Multi-Factor Statistical Modelling of Demand and Spot Price of Electricity

Hilary Green, Nino Kordzakhia* and Ruben Thoplan
Department of Statistics, Faculty of Science, Macquarie University, Australia

(Received March 2012, accepted March 2013)

Abstract: The deregulation of electricity markets in different parts of the world has exposed consumers to irregularities in electricity prices driven by the principle of supply and demand. Development of accurate statistical models contributes to establishing protective mechanisms and risk measurement policies for both suppliers, consumers. In this paper multi-factor modelling methodology, solely applied to the spot price of electricity or demand for electricity in earlier studies, is extended to a bivariate process of spot price of electricity and demand for electricity. The suggested model accommodates common idiosyncrasies observed in deregulated electricity markets such as cyclical trends in price and demand for electricity, occurrence of extreme spikes in prices, and a mean-reversion effect seen in the settling of prices from extreme values to the mean level over a short period of time. A time series model for de-seasonalised demand for electricity is used in combination with a linear regression model developed for logarithms of deseasonalised daily averages of electricity spot prices. The spiky behaviour of prices occurring in clusters, interpreted as ‘a post-spike’ effect, is addressed by a filtered Poisson (i.e. shot noise) factor of the model. The demand for electricity is found to be the primary stochastic factor driving the electricity prices. In the linear regression model for ‘de-seasonalised’ and ‘de-spiked’ spot prices the back-shifted variables play the role of exogenous variables. These variables capture the ‘price’ and ‘demand’ inter-dependence observed in practice. The historical data is obtained from the NSW section of Australian Energy Markets.

Keywords: ARMA, bivariate normal inverse Gaussian distribution, filtered Poisson process, Gaussian copula, t-copula.

1. Introduction

Modelling of commodity prices is linked to the notion of convenience yield viewed as a factor associated with the storability of a commodity. However, due to the scarce storability of electricity, there is not a convenience yield factor which may affect these prices directly. Electricity is transported through transmission lines which may be a part of highly interconnected network. The complexity of transmission systems, technical limitations on the capacity of local networks, aggravating weather conditions, or other circumstances, may translate into unexpectedly high demand at any time of a day. Difficulties associated with delivery of electricity arising during congestions are transferred by grid operators into electricity tariffs for power transmission; hence creating some similarity with the convenience yield factor observed in other commodity markets. Lucia and Schwartz [17], considered the two-factor model originally designed for commodity markets and extended the model to electricity markets by replacing the convenience yield factor with an unobservable additive stochastic factor which accommodates electricity transmission related constraints affecting spot prices. The spot price of electricity is often

* Corresponding author. E-mail: nino.kordzakhia@mq.edu.au
modelled as the continuous time mean-reverting Gaussian process. The model originated in view of its adaptability to risk-neutral pricing of electricity futures and derivatives where the risk premium of interest is calculated as the expected value of spot price or, its function, at the maturity of contracts under the risk-neutral measure rather than objective statistical measure. In practice the models that are used for pricing of electricity contracts are calibrated on the prices of traded derivatives and, consequently, the statistical distribution of the underlying spot price of electricity becomes irrelevant. However, it is impossible to underestimate the importance of statistical modelling of spot price of electricity in building a risk profile of derivatives portfolio. The electricity spot price is deemed to be a risk factor which may significantly impact the value of a portfolio. For example, when quantifying a portfolio market risk, one needs to know by how much the portfolio value will change when the spot price changes according to realistic stress testing scheme.

This paper aims to establish the statistical model for the spot price of electricity in the bivariate setup. The suggested multi-factor modelling approach is rather flexible to incorporating stylized facts specific to electricity markets. In previous studies the authors proposed models for the natural logarithms of the average daily spot prices which included a deterministic (seasonal) component and a mean reversion component based on the Gaussian Ornstein-Uhlenbeck process see, Clewlow and Strickland [8], Cartea and Villaplana [7], Hinz et al. [13]. The others, Geman and Roncoroni [11], Meyer-Brandis and Tankov [19] and Schmidt [24] proposed a new generation of models which incorporated an additional jump process component to model spikes in spot prices of electricity. Recent developments of stochastic models extend to generalized Ornstein-Uhlenbeck processes driven by Lévy processes and were used for the pricing of electricity contracts Benth et al. [2], Benth et al. [3].

The linear regression models have been used by Ružić et al. [23], Ramanathan et al. [22], to model consumer demand for electricity affected by weather conditions and other energy market variables such as oil and coal prices. In a recent study by Seifert and Uhrig-Homburg [25] the price of carbon emission vouchers has been included into a linear model as an exogeneous variable.

