LAPPEENRANTA UNIVERSITY OF TECHNOLOGY
Faculty of Technology
Department of Mathematics and Physics

DYNAMICS OF OIL AND ELECTRICITY SPOT PRICES IN ENSEMBLE STOCHASTIC MODELS

Sihoja Hamis

The topic of this Master's thesis was approved by the faculty council of the Faculty of Technology on April 24, 2012

The examiners of the thesis was: Prof. Matti Heillio
The thesis was supervised by: Prof. Tuomo Kauranne

Lappeenranta, April 24, 2012

Sihoja Hamis
Teknologiapuistonkatu 4 A 2
53850, Lappeenranta, Finland
+358465541251
Sihoja.Hamis@lut.fi

Abstract

Lappeenranta University of Technology
Department of Mathematics and Physics
Technomathematics

Sihoja Hamis

Dynamics of oil and electricity spot prices in ensemble stochastic models

Thesis for the Degree of Master of Science in Technology
2012
59 pages, 24 figures, 5 tables
Examiners: Prof. Matti Heilio.

Keywords: Ensemble, Stochastic models, Burgers' Equation, Least Squares, Linear Kalman Filter, Variational Ensemble Kalman Filter.

In markets, traders focus on gaining high profits, while consumers aim at buying at low prices, high quantities with high quality. Also, investors seek risk-free investments.

Volatility of prices is one of the major sources of risks in commodity markets. Stochastic models of various forms are commonly used for modelling volatility of commodity prices. Usually, modifications of the models are done with the aim of improving price simulations, which can later be used for other purposes such as forecasting and risk minimization. As many other alterations in stochastic models, ensemble modelling is a recent approach that has been employed in mean reversion models.

In this work, the dynamics of oil and electricity spot prices have been studied by the use of different stochastic models which involve ensemble modelling technique. The stochastic models used are ensemble coupled mean reversion, Jabłońska-Capasso-Bianchi-Morale and Kalman Dynamics models. The Least Squares and Maximum Likelihood methods have been used in data fitting in all models. The results show that among the stochastic models used in this research, simulated prices from all of them have closely reproduced the pure (original) prices. Also, Jabłońska-Capasso-Bianchi-Morale and Kalman Dynamics models produce similar results when compared to each other. Much better results are produced by electricity spot prices in all ensemble models than oil spot prices.

Acknowledgements

I would like to express my sincere thanks to LUT for giving me study opportunity and the Department of Mathematics for the scholarship. Surely, studies couldn't be easy without financial support.

Special thanks to my supervisor, Prof. **Tuomo Kauranne** for his patience, support, encouragement, closeness, constant guidance and supervision throughout my research work. **Matylda Jabłońska** and **Anna Shcherbacheva** for their insightful review and technical support. It was very great for me to have this team around for academic and social life in Finland, thank you very much and have blessed life.

I also thank my **husband and our lovely sons** for their patience all the time when I was away from home. They have showed me enough support accompanied with encouraging words, thank you my lovely family.

Special thanks go to my parents for their prayers and unconditional support at each step of my life and my utmost thanks to Prof. **Tuomo Kauranne (Arbonaut Company)** for his parental love. No way I can pay back your kindness, just receive glory and blessings from almighty God.

Sihoja Hamis

Contents

List of Tables

List of Figures

ABBREVIATIONS

3D-VAR	Three dimensional variational data assimilation
ACF	Autocorrelation Function
ARIMA	Autoregressive Integrated Moving Average
BFGS	Broyden, Fletcher Goldfarb and Shanno quasi-Newton Method
CA	California
cdf-emp	cumulative distribution function-empirical
EKF	Extended Kalman Filtering
EnKF	Ensemle Kalman Filtering
EU	European Union
EWURA	Energy and Water Utilities Regulatory Authority
GARCH	Generalized Autoregressive Conditional Heteroscedastic
Invcdf	Inverse cumulative distribution function
JCBM	Jabłońska-Capasso-Bianchi-Morale
LBFGS	Limited memory BFGS
MA	Moving Average
MAPE	Mean Absolute Percentage Error
OK	Oklahoma City
OPEC	Organization of Petroleum Exporting Countries
PACF	Partial Autocorrelation Function
VAR	Value-At-Risk
VEnKF	Variational Ensemble Kalman Filtering
vol	Volume

1 Introduction

1.1 General introduction

Stochastic processes are widely used in modelling in different research fields. Due to their big scope of applications, more and more new formulations of stochastic processes are being proposed. It has been a challenge to researchers for years to model real life data with the use of stochastic processes in an accurate way and get promising results for further analysis.

One of the research areas, where stochastic processes are widely used, is financial and commodity markets. Energy markets form one type of commodity markets that specifically deal with trade and supply of energy. This can be either electricity energy, crude oil or any other form of energy. Energy markets form an interesting laboratory for mathematical modelling as, for instance, electricity prices are known to be the most volatile of all markets. The main reason behind the high volatility is deregulation of many markets. This way of avoiding monopoly, when properly implemented, assures an overall decrease in price levels, but allows sudden price changes, including price spikes. Special bodies have been established for the purpose of controlling deregulated markets. Examples of these bodies are Australia Energy Market Commission in Australia, Energy Market Authority in Singapore, Energy Commission in Europe and EWURA in Tanzania. They seek for optimal policies to reduce the risks, origins and resulting consequences of sudden price spikes.

Energy prices are generally regarded as mean-reverting or anti-persistent prices [34]. This type of data is commonly modeled by the use of stochastic approaches, with the aim of risk assessment. In this work energy prices, specifically oil and electricity spot prices, have been studied using three differ-

ent stochastic models originated from a classical Ornstein-Uhlenbeck model. Ornstein-Uhlenbeck model was formally developed in 1930 by the two researchers that go with the model name [31]. This model has two parts, deterministic and stochastic. The deterministic part of the process is expressed as

$$dX_t = \alpha(X^* - X_t)dt$$

where α is the mean reversion rate which is always positive and X^* is the value around which X_t tends to oscillate. When $X_t > X^*$, the drift term is negative and, as a result, movement tends to pull the system down to the equilibrium (mean reversion level). When the drift term is positive, the movement is upwards. The greater α is, the faster the process is being pulled towards the equilibrium.

For a stochastic process, a stochastic term is added to the deterministic model. This term assures randomness in the process movements around the equilibrium. The Ornstein-Uhlenbeck process is a solution of the following stochastic differential equation

$$dX_t = \alpha(X^* - X_t)dt + \sigma X_t dW_t$$

where σ stands for the volatility, dW_t is Brownian motion or standard Wiener process, $W_t \sim N(0, \sqrt{t})$. This model is the basic approach which is often modified for modelling various types of mean-reverting processes. Modification can be either in the deterministic or stochastic part, depending on which factor one wants to model within the process.

Different model fitting approaches have been proposed with the aim of getting as good model performance as possible and, hence, reliable simulation results for further analysis such as future predictions. Model fitting can also be termed as data assimilation in which all the available information is used in determination of the accuracy of the model [5]. The idea is to make correction of the data that contain systematic error or non-Gaussian noise using

background information [28]. Least Squares method is one of the mathematical techniques that is used in model fitting for data which are linear. It proceeds by finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets ("the residuals") of the points from the curve. It provides the best estimate within the class of linear unbiased estimators [14]. Due to the increasing demand for real time accuracy determination of position and velocity, Kalman Filter techniques have been developed from the Least Squares method for the sake of accurate assurance. Kalman Filter is an important block of real data process and quality assurance procedure for dynamic systems [22]. These data fitting techniques have been applied to the models in this work.

1.2 Background literature review

As stated above, the tendency towards improvement of simulated results started long ago when researchers worked a lot in modelling of stochastic processes. In this section, recent history of Ornstein-Uhlenbeck model modifications is presented, together with other stochastic models that have been used in modelling of energy prices.

In [29], the model for forecasting palm oil prices of Thailand was studied in three types, that is farm prices, wholesale prices and pure oil prices for a five-year period. The intention was to investigate an appropriate ARIMA model for forecasting of the three types of oil prices by considering the minimum mean absolute percentage error (MAPE). They found that for forecasting farm prices ARIMA(2,1,0) is the right model, while for forecasting wholesale prices and pure oil prices, ARIMA(1,0,1) and ARIMA(3,0,0) are to be used respectively. These are among simple and common models that are used in studying the dynamics of the commodity prices.

In [27], a discussion on how oil prices can be forecasted was done by conducting a series of exercises and comparing performance of the models that use oil futures prices and spot prices. The best model was claimed to be the one which performed best in these exercises. Four models were formulated based on oil futures prices and oil spot prices and evaluated with the use of two criteria. First criterium was to estimate the model over the whole sample (mid 1980's to 2005), calculate its forecast for horizons that varied from one to 24 months and compare the forecasts with the actual oil prices over these months. The model with the smallest average prediction error was said to be the best "in-sample" fit. The second criterium was to conduct a more realistic "out-of-sample" forecasting exercise where the models were estimated using the data up to a given month, instead of a whole sample. Then a forecast was made for the futures months. The model with the smallest forecast error was considered to be the best or to have the most forecasting power. One of the models used was the "random walk" model which predicted that oil futures prices would stay at their current level. The other one was Hotelling's model which predicted that the futures prices would be adjusted by the interest rates. The third model was the Futures model which predicted that futures prices would be identical to the present prices. Lastly, the authors formulated "a futures spot spread" model, which used the spread between the current futures prices and the spot prices to predict the movement in the futures prices of oil. In both criteria used, they discovered that the futures spot spread model was the best in both cases followed by Hotelling's model. It was concluded that oil future prices contain important information about futures price movement especially for the closest future. In particular, taking into account the relationship between current spot prices and futures prices instead of considering only the raw future prices can significantly improve forecasting accuracy.

Factors that contribute to the oil price changes in addition to the demand and

supply for crude oil were discussed in [7]. The authors expanded the model described by [11] for crude oil prices by including refinery utilization rates, a non-linear effect of OPEC capacity and conditions in the future market as explanatory variables. These factors altogether allowed the model to perform relatively well in forecasting as implied by the far month contracts on the New York Mercantile Exchange and are able to account for much of the 26 percent rise in crude oil prices between 2004 and 2006. This brings the conclusion that it is not possible to account for the rapid rise of crude oil prices between the year 2004 and summer 2006 by just the usual, fundamentally related demand and supply. Other factors have to be included, such as future oil market conditions, changes in refining sectors, refining utilization rates. However, the supply and demand factors continue to be important when accounting for the non-linear relationship between OPEC spare capacity and oil prices.

Investigation of the impact of shocks to the world crude oil prices on retail gasoline in Turkey during 1991-2007 was done in [30] by comparing with U.S real prices with the use of Structural VAR method. To tackle this problem, the authors decided to separate data series into series based on the sign of the growth rate of the world crude oil prices in order to distinguish between increase and decrease of oil prices and to capture the possible price changes in world crude oil prices. They derived accumulated impulse to reach conclusion and observed that Turkish gasoline prices respond significantly only to the world crude oil prices increase but not decrease, whereas for U.S gasoline the responses were symmetric. Moreover, the results were the same even for other oil products, like diesel, in both countries. The response of Turkish crude oil prices gives us the picture of how crude oil prices behave in several other countries. As contribution to this work, American and European oil products' spot price data are going to be studied with reference to electricity prices.

In [19], a comparison of mean-reverting Ornstein-Uhlenbeck model with ARIMA-GARCH was done using the data from Nord Pool spot market. The models were tested for their capability to capture statistical properties of the real electricity price time series. The author calibrated model parameters using real prices and came to the conclusion that neither of the two models was capable of capturing the statistical characteristics of the real price series. In terms of spikes, the mean-reverting Ornstein-Uhlenbeck model can partly capture this behavior, though its classical form spikes too low and too often.

