare descriptive statistics of the model estimated on a dataset with that of the dataset itself. For instance, many authors (e.g. Geman and Roncoroni 2006, Benth et al 2012, Cartea et al 2009), simply compare the first four moments of their models compared to that of their data. This is a fairly simple way to see what the model tend to overestimate and underestimate respectively, but one should note that the goodness of fit of the model in this regard may depend on the method of estimation. For instance, (Cartea et al 2009) actually uses the squared sum of differences between the first four empirical and model moments as the objective function in their parameter estimation algorithm. Thus one should be careful when utilizing this method for model evaluation and comparison, but it may of course be a very useful tool for spotting model inadequacies.

While the above statistics are the most common within this method other statistics are used as well. For instance (Janczura et al 2010) use the Inter-Quartile and the Inter-Decile Range (IQR and IDR, respectively). These measures are the distances between the third and first quartiles, and ninth and first deciles respectively. The authors state robustness to outliers as reasons for using these measures. However, one should note that whatever the measure, there are no clear accept-reject criteria based on these, or a stated method of determining significance of differences. Rather, these statistics should probably be seen more as useful tools from a modelling standpoint to better understand the models.

These kinds of statistics will be discussed in the last section of the mathematical background, since there are some limit theorems that informs us of what we are really measuring when we measure models using order-invariant statistics such as sample moments, IQR and IDR of stationary processes.

3.2.3 Likelihood-related tests: AIC, BIC/SC, LR tests

For comparative purposes, many authors perform likelihood-related tests on different models. In their paper comparing Markov regime-switching models (Janczura et al 2010) supplements various tests with Likelihood statistics for each model. (Higgs and Worthington 2008) and (Rambharat et al 2005) use the well-known Akaike Information Criterion (AIC) in order to rank the different models under consideration. The AIC is given by the equation (11) and adjusts likelihood, L with the number of parameters used, k (Spanos 2010). Note that measures such as the AIC are only useful in the model comparison setting, as the AIC in and of itself gives no real information.

$$AIC = 2k - 2\ln(L) \tag{11}$$

Both these articles could be said to actually contain model validation of sorts, since they both compare more complex models with more standard, simpler “benchmark” models. For instance, (Rambharat et al 2005) compares a more complex TAR model with a mean-reverting jump-diffusion (MRJD) model and find a lower AIC for the TAR model than for the MRJD model. However, what constitutes a significant difference in AIC is not clearly explained in any case.

(Bierbrauer et al 2007) however, introduces significance into likelihood-based model selection. Using the Likelihood Ratio (LR) test, the authors test nested pairs of models or unrelated models via pair wise testing. Regarding the nested case, the authors find evidence for the more complicated models being more appropriate. For instance, a regime switching model with two regimes and pre-specified probability 1 to return to the normal regime is rejected at 1 % level in favor of the more general model where probabilities are not pre-specified. Furthermore, by pair wise testing, the authors conclude that in terms of this test and for the dataset consisting of daily prices from the EEX market, a number of different regime-switching models are superior to all the considered diffusion-models. However, it should be noted that even this more rigorous kind of comparison is an in-sample test.

3.2.4 Futures/Forwards-related investigations

Various authors (e.g. Cartea and Figueroa 2005, Benth et al 2012, Bierbrauer et al 2007) have, as a part of model evaluation, investigated the implied behavior of forward or futures prices according to the model or models under evaluation. The market price of a forward with maturity t at time t is defined as (Benth 2012)

$$F(t,T)^{\mathbf{Q}} = \mathbb{E}^{\mathbf{Q}}\left[S(T) \mid \mathcal{F}_t\right] \tag{12}$$

Where Q is an equivalent pricing measure and S(T) is the spot price at time t and $\mathcal{F}_t$ denotes the information up until time t (or rather, the σ-algebra representing the information up until time t in a filtered probability space). However, as the authors point out, the electricity market is incomplete and thus there exist several such measures Q. To identify such a Q, one usually restricts the choice to some parametric class which is then fitted to the data (Benth et al 2012). Generally, the qualitative behavior of the forward prices or the implied risk premium (market forward price minus predicted spot price) is observed and discussed.

Since this is done in several papers, it is worth mentioning in the literary review. However, since derivatives valuation is not the focus of this thesis, this topic will not be delved into any further. One might also note that in the sources covered for this work, no clear test statistic is derived for this kind of investigation and from a model validation standpoint, modelling Q introduces additional model uncertainty apart from the modelling of the spot price and thus clouds the main issue of the thesis somewhat.

3.2.5 Hypothesis testing ("other")

A number of authors also perform various hypothesis tests that in this thesis will be classified as "other" hypothesis tests as they are not Likelihood based or based on the probability integral transform, which will be described in more detail below.

First of all, there seems to be no consensus in the literature regarding spike distribution, thus in model selection different distributions are used in different papers. In their comparative paper, for instance, (Benth et al 2012) find evidence for choosing a Gamma distribution for jumps in two different diffusion models using the non-parametric Kolmogorov-Smirnov (K-S) test statistic. Authors like (Cartea and Figueroa 2005) simply assume normal jumps of de-seasonalized log prices.

More in line with model validation, in their paper comparing three different regime-switching models, (Janczura et al 2010) actually hypothesis test the implied distributions of the different regimes against data. This is done by transforming the data into a mixture of IID samples using smoothed inferences and testing the regimes individually as well as the whole dataset using the Kolmogorov-Smirnov test. In this way, by considering how many rejections occur, one can differentiate between different models in terms of adequacy. While no absolute accept-reject criterion is given, one could easily been construed. Although this approach is indeed a genuine model validation procedure, one should note that this specific method is only applicable to regime-switching models and furthermore that the test is an in-sample test. In any case the approach is similar to the proposed probability integral transform approach. However, it is difficult to determine exactly what these smoothed inferences can be intuitively thought of and what a rejection would mean for the modeler.

(Meyer et al 2015) perform tests very similar in spirit to the probability integral transform method below, in a cross-validation setting. Put concisely, the authors remove one month at a time from the dataset, estimate their models on the rest of the data and make comparisons between simulated model paths and the data for that period. The result is a variety of accepted and rejected months and models can be compared by for how many months they are rejected. The hypothesis test performed is a rank sum test, which is a novel approach in the electricity spot price literature. An interesting note that the authors make is that for spot price models that are to be used for option pricing, for instance, the dispersion of the price is more important than the exact price forecast that the model makes. This is the same view on spot price model adequacy that this thesis takes, namely that the entire distribution of the model should be, in some sense, correct. However, we should again note that while not technically an in-sample validation method, cross validation in a time series setting uses information from 'the future' to model data from 'the past', making it a somewhat unrealistic setting, similar to an in-sample test.

3.2.6 Out of sample Interval forecasts

Rigorously testing interval forecasts was notably discussed in the seminal article by (Christoffersen 1998). This approach will be described in-depth in the mathematical background, but intuitively, it is simply hypothesis testing of out-of-sample model implied confidence intervals where both coverage and dependence in time can be tested.

While only (Bierbrauer et al 2007) in the reviewed literature have tested electricity spot price models using out of sample confidence intervals, they do not perform any hypothesis test, but rather compares the number of exceedencs between different models. An advantage of performing these kinds of tests is that it is useful in the context of Value-at-Risk back testing (Berkowitz et al 2009) and thus it seems logical to evaluate a spot price model in this way if one of its purposes regards risk.

3.2.7 Probability Integral Transform

The probability integral transform (PIT) is an out of sample transformation which transforms data into IID U(0,1) variables, under the hypothesis that the data is generated by the model of interest. This can be seen as a distributional test, but it also possible (and indeed advisable) to test for dependence.

In the reviewed literature, only (Bierbrauer et al 2007) and (Escribano et al 2011) have used this test to validate models, but the general nature of this validation scheme along with its intuitive foundations makes this the primary focus of this thesis, and thus its mathematical background will be recapitulated in Chapter 4 below.

4 Mathematical Background

In this section the relevant mathematical concepts of the thesis will be presented as they are described in the reviewed literature. In the last section of the mathematical background, the mathematical arguments for the choice of PIT – based validation will be presented, as well as some informative limit theorems for a wide class of models concerning order-invariant measures commonly seen in the literature.

4.1 The Probability Integral Transform (PIT)

Below the theory regarding the main validation idea for electricity spot price models will be presented. The test concerns the entire out-of-sample density functions that are implied by the model under testing, rather than just some statistic from the model. This method reduces the test of a model specification to a relatively simple test for distribution and independence for Independent and Identically Distributed (IID) variables.

4.1.1 Density Forecasts and loss functions

The concept of density forecasts and loss function as described by (Diebold et al 1998) will now be defined. Let $\{y_t\}$ be a time series, and let $\Omega_t = \{y_{t-1}, y_{t-2}, \ldots\}$. Furthermore, let $f_t = f(y_t| \Omega_t)$ be the density function of y_t given the outcomes Ω_t up to time t – 1. Note that in general we cannot observe *f*, rather, the modeler assumes $\{y_t\}$ to be generated by some model yielding a joint distribution. Thus, let $p_t = p(y_t| \Omega_t)$ be the 1-step-ahead *density forecasts*, i.e. presuming the time series follow the model, the conditional densities will be given according to p_t.

The importance of making good density forecasts can be seen very clearly in the environment of decision theory. Assume that a density forecast *p(y)* of a random variable Y with density *f* is given. The forecast user is assumed to have a *loss function L(a,y)*, where *a* represents an action choice out of some feasible action set *A*. The action *a* is chosen so that the *expected loss* from the forecaster's perspective is minimized. Thus the action a* = a*(p(y)) satisfies:

$$a^*\left(p(y)\right) = \min_{a \in A} \int L(a, y)\, p(y)\, dy \tag{13}$$

Given an outcome Y = y the action choice will result in a loss L(a*,y). Note now that the 'true' expected loss is given by:

$$\mathbb{E}[L(a^*, Y)] = \int L(a^*, y)\, f(y) dy \tag{14}$$

Now, clearly, a* might not (and indeed probably will not) minimize (14), hence if *p(y)* resembles *f(y)*, the user will be more likely to achieve a low 'true' expected loss compared to if *p(y)* does not resemble *f(y)*. Especially, this means that a density forecast p that coincides with f is *always preferable* with regard to the 'true' expected loss E[$L(a^*, Y)$].

4.1.2 The Probability Integral Transform (PIT)

Let $\{y_t\}$, $\{f_t\}$ and $\{p_t\}$ be defined as above. The objective now is to investigate whether or not one can reject the null hypothesis $p_t = f_t$. Equivalently, this means to testing whether or not the observed time series can be seen as a typical realization of the model yielding the density forecasts $\{p_t\}$. Furthermore, let $\{P_t\}$ be the cumulative distribution functions associated with $\{p_t\}$. This can seem like a very difficult thing to determine, but the concept of the Probability Integral Transform allows the forecaster to approach this task. Define the Probability Integral Transform (PIT) of the values $\{y_t\}$ as (Diebold et al 1998):

$$z_t = \int_{-\infty}^{y_t} p_t(u)\,du = P_t(y_t) \tag{15}$$

Let us investigate the density function q_t of z_t. We assume that $\partial P_t^{-1}(x)/\partial x$ is continuous and non-zero over the support of y_t and recall the relationship $p_t(x) = \partial P_t^{-1}(x)/\partial x$. Then z_t has support on the unit interval with density:

$$q_t(x) = \left|\frac{\partial P_t^{-1}(x)}{\partial x}\right| f_t(P_t^{-1}(x)) = \frac{f_t(P_t^{-1}(x))}{p_t(P_t^{-1}(x))} \tag{16}$$

This equality holds for the entire unit interval. Now particularly, if $p_t(x) = f_t(x)$, for all x, then q_t will be equal to 1 on the unit interval, meaning that z_t is U(0,1) distributed. In fact this result can be extended: If $p_t(x) = f_t(x)$, for all x and all t, the time series $\{z_t\}$ is IID U(0,1) distributed. More formally, the result can be summarized by the following proposition with proof from (Diebold et al 1998), first studied in (Rosenblatt 1952):

PROPOSITION: *Suppose $\{y_t\}$ $(t = 1, \ldots m)$ is generated from $\{f_t(y_t|\ \Omega_t\ )\}$ $(t = 1, \ldots m)$, where $\Omega_t = \{y_{t-1}, y_{t-2}, \ldots\}$. If a sequence of density forecasts $\{p_t(y_t)\}$ $(t = 1, \ldots m)$ coincides with $\{ft(y_t|\Omega_t\ )\}$ $(t = 1, \ldots m)$, then under the usual condition of a nonzero Jacobian with continuous partial derivatives, the sequence of probability integral transforms of $\{y_t\}$ $(t = 1, \ldots m)$ with respect to $\{p_t(y_t)\}$ $(t = 1, \ldots m)$ is IID U(0,1).*

PROOF: The joint density of $\{y_t\}$ can be decomposed as follows:

$$f(y_m,...,y_1 \,|\, \Omega_1) = f_m(y_m \,|\, \Omega_m)\, f_m(y_{m-1} \,|\, \Omega_{m-1})...\, f_1(y_1 \,|\, \Omega_1) \tag{17}$$

Now, we use the change of variables formula to compute the joint density of $\{z_t\}$:

$$\begin{aligned} q(z_1,...,z_m \,|\, \Omega_1) &= \left| \left(\frac{\partial y_i}{\partial z_j} \right)_{i,j} \right| f_m(\mathrm{P}_m^{-1}(\mathrm{z}_m) \,|\, \Omega_m)...\, f_1(\mathrm{P}_1^{-1}(\mathrm{z}_1) \,|\, \Omega_1) \\ &= \frac{\partial y_1}{\partial z_1} ... \frac{\partial y_m}{\partial z_m}\, f_m(\mathrm{P}_m^{-1}(\mathrm{z}_m) \,|\, \Omega_m)...\, f_1(\mathrm{P}_1^{-1}(\mathrm{z}_1) \,|\, \Omega_1) \end{aligned} \tag{18}$$

The last inequality is due to the Jacobian being lower triangular, which follows from the decomposition in (17). Now we may obtain the following expression for the density:

$$q(\mathrm{z}_m,...,\mathrm{z}_1 \,|\, \Omega_1) = \frac{f_m(\mathrm{P}_m^{-1}(\mathrm{z}_m) \,|\, \Omega_m)}{p_m(\mathrm{P}_m^{-1}(\mathrm{z}_m) \,|\, \Omega_m)} \frac{f_m(\mathrm{P}_{m-1}^{-1}(\mathrm{z}_{m-1}) \,|\, \Omega_{m-1})}{p_{m-1}(\mathrm{P}_{m-1}^{-1}(\mathrm{z}_{m-1}) \,|\, \Omega_{m-1})} ... \frac{f_1(\mathrm{P}_1^{-1}(\mathrm{z}_1) \,|\, \Omega_1)}{p_1(\mathrm{P}_1^{-1}(\mathrm{z}_1) \,|\, \Omega_1)} \tag{19}$$

Now, if $p_t(x) = f_t(x)$ for all x and t, $\{\mathrm{z}_\mathrm{t}\}$ is IID distributed, since the density is a product of the marginal distributions (all having the value one).