A time series approach with heavy-tailed innovations to modelling and forecasting the demand for electricity has been reported and reviewed by Weron in [26] and Weron and Misiorek [27]. A comparative study of a number of time series and semi-parametric models has been conducted in Misiorek et al. [20] and included the demand and air temperature as the exogenous variables in the autoregressive model for the high frequency spot prices of electricity with pre-processed spikes. The multi-factor time series model with the lognormal innovations has been developed in Burger et al. [6] for the electricity spot prices and demand process and further applied to the derivatives pricing problem. The studies of reference have mainly focused on the European and North American electricity markets. In this study we found that the deregulated electricity markets in Australia exhibit similar features to these markets. However the spot price of electricity in Australia is subject to more frequent spiky behaviour. The model developed in the paper shows that the higher volatility can be explained by increasing demand which may be often caused by extreme weather conditions and drought.

In Section 2 a detailed description of the multi-factor model is given. A times series approach to modelling of de-seasonalised demand for electricity is used. The rising electricity spot prices are viewed as an implication of surging demands for electricity which have a delayed effect on prices. In the linear regression model for logarithms of daily
average price, the time delay effect is addressed by the inclusion of lagged spot price and demand variables as predictors.

The filtered Poisson process is used to address the occurrence of spike clusters in the logarithm of the daily average spot prices. The filtered Poisson process was introduced for modelling the dynamics of electricity forward curves in Schmidt [24], where the de-seasonalised spot prices of electricity were modelled using the Gaussian jump-diffusion process.

In Section 3 we describe specifics of the Australian electricity markets operations and relevant data. The results of statistical analysis are presented in Section 4. This section includes results of forecasting procedure developed for prediction of daily averages of demand for electricity. The bivariate Normal Inverse Gaussian (NIG) distribution is used to model dependence of residuals of multi-factor model fitted to bivariate data of demand and spot prices of electricity. As an alternative, the $t$- and Gaussian copulae with far fewer parameters than those of bivariate NIG were applied to reproduce an adequate dependence structure of residuals.

2. The Model

Let us consider a bivariate process $X_t = (X_t^{(1)}, X_t^{(2)}), \ t = 1, 2, \ldots$, of logarithms of the daily average spot prices of electricity and demand, as

$$
\begin{align*}
X_t^{(1)} &= f^{(1)}(t) + Z_t^{(1)}, \\
X_t^{(2)} &= f^{(2)}(t) + Z_t + Y_t, \\
Y_t &= \sum_{i=1}^{\infty} c_i Z_{t-i} + \sum_{i=1}^{\infty} d_i Y_{t-i} + \xi_t^{(2)},
\end{align*}
$$

(1)

where $Z_t^{(1)}$ is the linear stationary process given in (3), $\xi_t^{(2)}$ is a white noise and $Z(t)$ is a generalized filtered Poisson process observed at discrete times. The process $Z(t)$ can be described as follows,

$$
Z(t) = \sum_{k=1}^{N_t} \xi_k e^{-\lambda_k(t-\tau_k)},
$$

(2)

where $N_t$ is a Poisson process with rate $\nu$, $\tau_k$ is the time of arrival of the $k^{th}$ impulse such that the impulse inter-arrival times $\tau_{k+1} - \tau_k$ are exponentially distributed with the parameter $\nu$. We assume that the spikes arising at $\tau_k$ decline exponentially as $e^{-\lambda_k(t-\tau_k)}$ function for $t \geq \tau_k$. If ‘smoothing’ parameters $\lambda_1 = \lambda_2 = \lambda_3 = \ldots = 0$, then the process $Z_t$ is Markovian; in particular, if $\lambda_1 = \lambda_2 = \lambda_3 = \ldots = 0$, then $Z_t$ is a Poisson process. The spike, or jump, sizes $\xi_k$ are independent and identically distributed (iid). A generalized filtered Poisson Process has been used previously for modelling spikes in electricity markets by Schmidt [24].

Further we assume that the white noise $\xi_t^{(2)}$ follows a Normal Inverse Gaussian (NIG) distribution. The NIG distribution found to be suitable for modelling of daily averages of demand for electricity in Californian electricity markets in Weron [26].