The Ornstein-Uhlenbeck model was also considered by [18]. There, its stochastic term was modified by introducing *coloured noise* instead of white noise which is normally used. The model was then also used to simulate electricity prices. The author described different behavior of electricity spot prices such as high volatility, mean-reversion, spikes and seasonal patterns as a result of non-storability of electricity. The Maximum Likelihood methodology was used for parameter estimation of the model. With the estimated parameters, similar trends were found between the simulated price series and the real price series, but since the model was not deterministic, it is difficult to rely just on a single simulation. Instead, multiple simulations have to be done due to the fact described by [25] that using an ensemble of individual predictors performs better than a single predictor on the average. However, the simulation was not able to reconstruct spikes in the price.

In [20], identification and analysis of the most important features of electricity spot prices like non-spiky prices and prices with spikes were done. This work suggested an approach which treated separately the regular and the spiky regime of Nord Pool system price, by using separate mean-reversion rates. The mean-reversion model *with jump* was suggested, with jump probability depending on the process itself. The results show that it is possible to simulate the jumping mean reverting process with distribution close to real

data.

A modification to ensemble coupled mean-reversion model without jumps was done in [16] and used to investigate how much better it works in replicating the dynamics of real prices compared to other commonly used models. The author used this model with the other form of ensemble mean-reversion models and compared the simulations with each other. The other models were Ornstein-Uhlenbeck, power Ornstein-Uhlenbeck, Ornstein-Uhlenbeck with colored noise, and power Ornstein-Uhlenbeck with colored noise. The results showed that the ensemble coupled mean-reversion model performed best in reproducing behaviour of the real prices. The statistics of the simulated prices were very close to the real ones, although improvement of calibration of parameters could probably improve the results. Therefore, ensemble coupled mean-reversion model is one of stochastic models used in analysis of oil and electricity spot prices in this work. The other model is Jabłońska-Capasso-Bianchi-Morale model in which drift part of ensemble coupled mean-reversion has been equipped with the local interaction term. Also, Kalman Dynamics model is used, in which part of Jabłońska-Capasso-Bianchi-Morale model has been analyzed by Kalman Filtering techniques. The results of all models are then compared.

1.3 Ensemble

It is scientifically believed that several factors contribute to commodity price behaviour, but a common factor for different prices are traders. They introduce a psychological factor and it may have individual character for different commodities. This is related to the idea of "animal spirits", first introduced by Keynes in 1936 [13]. It has been widely discussed by now that many human actions are related to animal behaviour. If we consider an ant colony moving forward through various types of terrain, each single individual of

that population cannot know the shape of the whole column. But thanks to interactions with their neighbours and regulating the target density (different for each type of terrain), they remain sufficiently aggregated to form one colony, but keep repulsing to avoid overcrowding. Also, it is the joint influence of a number of leaders that drive the whole population in a specific direction. An analogical understanding can be considered in the markets. The way how traders' bids are related to their perception of the neighborhood, with the difference that the distance between them is measured by price. And it is a sufficiently big subgroup of trades bidding far from the other, that will cause the rest of the market to follow them.

When a group of traders are considered in model formulation and the mode of their prices is taken into account, this is termed as ensemble modelling. In one of the studies, an extension of Ornstein-Uhlenbeck model was proposed where one term of Burgers' equation (the momentum term) was added to the drift term of the basic Ornstein-Uhlenbeck model and formed an ensemble coupled mean-reversion model as well as Jabłońska-Capasso-Bianchi-Morale model [10].

1.4 Burgers' equation

The Burgers' equation is a one-dimensional form of the Navier-Stokes momentum equation which has no pressure term, neither any volume forces. It is widely used in applied mathematics like fluid dynamics or traffic flow [6].

$$u_t + \alpha u u_x + \beta u_{xx} = f(x,t) \tag{1}$$

The model shows interesting results in one of the recent studies in simulation of fluid pressure measurements that resemble electricity price realizations. The idea was transferred to the energy price analysis, started with electricity spot prices in [10] to oil spot prices in this work.

1.5 Definitions of some stochastic terms

(1) Stochastic process

A stochastic process is a parameterized collection of random variables $\{X_t\}_{t\in T}$ defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, assuming that the values are in R^n and $\mathcal{P}$ is a measurable function. T usually belongs to $[0, \infty)$ but sometimes can be in $[a, b]$, and $n \neq 0$. The process can also be regarded as a function of two variables $(t, \omega) \rightarrow X(t, \omega)$ where $t \in T$ and $\omega \in \Omega$ which is a natural point of view in stochastic analysis. Because it is crucial to have $X(t, \omega)$ jointly measured in (t, ω) [21], $t \longrightarrow X(\omega, t)$ represents a random variable in probability space Ω, and $X(\omega, t)$ is called a sample path or trajectory of the stochastic process. Some examples of stochastic processes are interest rates, currency exchange rates, commodity prices,etc.

(2) Stochastic differential equation

A stochastic differential equation is a differential equation that contains one or more stochastic term(s) and its solution is a stochastic process. It incorporates the noise term to model uncertainties. White noise is used and is conventionally taken to be the derivative of Wiener process or Brownian motion. The equation is basically formulated as

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dB_t \tag{2}$$

where μ is the drift, trend or damping coefficient, σ is the diffusion or volatility coefficient and B_t is Brownian motion. Stochastic differential equations (SDEs) are a natural choice to model the time evolution of dynamic systems which are subject to random fluctuations [21].

(3) Brownian motion

Brownian motion is a stochastic process defined on the probability space $(\Omega, \mathcal{F}, \rho)$: $B_{t,t\geq 0}$ which mathematically describes the position of a

particle at time t [17]. Stochastic process $B_{t,t\geq 0}$ is known as Brownian motion when the following conditions are satisfied

(i) $B_0 = 0$

(ii) For all $0 \leq t_1 < \cdots < t_n$, the increments $B_{t_2} - B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}}$ are independent random variables.

(iii) If $0 \leq s < t$, the increment $B_t - B_s$ are normally distributed with mean 0 and variance $t - s$.

(iv) $B_{t,t>0}$ has continuous trajectory or path.

1.6 Time series and their terms

A *time series* is a set of observations generated sequentially in equal intervals of time, e.g. hourly, daily, weekly, yearly, etc. It can be either discrete or continuous. The special features of time series are that the data are ordered with respect to time, and that successive observations are usually expected to be dependent [33]. The order of an observation is denoted by a subscript t. A single observation of a time series at time t is denoted by x_t, while the preceding observation is x_{t-1}, and the next observation is x_{t+1} [15]. The following are the terms associated with time series. When a time series has one or more of these features, it is considered to be nonstationary and one has to remove these features first before proceeding with analysis, i.e. classical analysis of time series needs stationary data.

- **Trend** component is defined as a long-term movement in the mean. It is a component that presents variation of low frequency in a time series, the high and medium frequency fluctuations having been filtered out.

- **Cyclical** component of a time series refers to regular or periodic fluctuations within normal movement excluding their irregular component,

revealing a succession of phases of expansion and contraction.

- **Seasonal** component refers to regular movement that goes along specific periods like days, weeks, years, etc. It is a part of the variations in time series representing fluctuations that are more or less stable period after period. Irregular movement is the time series component that remains after removing the three mentioned terms.

1.7 Distribution moments

Part of the analysis of time series data is basically done by graphical investigation of the data as well as computation of the basic statistics (distribution moments) which include mean, standard deviation, skewness and kurtosis. Here are mathematical representations of the basic statistics.

- *Mean* or expectation is the average of all observations in a data series and each value in a series is expected to be close to it. One can obtain mean value by the following equation

$$\mu = E[X] = \sum_{i=1}^{n} \frac{X_i}{N} \tag{3}$$

where N is the total number of observations.

- *Standard deviation* is the measure of variation or diversity of the observations from the mean (expected value); this can be found by taking the square root of variance (σ^2) which is given mathematically by the formula

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(X_i - E(X))^2} \tag{4}$$

for the discrete values. A low standard deviation indicates that observations tend to be very close to the mean, whereas high value of

standard deviation indicates that the data is spread over a large range of values.

- *Skewness* is the third standardized moment which is often called a measure of asymmetry. It is given by the formula;

$$skewness = \frac{E[X - E[X]]^3}{\sigma^3} \tag{5}$$

The skewness coefficient can either be negative or positive, where negative coefficient means that distributions are left skewed and positive coefficient means distributions are right skewed. The coefficient of skewness for a symmetric distribution is zero, which is the case for normal distribution.

- When skewness measures the type of departure from normality in terms of symmetry, *kurtosis* measures the extent of peakedness (or the degree of flatness near its center) in a distribution. It is the fourth standardized moment given by the ratio

$$kurtosis = \frac{E[X - E[X]]^4}{\sigma^4} \tag{6}$$

If the data is normally distributed, kurtosis equals 3. Kurtosis greater than 3 indicates that more values are in the neighborhood of the mean but in the same time the distribution has also many extreme observations. If the ratio is less than 3, then it is an indication that the probability curve is flatter than the normal one.

1.8 Series dependence

Series dependence shows how values are related to each other in time series. This can be done by the computation or drawing of autocorrelation and partial autocorrelation. Autocovariance (autocorrelation) determines how

each value X_{t_1} is related to its previous value X_{t_2} or X_i related to X_{i-1}. Autocovariance can be obtained by the formula

$$\gamma_1 = E[(X_{t_1} - \mu)(X_{t_2} - \mu)] \tag{7}$$

For stationary series, constant autocovariance between X_{t_1} and X_{t_2} with that of X_{t_k} and $X_{t_{k+1}}$ for any k are expected. Equation (7) can also be written as

$$\gamma_s = E[(X_t - E(X_t))(X_{t-s} - E(X_{t-s}))] \tag{8}$$

When $s = 0$, we obtain autocovariance at lag zero which is the autocovariance of X_t and X_t. Since the autocovariance depends on unit measurement and measures the relationship between two consecutive observations, we do not expect immediate interpretation of this value. It is more convenient to use the normalized autocovariance (autocorrelation) which is obtained by dividing Equation (8) with data series variance as

$$ACF(s) = \rho_s = \frac{\gamma_s}{\gamma_0} = \frac{E[(X_t - \mu)(X_{t+s} - \mu)]}{\gamma_0} \tag{9}$$

for $s = 1, 2, 3, \ldots$ and when $s = 0$, we get covariance which is equal to 1. If the series is of independent values, the true ratio is zero, otherwise the ratio takes the range within [-1,1]. For the series with trend, the value of γ_s will not come to zero because each value is followed by the next higher value on the mean side and in fact the meaningful ratio is that from stationary data [3].

Stationary series

The time series $\{X_t\}, t \in \mathbb{Z}$, where $\mathbb{Z}$ is the integer set, is said to be stationary if

(i) $E[X_t^2] < \infty, \forall t \in \mathbb{Z}$

(ii) $E[X_t] = \mu, \forall t \in \mathbb{Z}$

(iii) $\gamma(s,t) = \gamma(s+h, t+h), \forall t \in \mathbb{Z}$

In the other words, a stationary series $\{X_t\}$ must have three features; finite variation, constant first moment and that of second moment depend only on $(t-s)$ for series starts from s to t and not depends on s or t

1.9 Structure of the thesis

This work has been divided into four sections, starting with introduction in subsection 1.1 which describes background of the problem, review of literature that is related to this work in subsection 1.2. This is followed with model formulations and descriptions in section 2. Also, model fitting techniques are described in this part. Basic analysis of data in section 3 is followed by model fitting with analysis of the results. Discussion of results for each model is done in conclusion part section 4 which is finally followed by recommendation for future studies in subsection 4.2.

2 Model formation and description

2.1 Introduction

The stochastic models which have been used in this thesis originate in the Ornstein-Uhlenbeck model, where its drift term is modified by adding new terms. They aim at representing basic animal spirits typical for traders acting in the markets. Originally, the Ornstein-Uhlenbeck model has one mean-reversion level that stands for a long term mean of the prices. In the following models that level is moving, calculated at each time point for

specific historical horizon only. This term is related to traders' short-term thinking and it is one of the main trading biases.