□

To put it concisely, we can try to reject the null hypothesis of correct density forecasts by simply testing whether or not the sequence is IID uniform, which to be sure is a surprisingly simple task compared to what one might expect from the problem formulation.

4.2 Testing for distribution and independence

After a given time series $\{y_t\}$ is transformed using PIT, all validation procedures in the literature consist of testing the transformed values $\{z_t\}$ for independence and distribution in some manner. The most prolific of these methods are covered below.

4.2.1 The importance of testing both distribution and independence: An illustrative example

It might not be obvious what the significance of testing both distribution and independence of the PIT values of a time series. For that reason, we will here give an example of an incorrect density forecast which yields uniformity but not independence, illustrating the importance of testing both for distribution and independence. Afterwards, we will give an example of how this could occur in practice.

Consider a time series $\{X_t \mid t=1,...,n\}$, with realized values $\{x_t \mid t=1,...,n\}$, and the empirical distribution function F_n given by:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \le x\}} \tag{1.20}$$

This distribution can easily be smoothed so that it has a corresponding density function, but coincides with F_n on the values $\{x_t \mid t = 1,...,n\}$. Let us call the smoothed function F and its derivative f. Now consider the density forecast, which we in reality can only make ex post, of $f_i = f$ for i = 1,...,n. Further assume that the time series has some dependence, for instance autocorrelation, making these density forecasts incorrect. However, the values $z_i = F(x_i)$ will take the values:

$$F(x_i) = F_n(x_i) = \frac{1}{n}\sum_{j=1}^{n} 1_{\{x_j \le x_i\}} = \frac{o(x_i)}{n} \tag{1.21}$$

Where $o(x_i)$ denotes the order statistic of x_i i.e. the number of values in the time series that are less than or equal to x_i. Hence, $\{z_i \mid t = 1,...,n\} = \{i/n \mid i = 0,...,n\}$, meaning that the PIT values will be almost perfectly evenly distributed on the unit interval. Thus, we can expect $\{z_i \mid t = 1,...,n\}$ to pass any distributional test for uniformity, but presuming the underlying time series has for instance autocorrelation, the PIT values z_i will retain autocorrelation since the time series values are all transformed by a the same strictly increasing function.

In practice and without hindsight, if we consider data generated by an Ornstein-Uhlenbeck process and make the time-invariant density forecast given by the limiting/unconditional marginal distribution of the process, we can expect to have similar results to above; uniformly distributed, but not independent, PIT values. This is due to the limit theorems provided in the last section of the mathematical background: The empirical distribution of (amongst other) most Ornstein-Uhlenbeck processes tend towards the limiting marginal distribution of the process. Hence, even if we observe an Ornstein-Uhlenbeck process and make density forecasts out-of-sample, we could still estimate the limiting distribution of the process and make the constant density forecasts of the limiting distribution, yielding PIT values that will probably look uniform, but display a very clear autocorrelation. For some more examples of incorrect density forecasts yielding uniform looking histograms, see also (Balabdaoui et al 2007). One should note that this is primarily an example concerning properties of the PIT, as many electricity spot price models are not stationary (having for instance time-dependent volatility). This example might in other words not be directly relevant to electricity spot price models.

Hence, testing both distribution *and* independence actually tests the time-dependent dynamics of the model, rather than just its very large-scale distributional properties.

4.2.2 The Likelihood Ratio Test

A common PIT-based test of a density forecast model is to test the null hypothesis $\{x_t\}$ ~ IID N(0,1) in the Likelihood Ratio framework (see for instance Berkowitz 2001 and Bao et al 2007). What is commonly done for testing methods in the likelihood ratio framework is to write $x_t = \Phi^{-1}(z_t)$, where $\Phi(x)$ is the distribution function of the standard normal distribution. Then the objective is to test the null hypothesis $\{x_t\}$ ~ IID N(0,1). The reason for this additional transformation is that there exist formulas regarding the above mentioned hypothesis in the Likelihood Ratio framework. The method suggested by (Berkowitz 2001) is to assume $\{x_t\}$ follows an AR(k) model:

$$\begin{gathered} x_t - \mu = \rho_1\left(x_{t-1} - \mu\right) + \ldots + \rho_k\left(x_{t-k} - \mu\right) + \epsilon_t \\ \epsilon_t \sim N\left(0, \sigma^2\right) \end{gathered} \tag{22}$$

Writing $\vartheta = (\mu, \sigma, \rho_1, \ldots,\rho_k)$ one then tests the null hypothesis $\vartheta = (0,1,0,...0)$ using the likelihood ratio (LR) statistic:

$$LR = -2\left(L\left(0,1,0,\ldots,0\right) - L\left(\boldsymbol{\theta}^*\right)\right) \tag{23}$$

Where $L(\vartheta)$ is the log-likelihood function of the model and ϑ^* is the parameter vector that maximizes the log-likelihood. Under the null hypothesis, LR ~ χ^2(k+2), since there are k+2 restrictions. In their paper validating and testing different GARCH-style models, (Escribano et al 2011) uses the likelihood ratio test for an AR(1) model:

$$x_t - \mu = \rho\left(x_{t-1} - \mu\right) + \epsilon_t \tag{24}$$

A clear advantage with the above approach is that the forecaster may use textbook methods to calculate the likelihood ratio (Berkovitz 2001). The above AR approach can be extended to second- and third-order autoregression, and also the normality assumption can be relaxed (see e.g. Bao et al 2007). However, (Berkovitz 2001) points out that as the number of restrictions increases, the test will increasingly start to resemble a non-parametric test, in the sense that it "will, in principle, reject the null in the presence of *any* departure from iid normality in large samples".

4.2.3 Non-Parametric tests

In their paper on evaluating inflation density forecasting, (Diebold et al 1999) discuss evaluating whether or not the series $\{z_t\}$ is indeed IID U(0,1) using the Kolmogorov-Smirnov test statistic D_n. The test statistic D_n of whether a set of samples $\{s_1, \dots s_n\}$ are generated from a distribution governed by the cdf F(x) is (Kim and Whitt):

$$D_n = \sup_x \left\{ \left| F_n(x) - F(x) \right| \right\} \tag{25}$$

Where F_n denotes the empirical distribution of $\{s_1, \dots s_n\}$. However, as (Diebold et al 1999) point out, little is known regarding the impact on D_n when the samples depart from independence. Indeed, we note that the test statistic D_n is entirely independent of the order of $\{s_1, \dots s_n\}$, and thus the samples could easily be designed, when viewed as a time series, to be dependent in time (by making the series increasing, say). This suggests that using the K-S test alone for the joint hypothesis of IID uniformity might be unsuitable with regards to independence.

(Bierbrauer et al 2007) uses the Kuiper test statistic of (Stephens 1970) in order to test their PIT values. This statistic, called V, is a variation of the K-S test, and is given by:

$$\begin{aligned} D^{+} &= \max_{i=1,..n} \left\{ i/n - F_n(i) \right\} \\ D^{-} &= \max_{i=1,..n} \left\{ F_n(i) - i/n \right\} \\ V &= D^{+} + D^{-} \end{aligned} \tag{26}$$

In this case, the K-S test statistic would be given by max(D^+, D^-) in this case and that this test is one specifically for uniformity. Note, however, that this test too is order-independent and thus probably not good for testing independence.

The Kolmogorov-Smirnov test is not very sensitive to discrepancies in the tails (Mason et al 1983) and thus the Kuiper test statistic is often considered preferable (Tygert 2010).

4.2.4 Visualization of results: Histograms and Correlograms

(Diebold et al 1998) propose, in combination with more rigorous testing, that histograms and correlograms should be inspected for easy understanding by the practitioner. The argument for these graphical methods is that given that given that the model failed a test, the model user would want to know why. For instance, if the histogram of the $\{z_t\}$ series seems to be fatter at the ends, it would mean that a lot of outcomes have been placed in the tails of the density forecast; hence the user could surmise that the proposed model is too thin-tailed. Importantly, note that under the IID U(0,1) assumption, individual confidence interval for bin heights are easily constructed, since the bin heights will be binomially distributed. Similarly, confidence intervals for the correlograms may be construed (Diebold et al 1999).

This graphical approach, according to (Diebold et al 1998), is “less formal, but more revealing”. Since these graphics take almost no effort to produce compared to actually producing the PIT transformed values, there seems to be no reason for the user not to supplement formal test with this kind of graphical methods.

Another argument for visual evaluation is that each failure of a model to pass the PIT-test must be analyzed on a case-by case basis. As an example: If extreme values of $\{z_t\}$ cluster around some time period, that would imply that the model might not be very good for risk applications since several unexpected extreme events (from the model’s perspective) could happen consecutively, rather than spread out in time. This example is relevant since this could happen if one models the electricity spot price without accounting for price spike seasonality and predicts an even distribution of price spikes in time. Similarly, if a model for electricity spot price totally lacks spike modeling, one could reasonably expect the model to fail the test for uniformity and probably exhibit more values close to 0 or 1 (in the IID U(01) context) than would be expected for an IID U(0,1) sequence.

4.3 Ranking validated density forecasts using sharpness

(Balabdaoui et al 2007) propose a simple way to compare models yielding PIT values that are validated as uniformly distributed, but are maybe not independent as discussed in section 4.2.1. In addition to testing PIT values for IID behavior, the authors also suggest ranking the different density forecasts according to *sharpness,* which intuitively means that the density forecast should be as concentrated as possible around the outcome while still being valid, or as (Balabdaoui et al 2007) put it, calibrated. One way of measuring sharpness is to evaluate the average width of the two-sided 50% and 90% confidence intervals implied by the model. According to the sharpness principle that the authors propose, the narrower these average confidence intervals are, the better.

An advantage with this approach to ranking is that it is simple to evaluate and intuitively makes sense. Another advantage is that this pragmatic and flexible approach holds for k-step-ahead forecasts; i.e. forecasts where we condition on information available k-steps before the outcome. In this case, the PIT values are at most k − 1 dependent and should still display uniformity (Balabdaoui et al 2007). However, one can see that rigorously testing this dependence may be difficult, other than by inspecting autocorrelation, and thus the sharpness principle in concert with uniformity tests and inspection of autocorrelation allows us to compare density forecasts via sharpness conditionally on that they are correctly calibrated (i.e. pass the PIT-related tests).

A disadvantage of using this approach for the practitioner is that this method does not seem to be very well explored. First of all, the equivalence between sharpness and the preferability of forecast, and the exact nature thereof, is an open question. As far as is known, only when comparing very well-calibrated models does sharpness asymptotically prove equivalent to the ideal forecaster (Balabdaoui et al 2007), although useful counter-examples for the case of less well-calibrated models are unknown. Furthermore, there does not seem to exist any agreed upon way to assess sharpness in the literature, though looking at the widths of confidence intervals seems like a logical candidate.

These issues notwithstanding, when faced with two models which one cannot distinguish between using PIT, i.e. two validated models, the alternative to compare by sharpness should be considered.

4.4 Evaluating interval forecast performance of models

An even simpler, but still useful, way of validating statistical models is to evaluate their performance in terms of *interval forecasts*, i.e. out of sample one-step-ahead confidence intervals, which can be either one-sided or two-sided. (Christoffersen 1998) gives a detailed and rigorous testing framework for testing interval forecasts, with methods that are analogous to the approach regarding *density forecasts* described above. The main difference is that instead of testing the distribution, one simply tests for coverage, i.e. if the ex post price falls within the forecasted confidence interval or not.

Note first, however, that this is not model validation in the strictest sense; rather, it is a validation of the interval forecasts the model makes. In the words of the author: “The aim is to develop tests of the *forecasting methodology* being applied – regardless of what it might be – not of any hypothesized underlying true conditional distribution” (Christoffersen 1998). Thus a failure to reject a model based on its interval forecasts still does not mean the model is necessarily correctly specified. However, with regards to risk purposes, correct interval forecasts may be an integral part of the model purpose, thus a specifically designed test for this is potentially very useful.