The NIG distributions represent a subclass of mean-variance mixture distributions which are obtained using an independent inverse Gaussian variable as the mixing random variable, see [1],

$$
\xi = \mu + \beta \kappa + \sqrt{\kappa} \xi,
$$
where \( \xi \sim N(0, 1) \), and \( \kappa \sim IG(\alpha^2, \beta^2) \), \( |\beta| \leq \alpha \), follows an inverse Gaussian distribution with the probability density function (pdf)

\[
f(x) = \left( \frac{\alpha^2}{2\pi x^3} \right)^{1/2} e^{\alpha \sqrt{\beta^2 - \beta^2 x^2}} \left[ \frac{1}{2}(\alpha^2 x^2 - (\alpha^2 - \beta^2)x) \right], \quad x > 0.
\]

For bivariate NIG distribution see Definition A1 in the Appendix. The linear stationary process \( Z_t^{(1)} \) is given in the form of ARMA\((p, q)\) process

\[
Z_t^{(1)} + \sum_{i=1}^{p} \phi_i Z_{t-i}^{(1)} = \varepsilon_t^{(1)} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}^{(1)},
\]

with the NIG white noise \( \varepsilon_t^{(1)} \), \( t = 1, \ldots, N \). The Hannan-Rissanen estimate of variance of \( \varepsilon_t^{(1)} \) is

\[
\hat{\sigma}^2 = \frac{\sum_{t=p+q+1}^{N} \left( Z_t^{(1)} + \sum_{i=1}^{p} \phi_i Z_{t-i}^{(1)} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}^{(1)} \right)^2}{N - p - q},
\]

where \( \hat{\phi}_i, \hat{\theta}_i, i = 1, \ldots, p, j = 1, \ldots, q \), the least square estimates of parameters and the estimated residuals \( \hat{Z}_t^{(1)} \) are obtained via Hannan-Rissanen algorithm, Section 5.1.4 in [5]. The one-step forecasts of stationary and invertible ARMA\((p, q)\) process \( Z_t^{(1)} \), can be evaluated as follows,

\[
\hat{Z}_{t+1}^{(1)} = -\sum_{i=1}^{p} \hat{\phi}_i \hat{Z}_{t+1-i}^{(1)} + \sum_{i=1}^{q} \hat{\theta}_i \hat{\varepsilon}_{t+1-i}^{(1)},
\]

where \( \hat{\varepsilon}_t^{(1)} = Z_t^{(1)} - \hat{Z}_t^{(1)} \), \( t = N, N+1, \ldots \) are generated recursively. The \( h \)-step forward forecasts can be computed as follows

\[
\hat{Z}_{t+h}^{(1)} = \begin{cases} 
-\sum_{i=1}^{p} \phi_i \hat{Z}_{t+h-i}^{(1)} + \sum_{i=1}^{q} \theta_i \hat{\varepsilon}_{t+h-i}^{(1)}, & h = 1, 2, \ldots, q, \\
-\sum_{i=1}^{p} \phi_i \hat{Z}_{t+h-i}^{(1)}, & h = q+1, q+2, \ldots,
\end{cases}
\]

the forecasts beyond the moving average order \( q \) only depend on parameters of autoregressive part of order \( p \), Chapter 4, [12].

3. The Data

In 2009, the Australian Energy Market Operator (AEMO) was established to manage and integrate planning and trading activities across all energy transmission systems in Australia, and the National Electricity Market (NEM) in Australia now operates within this structure.

As in most markets, the spot price of electricity varies with demand; they vary according to time of day, type of day and the season of the year. Unforeseen circumstances such as breakdowns of generators, outages or extreme weather conditions can trigger spikes in the prices where a few sudden steep upward price jumps are followed by a slower downward move to a typical or pre-spike price. As well, spot prices tend to exhibit mean reversion behaviour; they tend to revert to a typical value after variations in the price. These reversion behaviours are partially due to the introduction of stand-by generators into the
system allowing demand to be met. Thus the NEM operates 24 hours a day every day of the week. The price and demand data by region, for each half hour of the day since December 1998 can be found on the AEMO website, at http://www.aemo.com.au/data/.

In this section we investigate only the NSW data for the years 2005 to 2008. We calibrate the multi-factor models which incorporate seasonality, spikes and random fluctuations in both the demand and prices of electricity. The data includes the time and date of the spot price and demand of the electricity prices in 6 regions of Australia reported for each half hour period for every day of the year. We extract the natural logarithms of the average daily prices and the average daily demands as well as the day and type of day for each day from January 1, 2005 to December 31, 2008. The variables of interest are $X^{(1)}_t = \ln(D_t)$ is the natural logarithm of the average demand for electricity in NSW on day $t$ and is referred to as demand. $X^{(2)}_t = \ln(P_t)$ is the natural logarithm of average spot price of electricity in NSW on day $t$ and is referred to as price.

