State space model approach for forecasting the use of electrical energy (a case study on: PT. PLN (Persero) district of Kroya)

Devi Kurniati¹, Abdul Hoyyi², Tatik Widiarih³
¹,²,³Department of Statistics, Faculty of Science and Mathematics, Diponegoro University
E-mail: devikurniati21@gmail.com, hoyyistat@live.undip.ac.id, widiharigh@gmail.com

Abstract. Time series data is a series of data taken or measured based on observations at the same time interval. Time series data analysis is used to perform data analysis considering the effect of time. The purpose of time series analysis is to know the characteristics and patterns of a data and predict a data value in some future period based on data in the past. One of the forecasting methods used for time series data is the state space model. This study discusses the modeling and forecasting of electric energy consumption using the state space model for univariate data. The modeling stage is began with optimal Autoregressive (AR) order selection, determination of state vector through canonical correlation analysis, estimation of parameter, and forecasting. The result of this research shows that modeling of electric energy consumption using state space model of order 4 with Mean Absolute Percentage Error (MAPE) value 3.655%, so the model is very good forecasting category.

1. Introduction
In this globalization and modernization era, electricity becomes one of vital needs in human’s daily life. The development of science and technology has recently made human’s life easier in every aspect of life. Inevitably, almost every sector in human’s life cannot be separated from the usage of electrical energy. It is, then, used to run various electronical devices functioning to ease human’s works. An economic growth and society’s welfare as well as density are more increasing following the development of the age itself. As consequence, the necessity of electrical energy tends to increase from time to time.

Increased demand for electrical energy needs to be balanced with the supply of electrical energy from power plants, so that the distribution of it to consumers can run optimally. Bearing in mind that the importance of service on consumer and a limited number of energy resources, it is required such better plan and projection to optimize the electrical distribution toward society by adjusting the availability of this limited energy resources. A forecasting is a means that can be used as consideration in arranging any plan or projection in the usage of electrical energy by consumers.

The state space model represents a time series data through auxiliary variables, are called state vector, which summarize all the information from the present and past values of the time series relevant to the prediction of future values of the series [5]. This model is general, since it can be applied to all modelling of time series by Box Jenkins method, and it is flexible as it can be used in the data of either univariate or multivariate of time series.
State space modeling for univariate time series has been discussed in [1]. There is also other research on this modeling as in [4]. Based on above review, it is, thus, used state space model for modelling the usage of electrical energy at PT. PLN (Limited) of District of Kroya. After this model is gained, it is continued with the forecasting of the usage of electrical energy in the PT. PLN (Limited) of District of Kroya for 9 (nine) upcoming periods.

2. Literature Review

Time series analysis is an analysis of a set data over a period of time that is beneficial to measure or predict upcoming period [6]. The rationale of time series is recent observation ($Z_t$) depending on one or some ($k$) previous observations ($Z_{t-k}$).

Stationarity is an important requirement in the mathematical modelling for the data of time series. Further, the stationer data is any data that show the mean, variance and auto-variance (in the lag variation) remain constant in any time the data is formed or used, that is, with stationer data the time series model can be said stable. If there is any non-stationer data used in this model, the data can be reconsidered of its validity and stability. In order to know the stationarity in the mean of any data in the time series, it is, thus, tested by Augmented Dickey Fuller test (ADF) at $H_0 : \gamma = 0$ (un-stationer data) and $H_1 : \gamma < 0$ (stationer data).

Hence, the non-stationer data in the time series can be stationerized by differencing process at $d$ degrees. If the process of differencing is non-stationer until the second order ($d=2$), it can be used other transformation. A common transformation used is power transformation or it is usually called as the Box-Cox transformation. In this transformation, it is resulted the value of Lambda ($\lambda$), which determines transformation that has to be performed [7].

The state space model represents a time series data through auxiliary variables, are called state vector, which summarize all the information from the present and past values of the time series relevant to the prediction of future values of the series [5].