2.2 Mean-reversion

Mean-reversion is a tendency of a stochastic process to remain near, or to return over time to a long-run mean level. Besides, not all processes termed as stochastic processes can exhibit mean-reversion. The processes like interest rates and implied volatilities tend to exhibit mean-reversion but currency exchange rates and stock prices tend not to. Energy spot prices and spot price returns are well known stochastic processes that exhibit mean-reversion [34]. Each stochastic process has its own rate of reverting to the global mean price level. These rates depend on the factors influencing prices of such commodity. These can be seasons, economic instabilities, weather and so forth that can cause high demand which leads to scarcity of a particular commodity, hence price increases. Low demand lowers the prices.

The Ornstein-Uhlenbeck process is the solution of the following stochastic differential equation

$$dX_t = \alpha(\mu - X_t)dt + \sigma dB_t \tag{10}$$

where

- $\alpha(\mu - X_t)$ is the drift term;
- μ – represents the equilibrium or mean value supported by fundamentals;
- α – is the rate by which the shocks dissipate and the variable reverts towards the mean;
- σ – is the degree of volatility;

- B_t – stands for Brownian motion which brings stochasticity into the system;

The drift term behavior of an Ornstein-Uhlenbeck process carries the name "mean-reverting" due to its dependency on the current price value of the process: the term can either be positive or negative depending on which side of the mean level the process is at each time instant. In other words, the mean μ acts as an equilibrium level for the process.

2.3 Ensemble coupled mean-reversion model

The drift term in equation (2) has been modified to get the new equation with two reversion levels, namely X_t^* and $f(k, \mathbf{X}_t)$ as presented in equation (11).

$$dX_t^k = \alpha_t[(X_t^* - X_t^k) + (f(k, \mathbf{X_t}) - X_t^k)]dt + \sigma_t dB_t^k \tag{11}$$

where

- $k = 1, 2, 3, \ldots, N$ is the index of a given trader in an ensemble;
- α_t is the moving mean-reversion rate at time t;
- X_t^* is the moving mean-reversion level which models the short-term thinking of the traders about the market behaviour;
- X_t^k is the trader (k) price series in the time interval t;
- $f(k, \mathbf{X}_t)$ is the momentum value of bids which models the most common bids in the market;
- σ_t is the degree of volatility;
- N stands for the maximum number of traders (ensemble size);

2.4 Model description

Considering the simulation window of six months, X_t^* is the moving mean of prices in the present interval of time t, X_t^k is the trader (k) price series in the time interval t, $f(k, \mathbf{X}_t)$ is momentum value of the traders in the previous time interval $t-1$. The new simulation window is obtained by leaving a day behind and adding one day ahead.

The term $f(k, \mathbf{X}_t)$ stands for a price level which is obtained by multiplying the difference between mean and mode of the ensemble traders' prices with the mode of the ensemble traders' prices in which both mean and mode are calculated in the time interval $t-1$, so as to produce spikes, as

$$f(k, \mathbf{X}_t) = ens\_mode \times (ens\_mean - ens\_mode)$$

ens stands for ensemble which is a group of traders.

This concept is derived from the Burgers' equation

$$u_t + \alpha u u_x + \beta u_{xx} = f(x, t)$$

which is mainly used in modeling of fluid dynamics. The terms in this model are related to dynamics of commodity prices in the spot markets as

- $u = u(x, t)$ corresponds to the mean of prices in a commodity market;
- u_x corresponds to the spread between mean and mode;
- uu_x corresponds to the momentum of traders toward higher prices;
- βu_{xx} corresponds to the diffusion term (spot market tend to reach equilibrium price)in normal market model;
- $f(x, t)$ is a fundamental of a periodic character;

The term uu_x is the one added to our basic model (11), and is considered as competition of traders in commodity markets towards higher prices.

In Equation (11), prices are expected to revert to either of the two price levels, X^* which is the global mean-reversion level at time t or bids price $f(\mathbf{k}, \mathbf{X}_t)$, depending on the forces acting on the prices and the *Brownian motion* increment. We have two types of forces to be considered, the general side and the trader side. Each traders' bid at a given timepoint depends on the forces acting from those two sides. The model in this work includes both market situation and the individual traders by considering their neighbors. The mean price is calculated after each given time interval and in this work we have considered three months. Equation (11) has been named as *ensemble coupled mean-reversion model.*

2.5 Jabłońska-Capasso-Bianchi-Morale (JCBM) model

In this model, two more trade interactions scenarios have been added to improve the model performance. In the ensemble coupled mean-reversion model only global interaction was considered, whereas in JCBM also the local interactions are included.

$$dX_t^k = 3[(\frac{\gamma_t}{3}X_t^* + \frac{\theta_t}{3}h(k, X_t) + \frac{\xi_t}{3}g(k, X_t)) - X_t^k]dt + \sigma_t dB_t \qquad (12)$$

where

- X_t^* – 6 months moving average;
- $h(k, X_t)$ – global interaction;
- $g(k, X_t)$ – local spread;
- $g(k, X_t) = max_{k \in I}\{X_t^k - X_t\}$;

- $I = \{k|_{X_t^k \in N_{p\%}^k}\}$;
- $N_{p\%}^k$ is the neighborhood of the k^{th} particle formed by $p\%$ of the particles;
- $h(k, X_t) = M(X_t)[E(X_t) - M(X_t)]$; where $M(X_t)$ is the ensemble mode and $E(X_t)$ is the ensemble mean.

The remaining notations are the same as in the previous model.

2.6 Kalman Dynamics model

The Kalman Dynamics model requires a set of preliminary detrended and deseasonalized spot prices together with the same set of data with removed spikes. The first H values from the original data are used as initial values for a set of particles $\mathbf{s}_i^k$, where $i = 1, \ldots, n$, so that

$$\mathbf{s}_i^k = original\_prices(k), \quad k = 1, \ldots, H \tag{13}$$

Spikeless series are then employed to estimate a mean-reversion level via the Least Squares.

$$\begin{pmatrix} 1 & x_{n-1} \\ 1 & x_{n-2} \\ \vdots & \vdots \\ 1 & x_{n-H} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} dx_{n-1} \\ dx_{n-2} \\ \vdots \\ dx_{n-H} \end{pmatrix} \tag{14}$$

where

$$x_k = original\_prices(k), \quad k = n - H, \ldots, n - 1;$$
$$dx_n = original\_prices(n + 1) - original\_prices(n).$$

The system (14) can be equivalently rewritten using vector-matrix representation:

$$\mathbf{X} \cdot \mathbf{p} = \mathbf{dx} \tag{15}$$

A simulation procedure includes subsequent application of evolution and filtration operators to the state vector. The evolution operator is based on the JCBM model for a group of traders in the market. The state vector represents a set of particles or bids from market participants that composes a distribution, where each price has a corresponding volume or frequency measured in [KW/h]. Each evolution step implies resampling of a state vector from the current distribution that is induced by the nonlinearity of an evolution operator.

2.7 "Kalman Dynamics" operator

The Kalman Dynamics model uses the Variational Ensemble Kalman Filter (VEnKF) by Solonen et al (2011) as,

for t=1:N_steps

1. for j=1:$n\_periods$

   (a) $k = (t-1) \cdot (n\_periods + 1) + j$

   (b) estimate a current mean-reversion level from the spikeless series:

$$\mathbf{p} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{dx};$$
$$rev\_rate = -\mathbf{p}_2;$$
$$rev\_level = \mathbf{p}_1/rev\_rate;$$

   (c) estimate the sample standard deviation σ_s;

   (d) assign values to the moving average, the global term coefficient

and the standard deviation:

$$\begin{aligned} \mathrm{MA}_H &= rev\_level; \\ \gamma' &= \gamma \cdot rev\_rate; \\ \sigma' &= \sigma \cdot \sigma_s; \end{aligned}$$

(e) apply the JCBM operator to the state vector:

$$\mathbf{x}^k = \mathbf{JCBM}(\mathbf{y}_0, model\_parameters), \tag{16}$$

where $model\_parameters = (\gamma', \theta, \xi, \mathrm{MA}_H, \sigma')$ and σ' is a standard deviation of the stochastic term;

(f) build a distribution out of the resulting state vector $\mathbf{x}^k$:

$$\begin{aligned} [\mathbf{vol}, \mathbf{price}] &= hist(\mathbf{x}^k, n\_bins); \\ Total\_volume &= sum(\mathbf{vol}); \\ \mathbf{cdf\_emp} &= cumsum(\mathbf{vol}/Total\_Volume) \\ \mathbf{cdf\_loc} &= \mathbf{price}, \end{aligned}$$

where **cdf_emp** are the CDF values and **cdf_loc** are the corresponding points where the numerical CDF values are given.

(g) sample the initial vector $\mathbf{y}_0$ for the next iteration from the distribution composed out of the resulting state vector $\mathbf{x}^k$:

$$\mathbf{y}_0 = \mathbf{invcdf}(\mathbf{cdf\_loc}, \mathbf{cdf\_emp}). \tag{17}$$

2. Perform procedures from items 1.(b),1.(c),1.(d);

3. assign a value to the prior state vector $\mathbf{x}^p$ and specify the observations $\mathbf{x}^o$:

$$\begin{aligned} \mathbf{x}^p &= \mathbf{x}^k + p \cdot \epsilon, \\ \mathbf{x}^o &= \mathbf{x}^j + q \cdot \epsilon, \quad j = t \cdot n\_periods, \end{aligned}$$

where $\epsilon \sim \mathbf{N}(\mathbf{0}, \mathbf{1})$;

4. assign the initial covariance: $\mathbf{Cest0} = cov(\mathbf{x})$;

5. generate the set of particles:

$$\mathbf{s}^j \sim \mathbf{N}(\mathbf{x}^p, \mathbf{Cest0}), \quad j = 1, \ldots, n\_particles;$$

6. specify an observation operator:

$$\mathbf{K}_i = \mathbf{exp}\left(\frac{(\mathbf{x}^k - \mathbf{x}_i^k \cdot \mathbf{I})^2}{\sigma_1^2}\right), \quad i = 1, \ldots, n\_participants; \tag{18}$$

7. apply the VEnKF to the state vector $\mathbf{x}^p$:

$$\mathbf{y}_0 = \mathbf{VEnKF}(\mathbf{x}^p, \mathbf{s}). \tag{19}$$

2.8 Model parameters

- $n\_periods$ – a serial number of evolution model propagation steps performed before applying a filtration operator;
- $\gamma, \theta, \xi, \sigma$ – evolution operator parameters;
- σ_1 – observation operator parameter;
- $n\_particles$ – number of particles generated for filtration (VEnKF parameter);
- H – assimilation window length required for computing Moving Average;
- T_r – number of steps for preliminary running up the model employed for specifying an initial state and covariance;
- N_steps – number of filtration operator steps accomplished;
- p – standard deviation of model error;

- *noise_ratio_obsmo* $= q/p$, where q is a standard deviation of observation error – standard deviations ratio;
- *n_participants* – number of the state vector components;
- *n_bins* – number of bins employed for building up a histogram out of the state vector components;

2.9 Observation operator

Consider a prior state vector $\mathbf{x}_s^p \in \mathbb{R}^n$ for a fixed s-th step. An observation operator $\mathcal{K}_s$ then composes a matrix with each i-th row represented by a symmetric Gaussian function:

$$\mathcal{K}_s(i,j) = e^{\frac{-\left(\mathbf{x}_s^p(j) - \mathbf{x}_s^p(i)\right)^2}{2\sigma^2}} \tag{20}$$

For easy fitting of the model to the data, an appropriate technique is needed. At this part, the Least Squares method as well as Kalman Filtering development is summarized.