The framework introduced by (Christoffersen 1998) is as follows (with a slight difference in notation): Given a sample path of a time series $\{ y_t \mid t = 1, ... t \}$, let $\{(U_t, L_t) \mid t = 1, ... T\}$ be a set of out of sample interval forecasts and let $\{I_t \mid t = 1, ... T\}$ be the indicator functions of the events $y_t \in (U_t, L_t)$. Note that if we have PIT values, the $\{I_t\}$ values are easily generated.

Furthermore let these interval forecasts be such that $P(y_t \in (U_t, L_t)) = p$ for some confidence level p. One could also say that the interval forecasts have coverage $1 - p$. The interval forecasts are said to be *efficient* if $E[I_t \mid \{I_{t-1} , ... I_1\}]= p$ for all $t = 1, ... T$. Note that making interval forecasts for time t conditional on $\{y_{t-1} , ... y_1\}$ will yield efficient interval forecasts, since I_t will then be independent of $\{ y_{t-1} , ... y_1\}$ and by extension independent of $\{I_{t-1} , ... I_1\}$.

What (Christoffersen 1998) concludes is that if interval forecasts are *efficient* and have *coverage p*, then the indicator functions I_t are IID Bernoulli(p) distributed variables. Now, similarly to the PIT case one can now test for distribution (or coverage) and independence. Independence in this case can be very important in the risk evaluation case: dependence between forecasts can indicate that extreme events cluster around certain periods, thus posing a more concentrated risk than is indicated by just the coverage. This is also called testing for correct conditional coverage by the author.

(Christoffersen 1998) proposes as the basic framework likelihood ratio test (again, similar to the approach of (Berkovitz 2001)), testing the hypothesis of IID Bernoulli(p) distributed variables against a Markov chain alternative as the simplest test that one can perform.

4.5 Motivation for the use of PIT-based validation methods

In this section, further and more detailed arguments for and discussion of PIT-based validation techniques will be laid out. Furthermore, some mathematical framework is provided that show the potential error in comparing order-invariant statistics of data with those of models, such as for instance the sample moment comparison made by several authors. These limit theorem results are however only relevant for stationary processes and given that many spot price models have for instance time-varying volatility, these limit theorems should be seen more as illustrative examples of where one could go wrong in practice when using order-invariant statistics. This also provides a useful general example of a constant density forecast yielding uniform, though not independent, PIT values as was discussed in section 4.2.1.

4.5.1 Limit theorems for some classes of AR(1) processes

We will first present results regarding the empirical distribution function of stationary processes with weak dependence. Put simply, the results say that if we look at the empirical distribution function of a stationary and weakly dependent time series, in the same way as we would with IID variables, the empirical distribution would tend to the limiting or unconditional marginal distribution of the elements in the time series. This in turn implies that if we look at things like sample moments and other order-independent statistics of a process with these qualities, all we really do is measure statistics of the limiting distribution of the process, rather than the time series itself in some sense. This result is hopefully useful for practitioners as information of what looking at the entire empirical distribution of a time-series really means.

To start with, we define the necessary weak dependence conditions and show that all ARMA processes under certain conditions are weakly dependent in that way (Marquardt et al 2007):

DEFINITION (strong (α-) mixing*)*: *Consider a set of random variables $\{X(t) \mid t \in\mathbb{Z}\}$ on a probability space (Ω, ,P). Define $\mathcal{F}_n{}^m = \sigma\{X_t \mid n \leq t \leq m\}$ be the sigma field generated by the random variables $\{X_n, \dots X_m\}$. Define*

$$\begin{gathered}\alpha(m) = \sup |\,\mathrm{P}(\mathrm{E}\cap \mathrm{F}) - \mathrm{P}(\mathrm{E})\,\mathrm{P}(\mathrm{F})\,| \\ n \in \mathbb{Z}, E \in \mathcal{F}_{-\infty}^{n}, F \in \mathcal{F}_{n+m}^{\infty}\end{gathered} \tag{27}$$

$\{X(t)\}$is said to be strongly mixing or a-mixing if a(m) tends to zero as m tends to infinity.

Furthermore, (Athreya et al 1986) state that under some mild conditions, any AR(1) process is mixing:

THEOREM: *Let Y_t be an autoregressive process given by $Y_t = \rho Y_{t-1} + e_t$, $t = 1, 2, \ldots$ where $|\rho| \leq 1$ and $|e_t|$ are IID random variables independent of Y_0. Assume that*

(a) $E[\{\log|e_1|\}^+]$ is finite, and

(b) e_1 has a non-trivial absolutely continuous component.

Then, for any initial distribution Λ of Y_0, $\{Y_n\}$ is strong mixing.

Under the further conditions of the characteristic equation having roots less than one in modulus and Y_0 being independent of $\{e_j\}$, this result actually holds for general ARMA processes.

Mixing implies ergodicity, i.e. the sample moments of the process will converge to that of the unconditional marginal distribution (Stelzer 2011). In the case of an AR(1) process, this distribution is given by the random variable

$$\sum_{k=0}^{\infty} \rho^k \varepsilon_k \tag{28}$$

Here $\{\varepsilon_k\}$ are IID and distributed as the innovations of the AR(1) process and ρ is the autoregression parameter of the AR(1) process. Note that the unconditional marginal distribution is also the limiting distribution of the AR(1) process, as these terms can be used more or less interchangeably.

This gives us an interesting insight: When authors compare sample moments of data to the sample moments of a model, what they are actually comparing is the *limiting distribution* of the process to that of the data. However, depending on what dependence structure a models display, this limiting distribution is not unique up to a model and may in any case not be of primary importance for practitioners.

Under some more conditions, we may have an even stronger result: If the infinite sum of mixing parameters converges, the empirical distribution converges to the unconditional marginal distribution (Rio 2013).

THEOREM: *Let {X(t) | t ∈ℤ} be a strictly stationary sequence of real-valued random variables and let {a(k)} denote the sequence of strong mixing coefficients defined by (27). Suppose that the common distribution function F of the random variables is continuous. Define by* F_n *the empirical distribution function of {X(1), … X(n)} and let* $\nu_n = n^{1/2}(F_n(x) - F(x))$ *then*

$$\mathrm{E}[\sup_{x\in\mathbb{R}} |\nu_n|^2] \le \left(1+4\sum_{k=0}^{n-1}\alpha_k\right)\left(3+\frac{\log(n)}{2\log(2)}\right)^2 \tag{29}$$

Now, interestingly for us, (Marquardt et al 2007) show that a class of processes called CARMA processes, which include AR(1) processes with a driving Levy process, have geometrically declining mixing coefficients (which thus have a converging infinite sum). This means that all such processes have empirical distributions that converge to their limiting/unconditional marginal distributions as described in (29). Furthermore we note that Levy-driven mean reverting processes include most mean reverting jump processes as well as regular Brownian motion driven mean reverting processes. For a more detailed description of Levy-driven mean reverting processes, see for instance (Barndorff-Nielsen et al 2001).

Since all the time-homogeneous jump diffusion processes we consider here can be seen as simply the sums of independent Levy-driven mean reverting processes, our final result is that the empirical distributions for this class of models converge as in (29) to the limiting/unconditional marginal distributions. This means that almost any measure on a model-generated dataset which is not dependent on the *order* of the data (and hence is given entirely by the empirical distribution function) is only a measure on the limiting/unconditional marginal distribution of the model if it falls into this category. This of course includes sample moments.

Another consequence of this limit theorem is that if one calculates the unconditional marginal distribution for a model and makes the erroneous density forecast as described in section 4.2.1, we will get PIT values that look uniform but displays dependence.

4.5.2 Motivation for the use of PIT-based validation

Of what has been observed in the literature, the model testing performed by different authors roughly fall in one of three categories:

(1) They are insufficient for model validation, i.e. they measure the wrong thing and/or lack any acceptance/rejection criteria
(2) They are based on model comparison, often via likelihood and are not freestanding
(3) They closely resemble the PIT-methodology or actually involve the Probability Integral Transform

With regards to (1), a common special case that warrants some deeper discussion is the Monte Carlo estimation of sample moments and comparison to the moments of the data. Due to the results in the section above, the empirical cumulative distribution function (ecdf) converges uniformly, in a sense, to the limiting distribution of Levy-driven AR processes, and thus when we measure any statistic that depends on the ecdf alone, what we really do is compare the *limiting distribution* of our model to the distribution of the data. If the limiting distribution is all that the modeler is interested in, these measures are of course not irrelevant, but it is imperative that these types of statistics are understood as statistics of the limiting distribution of the model. However, if we are interested in the dependency structure of the time series in any way, any order-invariant statistics should be seen as of at most secondary importance. This same argument of course holds for other statistics that treats dependent time series data as IID variables, such as the inter-quartile and inter-decile ranges.

With regards to (2), (Spanos 2010) argues that model selection based on the AIC or similar measures can lead to erroneous model selection choices since the step of validating any of the models compared is ignored. As mentioned, he instead argues for model selection based on statistical adequacy. Furthermore, there does not seem to be any description on what constitutes a significant difference in AIC. The only proper hypothesis tests that falls into category (2) are the pair wise Likelihood Ratio tests performed by (Bierbrauer et al 2007). However, these are not performed as a model validation tool in the article; rather, it works as a pure model selection tool, and the selected models are themselves *validated* with PIT methodology afterwards. Furthermore, as observed in the literature overview article by (Weron 2014), it is common that authors compare their favored models against simpler models from some other class, making the comparison between the models biased towards their preferred model. This is another good motivation for a freestanding validation scheme such as the one given by the PIT - approach. Finally, if we recall the definition of statistical adequacy, the AIC does not really inform us of whether or not the data can be seen as a typical realization of the model of interest.

This brings us to category (3), of which PIT seems to be the most rigorous and model non-specific version. If we consider the methodology of (Janczura et al 2010) for instance, they identify what they call "smoothed inferences" of MRS models, which should be IID. However rigorous, this concept is of course restricted to MRS models and is difficult to generalize. Furthermore, as mentioned in section 3.2.5, rejection of smoothed inferences is difficult to interpret, unless one is intimate with the estimation process of the model. The PIT values, however, can be readily interpreted, as discussed in for instance (Diebold et al 1998). For instance, from the histogram one can see whether volatility seems to be over- or underestimated.

For practitioners, the fact that PIT is relatively easy to calculate, understand and interpret is a very important advantage of the method over many others.

The PIT-based approach, however, only requires that a model implies at time t-1 a density function at time t, which any stochastic model does. Another PIT-related approach is to investigate residuals of a process, as in for instance (Collet et al 2006). With residuals means the outcome of the process at time t minus the expected value of the process at time t-1, given by R_t in (30).

$$R_t = X_t - \mathrm{E}[\mathrm{X}_t \mid \mathrm{X}_{t-1}, \mathrm{X}_{t-2} ...] \tag{30}$$

These residuals are also commonly analyzed in the form of mean-squared error in many papers concerning point forecasts. However, the PIT values of a process can be seen as the generalized version of residuals. As a matter of fact (Yun 2014) refer to PIT values as "generalized residuals" in an article on density forecasting of jump-diffusion models. This is because for an AR(1) process with known parameters, for instance, the residuals are in fact the IID innovation terms and can thus be tested for independence and distribution much like PIT values. But because some models are more complicated than this that we introduce the notion of PIT, since this case residuals might not be IID.

As we have argued for the advantages of PIT-based validation over some other methods above, it is also worth considering the merits of this methodology in and of itself. The main merit of PIT-based validation is that this method is the most intuitive and well documented of all methods of evaluation of density forecasts in the literature. Thus, insofar as an electricity spot price model can be viewed as a density forecast, the use o PIT is well-grounded in the density forecast literature. If the sole purpose of a model is to make point forecasts, the PIT methodology might indeed be unnecessary, but as (Meyer et al 2015) note, if we are interested in for instance option pricing we are interested in the entire dispersion of the price that the model indicates. Hence the suggested methodology is not excessively complicated if the purpose of the model is at all dependent on the implied distributions in their entirety. Finally, the PIT methodology tries to answer the question, as (Spanos 2010) puts it, if the model "*accounts for the regularities in the data*" i.e. if the regularities are embedded in the model, the PIT values should be IID.

Furthermore, on a related note, the manner in which the PIT values are tested can be adapted to model usage. If, for instance, the model is primarily used for VaR purposes, we test the PIT-values for exceedences of confidence intervals in the manner of (Christoffersen 1998). Similarly, (Berkowitz 2001) elaborates of various ways to put emphasis on extreme values in the LR testing framework.

5 Methodology

In this section the set up of the tests to be performed is outlined. Overall, the tests are structured as follows: For each combination of model and dataset, the model is estimated on the first half of the dataset and out-of-sample probability integral transforms are calculated on second half. Each dataset spans four years. Following this, various tests on the PIT data are performed.

5.1 Data description

Below the markets from which the data is gathered are presented, as are the specific periods chosen from the markets. The set-up of the testing is as follows: Two four-year periods are chosen from each market. For each such period, the first two years are used to estimate the models and the two last years are used for extracting PIT values.

Since, as can be seen from the datasets, market fundamentals can change over time, it seems reasonable to try to pick datasets which display some degree of constant behavior during the selected period. Otherwise, p-values from the validation procedure will probably be quite low for all models, making ranking difficult. However, selection is done purely based on ocular inspection, so some of the pairs of estimation/validation datasets can be expected to be worse than others.

A further note on the data is that by “daily spot prices”, we refer to the daily baseload price, which is the arithmetic mean of all hourly spot prices from day-ahead auctions for each respective market.