Based on Wei [7] and Akaike (1976) referred in [5], the state space model represents a stochastic process of stationered $Z_t$. It is defined as state transition equation:

$$Z_{t+1} = FZ_t + Ga_{t+1}; \quad t = 1,2,...,T$$

and output equation:

$$X_t = HZ_t \text{ or } X_t = [I_r, 0]Z_t$$

where:

- $X_t$ : is an observation vector of dimension $r$.
- $Z_t$ : is a state vector of dimension $s$.
- $F$ : is an $s \times s$ transition matrix.
- $G$ : is an $s \times r$ input matrix.
- $H$ : is an $r \times s$ output matrix.
- $a_t$ : randomly residual vector and normally distributed by $r$ dimension with average vector of 0 and $\Sigma$ of covariant matrix.
- $r$ : number of observable variable
- $s$ : number of significant variable based on analysis of canonic correlation
- $n$ : number of observation (a length of time series)

State vector, is formed through canonical correlation analysis of data space $D_t = (Z_t^T, Z_{t-1}^T, \ldots, Z_{t-p}^T)^T$ and predictor space $F_t = (Z_t^T, Z_{t+1}^T, \ldots, Z_{t+p+1}^T)^T$. Variable with real canonical correlation is added into state vector. Meanwhile, the unreal variable is excluded from state vector [5]. For each process of a series of analysis of canonic correlation, it is then tested by significant test at the smallest canonic correlation of $\rho_{min}$, by hypothesis that $H_0 : \rho_{min} = 0$ ($\rho_{min}$ is
not significantly different with zero) and $H_1 : \rho_{\text{min}} \neq 0$ ($\rho_{\text{min}}$ is significantly different with zero) with statistical test of $\chi^2_{\text{count}}$.

For a series of observation at $Z_1, Z_2, \ldots, Z_n$, with $B$ as backshift operator and $E$ is residual matrix by $r \times n$ of dimension,

$$Z_t = (I - FB)^{-1} G a_t,$$
$$X_t = H(I - FB)^{-1} G a_t,$$
$$a_t = [H(I - FB)^{-1} G^{-1}]^t X_t.$$

The parameter measurement are done by Maximum Likelihood Estimation approach (MLE). The function of ln-likelihood formed is:\[ L(\theta) = -\frac{n}{2} \ln|\Sigma| - \text{trace} (\Sigma^{-1} EE^t) \]

The first partial derivation of the function of ln-likelihood estimated is non-linear function, so it requires iterated method to obtain its parameter estimation, which one of it is Newton-Raphson method.

Analysis of time series assumes that a residual resulted by this model should follow the White Noise process. This process is a stationer process where $a_t \sim \text{NID}(0, \sigma_a^2)$, meaning that the residual is distributed independently normal by mean of 0 and constant variant. To test whether or not time series follow the White Noise process, it is then tested by Kolmodorov-Smirnov test to examine residual’s normality and Ljung-Box test for residual’s independency.

For the forecasting user, any precision of forecasting result for the future is highly important. One of statistical measures that can be used as precision measurement of forecasting is mean absolute percentage error, or MAPE \[3\].

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{Z_t - F_t}{Z_t} \right| \times 100$$

where $Z_t$ is actual data for $t$ period and $F_t$ is forecasting (fitted value) for similar period.

### 3. Methodology

The data used in this research is secondary data of the monthly report of electrical energy selling of PT. PLN (persero) of District of Kroya on January 2007 - January 2016. It is 109 periods, comprising of 100 periods of training data and 9 periods of testing data.

A modelling of the usage of electrical energy is performed with SAS 9.1.3 software and supported with MINITAB 17 and EVIEWS 9. The steps of data analysis for this modelling of the usage of electrical energy in the PT. PLN (persero) of District of Kroya is began with stationary test and followed with a selection of optimal autoregressive (AR) order ($p$). The optimal AR order obtained is determined the selection process of state vector by canonical correlation analysis. After the state vector is formed, then estimation of parameter is performed to gain equation model. The next step is performing significant parameter test and diagnostic model test to observe whether the model is qualified or not for forecasting. The accuracy of this forecasting result can be viewed based on the value of out-sample MAPE gained.