2.10 Least Squares method

The Least Squares technique is a procedure to determine the best fit of a line to the data; it is the approach used to approximate the unknown parameter/s in a given model. This method does not solve the equations exactly. Instead, it seeks minimization of sum of the squared residuals. It usually produces the Maximum Likelihood estimate of the parameter/s [2], [12]. In brief, to apply the Least Squares fit, one should have data as $x_i, i = 1, 2, 3, \ldots, n$. These can be fitted in a linear model as

$$y = b_0 + \sum_{i=1}^{n} b_i x_i$$

To get unknown parameters $b_0, \ldots, b_n$, the system has to written in the following form

$$y = Xb + \epsilon \tag{21}$$

where

- y is the measured data;
- X is the extended matrix having ones as its first column;
- ϵ is the measured noise;

The Least Squares solutions are found under the objective function that minimizes the sum of the squared residuals as

$$L(b_0, b_1, \ldots, b_n) = \sum_{i=1}^{n} (y_i - (x_i)^{'} b)^2 = \|y - Xb\|$$

The solution of Least Squares estimate is the solution of the normal equation

$$X^{'} X b = X^{'} y$$

which can be obtained as

$$b = (X^{'} X)^{-1} X^{'} y$$

If the measurements are independent with covariance matrix $cov(b) = \delta^2 I$ then for the estimated parameters their covariance can be obtained as

$$cov(b) = \delta^2 (X^{'} X)^{-1}$$

For JCBM model, more advanced techniques of model fitting were needed due to its complexity and considerable number of parameters. The preferred

model fitting method was Variational Ensemble Kalman Filter which is a modified Least Squares method. Kalman Filter is a set of mathematical equations that provide an efficient computational (recursive) means to estimate the state of the process, in a way that minimizes the mean of the squared error [35].

2.11 Kalman Filtering technique

Getting to understand this approach one should recall the idea of Bayes formula as

$$\prod(b) \approx p(y \mid b)p(b)$$

where $p(y|b)$ is the likelihood function and $p(b)$ is the prior probably.

From the linear model (21), assume the measurement error and prior distribution are Gaussian with covariance matrices S_ϵ and S_a respectively and if the center of prior is assumed to be b_a, then

$$p(b) \approx e^{\frac{-1}{2}(b-b_a)' S_a^{-1}(b-b_a)}$$

$$p(y|b) \approx e^{\frac{-1}{2}(y-Xb)' S_\epsilon'(y-Xb)}$$

Product of these and logarithm gives

$$-2log(\prod(b)) = (b-b_a)' S_a^{-1}(b-b_a) + (y-Xb)' S_\epsilon'(y-Xb) \qquad (22)$$

This can be reduced to a simple form decomposing covariances e.g. the Cholesky decomposition, for positive definite matrices $S_\epsilon^{-1} = K_\epsilon' K_\epsilon$ and $S_a^{-1} = K_a' K_a$ This substituted into equation (22) gives

$$-2log(\prod(b)) = (K_a b - K_a b_a)'(K_a b - K_a b_a) + (K_\epsilon y - K_\epsilon X b)'(K_\epsilon y - K_\epsilon X b)$$

$$-2log(\prod(b)) = \|K_a b - K_a b_a\|^2 + \|K_\epsilon X b - K_\epsilon y\|^2 \tag{23}$$

Equation (23) is the norm of a Least Squares problem $\tilde{y} = \tilde{X}b$, whereby

$$\tilde{y} = \begin{pmatrix} K_\epsilon y \\ K_a b_a \end{pmatrix} \text{ and } \tilde{X} = \begin{pmatrix} K_\epsilon X \\ K_a \end{pmatrix}$$

with $N(0,1)$ as the covariance which leads to

$$\tilde{b} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \text{ and } cov(\tilde{b}) = (\tilde{X}'\tilde{X})^{-1}$$

as the solution of parameters and covariance when Least Squares is used. If we then write

$$\tilde{X}'\tilde{X} = \begin{pmatrix} K_\epsilon X & K_a \end{pmatrix} \begin{pmatrix} K_\epsilon X \\ K_a \end{pmatrix} = X' K_\epsilon' K_\epsilon X + K_a' K_a = X' S_\epsilon^{-1} X + S_a^{-1}$$

and

$$\tilde{X}'\tilde{y} = \begin{pmatrix} K_\epsilon X & K_a \end{pmatrix} \begin{pmatrix} K_\epsilon y \\ K_a b_a \end{pmatrix} = X' K_\epsilon' K_\epsilon y + K_\epsilon' K_a b_a = X' S_\epsilon^{-1} y + S_a^{-1} b_a$$

These give

$$\tilde{b} = (X' S_\epsilon^{-1} X + S_a^{-1})^{-1}(X' S_\epsilon^{-1} y + S_a^{-1} b_a) \tag{24}$$

$$cov(\tilde{b}) = (X' S_\epsilon^{-1} X + S_a^{-1})^{-1} \tag{25}$$

Let us introduce $I = S_a^{-1} S_a$ in equation (24) as

$$\tilde{b} = (X' S_\epsilon^{-1} X + S_a^{-1})^{-1} S_a^{-1} S_a (X' S_\epsilon^{-1} y + S_a^{-1} b_a)$$

$$
\begin{aligned}
\tilde{b} &= (X^{'}S_{\epsilon}^{-1}X + S_{a}^{-1})^{-1}S_{a}^{-1}S_{a}(X^{'}S_{\epsilon}^{-1}y + S_{a}^{-1}b_{a}) \\
&= (X^{'}S_{\epsilon}^{-1}S_{a}X + I)^{-1}(X^{'}S_{\epsilon}^{-1}S_{a}y + b_{a}) \\
&= (X^{'}S_{\epsilon}^{-1}S_{a}X + I)^{-1}(X^{'}S_{\epsilon}^{-1}S_{a}y - X^{'}S_{\epsilon}^{-1}S_{a}Xb_{a} + b_{a} + X^{'}S_{\epsilon}^{-1}S_{a}Xb_{a}) \\
&= (X^{'}S_{\epsilon}^{-1}S_{a}X + I)^{-1}((y - Xb_{a})X^{'}S_{\epsilon}^{-1}S_{a} + (I + X^{'}S_{\epsilon}^{-1}S_{a}X)b_{a}) \\
&= (X^{'}S_{\epsilon}^{-1}S_{a}X + I)^{-1}((y - Xb_{a})X^{'}S_{\epsilon}^{-1}S_{a}) + b_{a} \\
&= b_{a} + (X^{'}S_{\epsilon}^{-1}S_{a}X + S_{a}^{-1}S_{a})^{-1}((y - Xb_{a})X^{'}S_{\epsilon}^{-1}S_{a}) \\
&= b_{a} + (X^{'}S_{\epsilon}^{-1}X + S_{a}^{-1})^{-1}X^{'}S_{\epsilon}^{-1}(y - Xb_{a})
\end{aligned}
\tag{26}
$$

setting

$$
M = X^{'}S_{\epsilon}^{-1}X + S_{a}^{-1})^{-1}X^{'}S_{\epsilon}^{-1} \tag{27}
$$

$$
\tilde{b} = b_{a} + M(y - Xb_{a}) \tag{28}
$$

and for equation (25) as

$$
\begin{aligned}
cov(\tilde{b}) &= (X^{'}S_{\epsilon}^{-1}X + S_{a}^{-1})^{-1}S_{a}^{-1}S_{a} \\
&= (X^{'}S_{\epsilon}^{-1}XS_{a} + I)^{-1}S_{a} \\
&= (X^{'}S_{\epsilon}^{-1}XS_{a} + I)^{-1}(X^{'}S_{\epsilon}^{-1}XS_{a} + I)S_{a} - S_{a}X^{'}S_{\epsilon}^{-1}XS_{a} \\
&= S_{a} - ((X^{'}S_{\epsilon}^{-1}XS_{a} + I)^{-1}S_{a}X^{'}S_{\epsilon}^{-1}XS_{a}) \\
&= S_{a} - (X^{'}S_{\epsilon}^{-1}X + S_{a}^{-1})^{-1}S_{a}X^{'}S_{\epsilon}^{-1}X
\end{aligned}
\tag{29}
$$

Applying equation (27)

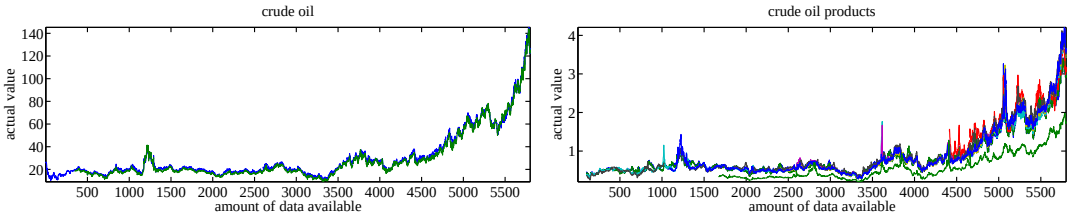
$$
cov(\tilde{b}) = S_{a} - MS_{a}X \tag{30}
$$

This is what is called Linear Kalman Filter which is normally used in estimation of linear stochastic processes resulted from measurements that relate linearly to the process. When the measurement relationship to the process is nonlinear, the Kalman Filter that linearizes current mean and covariance is referred to as Extended Kalman Filter or EKF [35]. Ensemble Kalman

Filter or EnKF is used to estimate optimal mean state and covariance to the problems that use ensemble models. The EnKF is a recursive filter suitable for problems with a large number of variables, essentially, the covariance matrix is replaced by the sample covariance. Variational Ensemble Kalman Filter is a more advanced technique of minimizing root mean square error which uses unified 3D-Variational method (3D-VAR) together with LBFGS method to find optimal results [26]. Due to complexity of ensemble models, an appropriate technique had to be used for model fitting and the results are compared in the next section.

3 Data analysis

In this work, secondary data downloaded from the Internet are used. The data comes from six American and European spot crude oil markets, namely Mont Belvieu, U.S Gulf Coast, New York Harbour, Los Angeles CA, European Brent and Cushing OK. These are weekday oil product spot prices collected from years 1986-2010 (the weekend prices were not included). There are two crude oil data series (OK and EU) and a number of crude oil products, namely propane, kerosene, sulphur low diesel, heating oil, regular gasoline, conversional gasoline. Based on oil products and the spot prices market where the data have been collected, there are eleven data series to be studied. To reduce the biases, crude oil data have been used to study the dynamics due to the fact that prices of commodities are influenced by several factors including the production costs to the final product, and the fact that production costs differ from one commodity to the other. Figure 1 shows the time line plot of crude oil and oil products prices from our data set. As it can be observed from Figure 1, the data are similar, therefore it is enough for one to just take one of these and proceed with the studies. Crude oil OK data series Figure 1(a) in (in blue) has been taken to represent oil. The electricity



(a) (b)

Figure 1: Time line plot of crude oil prices data and crude oil products.

data used in this case were collected from Nord Pool spot market and are studied along with oil prices.