5.1.1 Six Datasets

The data is gathered, via Vattenfall, from the European Energy Exchange (EEX), the Scandinavian/Baltic Nordpool Market and Netherlands power data from the APX NL market.

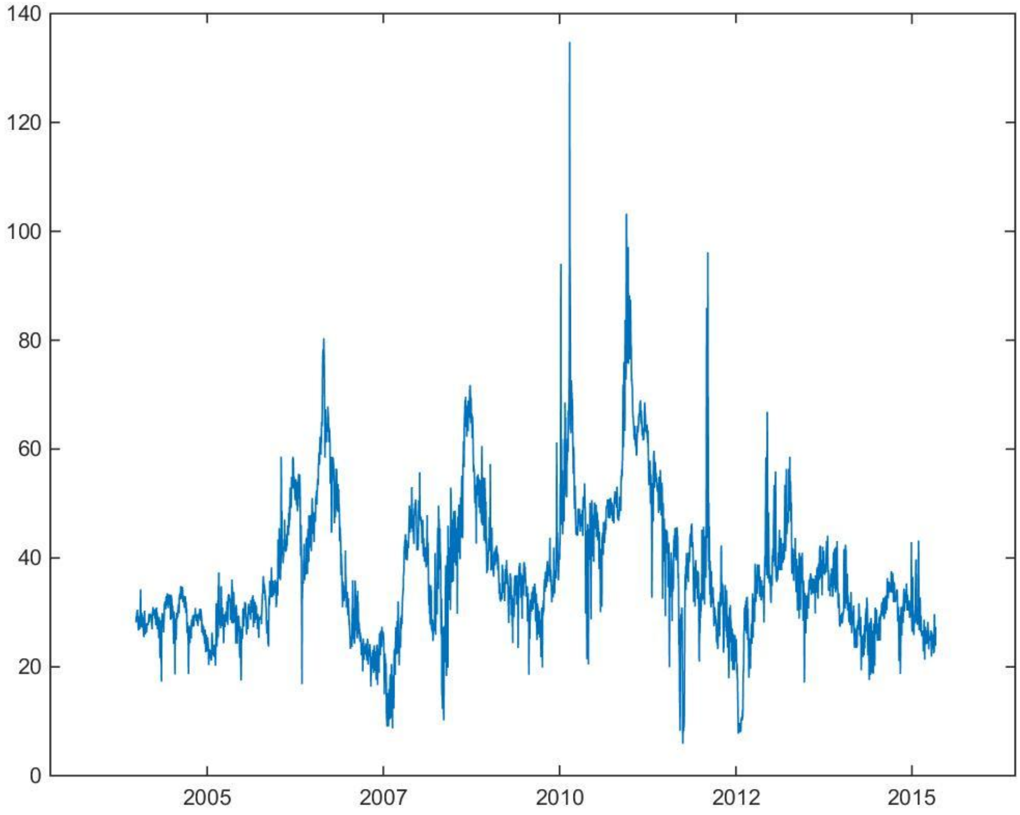

Figure 2: The daily baseload price of the System price of the Nordpool market. The prices are quoted in EUR.

From the EEX market, the periods 2005-01-01 – 2008-12-31 and 2009-06-02 – 2013-06-01 are chosen. These datasets will be denoted EEX1 and EEX2, respectively. From the Nordpool System spot prices, the periods 2006-01-01 – 2009-12-31 and 2010-01-01 – 2013-12-31 are chosen and denoted SYS1 and SYS2, respectively. From the APX NL market, the periods 2001-01-01 – 2004-12-31 and 2011-01-01 – 2014-12-31 are chosen and denoted NL1 and NL2, respectively. All prices are quoted in EUR/MWh.

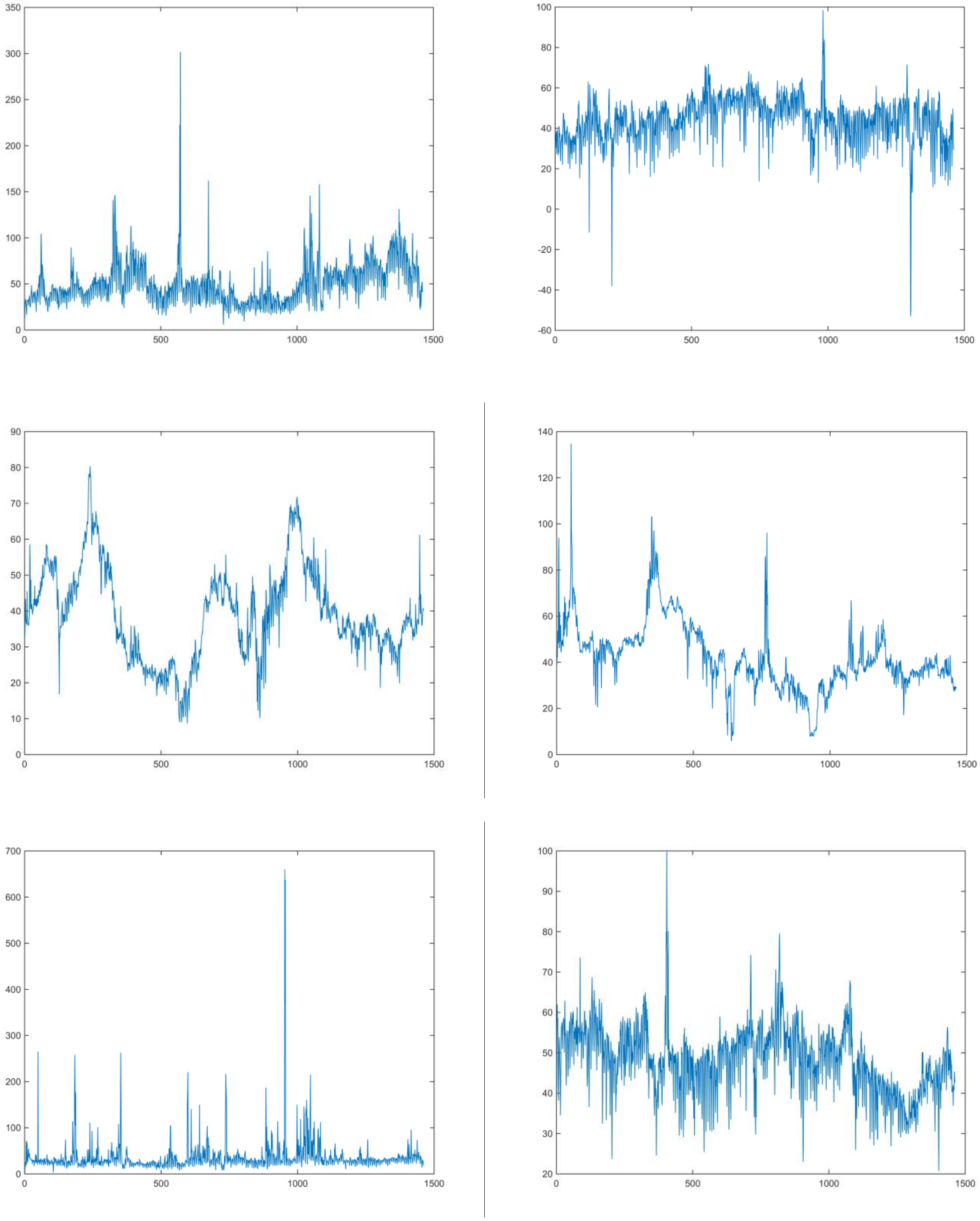

Figure 3: From top-right to bottom-left: The datasets EEX1, EEX2, SYS1, SYS2, NL1 and NL2.

5.2 Deterministic Component

5.2.1 Model assumptions

The modeling of the stochastic and deterministic component will follow that of for instance (Benth et al 2012, Bierbrauer et al 2007) and be divided geometrically:

$$\ln(P(t)) = f(\mathrm{t}) + \mathrm{X}(\mathrm{t}) \tag{31}$$

Where, as before, P(t) denotes the spot price, f(t) is the deterministic component and X(t) is the stochastic component. Since one of the datasets, EEX2, display negative prices during some downward spikes, making the geometric model problematic. This is solved using a simple affine transformation, analyzing the transformed price P'(t) = P(t)+100 for this dataset. This value is arbitrarily chosen, but shouldn't make a great difference to the analysis. It is in any cased preferred here to just removing negative spike values, as is the approach in (Meyer et al 2015).

In order to account for regularities down to daily granularity (since the data tested will be daily average spot prices) the deterministic component of (Bierbrauer et al 2007) is chosen:

$$f(t) = \alpha + \beta t + \mathbf{d}^{\mathrm{T}}\mathbf{D}_{\text{day}} + \mathbf{m}^{\mathrm{T}}\mathbf{D}_{\text{month}} + \mathbf{y}^{\mathrm{T}}\mathbf{D}_{\text{year}} + \gamma \sin\left((t+\tau)\frac{2\pi}{365}\right) \tag{32}$$

The vectors D_{day} and $\mathrm{D}_{\text{month}}$ are vectors of indicator functions for different days and months. The addition made to the trend function of (Bierbrauer et al 2007) is the introduction of a dummy variable for the year, D_{year}, in the cases of varying yearly market condition not showing any linear or cyclic occurrences.. Note though that this last addition can only be used in the case of in-sample fitting. To avoid issues with colinearity, the dummy variables do not contain the first day, month or year. The vectors d, m and y are parameter vectors. The parameters are estimated on the entire datasets.

An important note on the modeling of the seasonal component is that *f*(t) in this thesis is estimated on the entire dataset as opposed to X(t) which is estimated on a part of the data set. The reason for this is that the primary goal of these tests is to validate the modeling of X(t). Hence, the choice in this thesis is to make the conditions for the stochastic component X(t) as good as possible, so that acceptance/rejection of the modeling of X(t) should be affected by *f*(t) as little as possible. In any case, since this is an active choice (and not an obvious one) which may differ a lot from estimating *f*(t) on the same dataset as X(t), this choice should be duly noted.

5.2.2 Parameter estimation

The parameters of *f*(t) are estimated using non-linear least squares estimation. However, following the recommendation of (Bierbrauer et al 2007) the spikes are be filtered out using the "three sigma rule". This simple procedure consists of removing any log return whose absolute value exceeds three standard deviations of the log returns. The data point causing the spike is then replaced via a "similar day" method, which here denotes replacing the extreme data point the median of all data points of the same day and month. This filtering procedure is then repeated five times or until such time no more spikes are filtered out. An important technical note is that the log-returns must be recalculated after *each* data point replacement; else the log return following a spike might be classified as a spike as well.

The parameters of *f*(t) are then estimated on the filtered data, using on-linear least squares. Least squares estimation means choosing parameters of *f*(t) = *f*(t; α, β, d, m ,y, γ,τ) such that the following optimization problem is solved:

$$\min_{\alpha,\beta,\mathbf{d},\mathbf{m},\gamma,\tau} \sum_{t=0}^{T} (f(t;\alpha,\beta,\mathbf{d},\mathbf{m},\mathbf{y},\gamma,\tau) - P_{\text{filtered}}(t))^2 \tag{33}$$

Here P_{filtered} denotes the filtered spot price data and t = 0, 1 ... t is the observed days. This optimization problem can for instance be solved using MATLAB, as is done in this thesis.

5.3 Diffusion Models

As noted in the literature review, mean reverting jump-diffusion models are a popular alternative to model X(t) in equation (21). In order to test the appropriateness of validating models using PIT values as well as investigating the suitability of this model, diffusion models of increasing complexity and suitability (according to the literature) are tested. Intuitively speaking, if the proposed method of validation is appropriate, the "bad" models of X(t) should be clearly rejected, and "good" models should not be, at least at low confidence levels. The initial assumptions of which models for the electricity spot price are "bad" and "good" are derived from the literature, as well as from whether or not they possess the basic properties X(t) should have (jumps and mean-reversion).

5.3.1 Brownian Motion (Model BM)

One of the simplest models for any stochastic process is one guided by a stochastic differential equation (SDE) with constant terms:

$$\mathrm{d}X(t) = \mu \mathrm{d}t + \sigma \mathrm{d}W(t) \tag{34}$$

This model will be referred to as model BM. This corresponds to a process X(t) with IID increment s X(t+1) – X(t) ~N(μ,σ^2). Note that this stochastic process neither possesses any mean reversion quality, nor any spike-behavior.

The density forecast of model BM at time t is as follows:

$$p_t(x) = \frac{1}{\sigma}\phi\left(\frac{x-(\mu+\mathrm{X(t-1)})}{\sigma}\right) \tag{35}$$

ϕ is the density function for the standard normal distribution function. This follows directly from the normal IID increment property.

5.3.2 Ornstein- Uhlenbeck model (Model OU)

An Ornstein-Uhlenbeck model is proposed by for instance (Lucia and Schwartz 2002) and is governed by the following SDE:

$$\mathrm{d}X(t) = -\lambda X(t)\mathrm{d}t + \sigma \mathrm{d}W(t) \tag{36}$$

Here λ > 0 and is called the mean reversion rate. This model will be referred to as model OU. Note that this special case of the Ornstein-Uhlenbeck process is reverting back to zero, since *f*(t) is supposed to be the mean price level. An important note is that (36) observed at times t = 0, 1, ... is equivalent to an AR(1) model:

$$\begin{gathered} X(t) = e^{-\lambda}X(t-1) + \sigma\sqrt{\frac{1-e^{-2\lambda}}{2\lambda}}\varepsilon_t \\ X(0) = X_0 \\ \varepsilon_t \sim \text{IID } N(0,1) \end{gathered} \tag{37}$$

This can be derived from the solution of the constant parameter Ornstein-Uhlenbeck process (see appendix). Note that this is not an approximation of (36) but an exact expression which follows from the solution of (36). If we change notation slightly and let σ instead denote the volatility factor of (37), the density forecast at time t of model OU is clear:

$$p_t(x) = \frac{1}{\sigma}\phi\left(\frac{x-\exp(-\lambda)X(t-1)}{\sigma}\right) \tag{38}$$

This model has the mean-reversion property, but for instance (Collet et al 2006) found the fit lacking due to the lack of spikes in the data.