4. Data Analysis

4.1. Demand Data

In (1) we consider the trend $f^{(1)}(t)$ which incorporates the mean, linear and non-linear terms as follows

$$f^{(1)}(t) = a^{(1)} + b^{(1)} t + \sum_{i=1}^{k} \gamma_i^{(1)} \sin \left( \frac{2\pi t}{365 s_i^{(1)}} \right) + \delta_i^{(1)} \cos \left( \frac{2\pi t}{365 s_i^{(1)}} \right),$$

where $k$ and frequencies $s_i, \ i=1,\ldots,k$ will be calibrated empirically.

The ‘periodogram’ function in Matlab has been applied to $X^{(1)}_t$ to identify the wave components with highest intensities discerning the significant sinusoids with periods of $365/s_i, \ i=1,\ldots,k$. Some of these seasonal patterns are intuitive, for example, the weekly cycles and the half yearly cycle. However, some of the seasonal components do not seem to have a practical basis, for example, the 19 week cycle. Nevertheless the twelve cycles listed in Table 1 have been included in the seasonal component $f^{(1)}(t)$ of $X^{(1)}_t$. Using ordinary least square regression we found the parameter estimates in (5). The number of cycles $k$ to include in the model has been determined based on the value of $R^2$. As illustrated in Figure 1 using $k=12$ sinusoids $R^2$ of 63.4% is achieved.

![Figure 1. Selection of optimal number of sinusoids k.](image-url)
Table 1. The periods of sinusoids for demand data.

<table>
<thead>
<tr>
<th>Period  $365/\xi^{(1)}$</th>
<th>Approximate part of year</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5000</td>
<td>Half-yearly</td>
</tr>
<tr>
<td>0.0191</td>
<td>Weekly</td>
</tr>
<tr>
<td>0.0096</td>
<td>Half-weekly</td>
</tr>
<tr>
<td>0.3333</td>
<td>Third of year</td>
</tr>
<tr>
<td>0.0192</td>
<td>Weekly</td>
</tr>
<tr>
<td>0.0096</td>
<td>Half-weekly</td>
</tr>
<tr>
<td>0.2500</td>
<td>Quarterly</td>
</tr>
<tr>
<td>2.0000</td>
<td>2 years</td>
</tr>
<tr>
<td>0.3636</td>
<td>19 weeks</td>
</tr>
<tr>
<td>0.0190</td>
<td>Weekly</td>
</tr>
<tr>
<td>0.1250</td>
<td>1/8 year</td>
</tr>
<tr>
<td>0.0064</td>
<td>1/3 weekly</td>
</tr>
</tbody>
</table>

According to (1), the residuals from de-trending in view of (5), will be referred to as $Z_t^{(1)}$. We investigate various linear time series models by fitting them to $Z_t^{(1)}$ data. The Akaike’s and Bayesian information criteria (AIC and BIC respectively) are used to select the best fitting model. The goodness-of-fit measures for the ARMA models in question are shown in Table 2 below.

Table 2. The goodness of fit results of ARMA($p$, $q$) models. In the second column the number of unknown parameters includes the variance parameter; the mean parameter is not included in the count.

<table>
<thead>
<tr>
<th>($p$, $q$)</th>
<th>Number of Parameters</th>
<th>Log-Likelihood</th>
<th>AIC</th>
<th>BIC</th>
<th>Number of lags</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(7, 10)</td>
<td>18</td>
<td>2880.4</td>
<td>-5722.8</td>
<td>-5622.4</td>
<td>40</td>
<td>0.3121</td>
</tr>
<tr>
<td>(9, 6)</td>
<td>16</td>
<td>2879.0</td>
<td>-5724.0</td>
<td>-5634.2</td>
<td>40</td>
<td>0.2724</td>
</tr>
<tr>
<td>(9, 8)</td>
<td>18</td>
<td>2874.9</td>
<td>-5711.7</td>
<td>-5611.3</td>
<td>40</td>
<td>0.0769</td>
</tr>
<tr>
<td><strong>(9, 9)</strong></td>
<td><strong>19</strong></td>
<td><strong>2899.1</strong></td>
<td><strong>-5758.1</strong></td>
<td><strong>-5652.4</strong></td>
<td><strong>40</strong></td>
<td><strong>0.0644</strong></td>
</tr>
<tr>
<td>(12, 6)</td>
<td>19</td>
<td>2883.1</td>
<td>-5726.3</td>
<td>-5620.6</td>
<td>40</td>
<td>0.443</td>
</tr>
</tbody>
</table>

The model ARMA(9, 9) has the lowest value of AIC and BIC equal -5758.1 and -5652.40 respectively. However the most parsimonious contender ARMA(6, 9) is the second best with BIC of -5634.2. Residuals of ARMA(9, 9) process met the independency test criterion consistently, for example, the Ljung–Box statistic aggregated up to 40 ($\sqrt{4 \times 365} \approx 38$) lags has a P-value 0.0644.