From Figure 1, the important features of the series can be observed such as trend, outliers and discontinuities due to missing values. The data are clearly non-stationary and have a strong non-linear trend. As it was discussed before in section 1.6, time series data to be analyzed have to be free of features such as trend, seasonal or cyclic movement. The data observed in Figure 1 have trend in which one has to clean them so as to continue with analysis. This process of removing trend is known as data detrending which is easily done by fitting a trend line and subtracting it from the data series. Figure 2 presents time line plots for detrended oil and electricity spot price series. The oil

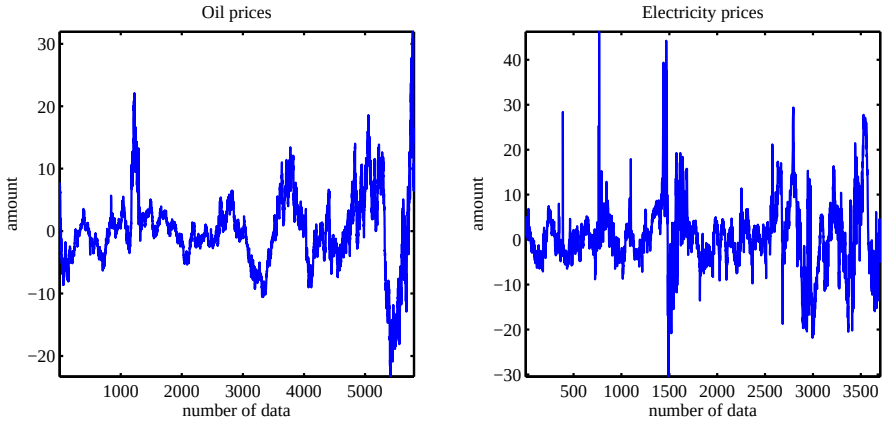


Figure 2: Time line of oil and electricity prices.

price series shows similar features to those of electricity prices. Histograms

for these are presented in Figure 3 in which similar features can be observed. Both are slightly positively skewed and with leptokurtic structure. Basic

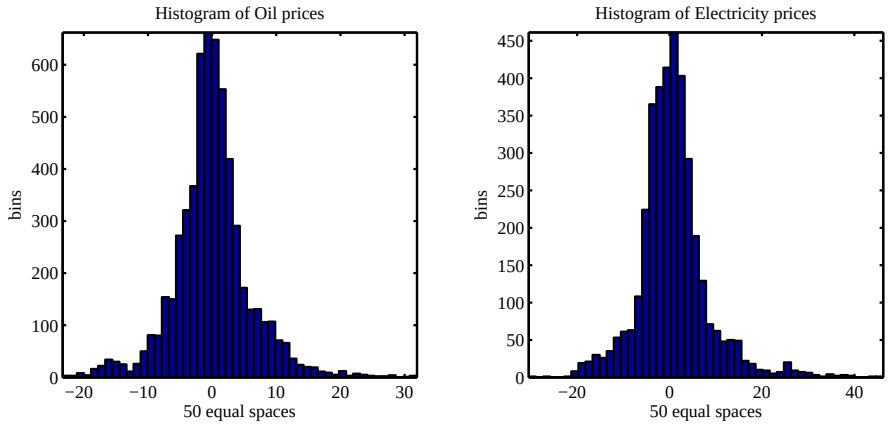


Figure 3: Histograms of oil and electricity prices.

statistics for these data have been summarized in Table 1 which contains basic statistics, such as mean, standard deviation, skewness and kurtosis.

Table 1: Table for basic statistics of oil and electricity data.

	Oil prices	Electricity prices
mean	0	0.7286
Standard Deviation	5.9098	7.4742
Skewness	0.3052	0.9231
Kurtosis	5.7845	6.9756

From Table 1, kurtosis coefficients are greater than 3 in both oil and electricity series which implies that our distributions are more fat-tailed [15]. This statement is supported by Figure 3 which presents the histograms of oil and electricity series. The statistics defined as skewness and kurtosis can be used to determine whether the given data follows normal distribution or not. Similarly, the plot of data can explore normality as well as stationarity. The series are neither normally distributed nor stationary as their means seem not to be constant in Figure 2.

Autocorrelation or autocorrelogram is a plot of autocorrelation functions (ACFs) against time lags $s = 0, 1, 2, \ldots, n-1$ and for these data the ACF is represented by Figure 4. The autocorrelation plots are used for checking randomness of the data set, the randomness ascending by computing autocorrelation at various time lag. For random data set, the autocorrelation should be near zero and for non random data sets, the autocorrelation coefficient should be significantly non zero.

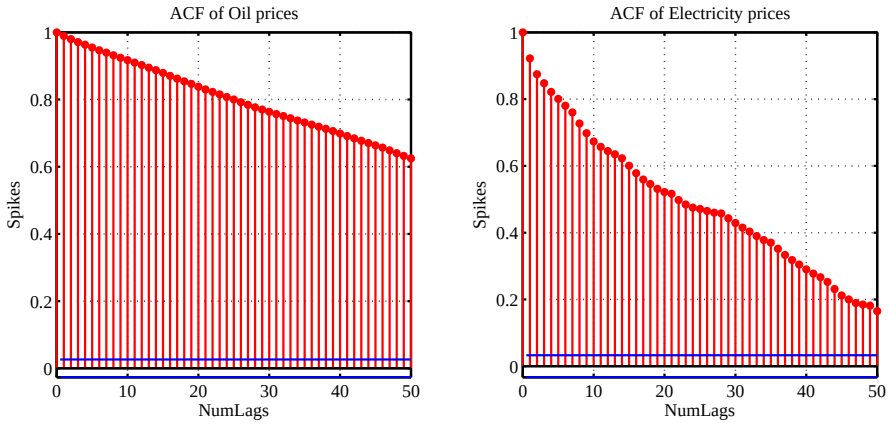


Figure 4: ACFs for oil and electricity prices.

By referring to the Box-Jenkins concept, if autocorrelations start high and decline slowly, then the series is not stationary, for stationary data spikes die out faster. The subplots in Figure 4 reveal that our series have non stationarity behavior since their spikes are dying out very slowly [33]. This is in both oil and electricity price autocorrelation plots.

We also need to discuss the partial autocorrelation of the series which is an extension of autocorrelation, where the dependence on the intermediate elements (those within the lag) are removed. It basically measures the degree of association between the random variable with the effect of a set of controlling random variables removed, i.e it measures the correlation between an observation of k time steps ago and the current one, that is the correlation between x_t and x_{t-k} after removing the influence of all observation in

between.

The autocorrelation and partial autocorrelation concepts are similar, except that when calculating partial autocorrelation, all the elements within the lag are eliminated. This is a commonly used tool for model identification in Box-Jenkins methodology. Specifically, it is used in order identification of ARMA(p,q) model. ACF is for identifying p order in MA(p) model and PACF is for q identification in AR(q) model. The partial autocorrelation function is given by Equation (31) ([33]) and Figure 5 shows the plot of PACF.

$$\rho_{kk} = \frac{\rho_k - \rho_{k-1}^2}{1 - \rho_k^2} \tag{31}$$

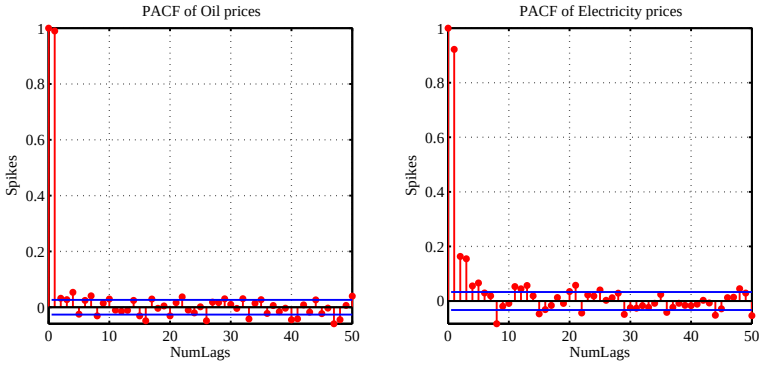


Figure 5: PACFs for oil and electricity prices.

From Figure 5, there are significant spikes at lag 1 at each PACF plot. Non stationary series are unpredictable and can not be modeled or forecasted and the results obtained by using non stationary data may be spurious results leading to poor understanding and forecasting [33].

3.1 Model fitting

3.2 Ensemble coupled mean-reversion model

In this section, model fitting results are going to be analyzed in which ensemble coupled mean-reversion model is discussed first.

$$dX_t^k = \alpha_t[(X_t^* - X_t^k) + (f(k, \mathbf{X_t}) - X_t^k)]dt + \sigma_t dB_t^k$$

The analysis is in the form of statistical figures as in section 3.

Figure 6 presents pure price in black color together with simulated prices in red color for oil and electricity prices. These prices can also be viewed in the form of histograms as shown in Figure 7 and Figure 8.

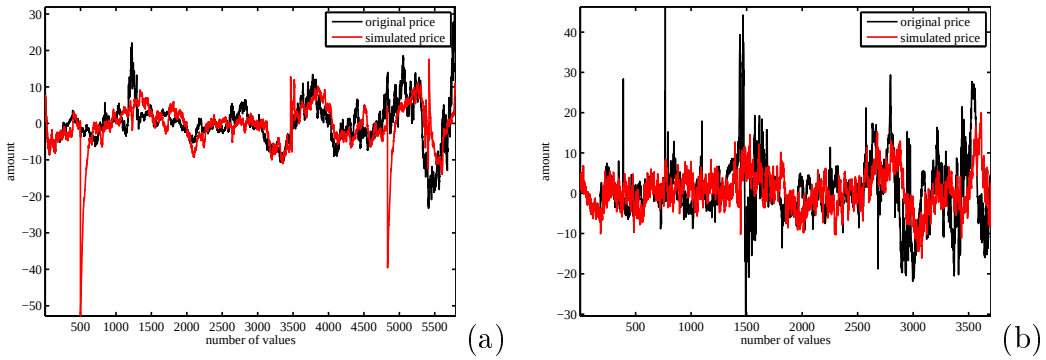


Figure 6: Time line plot for (a) oil prices and (b) electricity prices (Ensemble coupled mean reversion model).

The simulated prices tend to attain symmetric structure with lower kurtosis for simulated electricity prices compared to pure prices. For the case of oil prices, the simulated price histogram is more positively skewed and has higher kurtosis than its corresponding original prices.

All features described in histograms can also be observed via numerical values of basic statistics values in Table 2 and Table 3. Mean value is negative for oil prices instead of zero as in original prices. Standard deviation is higher

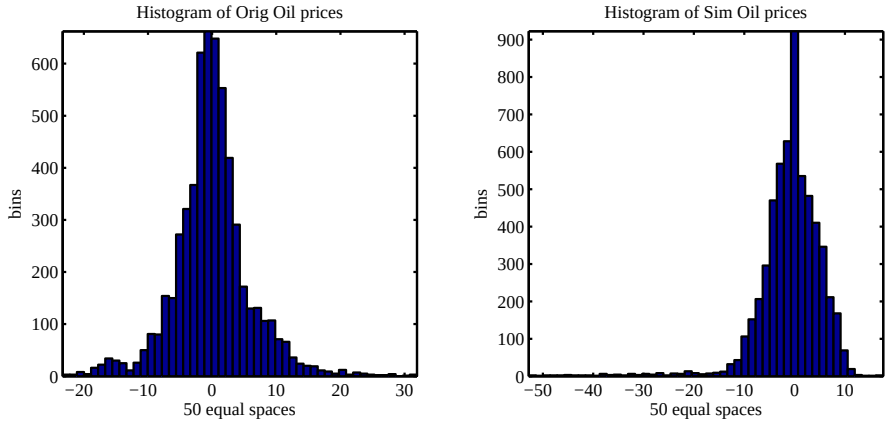


Figure 7: Histograms of original and simulated oil prices (Ensemble coupled mean reversion model).

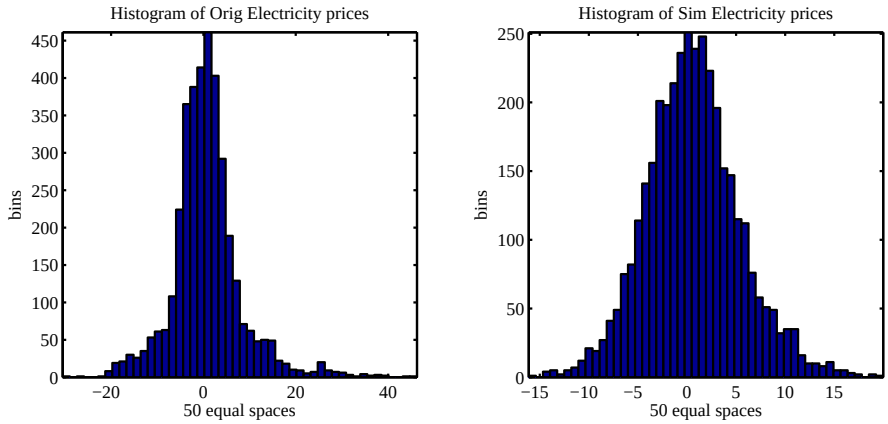


Figure 8: Histograms of original and simulated electricity prices (Ensemble coupled mean reversion model).

Table 2: Table for statistics of original and simulated oil prices (Ensemble coupled mean reversion model).

	Orig oil prices	Sim oil prices
mean	0	−0.8038
Standard Deviation	5.9098	6.2434
Skewness	0.3052	−2.3866
Kurtosis	5.7845	16.0882

Table 3: Table for statistics of original and simulated electricity prices (Ensemble coupled mean reversion model).