5.3.3 Simple Jump-diffusion model (Model JD)

One of the simplest conceivable models with both the mean reversion and spike property is the following jump-diffusion model (called model JD):

$$\mathrm{dX}_t = -\lambda X_t \mathrm{d}t + \sigma \mathrm{d}W_t + \mathrm{d}q_t \tag{39}$$

Here, q_t denotes a compound Poisson process, i.e. shocks arrive at exponentially distributed time intervals which are IID. Mathematically, let {N(t) | t = 1, 2 ...} be IID random variables and let Po(t) be a Poisson process. Then the compound Poisson process q_t can be described by:

$$q_t = \sum_{k=0}^{Po(t)} N(k) \tag{40}$$

The Poisson process is here assumed to have constant intensity *l*, making the discretization of (27) as follows (with the same abuse of notation as above, letting σ denote the volatility term in both cases, even though it differs):

$$\begin{cases} X(t) = e^{-\lambda} X(t-1) + \sigma\varepsilon_t & \text{with probability } 1-l \\ X(t) = e^{-\lambda} X(t-1) + \sigma\varepsilon_t + J(t) & \text{with probability } l \end{cases} \tag{41}$$

Here J(t) (t=1,2 ...) has the same distribution as the variable N(t), the discretization turns the Poission process into a Bernoulli process. For this simple model, jumps are assumed to be normally distributed J(t) ~ $N(\nu,\tau^2)$.

Density forecasts for this type of models are more easily calculated using the cumulative distribution rather than the density function. Since the PIT actually is formulated as $P_t(X(t))$, where P_t is the cumulative distribution forecast at time t of the process X(t), this is no problem. Using basic probability laws we get (denoting the probability measure conditional on observations up to time t-1 by $P^{(t)}$):

$$\begin{aligned} P_t(x) = P^{(t)}(X(t) < x) &= P^{(t)}(\text{Jump})\, P^{(t)}(X(t) < x \mid \text{Jump}) + P^{(t)}(\text{No Jump})\, P^{(t)}(X(t) < x \mid \text{No Jump}) \\ &= l\Phi\left(\frac{x - (\exp(-\lambda)\, X(t-1) + \nu)}{\sqrt{\sigma^2 + \tau^2}}\right) + (1-l)\Phi\left(\frac{x - \exp(-\lambda)\, X(t-1)}{\sigma}\right) \end{aligned} \tag{42}$$

Here “Jump” and “No Jump” denotes the events of a jump occurring or not occurring at time t. Note that it is due to the normality of J(t) that one can express the two summands in (30) so concisely. If J(t) is not normally distributed one must convolute the density functions in of the innovation variable ε_t and the jump variable J(t). In the general jump distribution case, if a jump occurs and letting *g* denote the density function of the underlying AR process density forecast, and letting *h* denote the jump density function, we get the following density function in the case of a jump (Gut 2009):

$$(g*h)(z)=\int_{-\infty}^{\infty} g(z-y)h(y)\mathrm{d}y \tag{43}$$

5.3.4 Factor model (Model 2F)

A way to increase the suitability of model JD is to give jumps time-varying intensity and give the jumps a mean-reversion coefficient separate from that of the regular market fluctuation. To do this, a two-factor model is introduced, denoted model 2F:

$$\begin{aligned} X(t) &= Y(t)+Z(t) \\ \mathrm{d}Y(t) &= -\lambda_1 Y(t)\mathrm{d}t+\sigma \mathrm{d}W(t) \\ \mathrm{d}Z(t) &= -\lambda_2 Z(t)\mathrm{d}t+\mathrm{d}q_t \end{aligned} \tag{44}$$

Unlike the setup of model JD, here $\mathrm{dq_t}$ is a compound Poisson process with not necessarily normally distributed jumps. This type of model, although with non-Gaussian driving processes, is proposed by (Benth et al 2012) and is also the "Jump diffusion"-style model in (Bierbrauer et al 2007) which performs the best. In (Bierbrauer et al 2007) though, jumps are assumed to be normally distributed. Furthermore, this class of models, although with GARCH-style volatilities, is explored by (Meyer et al 2015). Due to the large numbers of models to be tested, we will only extend model JD by introducing separate mean reversion rates and introducing different jump distributions.

With regards to jumps, we will follow the lead of (Meyer et al 2015) and introduce a mixture distribution for jumps. Given that a jump occurs, with probability l as in model JD, the jump will be positive with probability *w* and negative with probability 1-*w*. This approach differs slightly from that of (Meyer et al 2015), who technically allow for a positive and a negative jump to occur simultaneously, which is not allowed here. For the two factor model, as is done by (Meyer et al 2015), we will use a mixture of log-normal jumps. We assume the following discrete dynamics, given a jump occurrence (with probability *l*):

$$\begin{cases} \ln(q(t+1)-q(t)) \sim N(\nu^{+},\tau^{+}) & \text{w.p. } w \\ \ln(-(q(t+1)-q(t))) \sim N(\nu^{(-)},\tau^{(-)}) & \text{w.p. } (1-w) \end{cases} \tag{45}$$

Where the + and (–) indicate positive and negative jump directions. With regards for density forecasting, the only difference from model JD is that since the mean-reversion rates of Y and Z differ, these processes must be separated for each density forecast, using the same methodology as will be described below for parameter estimation. The variable X(t+1) will be forecasted as:

$$X(t+1) = e^{-\lambda_1}\hat{Y}(t) + e^{-\lambda_2}\hat{Z}(t) + \int_t^{t+1} \sigma dW(t) + dq(t) \tag{46}$$

Here $\hat{Y}, \hat{Z}$ denotes the estimates of the processes Y and Z, made at time t, since these processes cannot be directly observed. The random part of the process is discretized and forecasted just as in model JD.

5.4 Methodology for model validation

In this thesis, all tests performed on the models of choice will be based on the transformed values, i.e. PIT values that the models yield. This will be done as follows: for the first halves of the datasets, the models will be estimated. Then on the second halves of the datasets, the PIT values will be calculated. After that various tests for distribution and independence will be performed on these.

5.4.1 Visual Methods

First off, the visual methods of (Diebold et al 1998), i.e. histograms and correlograms of the PIT values are calculated and displayed. Under the IID U(0,1) assumption, individual confidence interval for bin heights are easily constructed, since the bin heights will be binomially distributed (although the bin heights of a histograms are obviously not independent). Similarly, confidence intervals around zero of the autocorrelation function the PIT values are easily construed (as is done by MATLAB's autocorrelation function, for instance). As has been noted, these visual methods do not make up a proper model validation scheme by themselves, but they are very simple to calculate and offer some graphical depiction of the results, as opposed to just displaying the values of the test statistics used.

5.4.2 Non-parametric tests

Since it is very easily done in various programming languages, the Kolmogorov-Smirnov, Kuiper and the Anderson-Darling test statistics of the normalized PIT values against their presumed normal distribution will be calculated. As previously laid out, these tests are not effective for testing independence since they are order-invariant. However, since at least the simple Likelihood Ratio (LR) tests contain some assumption of normality even in the alternate hypothesis, testing just the distribution can be seen as a complement to the LR tests. It also gives us the chance to investigate which of the two tests can be said to be “stricter” than the other.

5.4.3 AR(1) Likelihood Ratio test

As detailed in the mathematical background, the simplest LR test one can perform is to test the PIT values (transformed to hypothesized IID normal random variables via the further transformation $x_t = \Phi^{-1}(z_t)$), against an AR(1) model:

$$\begin{gathered} x_t - \mu = \rho\left(x_{t-1} - \mu\right) + \sigma\epsilon_t \\ \epsilon_t \sim N(0,1) \end{gathered} \tag{47}$$

If we by ϑ denote the parameter vector (μ,σ,ρ) and letting ϑ^* denote the maximum likelihood estimate of the parameters given by the normalized PIT values, the likelihood ratio is given by:

$$LR = -2\left(L\left(0,1,0\right) - L\left(\boldsymbol{\theta}^*\right)\right) \tag{48}$$

And the specific likelihood function, although somewhat long, is given explicitly in for instance (Berkowitz 2001). The test statistic is in this case $\chi^2(3)$ distributed under the null hypothesis $\vartheta = (0,1,0)$. Since this test assumes normality due to the assumptions in (47), this test is paired with the non-parametric tests for distribution of the PIT values.

This means that for a Type II error to occur, i.e. for a badly specified model to pass the tests, it must be one that both a) produces uniform PIT values and b) Whose normalized PIT values are fairly close to one variance, zero mean and zero first-order autocorrelation. Thus, if the case would occur that separate models both pass the proposed tests, it would be prudent to look into further dependence-related testing. For sufficient sample sizes, the Type I error probabilities are roughly given by the p-values and can be handled by testing the models for several datasets. Considering to the examples and comments in (Berkowitz 2001), two years of testing data should be enough.

5.5 Parameter estimation

Below, the parameter estimation for the different models is described as exhaustively as possible. The aim is to clarify all the modelling choices made, as well as allowing for replication. Note that in the parameter estimation description we will only be working with the discretized versions of all the models. Furthermore, we assume that the estimation dataset consists of the values $\{X(t) \mid t = 0, ..., n\}$.

5.5.1 Model BM

The parameters μ and σ of model BM are estimated by taking sample mean and sample standard deviation of the increments $\Delta X(t) = X(t+1) - X(t)$. Leading to the following estimates:

$$\begin{cases} \hat{\mu} = \frac{1}{n}\sum_{t=0}^{n-1} \Delta X(t) \\ \hat{\sigma}^2 = \frac{1}{n-1}\sum_{t=0}^{n-1} (\Delta X(t) - \hat{\mu})^2 \end{cases} \tag{49}$$

5.5.2 Model OU

Let b = exp(-λ) and recall the discretization of the zero-reverting Ornstein-Uhlenbeck process:

$$X(t+1) = bX(t) + \sigma\varepsilon_t \tag{50}$$

With this in mind, we may estimate b and σ via linear least-squares regression. Letting X_t = (X(1),...X(n)) and X_{t-1} = (X(0),...X(n-1)) we get the following estimates:

$$\begin{cases} \hat{b} = (X_{t-1}{}^T X_{t-1})^{-1} X_{t-1}{}^T X_t \\ \sigma = std(X_t - \hat{\text{b}}\, X_{t-1}) \end{cases} \tag{51}$$

Here, for brevity, std() stands for the sample standard deviation of a vector.

5.5.3 Model JD

The main difficulty of estimating the parameters of model JD versus model OU is to estimate the jump-related components. Here a very simple to implement and robust iterative method is suggested. First, we estimate the mean reversion factor b = exp(-λ) and the standard deviation σ of the innovations as for model OU. Afterwards, jumps are identified as returns exceeding three standard deviations, are filtered out and we re-estimate b and σ. This is repeated until new jumps stops being identified, or otherwise five times. The filtering procedure works as follows: If X(t) is identified as a jump it is replaced by bX(t-1). From the filtered jumps, the jump mean and standard deviation is estimated. The daily jump probability *l* is estimated simply as the number of jumps identified divided by the number of days in total.

Note that in this procedure we do not remove the jump effect on the time series as the jump value exponentially declines in time, but only the jump value itself. This differs from the estimation of the model 2F, as will be outlined below.

5.5.4 Model 2F

Since this model actually consists of two summands with different speeds of mean reversion (Y for the regular process and Z for the jumps), this model is slightly less straight-forward to estimate than model JD. The iterative procedure that is used here is inspired by (Bierbrauer et al 2007) and (Meyer et al 2015) and works very similarly to that of model JD: First, the parameters of a Model OU, with initial guess $\lambda_2 = \lambda_1$, are determined. Then, as for model JD, we identify jumps using the "three sigma rule" as above. Then λ_2 is re-estimated by fitting the realization of Z to the data via least squares. This can be done since Y and Z are assumed to be independent. After this, Y can be identified and the values of λ_1 and σ can be re-estimated. After all this is done, the jump related variables attached to the process Z can be estimated from the identified jump occurrences just as in model JD.

The exact estimation procedure used in this thesis is modeled after that of (Meyer et al 2015) and relies on the following expression for $\Delta X(t)$ (letting $b_i=\exp(-\lambda_i)$ and $a_i=1-b_i$):

$$\begin{aligned}\Delta X(t) \;=\; \Delta Y(t) \;+\; \Delta Z(t) &= -a_1 Y(t) + \sigma\Delta W(t) - a_2 Z(t) + \Delta \mathrm{q(t)} \\ &= (-a_1 Y(t) - a_2 Z(t)) + (\sigma\Delta W(t) + \Delta \mathrm{q(t)})\end{aligned} \tag{52}$$

Since the processes W(t) and q(t) both have independent increments, this allows us to treat the sum of these increments as IID random variables, given that we have separated Y(t) and Z(t) and have estimates of a_1 and a_2. The iterative procedure works as follows:

We start out by an initial estimate of the model parameter and an initial estimate of the process $(\sigma\Delta W(t)+ \Delta q(t))$. These initial values are obtained by making the initial assumption $a_1 = a_2$, and estimating a_1 from the following linear regression that arises from (52):

$$\Delta X(t) = -\mathrm{a}_1 \, X(t) + \varepsilon(t) \tag{53}$$

The initial estimate of σ is obtained by taking the sample standard deviation of the random error term in (53). Thus from this we gain initial model parameter estimates and an initial estimate of the increments $(\sigma\Delta W(t)+ \Delta q(t))$.