The standard conditions of stationarity and invertability of the ARMA(9, 9) model were verified via analysis of roots of characteristic polynomials of autoregressive and moving average parts of the model.

The NIG distribution was fitted to residuals of ARMA(9, 9) model with P-value 0.395 for the Kolmogorov-Smirnov test (‘kstest’ in Matlab). The ML estimates of parameters obtained in Matlab were $\mu = 0$, $\alpha = 21.5521$, $\beta = 2.1927$, $\delta = 0.0235$.

**Forecasting Demand for Electricity**

The forecasting of future values of demand was carried out within the sample. The predicted values of demand were computed from Friday, December 5, 2008 to Wednesday, December 24, 2008 using the ARMA(9, 9) model established for the detrended demand.
The NIG confidence bounds were constructed by drawing innovations from the NIG distribution. For comparison the confidence limits were also evaluated for the predicted values using the Normal innovations. The sampled values of $Z_t^{(i)}$ and their predictions $\hat{Z}_t^{(i)}$ introduced in (6) are reported in the first two columns of Table 3 with the weekend data highlighted. The data provided in Table 3 is displayed in Figure 2.

Table 3. True and predicted values of logarithms of daily average demand with NIG and Normal 95%-Confidence Limits (CLs).

<table>
<thead>
<tr>
<th>Date</th>
<th>True demand</th>
<th>Predicted demand</th>
<th>NIG Lower CL</th>
<th>NIG Upper CL</th>
<th>Normal Lower CL</th>
<th>Normal Upper CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-Dec-08</td>
<td>8.962</td>
<td>8.967</td>
<td>8.894</td>
<td>9.162</td>
<td>8.902</td>
<td>9.033</td>
</tr>
</tbody>
</table>

Figure 2. Graph of predicted values of logarithms of demand true future values, NIG and Normal 95%-Confidence Limits (CLs).

The empirical coverage analysis shows that the NIG confidence interval outperforms the Normal confidence interval. Over weekends the true future values have a tendency to exceed the nominal threshold of 5% set by the Normal Confidence Limits $\hat{Z}_t^{(i)} \pm 1.96\hat{\sigma}(h)$.
with the square root \( \hat{\sigma}(h) \) of the mean-squared prediction error of \( \hat{Z}_{t+h} \), see also Section 3.3, Chapter 3, [5].

### 4.2. Spot Prices of Electricity

Our model for the logarithm of the Spot Price of the electricity data, hereafter referred to as “price”, consists of a seasonal component, a spike component and a time series process of the remaining residuals. The trend component for price is

\[
j^2(t) = a^2 t + b^2 t + \sum_{i=1}^{k} \gamma^{2}(i) \sin \left( \frac{2\pi t}{365 s^{2}(i)} \right) + \delta^{2}(i) \cos \left( \frac{2\pi t}{365 s^{2}(i)} \right),
\]

where \( k \) and \( s_i \), \( i = 1, \ldots, k \) are unknown parameters. As in Section 4.1, spectral analysis carried out in Matlab using ‘periodogram’ function for price data revealed the most prominent seasons which are shown in Table 4.

<table>
<thead>
<tr>
<th>Period  ( 365/s^{2}(i) )</th>
<th>Approximate part of year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3333</td>
<td>1 and a third years</td>
</tr>
<tr>
<td>1.0000</td>
<td>Yearly</td>
</tr>
<tr>
<td>2.0000</td>
<td>2 yearly</td>
</tr>
<tr>
<td>0.0191</td>
<td>Weekly</td>
</tr>
<tr>
<td>0.2222</td>
<td>Weekly</td>
</tr>
<tr>
<td>0.8000</td>
<td>Weekly</td>
</tr>
<tr>
<td>0.4444</td>
<td>Weekly</td>
</tr>
<tr>
<td>0.5000</td>
<td>Half-yearly</td>
</tr>
<tr>
<td>0.6667</td>
<td>Two-thirds year</td>
</tr>
<tr>
<td>0.2000</td>
<td>Two-thirds year</td>
</tr>
</tbody>
</table>

The parameters of the seasonal component of price were estimated via multiple regression routine in Matlab. The value of the goodness-of-fit statistic, \( R^2 = 0.4142 \), is only an indication of the amount of variation in price that is explained by the seasonal model.