### 4. Result and Discussion

The usage of electrical energy in the District of Kroya seen in Figure 2 fluctuates in each period. It is indicated by increasing tendency, following the growing period of time. The lowest usage of electrical energy was occured in March 2007, which was 7.982315 million kWh, and the highest was occured in March 2015 (16.916525 million kWh). Surprisingly, the extreme usage of electrical energy was occurred on January to February 2013, where there was 12.181961 million kWh of electrical usage on
January and it turned as 15.582159 million kWh on February. In the following month, there was an extreme decreasing as 9.982972 million kWh on March 2013.

The test of stationarity in the mean of the data usage of electrical energy in the PT. PLN (persero) of District of Kroya by ADF test showed that the data is non-stationer by the value of statistical test of 4.686 and p-value of 1.00. Then, the process of two-times differencing process was performed, until the stationer data was gained. The result of ADF test on the data of second differencing process was gained the value of statistical of -5.961 and p-value of 0.00.

Moreover, the selection of optimal AR order was conducted by viewing the smallest value of AIC on the AR model, which was the order of 4 \((p=4)\) and the AIC value was 2652.314. Based on the optimal order \((p=4)\), analysis of canonic correlation was done between data space \((Z_t, Z_{t-1}, ..., Z_{t-4})\) and predictor space \((Z_t, Z_{t+1|t}, ..., Z_{t+4|t})\). Variable, possibly as component of state vector, was \(Z_{n+1|n}\) up to \(Z_{n+4|n}\), but it was not all \(Z_{n+i|n}\) for \(i = 1, 2, 3, 4\) as the component of state vector.

**Table 1. Analysis of Canonic Correlation**

| State Vector | Correlation | IC | \(\chi^2\) | db | \(\chi^2_{db,0.05}\) |
|--------------|-------------|----|-----------|----|----------------|
| \(Z_t, Z_{t+1|t}\) | 1 0.560 | 28.841 | 36.090 | 4 | 9.488* |
| \(Z_t, Z_{t+1|t}, Z_{t+2|t}\) | 1 0.573 0.384 | 9.657 | 15.417 | 3 | 7.815* |
| \(Z_t, Z_{t+1|t}, Z_{t+2|t}, Z_{t+3|t}\) | 1 0.648 0.414 0.145 | -1.912 | 2.066 | 2 | 5.991 |

*significant

According to significant test of analysis of canonic correlation on the Table 1, it was obtained a significant component of state vector, which was \(Z_t, Z_{t+1|t}\), and \(Z_{t+2|t}\). In addition, the last component of state vector can be noted down, as follows:

\[
\begin{bmatrix}
Z_t \\
Z_{t+1|t} \\
Z_{t+2|t}
\end{bmatrix}
\]

While, parameter estimation was performed with Maximum Likelihood Estimation (MLE) method and the equation of state space model was gained, as follows:
\[
\begin{bmatrix}
Z_{t+1|t+1} \\
Z_{t+2|t+1} \\
Z_{t+3|t+1}
\end{bmatrix}
= 
\begin{bmatrix}
0 & 1 & 0 \\
0 & 0 & 1 \\
-0.05724 & -0.12838 & -0.494
\end{bmatrix}
\begin{bmatrix}
Z_{t+1|t} \\
Z_{t+2|t} \\
Z_{t+3|t}
\end{bmatrix}
+ 
\begin{bmatrix}
1 \\
-0.26675 \\
-0.62108
\end{bmatrix} \sigma_{t+1}
\]

where \(\sigma_{t+1} = 0.50564\).

After parameter estimation was gained, the significant test was performed and model parameter was tested by following hypothesis:

\(H_0: \theta_j = 0\) (non-significant parameter)

\(H_1: \theta_j \neq 0\) (significant parameter)

**Table 2. Significance of Parameter in the State Space Model**

| Parameter   | Estimation | Error Standard | t-value |
|-------------|------------|----------------|---------|
| \(F (3,1) = \phi_3\) | -0.05724 | 0.148927 | -0.38 |
| \(F (3,2) = \phi_2\) | -0.12838 | 0.193952 | -0.66 |
| \(F (3,3) = \phi_1\) | -0.49400 | 0.200587 | -2.46* |
| \(G (2,1) = \psi_1\) | -0.26675 | 0.100998 | -2.64* |
| \(G (3,1) = \psi_2\) | -0.62108 | 0.100958 | -6.15* |