	Orig electricity prices	Sim Electricity prices
mean	0.7286	0.6281
Standard Deviation	7.4742	4.7744
Skewness	0.9231	0.2614
Kurtosis	6.9756	3.5605

than original one which is followed with negative skewness and much higher kurtosis. In electricity prices, the kurtosis value is very close to three which is the value of normally distribution. Autocorrelation plot for original and simulated prices in both cases are in Figure 9 and Figure 10.

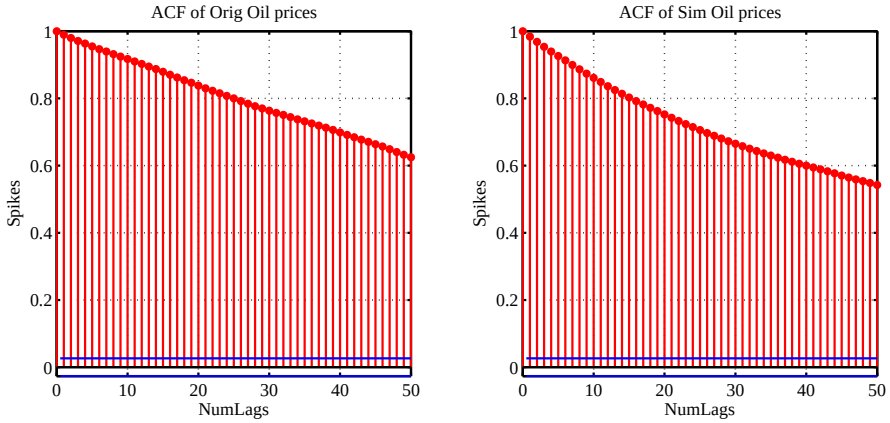


Figure 9: ACFs for original and simulated oil prices (Ensemble coupled mean reversion model).

Close relation between original and simulated prices can be seen, this is in terms downward spikes occurrence, though there is a big variation for electricity prices by having low downward spikes occurrence compared to that of original electricity prices. There is no a big difference between the two ACF figures of the oil original and simulated oil prices.

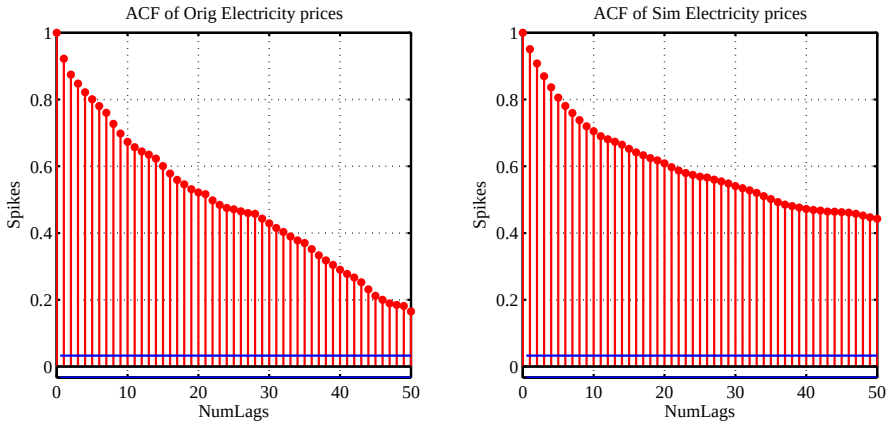


Figure 10: ACFs for original and simulated electricity prices (Ensemble coupled mean reversion model).

Partial autocorrelation figures for these prices are presented in Figure 11 and Figure 12. The situation is the same as it was in ACF figures, oil simulated

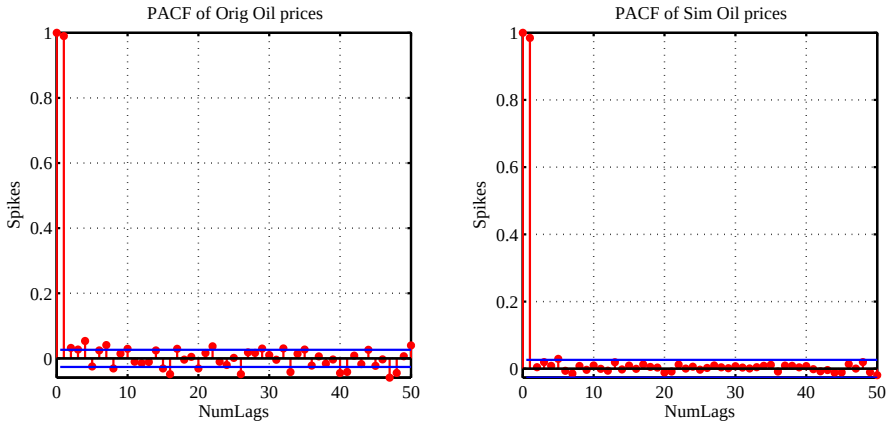


Figure 11: PACFs for original and simulated oil prices (Ensemble coupled mean reversion model).

prices produced very similar figure to that if is corresponding original prices

by having only one significant spike at the first lag, while the rest are almost in the significant level, though there are some spikes at the lower part especially for original price but they are not well identified. For the case of electricity prices, the PACF of the simulated prices closely replicates the features of pure electricity prices, the first lag spike is identified well for the real and simulated prices. More spikes are observed out of significant level in original

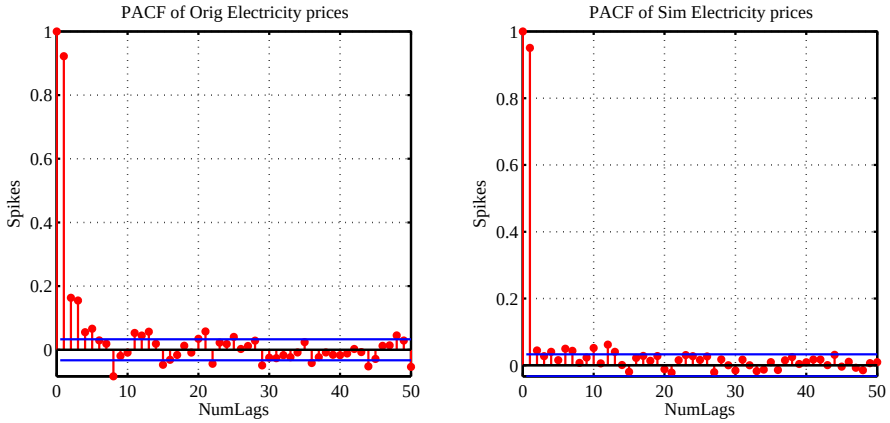


Figure 12: PACFs for original and simulated electricity prices (Ensemble coupled mean reversion model).

prices than in simulated prices. After all the explanations and comparison given, one can summarize by saying that simulation using ensemble coupled mean-reversion model was able to reproduce the features of original prices to some extent.

3.3 Jabłońska-Capasso-Bianchi-Morale (JCBM) and Kalman Dynamics models

Next is the study of dynamics of oil prices in JCBM model

$$dX_t^k = 3[(\frac{\gamma_t}{3}X_t^* + \frac{\theta_t}{3}h(k, X_t) + \frac{\xi_t}{3}g(k, X_t)) - X_t^k]dt + \sigma_t dW_t$$

In this analysis, Kalman Dynamics model is also added for comparison with JCBM model. The same comparison in terms of figures and statistical pa-

rameter is done. For Kalman Dynamics model analysis, oil data series had to be detrended, deseasonalized and spikes had to be removed first before being able to apply this model. Oil data fitting in this model is done in short term series where 500 consecutive values in a series are selected randomly. Since similar results were obtained in each model run, only two of them have been presented in this section to represent the rest. The results for both real and simulated prices for Kalman Dynamics and JCBM model are presented in figures and tables as follows. Figure 13 and Figure 14 depict the simulation results of the first two short term trend values in form of time line plot.

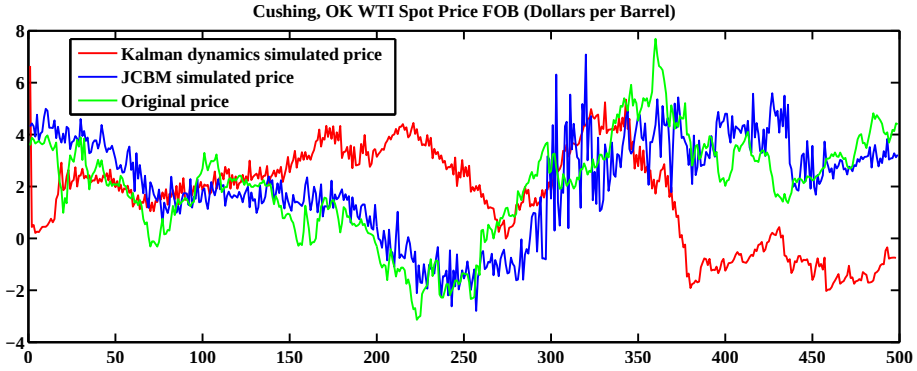


Figure 13: Time line plot for original and simulated oil prices (Kalman Dynamics and JCBM model).

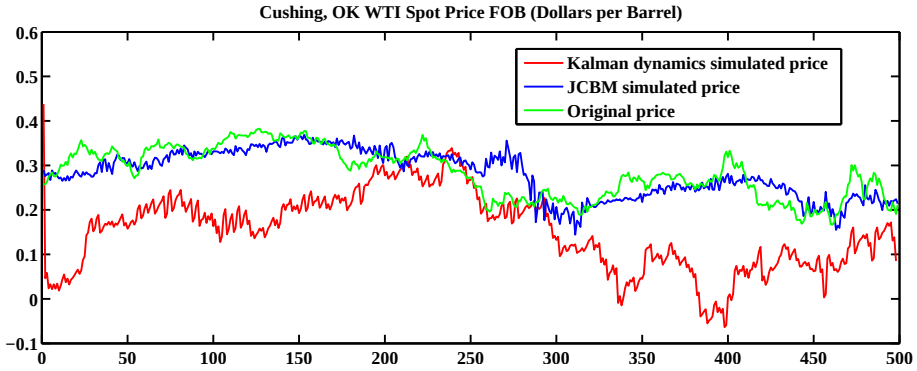


Figure 14: Time line for original and simulated electricity prices (Kalman Dynamics and JCBM model).

In these figures, it can be seen that among the two models, JCBM model

shows more similarity to the real prices compared to Kalman Dynamics model.

Histograms for these prices are presented in Figure 15 and 16 in which it is difficult to relate the histogram features with original prices due to irregular structures that are observed in real and simulated prices.

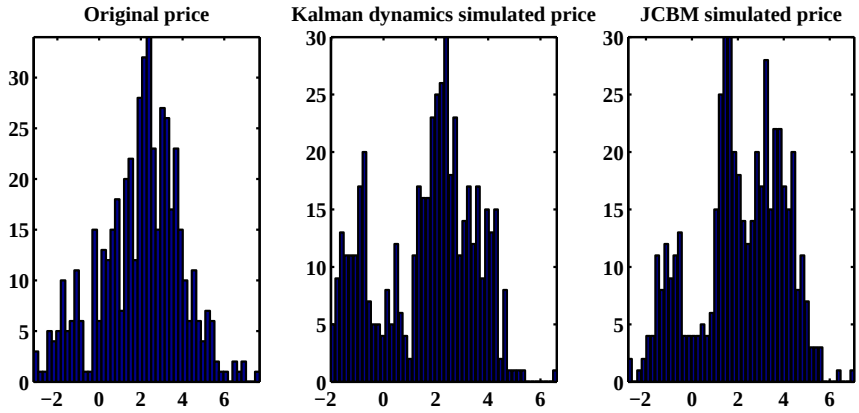


Figure 15: Histogram for original and simulated electricity prices (Kalman Dynamics and JCBM model).