The general seasonality of the data has been captured by this model; however, it is clear, from Figure 3, that there are many unexplained spikes in the data. The detrended prices will henceforward be referred as \( X^{(2)}_t \).

**Dealing with Spikes**

The number of crossings of a high threshold by stationary Gaussian process asymptotically follow the Poisson process, Lindgren et al. [16]; the result has been extended to a general class of Markov processes, Borovkov and Last [4], including filtered Poisson processes. Intuitively, the upcrossings of sufficiently high threshold by the process \( X^{(2)}_t \) occur at exponentially distributed inter-arrival times and comprise upcrossings by the fPp, \( Z_t \),

\[
Z_t = \sum_{k=1}^{N(t)} \xi_k e^{-\lambda(t-t_k)}.
\]
We define the threshold level $h$ on the detrended data price $X_1^{(2)} - f_1^{(2)}$, defined in (1). Any value above this threshold will be considered a spike. The empirically selected threshold, $h = 0.7$, is shown in Figure 4.

Spikes are short lived and tend to last from one to a few days only. An individual spike or several consecutive spikes comprise a cluster. We identify $k = 1, 2, \ldots$ clusters of spikes in the series and record $\tau_k$, the time of occurrence of the maximum spike in cluster $k$, as well as $\xi_k$, the height (jump size) of this maximum spike in cluster $k$. We assume that an occurrence of spikes follows a homogenous Poisson process with rate $\nu = \nu(h)$, where

$$
\nu(h) = \frac{\text{The number of upcrossings}}{N} = \frac{34}{1461} = 0.0232, \quad (8)
$$

and $N$ is the sample size. The smoothing parameters $\lambda_k$, $k = 1, 2, \ldots$, are estimated at each $\tau_k$. These are the values which minimise the sums of squares of differences between the detrended prices and the fitted values. The maximum spike is the first and only spike identified in each cluster of spikes. In most cases spikes occur unexpectedly but decrease at
a slower rate. Both the height of the spike and the length of the cluster determine the rate of decay of the spike; the taller the spike, the faster the rate of decay, and the longer the cluster, the slower the rate of decay. There were two large spikes around consecutive days 741 and 742, and our process only detected the larger spike which occurred on day 742. Thus the first of the two remains a large outlier in the ‘post Poisson residuals’ data shown in Figure 4.

The Weibull distribution

\[ F_{\xi}(x) = 1 - e^{-\left(\frac{x-a}{b}\right)^{\alpha}}, \quad b \geq 0, \quad x \geq a, \]  

(9)

was fitted to the spikes data (‘wblfit’ in Matlab), \( \xi_k, \quad k = 1, 2, \ldots \), with the location, scaling and shape parameter estimates being \( a = 2.187, \quad b = 1.922 \) and \( \alpha = 1.966 \), respectively. The location parameter is biased due to presence of large spike on day 742. Using the goodness-of-fit test we obtained \( \chi^2 = 9.53 \) resulting in a P-value of 0.09. As seen from Figure 5, the Generalized Extreme Value (GEV) distribution applied to spikes data in the range \([1,3.5]\) showed a reasonable fit by the Gumbel distribution

\[ F(x) = \exp\left(-e^{-\left(\frac{x-a}{b}\right)}\right), \quad b \geq 0, \quad x \geq a, \]  

(10)

with adjusted location and scaling parameter estimates of \( a = 2.198, \quad b = 0.823 \) respectively. The reasonable fit to the tail of the spikes data by the Gumbel distribution, confirms the hypothesis that the data is likely to be drawn from Weibull distribution. Assuming \( \lambda_k = \lambda_0, \quad k = 1, \ldots, n \) the formula (A5) enables us to evaluate \( E(Z_i^{(2)}) \) which tends to a constant value as \( t \to \infty \), with \( Z_i \sim \text{Weibull}[a, b, \alpha] \), and

\[ E(\xi_i) = a + b\Gamma\left(1 + \frac{1}{\alpha}\right) = 3.891, \]

where \( \Gamma(x) \) is the Gamma function.

Figure 5. The QQ-plots for Weibull–(a) and GEV-(b) distributions fitted to spikes data and the tail of data respectively.
The “post Poisson residuals” introduced in (1), which resulted from detrending and “de-spiking” the price data will henceforward be referred to as $Y_t$.

