Senior Project Proposal 2102490 Year 2017

Recursive Estimation of Solar Forecasting at Chulalongkorn University

Tony Fang ID 5730212721

Advisor: Assist. Prof. Jitkomut Songsiri

Department of Electrical Engineering, Faculty of Engineering

Chulalongkorn University

Contents

1 Introduction

There are many methods used to forecast the solar irradiance. This study considers two methods: time-series model and Weather Research Forecasting (WRF) model.

1. Time series model
   Time series model is a model which is used to analyze the solar irradiance in the past and forecast the future solar irradiance. Moreover, this model is typically used in many fields such as engineering and science. In this study is mainly focused on time series model. A time-series model can be divided into 2 types which are a time-series model used in stationary process and a time-series model used in non-stationary process

   (a) Time series model that include stationary properties is the time series model which have a constant mean and autocorrelation depending on time gap for example AR, MA, ARMA, ARX, ARMAX

   (b) Time series model that include non-stationary properties are the time series model which either mean or autocorrelation depending on time example ARIMA, ARIMAX

   This report uses a time series model with non-stationary process for describing a trend of the solar irradiance which according to [4].

2. Weather Research Forecasting (WRF) is a regional forecasting model. WRF is a type of Numerical Weather Prediction(NWP) which is a method of weather forecasting that uses partial differential equation which is derived from a characteristic of meteorological parameters such as temperature, wind speed,relative humidity.

When the new solar irradiance data are measured, The model must be estimated by using both historical data and the new measured data. The model estimation sometimes spends too much time on an online implementation. A recursive identification method should be applied for model estimation. One advantage of this method is that this model uses only the new data and the previous estimated model. Thus, this model is compatible for an online implementation because of its estimation speed. Typically, the recursive method is applied in model estimation such as least-square method (LS), maximum likelihood estimation (ML), and predicton error method (PEM).

2 Objectives

The objectives of this study are

1. To apply recursive identification of time series model to forecast solar irradiance at Chulalongkorn University.

2. To implement the forecasting scheme for EE-CU solar forecasting center.

3 Background on Model Estimation and Model Validation

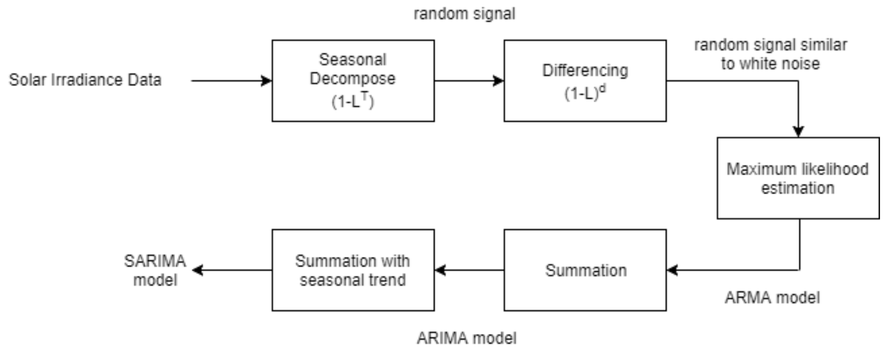


Figure 1: Model Estimation Flowchart

This section describes the time-series model estimation of which a flowchart is shown in Figure 1. First, a seasonal trend must be removed from the solar irradiance data. This means we could write the seasonal trend into fourier series. After the seasonal trend is removed, the data will be similar to the random signal. Before finding the parameters in ARMA model, we do the autocorrelation test. If the autocorrelation graph is not similar to white noise, we could differentiate the data and check the autocorrelation again. After differencing until the autocorrelation graph is similar to white noise, we use that set of differencing data to find the parameter in ARMA model by using maximum likelihood estimation (ML). Then this model use prediction error method (PEM) to find the estimated model and the estimated model is then used to forecast solar irradiance data.

From Figure 2, the solar irradiance data can be decomposed to the seasonal component and a random component. That means we can use same model to estimate GHI.

According to [4], the result shows that SARIMA$(2,2,4)(0,1,1)_{16}$ achieve the best performance to forecast the solar irradiance. Thus, this study uses Seasonal AutoRegressive Integrated Moving Average (SARIMA) model because this model is a general form of time-series model. SARIMA model is defined by.