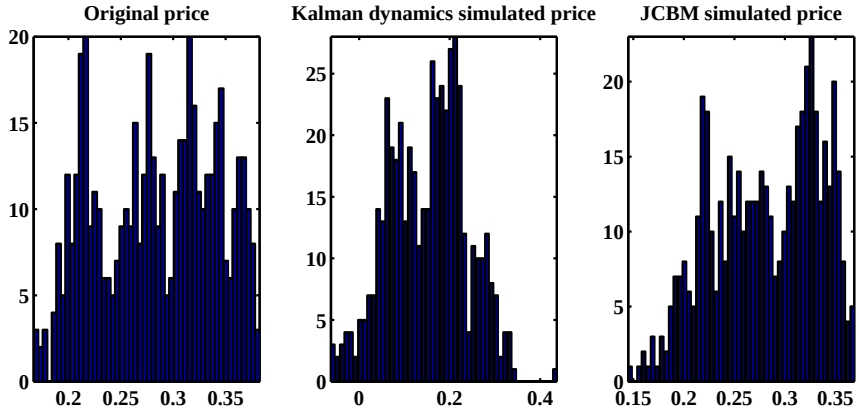


Figure 16: Histogram for original and simulated electricity prices (Kalman Dynamics and JCBM model).

Basic statistics for these figures are given in Table 4 where mean, standard deviation, skewness and kurtosis for original and simulated prices in each

model are collected. When the values are compared with the real prices, the values for JCBM model seem to be closer to those of the real prices, especially at kurtosis values. But generally, the two models are closely related to each other since there is not much deviation between the values.

Table 4: Table for statistics of original and simulated oil prices (Kalman Dynamics and JCBM model).

	Original prices	Kalman dynamics sim price	JCBM sim price
mean	0.2837	1.0129	1.4496
Std	0.0561	1.1342	1.8579
Skewness	−0.1265	−0.1582	−0.2871
Kurtosis	1.8553	2.3052	1.9905

Autocorrelation figures for this analysis are shown in Figure 17 and Figure 18 in the same arrangement starting with original prices, simulated prices by Kalman Dynamics model, then JCBM model. When original prices are compared with simulated prices in both models, more closely related features are seen between the original prices with that of JCBM' ACF figure.

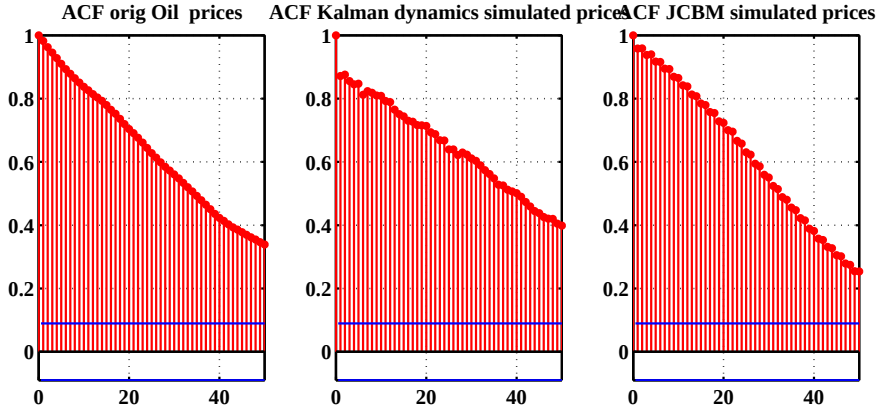


Figure 17: ACFs for original and simulated oil prices (Kalman Dynamics and JCBM model).

In evaluation of autocorrelation figures the best results is when spikes move down in faster rate (dying out quickly) then JCBM' results are better than

Kalman Dynamics. This can be observed more in Figure 17, and in the case of replicating real price dynamics the first (original) and last autocorrelation (JCBM) figures are more similar than the middle one, although the difference is not big.

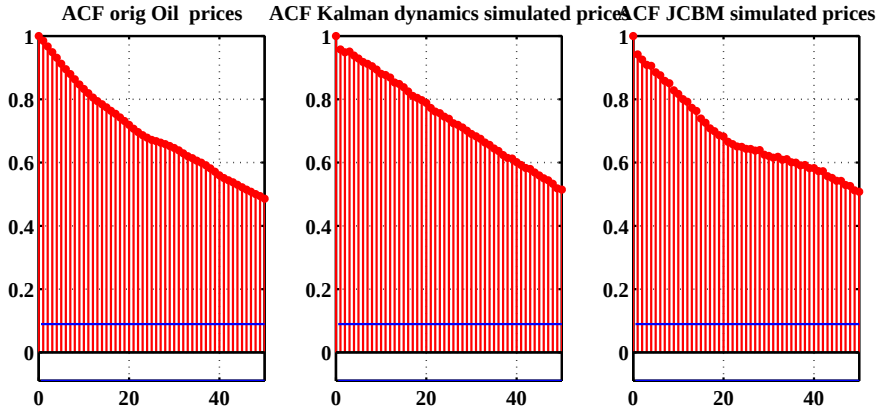


Figure 18: ACFs for original and simulated electricity prices (Kalman Dynamics and JCBM model).

The partial autocorrelation plots are presented in Figure 19 and Figure 20.

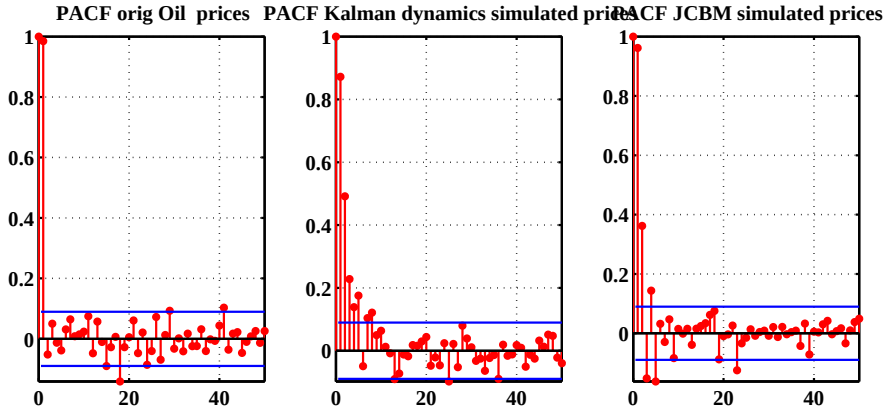


Figure 19: PACFs for original and simulated oil prices (Kalman Dynamics and JCBM model).

The study of these figures shows that there is a close relation between JCBM

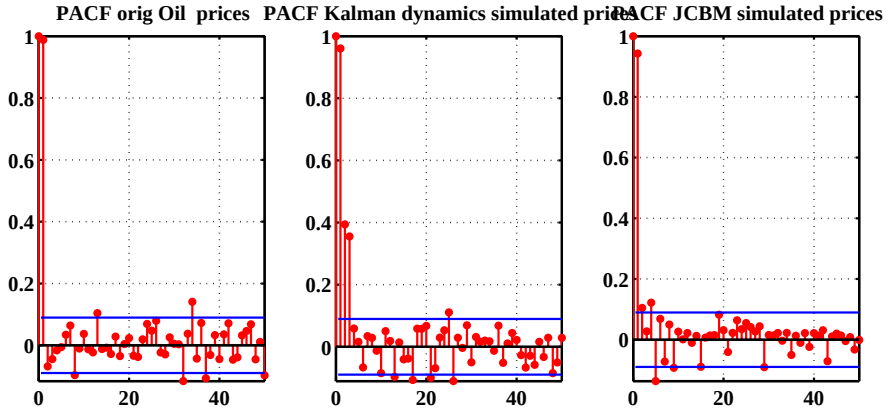


Figure 20: PACFs for original and simulated oil prices (Kalman Dynamics and JCBM model).

model and Kalman Dynamics model when one takes a look on spikes arrangement. Also, the relation to the real price PACF is not very far, in other words both JCBM and Kalman Dynamics have tried to follow the real price dynamics. Basic statistics for these prices can also support this statement due to closeness of mean, standard deviation, skewness as well as kurtosis of the three spot prices. There is no big difference between the three price series.

For the case of electricity prices, model simulations were applied to almost the whole data series. This is different from the case of oil prices, where model applied to the whole data set at once did not produce reasonable results. The results in time line plot is in Figure 21 where it can be seen that simulated prices for both models try to capture the movement of original prices up to a significant level. Also, the simulated results follow one another, especially before and after high spikes around 1200 to 1400 days. All models fail to generate the higher spikes which are observed before and after the days in between 1200^{th} day and 1400^{th}. Histograms for these series are presented in Figure 22, where arrangement of figures are as in oil prices analysis.

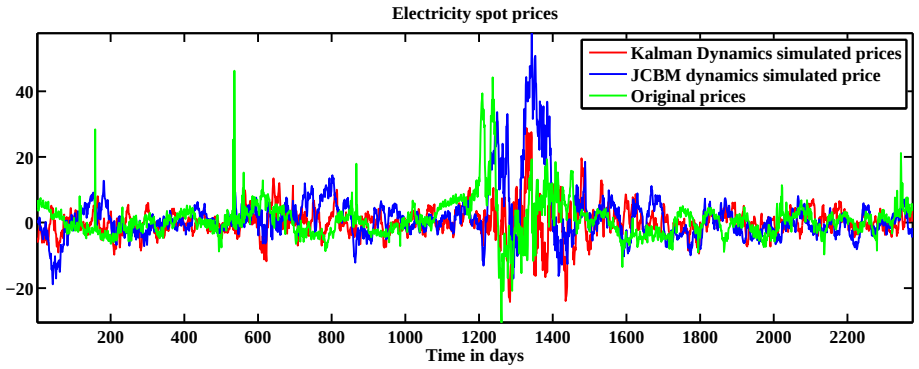


Figure 21: Time line plot for original and simulated electricity prices.

In these, comparable features are seen unlike in oil prices. Simulated price histograms have similar features to that of the real price. Kalman Dynamics histogram seems to be more symmetric than the other two, while JCBM' histogram is more right skewed than the original one. When the two simulated price histograms are compared to the original price, JCBM histogram can be regarded as more similar to the real data than Kalman Dynamics histogram.

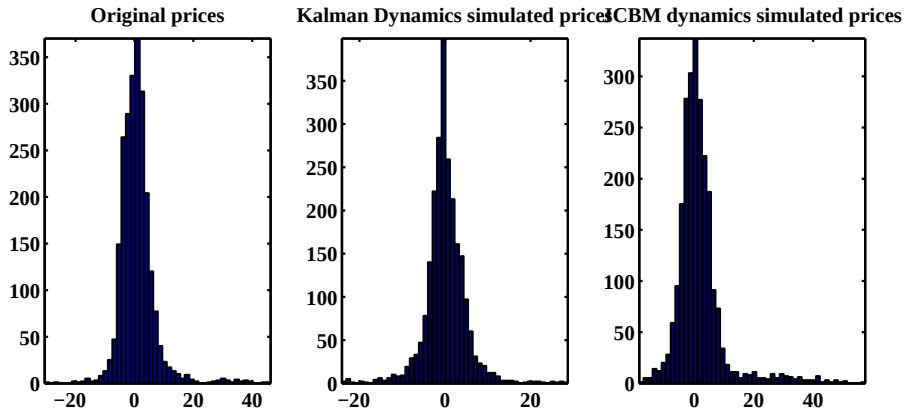


Figure 22: Histogram of original and simulated electricity prices (Kalman Dynamics and JCBM model).

The basic statistics for these prices are given in Table 5 where close statistical values can be seen between the real price and those of simulated prices from

each model. When the real price values are compared to those of Kalman Dynamics model and JCBM model, one can say that JCBM model values are closer to the real ones in each moment of distribution. Histograms express the observed features as well.

Table 5: Table for statistics of original and simulated electricity prices (Kalman Dynamics and JCBM models).