**Analysis of ‘post Poisson’ Residuals of Price**

In this section we proceed with time series analysis of “post Poisson residuals” $Y_t$. We find these residuals from the detrended and de-spiked price data $Y_t$ are positively correlated with the detrended demand data $Z_t^{(1)}$, as seen in Figure 6.

Autoregression analysis on $Y_t$ with lagged exogeneous variable $Z_t^{(1)}$ was carried out

$$Y_t = \sum_{i=1}^q c_i Z_{t-i}^{(1)} + \sum_{i=1}^q d_i Y_{t-i} + \varepsilon_t.$$  

(11)

The first nine lags of both $Y_t$ and $Z_t^{(1)}$ were found to be significant, i.e. $m=n=9$. The inclusion of extra lags did not produce extra significant predictors, and the improvement on the goodness of fit, $R^2$, with extra lags, was minimal. The Ljung-Box test applied up to 45 lags confirmed that the residuals obtained from regression analysis are uncorrelated at the significance level of 5%. The NIG distribution fits the residuals and the P-value from the Kolmogorov-Smirnov test (“kstest” in Matlab) of the fit of this distribution is 0.5569. The values of parameter estimates of the NIG distribution, using the likelihood function developed in Matlab, are $\mu = -0.0454, \alpha = 4.7232, \beta = 1.1658, \delta = 0.1782$.

**4.3. Analysis of Bivariate Residuals**

In this section we provide a statistical analysis of distribution of the bivariate residuals, $\varepsilon_t = (\varepsilon_t^{(1)}, \varepsilon_t^{(2)})$, obtained from fitting the models defined in (1-3) to demand and spot prices of electricity. In Sections 4.1 and 4.2, the NIG distributions, $F_i(x), i=1,2$, were fitted to the marginal random variables, $\varepsilon_t^{(i)}$. Let $\eta = (\eta_1, \eta_2)$ be a bivariate NIG distributed random vector introduced in Definition A1 in the Appendix. Then the covariance matrix of $\eta$ is

$$E\left( (\eta - E\eta)(\eta - E\eta)^T \right) = Var(\tau)\Delta \beta \beta^T \Delta + E(\tau)\Delta = \frac{\delta}{\gamma} \left( \Delta + \frac{1}{\gamma^2} \Delta \beta \beta^T \Delta \right).$$  

(12)
where \( \gamma^2 = \alpha^2 - \beta^T \Delta \beta \). For any \( \mathbf{e} = (\varepsilon^{(1)}, \varepsilon^{(2)}) \) with NIG distributed marginals the parameterisation in the form (A1) may not exist. Indeed, in (12), the covariance \( \text{Cov}(\eta_1, \eta_2) > 0 \) due to \( \text{Var}(r) > 0 \) and (A2), thus making generation of independent random variables impossible in this setup. A discussion on modelling issues relating to the bivariate NIG process can be found in Chapter 5, [9]. To resolve this problem, Lillestøl [15] extended the model (A1) using two correlated mixing variables, \( \tau_1 \) and \( \tau_2 \).

The EM algorithm developed in [20] has been implemented in Matlab and used to fit the bivariate NIG distribution to residuals data. The parameter estimates, \( \mathbf{\mu}^T = (-0.0017, -0.05365), \alpha = 4.518, \beta^T = (-0.9284, 2.9953), \Lambda = ((1.326, 0.4707),(0.4707, 0.9212)) \) resulted in the Akaike Information Criterion, \( \text{AIC} = -5512 \).

The limitations of the bivariate NIG setup in the form (A1) suggest considering a copula approach. To capture the data pertinent tail-dependence we chose a \( t \)-copula, see [14]. The sample estimate of the Pearson's linear correlation coefficient \( \hat{\rho} = 0.45 \). Using "copulafit" function in Matlab the \( t \)-copula was fitted to transformed data \( \mathbf{U}_i = (F_1(\varepsilon_i^{(1)}), F_2(\varepsilon_i^{(2)})) \) with uniformly distributed marginals, where \( F_i(x), i=1,2 \) are the NIG distribution functions fitted to the demand and spot prices of electricity above. The estimates of correlation and degrees of freedom parameters of the \( t \)-copula are \( \hat{\rho}_t = 0.48, \nu = 6.96 \) respectively with \( \text{AIC} = -4023 \).

Fitting the Gaussian copula to the data resulted in \( \text{AIC} = -3852 \) with the correlation coefficient estimate \( \hat{\rho}_G = 0.47 \). However, in this setup the Gaussian copula was considered the least favourable distribution as the asymptotic tail dependence coefficient for the Gaussian copula is zero, [10]. According to the model selection Akaike criterion the bivariate NIG distribution outperformed both the Gaussian and \( t \)-copulae.