First $\sigma\Delta W(t)$ and $\Delta q(t)$ are separated from each other using the three sigma rule. This means that if $(\sigma\Delta W(t)+ \Delta q(t))>3\sigma$, put $\Delta q(t)= (\sigma\Delta W(t)+ \Delta q(t))$ and $\sigma\Delta W(t)=0$. After this, we re-estimate σ from the new separation of W(t) and q(t). From these new estimations of W(t) and q(t), we calculate Y(t) and Z(t). Here it is assumed that Y(0) = X(0), Z(0) = 0.

After the processes Y(t) and Z(t) are separated, we re-estimate a_1 and a_2 via least-squares from (52) and thus completes our re-estimation of all model parameters, and we may start over again.

For the tests performed in this thesis, 1000 iterations were deemed to be sufficient for reasonably stable parameter estimates.

Note that for density forecast purposes, one can simply perform these iterations without actually re-estimating parameters, as the only purpose in this case is to separate the two processes Y(t) and Z(t) in order to calculate the conditionally deterministic part in (46), i.e., the mean reversion effects from previous price movements. The PIT transformed is then calculated via Monte Carlo simulation using 10^6 simulations, as this was empirically deemed sufficient.

6 Results

In this section, the results of the parameter estimation and testing are presented. The results and their implications will then be discussed in the Discussion part of this thesis.

6.1 Parameter estimation results

6.1.1 Deterministic part

Below, the result of the parameter estimation of *f*(t), as described in (32), are presented graphically due to the large number of parameters included. For full results, see Appendix A1.

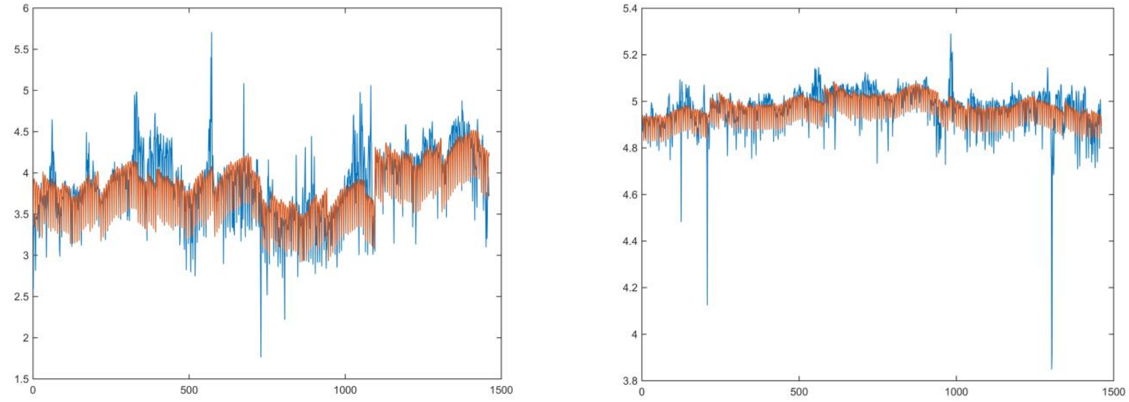

Figure 4: Trend curve (orange) and log dataset (blue) of EEX1 (left) and EEX2 (right)

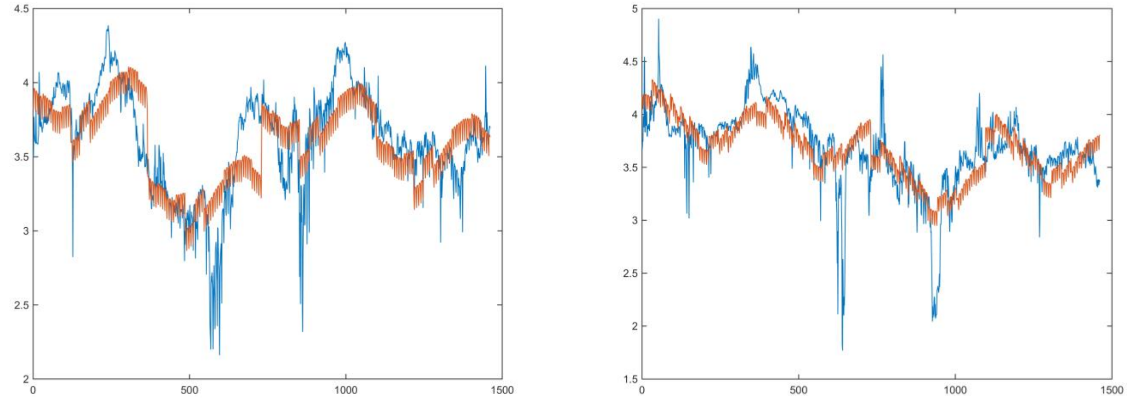

Figure 5: Trend curve (orange) and log dataset (blue) of SYS1 (left) and SYS2 (right)

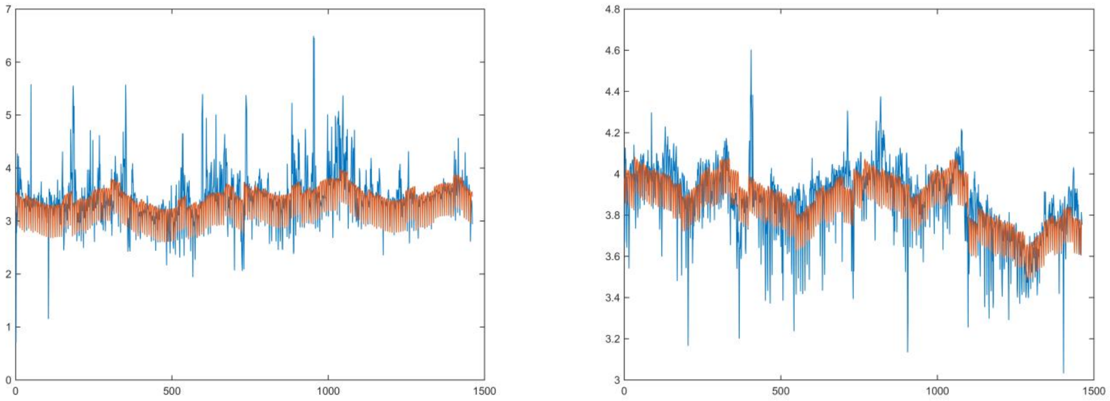

Figure 6: Trend curve (orange) and log dataset (blue) of NL1 (left) and NL2 (right)

In order to assess the effect of removing the deterministic part of the data, it is informative to compare the autocorrelation functions of the data P(t) and X(t), as defined in (31). For the datasets EEX1-2 and SYS1-2, most of the weakly seasonality seems to be captured by the procedure, but especially for the dataset NL2 we can see that there seems to be a lot of weakly seasonality left in the autocorrelation:

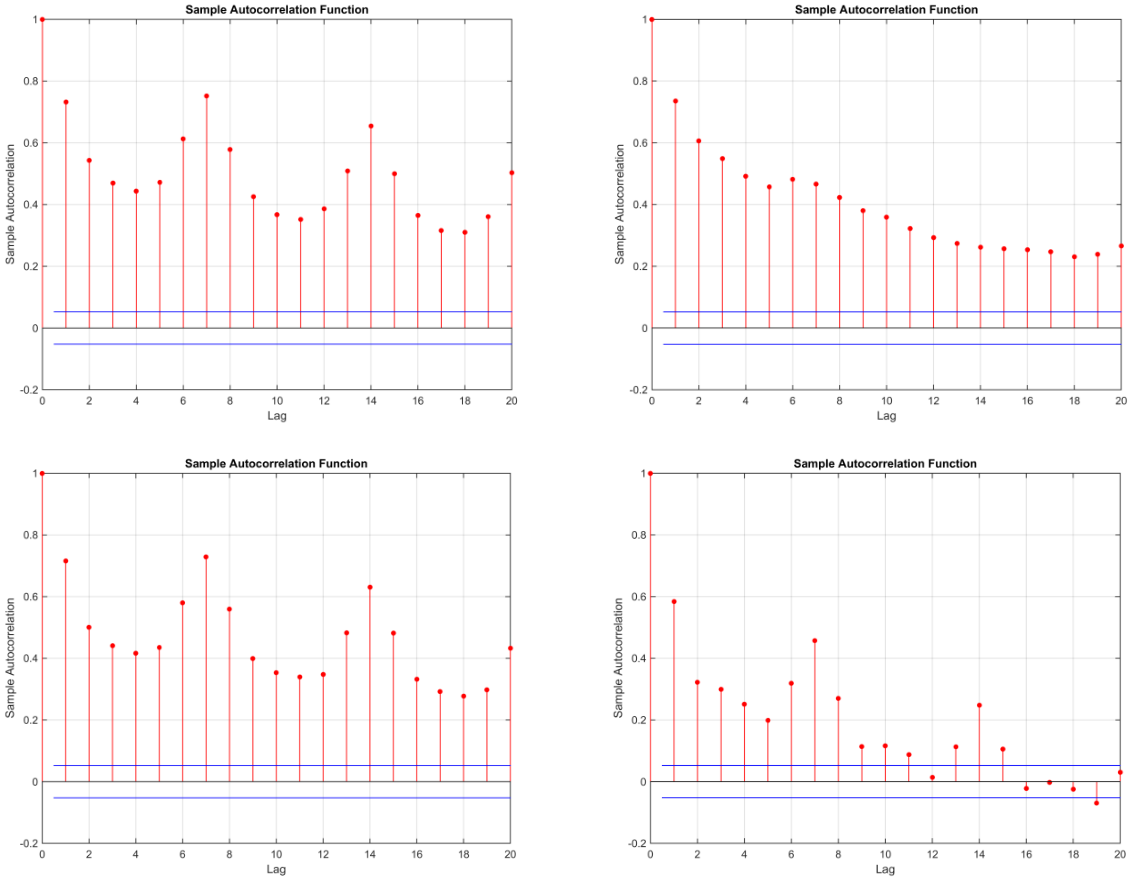

Figure 7: Autocorrelation functions, from top left to bottom right: The logarithm of the dataset EEX1, the stochastic part of EEX1, The logarithm of the dataset NL2, the stochastic part of NL2

Finally, the estimated components X(t) are presented for each of the six datasets. This is also informative for purposes of visually discerning in what ways the first half of the dataset may differ from the second. Especially note the apparent difference in daily volatility of the first and second halves of the dataset NL2.

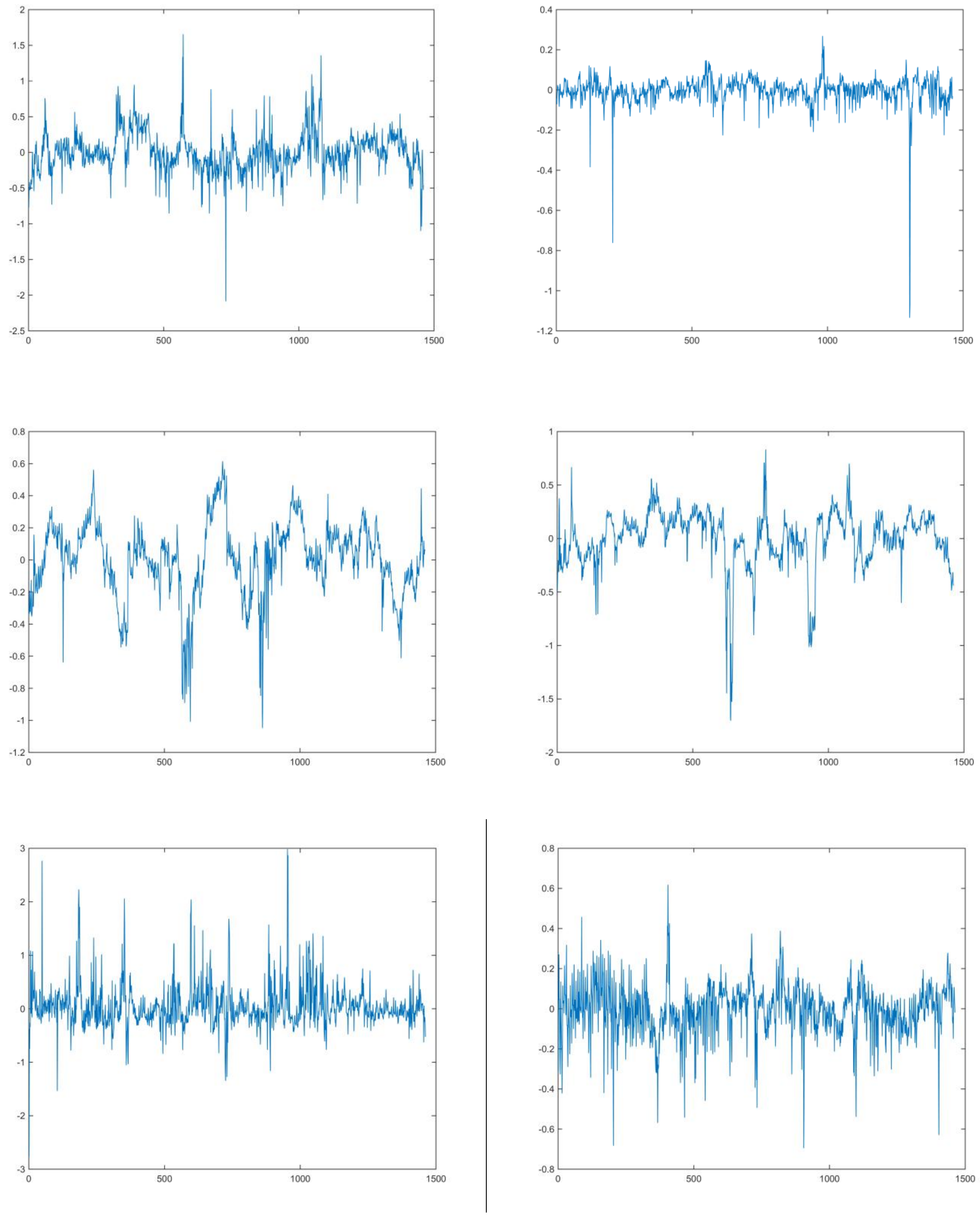

Figure 8: Stochastic part of datasets. From top left to bottom right: Estimation of X(t) from EEX1, EEX2, SYS1, SYS2, NL1 and NL2.