	Original prices	Kalman Dynamics sim price	JCBM sim price
mean	0.8640	0.0976	1.6687
Std	5.8950	4.6760	8.1299
Skewness	2.1168	0.2222	2.5797
Kurtosis	15.1597	9.2935	13.1890

Autocorrelation plots for these are in Figure 23, where a clear difference is seen between original and that of Kalman Dynamics, though in Kalman Dynamics results seem to be better since memory for previous features are kept more than that of JCBM model.

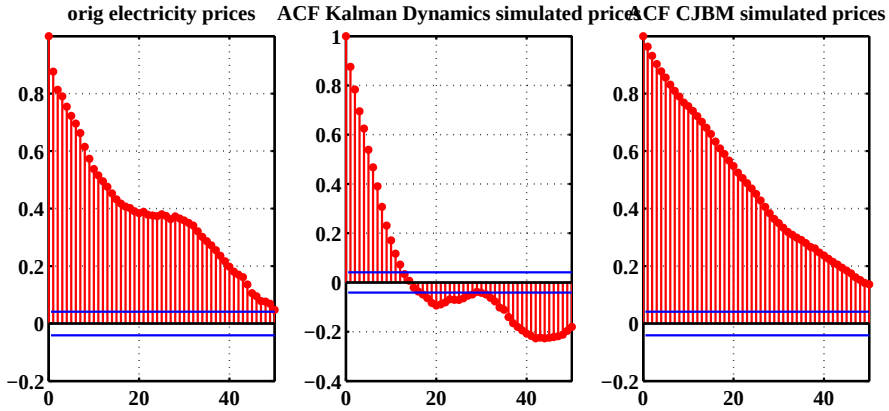


Figure 23: ACF of original and simulated electricity prices (Kalman Dynamics and JCBM model).

The real price autocorrelation figure is similar to that of JCBM model in spikes arrangement with a slight difference at the end where the last spike lies. PACF figures are as in Figure 24. It is difficult to tell the differences

and similarities among these figures but when one bases on the first few lags, Kalman Dynamics show somehow similar features to that of pure prices. There are number of spikes out of the significance level for each but Kalman Dynamics is closer to the real ones.

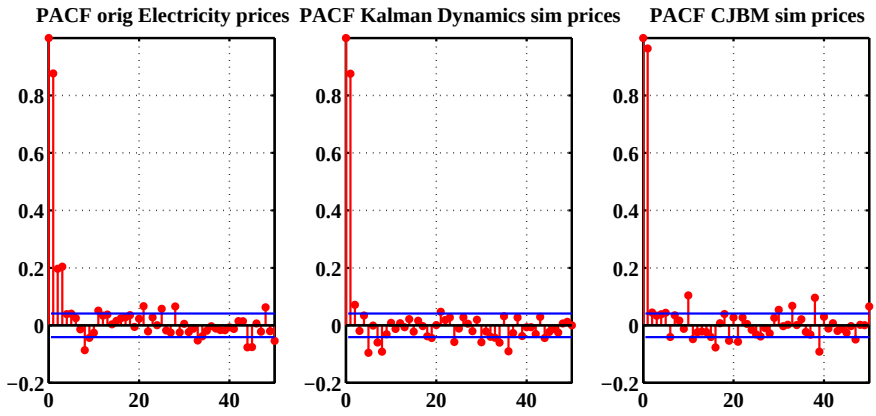


Figure 24: PACF of original and simulated electricity prices (Kalman Dynamics and JCBM model).

Generally, the two models seem to perform better in electricity prices than oil prices. This statement can also be applied to the ensemble coupled mean reversion model applied before, refer Table 2 and 3 for more clarifications. However, the performance of these models in both oil and electricity prices seem to be much better for Kalman Dynamics and JCBM models. When Kalman Dynamics is compared with JCBM model regarding the results obtained in this work, JCBM model performs better than Kalman Dynamics in many aspects.

4 Conclusion and recommendation

4.1 Discussion and conclusion

The ensemble coupled mean-reversion, JCBM and Kalman Dynamics models have been used in this study to compare the dynamics of electricity and oil prices. The work started by studying the behavior of oil and electricity spot prices visually and statistically and the results were presented in figures and tables. Figures 2-5 present visual results and Table 1 collects statistical analysis, where basic statistics for original oil and electricity prices are presented and compared. Series features for oil and electricity prices were similar, thus one could consider fitting and comparing the same model to both series. The model simulations, figures, as well as statistical values calculations were done by MATLAB software. The Least Squares and Maximum Likelihood techniques have been used in fitting models to the data. The original and simulated prices using the ensemble coupled mean-reversion model were compared visually and statistically. Figures 6-12 are for visual comparison of original and simulated prices, while Tables 2 and 3 are for their corresponding statistical analysis, where basic statistics are compared. Similar analysis was done for JCBM and Kalman Dynamics models using oil prices which had trend, and results were presented in Figures 13-20 and Table 4. Figure 21-24 present visual results for electricity prices while table 5 is for the their basic statistics.

Generally, the results show that ensemble models used in this work are capable of replicating the dynamics of real prices by producing similar results to that of real prices. Ensemble coupled mean-reversion model produces simulated results which are comparable with the original prices. Also, the pre-treated data in JCBM and Kalman Dynamics models replicate the dynamics of real prices to a great extent. Some deviations have also been seen

in JCBM model when compared with Kalman Dynamics model, though the difference was not excessive. Much better results have been seen when electricity prices are applied to Kalman Dynamics model and JCBM model. One can say the last two models have the same ability in dealing with real data in spite of the fact that Kalman Dynamics uses JCBM model. However, the study of ensemble modelling with ensemble model produce better results in electricity spot prices than oil spot prices.

4.2 Recommendation

Generally, plots as well as the descriptive statistics have shown that the models were able to replicate the dynamics of oil prices as well as electricity prices up to comparable level. More improved technique of parameter calibration are encouraged to capture the physical behavior of these prices and more for oil spot prices.

The failures which have been observed might be the results of weakness of the fitting technique which have been used which is the Least Squares method. Least Squares method works properly with Gaussian ensemble stochastic model with Gaussian-distributed data. The models are Gaussian in nature but real data were not completely Gaussian. We recommend the use of other models which are not Gaussian in studying non-Gaussian data and other techniques of parameter estimation.

Euler scheme with half order was used in this simulation; this scheme is of less accuracy and can also be the source of weakness observed in this work, other improved scheme such as Heun, Runge-Kutta of higher order and accuracy can be used for improvement of the results.

REFERENCES

References

[1] Bishwal, P.: (2008), *Parameter Estimation in Stochastic Differential Equations*, Berlin, Springer-Verlag Berlin Heidelberg.

[2] Böjorck, A.: (1997), *Numerical method for Least Squares problems*, SIAM, Philadelphia University.

[3] Chartfield, C.: (2005), *The Analysis of Time Series*, New Zealand, Chapman & Hall/CRC.

[4] Chris, B.: (2008), *Introductory Econometrics For Finance*, Cambrige, Cambrige University.

[5] Christopher, K., Wikle, L., Barlener, M.: (2006), *Bayesian tutorial for Data Assimilation*, Department of statistics, University of Missouri, 146 Middlebush Hall, no 65211. Department of Statistics, The Ohio State University, United state.

[6] Cole, J.D.: (1951), On a quasilinear parabolic equation occuring in aerodynamics. *Quart. Appl. Math.* 9(3), 225-236.

[7] Dees, S.M.: (2008), Assessing the factors behind oil price changes, *Journal in economics*, Economic Journal series No 855.

[8] George, E.M.: (1976), *TIME SERIES ANALYSIS, Forecasting and Control*, Cambrige, Cambrige University.

[9] Haario, H.: (2011), *Statistical Analysis in Modelling Course*, Lappeenranta University of Technology, Lappeenranta, Finland.

[10] Jabłońska, M.: (2011), *From Fluid dynamics to human psychology. What drives finacial markets towards extream events*, Doctoral Dissertation, Lappeenranta University of Technology, Lappeenranta, Finland.

[11] Kaufmann, R.K, Dees, S., Kaladeloglou, P., and Sanches, M.: (2004), Does OPEC mater? an econometric analysis of oil prices, *The Energy journal*, 25(4), 67-90

[12] Kauranne, T.: (2011), *Lecture notes on Modelling methodology in process engineering*, Lappeenranta University of Technology, Lappeenranta Finland.

[13] Keynes, J.M.: (1936), *The General Theory of Employment, Interest and Money.* London: Macmillan.

[14] Koch, K. R.: (1988), *Parameter Estimation and Hypothesis Testing in Linear Model*, Springer Verlag, Berlin.

[15] Martinez, W.A.: (2002), *Computational Statistics Handbook with MATLAB*, New Zealand, Chapman & Hall/CRC.

[16] Mirau, S.: (2011), *Ensemble Modeling of Spot Market Time Series*, Master's Thesis, Lappeenranta University of Technology. Lappeenranta, Finland.

[17] Möbert, J.: (2007), Crude Oil Price Determinants, *Journal in economics*, Applied Research Economics, No 186, August 10, 2007.

[18] Mtunya, A.: (2010), *Modelling Electricity Spot Price Time Series Using Coloured Noise Forces*, Master's Thesis, University of Dar es Salaam, Dar es Salaam, Tanzania.

[19] Naeem, M.: (2010), *A comparison of electricity spot prices simulation using ARMA-GARCH and mean-reverting model*, Master's Thesis, Lappeenranta University of Technology, Lappeenranta, Finland.

[20] Nampala, H.: (2010), *A stochastic mean-reverting jump-diffusion model with multiple mean-reversion rates*, Master's Thesis, University of Dar es Salaam, Dar es salaam, Tanzania.

[21] Øksendal, B.: (2000), *Stochastic Differential Equations*, Springer-Verlag.

[22] Salzmann, M.: (1993), Least Squares filtering and Testing for Geodetic Navigation Applications, *Netherland Geogetic Commission* series No 37, Puplication on Geodesy.

[23] Sailon, I.: (2009), *MCMC Analysis Of Classical Time Series Algorithms*, Master's Thesis, Lappeenranta University of Technology, Lappeenranta,Finland.

[24] Simkwembe, E.: (2009), *Pricing of Energy by means of Stochastic Model*, Master's Thesis, University of Dar es Salaam, Dar es salaam, Tanzania.

[25] Sollich, P.: (1997), Statistical Mechanics of Ensemble Learning, *Physical Review E*, vol. 55, no. 1, pp. 811-825.

[26] Solonen, A.: (2011), *Bayesian methods for estimation, optimation and experimental design*, Doctoral Dissertation, Lappeenranta University of Technology, Lappeenranta, Finland.

[27] Tao, W., McCallum, A.: (2005), Do Oil Futures Prices Help Predict Future Oil Prices?, *FBRSF Economic Letter* No 2005-38, December 30, 2005.

[28] Thomas, W.S.: (2000), Variational assimilation of meteorological observations in the lower atmosphere: a tutorial on how it works. *Journal of Atmospheric and Solar-Terrestrial Physics*, Vol 62:pages 1057-1070, April 2000.

[29] Titida, N.R.: (2005), ARIMA model for forecasting palm oil prices, *Economic Paper*, Mongkuts Institute of Technology Ladkrabang and Assumption University Huamark, Thailand.

[30] Torul, O.C.: (2009), Asymmetric adjustment of retail gasoline prices in turkey to world crude oil price changes: the role of taxes, *Economic paper*, Economics Bulletin, Vol. 29 No.2 pp. 775-787.

[31] Uhlenbeck, G.E. and Ornstein, E.S.: (1930), On the theory of Brown motion, *Physical Review*, Vol.36, 823-841.

[32] USA Today.: (2010), *Oil Briefly Spurts Near* 104 *per Barrel*, www.USAtoday.com, Article retrieved on March 12, 2010.

[33] Vandaele, W.: (1983), *Applied time series and Box-Jenkins model*, Academic Press, INC, United States.

[34] Weron, R.: (2005), Heavy tails and electricity prices, The Deutsche Bun-desbank's (2005), Annual Fall Conference (Eltville, 10-12 November 2005).

[35] Wilch, G. and Bishop, G.: (2006), *An introduction to Kalman Filter*, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3175,2006.