5. Conclusion

The objective of the paper to build a statistical model which captures seasonality and irregularities in electricity markets has been achieved by developing a multi-factor model for the bivariate variables of demand for electricity and spot prices of electricity. The model was applied to the regional Australian electricity markets’ data. In our model the demand for electricity is considered to be a primary variable forming a basis for demand-price hierarchy. The seasonal time series model with the NIG innovations was developed for the demand variable. In model validation purposes within the sample forecasting was carried out. We conclude the model can be used autonomously to forecast the future demand for electricity. In the multi-factor model, autoregression with the exogenous demand variable is used to fit detrended and “de-spiked” spot prices of electricity. These regression residuals followed NIG distribution. In addition this paper provides detailed statistical analysis of bivariate residuals of the multi-factor model fitted to bivariate data of demand and spot prices of electricity. The bivariate NIG distribution was found to be superior to the \( t \) - and Gaussian copulae in its ability to fit the data of residuals.

Acknowledgements

The authors are thankful to the editor and anonymous referees for their helpful suggestions and comments.

References


**Appendix**

**Definition A1.** [9, 21]. The bivariate vector \( \eta^T = (\eta_1, \eta_2) \)

\[
\eta = \mu + \tau \Delta \beta + \sqrt{\tau} \Delta^{1/2} W ,
\]

(A1)

follows a bivariate NIG distribution, where \( \mu, \beta \in \mathbb{R}^2 \), the matrix \( \Delta \) is a positive semidefinite symmetric matrix, \( W \) is a standard bivariate Gaussian vector, \( \tau \) has the inverse Gaussian distribution, i.e. \( \tau \sim IG(\delta^2, \alpha^2 - \beta^T \Delta \beta) \), \( \alpha > 0, \alpha^2 - \beta^T \Delta \beta > 0 \) and \( \tau \) is independent of \( W \).

The expected value and variance of \( \tau \sim IG(\delta^2, \alpha^2 - \beta^T \Delta \beta) \) are

\[
E(\tau) = \frac{\delta}{\gamma}, \quad Var(\tau) = \frac{\delta}{\gamma^2},
\]

(A2)

where \( \gamma = \alpha^2 - \beta^T \Delta \beta > 0 \). Note that, the conditional distribution of \( \eta \) given \( \tau = \tau_t \) is Gaussian \( \mathcal{N}(\mu + \tau \Delta \beta, \tau \Delta) \). The filtered Poisson process defined in (2.2) can be represented as a solution to the following stochastic differential equation, e.g. [18],

\[
dZ_t = -\lambda Z_t dt + dL_t ,
\]

(A3)

where \( Z_0 = 0 \) and the process \( L_t \) is a compound Poisson process with independent increments. Hence a stationary solution of (A3) can be presented in an explicit form

\[
Z_t = e^{-\lambda t} \int_0^t e^{\lambda s} dL_s .
\]

(A4)

**Lemma A1.**

\[
EZ_t = (1 - e^{-\lambda t}) \frac{\mu}{\lambda} m_t , \quad m_t = E\xi_t .
\]

(A5)

The compound Poisson process \( L_t , \ L_0 = 0 \) in (A4) can be represented as the sum of the martingale \( M_t \) and compensator \( A_t = Av_t, \ A_t = m_t \). As \( EM_t = EM_0 = 0 \) from (A4) we have

\[
EZ_t = Ee^{-\lambda t} \int_0^t e^{\lambda s} dL_s = Av e^{-\lambda t} \int_0^t e^{\lambda s} dL_s = (1 - e^{-\lambda t}) \frac{\mu}{\lambda} m_t .
\]
Authors’ Biographies:

**Hilary Green** has been a lecturer in the Department of Statistics at Macquarie University in Australia for more than 20 years. Her research interests in statistics span from development of 3D statistical graphics to the construction of multi-factor stochastic models with applications to Australian electricity markets and climate change problems.

**Nino Kordzakhia** is a lecturer at the Department of Statistics, Macquarie University, Australia. Nino’s research strength in Mathematical Finance is backed by her professional experience as a Quantitative Analyst in the financial industry. Her research interests also include asymptotic theory of estimators in Robust Statistics and development of numerical algorithms for pricing of financial derivatives.

**Ruben Thoplan** in Statistics at the University of Mauritius since August 2009. He holds a Master of Applied Statistics degree from the Macquarie University, Australia. He is presently research active and his research interest covers the areas of Data Mining, Forensic Statistics and Statistical Graphics.