6.1.2 Model parameter results

In tables 1 – 3, the parameter estimation results for models BM, OU, JD and 2F are presented. Note especially the low rate of mean reversion for the datasets SYS1 and SYS2 and the high rate of jumps for model JD for these. The reason for this is that the mean-reversion effect of jumps is not taken into account in the model JD parameter estimation procedure, which conflicts with the persistent nature of the market movements of Nordpool for these periods.

	Model BM		Model OU	
	μ	σ	λ	σ
EEX1	-0.00181	0.197	0.235	0.188
EEX2	0.000115	0.0551	0.628	0.0483
SYS1	0.00102	0.0811	0.0491	0.0802
SYS2	0.000172	0.111	0.0829	0.109
NL1	0.00261	0.404	0.611	0.351
NL2	0.0000693	0.129	0.0989	0.591

Table 1: Parameter estimation results for models BM and OU

	Model JD				
	λ	σ	ν	τ	l
EEX1	0.188	0.127	0.0542	0.818	0.0548
EEX2	0.330	0.0284	-0.0424	0.187	0.0452
SYS1	0.0203	0.0468	-0.184	0.325	0.179
SYS2	0.0426	0.0558	-0.230	0.527	0.136
NL1	0.571	0.227	0.788	1.141	0.0547
NL2	0.591	0.0989	-0.102	0.459	0.0205

Table 2: Parameter estimation results for model JD

	Model 2F								
	λ_1	λ_2	σ	ν^+	$\nu^{(-)}$	τ^+	$\tau^{(-)}$	l	w
EEX1	0.166	0.364	0.129	-0.554	-0.661	0.353	0.296	0.0479	0.371
EEX2	0.283	1.281	0.028	-2.241	-1.975	0.320	0.550	0.0384	0.393
SYS1	0.0235	0.254	0.051	-1.437	-1.502	0.387	0.328	0.0521	0.447
SYS2	0.0133	0.181	0.053	-1.357	-1.303	0.388	0.430	0.0795	0.362
NL1	0.611	0.611	0.224	0.0792	-0.113	0.334	0.297	0.0547	0.675
NL2	0.481	1.277	0.097	-1.077	-1.011	0.147	0.192	0.0233	0.412

Table 3: Parameter estimation results for model 2F

6.2 Visual results of PIT transforms

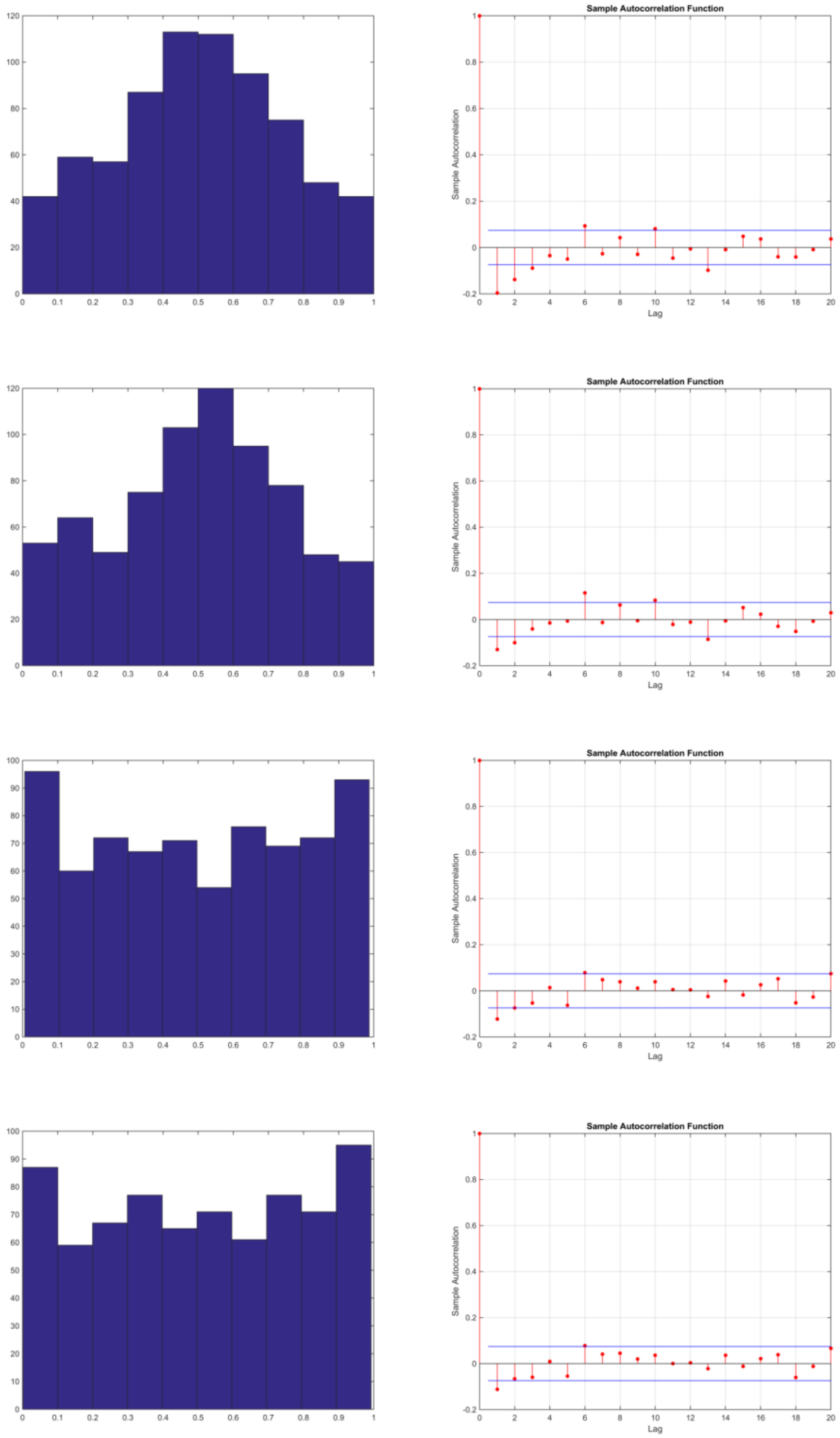

Figure 9:Histograms and autocorrelation functions for the PIT transformed values of the EEX1 dataset. From top to bottom: Model BM, OU, JD and 2F.

As an example of visual validation, above histograms and autocorrelation functions for the EEX1 dataset for the models are presented. Note the increased evenness as model complexity/suitability increases.

6.3 Test statistics

In tables 4 – 6, the validation results for models BM, OU, JD and 2F are presented, grouped by market. The tests are denoted K-S for Kolmogorov-Smirnov, Kuiper for the Kuiper test statistic and LR AR(1) for the Likelihood Ration test based on an AR(1) model as described in (47) and (48). Please note the very low p-values for the models BM and OU and the general increase in p-values as the model complexity and parameter estimation sophistication increases. Bolded figures indicate failure to reject at 5% confidence for the K-S and LR AR(1) test and 1% for the Kuiper test statistic.

	EEX1			**EEX2**		
	K-S	Kuiper	LR AR(1)	K-S	Kuiper	LR AR(1)
Model BM	0.00194	2.21E-10	5.17E-04	5.06E-06	1.04E-20	5.38E-08
Model OU	0.00366	2.54E-09	0.102	0.000177	1.19E-14	1.97E-12
Model JD	0.225	0.0479	0.0209	0.00748	3.12E-07	0.000271
Model 2F	0.337	0.0321	0.0574	0.000670	1.38E-06	0.00123

Table 4: Validation results of the models BM, OU, JD and 2F for the markets EEX1 and EEX2.

	SYS1			**SYS2**		
	K-S	Kuiper	LR AR(1)	K-S	Kuiper	LR AR(1)
Model BM	3.53E-07	6.00E-19	0.000380	2.22E-13	1.03E-45	4.41E-09
Model OU	0.00468	6.08E-07	0.000998	8.46E-08	3.69E-27	2.31E-09
Model JD	0.00115	0.00580	0.000201	0.00200	0.000429	0.000755
Model 2F	0.393	0.116	0.228	0.0513	0.0128	0.759

Table 5: Validation results of the models BM, OU, JD and 2F for the markets SYS1 and SYS2.

	NL1			**NL2**		
	K-S	Kuiper	LR AR(1)	K-S	Kuiper	LR AR(1)
Model BM	3.07E-18	1.08E-63	1.17E-07	1.48E-12	1.84E-45	6.22E-15
Model OU	1.62E-14	1.28E-53	0.000303	2.05E-10	1.04E-33	3.89E-10
Model JD	4.64E-09	3.03E-22	0.0374	3.89E-07	3.42E-21	1.30E-08
Model 2F	1.09E-07	5.75E-21	0.0127	2.29E-06	2.71E-21	1.19E-08

Table 6: Validation results of the models BM, OU, JD and 2F for the markets NL1 and NL2.

7 Discussion

Since the results are presented plainly, a more in-depth discussion of the results is in order. First off, we should note the difference in the order of magnitude of model p-values for the different datasets from the different markets. This can be seen as a consequence of what was touched on previously: That out-of-sample testing is only appropriate insofar the estimation dataset can reasonably be used to model the testing dataset. In this case, at least without the use of exogenous variables to explain yearly differences, the electricity markets visually seem to change fundamental behavior over time and thus make the effectiveness of out-of-sample testing questionable. Since the datasets were chosen purely based on ocular inspection, it is reasonable to expect that not all fundamental market behavior changes can be caught in this manner. Rather, some statistical measure of the consistency of market behavior seems in order for us to ensure p-values that are not in the order of 10^{-7}. However, these remarks mainly concern on the NL datasets, for which me must surely reject all suggested models, since it is conceivable that a more suitable model could still pass the hypothesis testing for the EEX2 dataset, for which all the models suggested here must also be rejected. Also we should particularly note that since the SYS1 and SYS2 dataset show a very slow mean reversion both with regards to spikes and regular movements, the parameter estimation procedure of model JD leads to an overestimation of jump probability and thus very low test results. This shows the importance of cleaning up jump reversion effects rigorously in the parameter estimation procedure, rather than just removing the jump itself.

An interesting note from the histograms is that due to the absence of spike modelling, the volatility of “regular” market movements is overestimated in the models BM and OU. This makes these models seem to be excessively thick tailed, which is represented by the fact that most of the actual price movements are grouped in the middle of the histogram. However, since the actual spike occurrences are deemed very unlikely in model BM and OU, this makes them simultaneously thin-tailed with regards to extreme price events.

These points aside, if we recall that the models BM, OU, JD and 2F are chosen according to increasing adequacy according to the literature, there is a clear pattern of increasing p-values as model adequacy increases. Actually, the p-values are almost across the board increasing with few exceptions. With our notion of validity as the representativeness of the data, this clearly points to the model 2F being more valid, both in the testing performed in this thesis *and* the literature. Even for data yielding very low p-values, the data is a *more likely* realization of model 2F than the rest of the models, even though it is an unlikely realization of model 2F. This reasoning actually allows comparison of models based on p-values, even if all models are rejected at some confidence level. However, model comparison based on p-values does not seem to be thoroughly grounded in the literature and caution should of course be observed when making these kinds of arguments. In any case, at least as far as this testing and these models are concerned, the PIT based validation approach does indeed point us towards the better models.

Consistent with being more focused on tails, the Kuiper test statistic consistently produces lower p-values than does the K-S test. This can be expected, since spikes tend to vary in size and are difficult to model, with several researchers trying several different distributions, for instance in (Benth et al 2012). Furthermore, since spikes are rare, we can also expect worse spike modeling due to the rather small spike-dataset that we use to estimate spike distribution parameters, to model a small number of future spikes. To put it simply, we are using something along the lines of 40 spikes to model some similar number of future spikes, thus we cannot expect the tail fit of the models to be as adequate as the fit of regular price movements. However, since spikes are an integral part of electricity price modelling, it is still appropriate to use a test that is somewhat tail focused. Due to its base in the literature, the Kuiper test statistic is thus a reasonable choice of non-parametric distributional test. Regarding the difference in "strictness" between the different tests we perform, the Kuiper test seems to be the most difficult to get a model past, followed by the LR AR(1) test of (Berkowitz 2001) and then by the K-S test. Since both the Kuiper test and the K-S test are non-parametric distributional tests, the Kuiper test, in light of its strictness in this setting, could conceivably replace the K-S test entirely, as is done in (Bierbrauer et al 2007).