*significant

The significant parameter estimation was parameter having the value of \(|t| > 1.984\). On the Table 2, the significant parameter estimation was \(\phi_1, \psi_1, \) and \(\psi_2\), whereas \(\phi_3, \) and \(\phi_2, \) were non-significant since \(|t| < 1.984\). Hence, assumption of on-significant parameter is still used in the model, since the value remains constant providing the effect of forecasting result [4].

Kolmogorov-Smirnov test showed that residual was distributed normally, where the value of statistical D test was 0.088215 and p-value was 0.0599. Similarly, in the Ljung-Box test, it showed that there was no residual correlation of inter-lag, where the value of statistical test on the lag Q was 30.196 and p-value was 0.456. Therefore, according to Kolmogorov-Smirnov and Ljung-Box test, it can be concluded that residual follows the White Noise process.

The forecasting result of the usage of electrical energy in the PT. PLN (Limited) District of Kroya in the 9 (nine) upcoming periods tends to increase. Comparison of forecasting values and actual values of the usage of electrical energy can be seen in Table 3:

**Table 3. Value of forecasting and actual data of the usage of electrical energy in the PT. PLN (persero) Rayon Kroya**

| Period     | Actual | Forecasting | APE |
|------------|--------|-------------|-----|
| May 2015   | 15,4037 | 15,6444 | 1,5624 |
| June 2015  | 14,9489 | 15,2536 | 2,0383 |
| July 2015  | 17,8727 | 15,8140 | 11,5185 |
| August 2015| 15,0211 | 15,4488 | 2,8474 |
| September 2015 | 16,3622 | 15,9270 | 2,6599 |
| October 2015 | 16,2900 | 15,6096 | 4,2296 |
| November 2015 | 15,5713 | 16,0732 | 3,2232 |
| December 2015 | 16,0488 | 15,7616 | 1,7895 |
This forecasting result was obtained the value of out-sample MAPE as 3.6548%. It is classified as very good according to (Chang, Wang, & Liu, 2007) in [2] that is MAPE < 10% showing a better forecasting ability.

Figure 2. The Plot of Actual and Forecasting Value of The Usage of Electrical Energy in The PT. PLN (Limited) of District of Kroya on January 2007-January 2016

5. Conclusion
The projecting result of the usage of electrical energy in the PT. PLN (Limited) of District of Kroya for the 9 (nine) upcoming periods tends to increase every time period. This Forecasting result can be classified as very good, with the value of out-sample MAPE obtained was 3.6548%.

References
[1] Hafli, M.R. 2008. “Model State Space untuk Data Deret Waktu Peubah Tunggal”. Bogor: Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor.
[2] Halimi, R. Anggraeni, W. dan Tyasnurita, R. 2013. “Pembuatan Aplikasi Peramalan Jumlah Permintaan Produk dengan Metode Time Series Exponential Smoothing Holts Winter di PT. Telekomunikasi Indonesia Tbk”. Jurnal Teknik POMITS 1(1), 1-6.
[3] Makridakis, S. Wheelwright, S.C. dan McGee, V.E. 1999. Metode dan Aplikasi Peramalan. Edisi 2 Jilid 1. Jakarta: Penerbit Erlangga.
[4] Nurlaila, I. 2007. “Penerapan Model ARIMA dan Model State Space pada Indeks Harga Saham Gabungan dan Nilai Tukar Rupiah terhadap Dolar Amerika”. Bogor: Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor.
[5] SAS Institute Inc. 2003. SAS/ETS User’s Guide, Version 9.1. Cary, NC, USA: SAS Institute Inc. hal 997-1047.
[6] Soejoeti, Z. 1987. Materi Pokok Analisis Runtun Waktu. Jakarta: Karunika.
[7] Wei, W.W.S. 2006. Time Series Analysis: Univariate and Multivariate Methods. Edisi 2. Pearson Education Inc. USA. hal 463 – 487.