A non-numeric key result of this thesis is the survey of what methods of testing and measuring electricity spot price models are commonly used in practice apart from the PIT-based ones. Since these methods are discussed at length in sections 3.2 and 4.4, a non-rigorous, very brief summary of these results is in order. First off, different authors seem to put varying amounts of effort into actually testing the models they propose. The utility as validation methods of some of the measures used is lacking as they do not possess any accept-reject criteria and/or are measuring statistics of questionable usefulness; for instance, order-independent measures, such as sample moments, for mean reverting models asymptotically measure the limit distribution of the process and does not in any case measure any dependence factors of the data such as autocorrelation. From this background, PIT-based validation emerges as the best alternative, partly due to the unsuitability of some methods for model validation, and partly because it resembles or generalizes other validation method's used by other authors, for instance the smoothed inferences of (Janczura et al 2010) and the testing of (Meyer et al 2015). Furthermore, in (Bierbrauer et al 2007) PIT values are tested for distribution, while in (Escribano et al 2011) PIT values are tested according to the LR AR(1) test as described by (Berkowitz 2001). In this thesis both tests are performed, as the LR AR(1) test presumes some degree of normality. A final argument for the PIT transform is that insofar we consider electricity spot price models as density forecasts, the PIT approach is the most developed testing method of density forecast testing in the literature. Put simply, it is the most commonly used method to test a set of non-IID random variables on a set of presumed outcomes of these.

There are a lot of opportunities for future research in this area, as this thesis merely scratches the surface of potential model complexity and possible testing procedures. Below, four key points that would be interesting future research are presented:

- In future model development papers, the PIT framework presented here should be used as a model validation scheme, so that different articles may be compared more readily. Especially, it would be interesting to see even more complicated models tested than those that the scope of this thesis permitted. For instance, no Markov Regime Switching models were tested here, but since these form a sizeable part of the spot price literature, research in which these are tested using PIT would be of great value.

- So far, we have only observed PIT in the setting where parameter estimation is made in an out-of-sample fashion. PIT used in an in-sample or cross-validation scheme could perhaps be a useful tool to measure goodness-of-fit in a validation sense. In any case, it is not an explored area.

- The impact of the choice of deterministic component, as well as an out-of-sample setting for the deterministic component has not been heavily researched as far as the reviewed literature is concerned. Here the PIT-validation approach could be used to compare models with the *same* modelling of the stochastic component X(t), but *different* model approaches of the deterministic component *f*(t).

- If a number of models are can be fairly consistently validated using the PIT, a comparison of density forecasts using sharpness, as is briefly described in section 4.3 could be in order. Also, research of other methods to distinguish between the validated models' PIT values would then be of importance.

8 Conclusion

In this thesis, the need for a comprehensive and general model validation scheme has been established, both due to the lack of such a scheme in the literature surveyed and the request of practitioners at Vattenfall's Models and Methodology unit. Such a validation scheme, based on the Probability Integral Transform (PIT) is suggested and motivated. Finally some models, for which we can be fairly sure of the order of adequacy based on previous research, are evaluated using the PIT approach. It is found that the more adequate models are indeed consistently ranked higher by the proposed tests, although the model seems not be valid for some of the datasets, depending on confidence level. Finally, it is firmly recommended that researchers and practitioners alike to use this suggested validation scheme in their evaluation of electricity spot price models, due to the mathematical soundness of doing so and so that models and articles can, in the future, be more readily compared.

9 References

Athreya, K. B. & Pantula, S. G. (1986). *A note on strong mixing of ARMA processes.* Statistics & Probability Letters 4, 187-190.

Balabdaoui, F., Gneiting, T., & Raftery, A. E. (2007). *Probabilistic forecasts, calibration and sharpness.* J. R. Statist. Soc. B 69, Part 2, 243–268.

Bao, Y., Lee, T.-H., & Saltoglu, B. (2007). *Comparing density forecast models.* Journal of Forecasting, 26, 203–225.

Benth, F. E., Kiesel, R., & Nazarova, A. (2012). *A critical empirical study of three electricity spot price models.* Energy Economics, 34(5), 1589–1616.

Benth, F. E., Benth, J. S., & Koekebakker, S. (2008). *Stochastic modeling of electricity and related markets.* Singapore: World Scientific.

Benth, F. E., Kallsen, J., & Meyer-Brandis, T. (2007). *A non-Gaussian Ornstein–Uhlenbeck process for electricity spot price modeling and derivatives pricing.* Applied Mathematical Finance, 14(2), 153–169.

Berkowitz, J., Christoffersen, P. & Pelletier, D. (2009). *Evaluating Value-at-Risk Models with Desk-Level Data.* CREATES Research Paper 2009-35.

Berkowitz, J. (2001). *Testing density forecasts with applications to risk management.* Journal of Business and Economic Statistics, 19, 465–474.

Bierbrauer, M., Menn, C., Rachev, S. T., & Trück, S. (2007). *Spot and derivative pricing in the EEX power market.* Journal of Banking and Finance, 31, 3462–3485.

Cartea, A., & Figueroa, M. (2005). *Pricing in electricity markets: a mean reverting jump diffusion model with seasonality.* Applied Mathematical Finance, 12(4), 313–335.

Cartea, A., Figueroa, M., & Geman, H. (2009). *Modelling electricity prices with forward looking capacity constraints.* Applied Mathematical Finance, 16(2), 103–122.

Christoffersen, P. (1998). *Evaluating interval forecasts.* International Economic Review, 39(4), 841–862.

Collet, J., Duwig, V., Oudjane, N. (2006) *Some non-Gaussian models for electricity spot prices.* 9th International Conference on Probabilistic Methods Applied to Power Systems. KTH, Stockholm, Sweden - June 11-15, 2006.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). *Evaluating density forecasts with applications to financial risk management.* International Economic Review, 39, 863–883.

Escribano, A., Pena, J. I., & Villaplana, P. (2011). *Modelling electricity prices: International evidence.* Oxford Bullentin of Economics and Statistics, 73(5), 622–650.

Geman, H., & Roncoroni, A. (2006). *Understanding the fine structure of electricity prices.* Journal of Business, 79, 1225–1261.

Gut, A. (2009). *An Intermediate Course in Probability, Second Edition.* Springer Science + Business Media.

Higgs, H., & Worthington, A. (2008). *Stochastic price modeling of high volatility, mean-reverting, spike-prone commodities: the Australian wholesale spot electricity market.* Energy Economics, 30, 3172–3185.

Janczura, J., & Weron, R. (2010). *An empirical comparison of alternate regime-switching models for electricity spot prices.* Energy Economics, 32, 1059–1073.

Mason, David M. & Schuenemeyer, John H. (1983). *A modified Kolmogorov-Smirnov test sensitive to tail aleternatives.* The Annals of Statistics, 11(3), 933–946.

Marquardt, T. & Stelzer, R. (2007). *Multivariate CARMA processes.* Stochastic Processes and their Applications 117, 96–120.

Mayer K., Schmid, T. & Weber, F. (2015). *Modeling electricity spot prices: combining mean reversion, spikes, and stochastic volatility*, The European Journal of Finance, 21(4), 292–31.

Rambharat, B. R., Brockwell, A. E., & Seppi, D. J. (2005). *A threshold autoregressive model for wholesale electricity prices.* Journal of the Royal Statistical Society, Series C, 54(2), 287–300.

Rio, E. *Inequalities and limit theorems for weakly dependent sequences.* 3ème cycle. 2013, pp.170.
(HAL Id: cel-00867106)

Rosenblatt, M. 1952. *Remarks on a Multivariate Transformation.* Annals of Mathematical Statistics 23 470–472.

Sapio, S. (2012). *Modeling the distribution of day-ahead electricity returns: a comparison.* Quantitative Finance, 12(12), 1935–1949.

Schwartz, E. S. (1997). *The Stochastic Behavior of Commodity Prices: Implications for Valuation and Hedging*. The Journal of Finance, 52(3), 923–973.

Spanos, A. (2010). *Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification.* Journal of Econometrics 158, 204–220.

Stelzer, R. J. (2011). *CARMA processes driven by non-Gaussian noise.* TUM-IAS Primary Sources –Essays in Technology and Science, 1(1).
arXiv:1201.0155 [math.PR]

Stephens, M. A. (1970). *Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics Without Extensive Tables.* Journal of the Royal Statistical Society. Series B (Methodological), 32(1), 115–122.

Tay, A. S., & Wallis, K. F. (2000). *Density forecasting: a survey.* Journal of Forecasting, 19, 235–254.

Tygert, M. (2010). *Statistical tests for whether a given set of independent, identically distributed draws comes from a specified probability density.* PNAS, 107(38) , 16471–16476.

Wallis, K. F. (2003). *Chi-squared tests of interval and density forecasts, and the Bank of England fan charts*. International Journal of Forecasting, 19, 165–175.

Weron, R. (2014). *Electricity price forecasting: A review of the state-of-the-art with a look into the future*. International Journal of Forecasting 30, 1030–1081.

Weron, R. (2006). *Modeling and forecasting electricity loads and prices: a statistical approach.* Chichester: Wiley.

Weron, R., Bierbrauer, M., & Trück, S. (2004). *Modeling electricity prices: jump diffusion and regime switching*. Physica A, 336, 39–48.

Yun, J. (2014). *Out-of-sample density forecasts with affine jump diffusion models.* Journal of Banking & Finance 47, 74–87.

10 Appendix A1: Tables

Below the full parameter estimation results for the deterministic function, defined as in(32).

EEX1

	alpha	*beta*	*gamma*	*to*
	3.144	0.000579	0.3849	146.260

	2006	*2007*	*2008*
y	-0.146	-0.646	-0.259

	feb	*mar*	*apr*	*may*	*jun*	*jul*	*aug*	*sep*	*oct*	*nov*	*dec*
m	0.2468	0.3261	0.4274	0.308	0.393	0.257	-0.097	-0.097	-0.164	-0.142	-0.288

	Sun	*Mon*	*Tue*	*Wed*	*Thu*	*Fri*
d	0.48981	0.57281	0.5784	0.5438	0.490	0.2713

Table 7: Parameter estimation results for the deterministic part of EEX1

EEX2

	a	β	γ	τ
	4.861	-0.000366	0.0587	271.244

	2010	*2011*	*2012*	*2013*
y	0.183	0.361	0.440	0.551

	feb	*mar*	*apr*	*may*	*jun*	*jul*	*aug*	*sep*	*oct*	*nov*	*dec*
m	0.0532	0.0647	0.0987	0.121	0.136	0.135	0.121	0.126	0.126	0.108	0.105

	Fri	*Sat*	*Sun*	*Mon*	*Tue*	*Wed*
d	0.00331	-0.00244	-0.0101	-0.0588	-0.105	-0.0158

Table 8: Parameter estimation results for the deterministic part of EEX2

SYS1

	a	*β*	*γ*	*τ*
	3.788	0.000488	-0.176	353.812

	2007	*2008*	*2009*
y	-0.773	-0.462	-0.850

	feb	*mar*	*apr*	*may*	*jun*	*jul*	*aug*	*sep*	*oct*	*nov*	*dec*
m	0.0356	-0.00662	0.0299	-0.228	-0.098	-0.252	-0.243	-0.192	-0.181	-0.144	-0.188

	Mon	*Tue*	*Wed*	*Thu*	*Fri*	*Sat*
d	0.14304	0.14746	0.1488	0.1400	0.118	0.0509

Table 9: Parameter estimation results for the deterministic part of SYS1

SYS2

	a	*β*	*γ*	*τ*
	3.735	-0.001200	0.4499	73.442

	2011	*2012*	*2013*
y	0.278	0.301	1.002

	feb	*mar*	*apr*	*may*	*jun*	*jul*	*aug*	*sep*	*oct*	*nov*	*dec*
m	0.1773	0.2321	0.4204	0.579	0.713	0.601	0.601	0.610	0.472	0.429	0.370

	Sat	*Sun*	*Mon*	*Tue*	*Wed*	*Thu*
d	-0.08140	-0.11665	0.0178	0.0208	0.022	0.0139

Table 10: Parameter estimation results for the deterministic part of SYS2

NL1

	a	β	γ	τ
	3.359	-0.000692	0.2941	159.860

	2002	*2003*	*2004*
y	0.171	0.691	0.851

	feb	*mar*	*apr*	*may*	*jun*	*jul*	*aug*	*sep*	*oct*	*nov*	*dec*
m	0.1437	0.2094	0.2185	0.233	0.367	0.117	0.029	0.093	0.083	0.267	0.149

	Tue	*Wed*	*Thu*	*Fri*	*Sat*	*Sun*
d	0.07738	0.07172	0.0938	0.0107	-0.250	-0.5257

Table 11: Parameter estimation results for the deterministic part of NL1

NL2

	a	β	γ	τ
	3.900	-0.000220	0.085	115.783

	2012	*2013*	*2014*
y	-0.008	0.158	0.006

	feb	*mar*	*apr*	*may*	*jun*	*jul*	*aug*	*sep*	*oct*	*nov*	*dec*
m	0.0904	0.1076	0.1369	0.154	0.122	0.053	0.067	0.108	0.076	0.111	0.059

	Sun	*Mon*	*Tue*	*Wed*	*Thu*	*Fri*
d	-0.06950	-0.11783	0.0347	0.0497	0.054	0.0339

Table 12: Parameter estimation results for the deterministic part of NL2