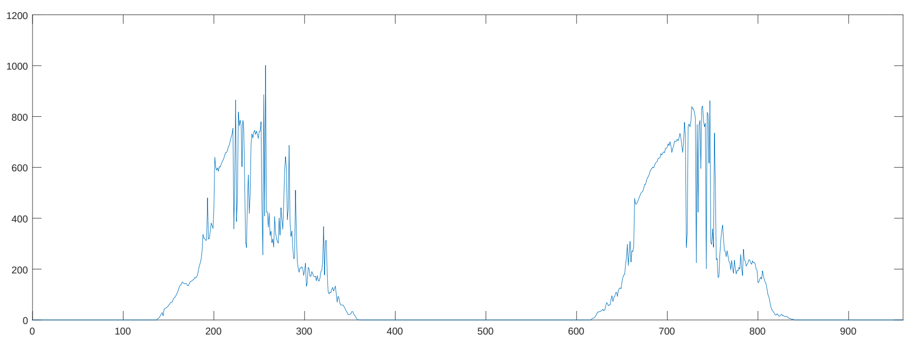


Figure 2: GHI in 1st-2nd January 2017

$$\tilde{A}(L)(1-L^T)^D A(L)(1-L)^d I(t) = \tilde{C}(L)C(L)v(t) \tag{1}$$

From (1) can be written as SARIMA$(p,d,q)(P,D,Q)_T$ where

$$\begin{aligned} A(L) =& 1-(A_1L + A_2L^2 + \cdots + A_pL^p) \\ C(L) =& 1+(C_1L + C_2L^2 + \cdots + C_qL^q) \\ \tilde{A}(L) =& 1-(\tilde{A}_1L^T + \tilde{A}_2L^{2T} + \cdots + \tilde{A}_PL^{PT}) \\ \tilde{C}(L) =& 1+(\tilde{C}_1L^T + \tilde{C}_2L^{2T} + \cdots + \tilde{C}_QL^{QT}) \end{aligned} \tag{2}$$

L is a lag operator, T is a seasonal period, d is an integrated non-seasonal order, D is an integrated seasonal order, $A(L)$ is AutoRegressive (AR) polynomial, $C(L)$ is Moving Average (MA) polynomial, $\tilde{A}(L)$ is Seasonal AutoRegressive (SAR) polynomial and $\tilde{C}(L)$ is Seasonal Moving Average (SMA) polynomial.
From Figure 2 considers the solar irradiance data as an additive seasonal trend which is written as 3

$$A(L)I(t) = s(t) + \alpha + C(L)e(t) \tag{3}$$

where $s(t) = s(t-T)$ is a seasonal component, α is constant and $e(t)$ is a noise. According to Figure 1, given that $I(t)$ is solar irradiance data. Then $I(t)$ is subtracted by $I(t-T)$ for doing a seasonal decompose. We get

$$\begin{aligned}
A(L)I(t) - A(L)I(t-T) &= s(t) + \alpha + C(L)e(t) - s(t-T) - \alpha - C(L)e(t-T) \\
&= C(L)e(t) - C(L)e(t-T) \\
A(L)(1-L^T)I(t) &= C(L)(1-L^T)e(t) \qquad (4) \\
A(L)(1-L^T)I(t) &= C(L)\tilde{C}(t)e(t) \qquad (5)
\end{aligned}$$

Substitute $(1-L^T)I(t) = z(t)$ and $\tilde{C}(t)e(t) = \eta(t)$. Then differencing $z(t)$ for d time until autocorrelation function is similar to white noise. We get

$$\begin{aligned}
A(L)(1-L)^d z(t) &= C(L)(1-L)^d \eta(t) \\
A(L)(1-L)^d z(t) &= C(L)v(t) \qquad (6)
\end{aligned}$$

Substitute $y(t) = (1-L)^d z(t)$. We get

$$A(L)y(t) = C(L)v(t) \tag{7}$$

After using ML to find the parameter in ARMA model, we can find the estimate ARMA model in 8.

$$\begin{aligned}
C(L)\hat{y}(t|t-1) &= (C(L) - A(L))y(t) \\
(C(L)-1)\hat{y}(t|t-1) + \hat{y}(t|t-1) &= (C(L)-1)y(t) - (A(L)-1)y(t) \\
\hat{y}(t|t-1) &= (C(L)-1)(y(t) - \hat{y}(t|t-1)) - (A(L)-1)y(t) \\
\hat{y}(t|t-1) &= (C(L)-1)e(t) - (A(L)-1)y(t) \\
\hat{y}(t|t-1) &= (C_1 L + C_2 L^2 + \cdots + C_q L^q)e(t) + \\
&\quad (A_1 L + A_2 L^2 + \cdots + A_p L^p)y(t) \qquad (8)
\end{aligned}$$

Then we can find the estimate ARIMA model in 9.

$$\begin{aligned}
\hat{z}(t|t-1) &= \hat{y}(t|t-1) + \hat{z}(t-1|t-1) \\
\hat{z}(t|t-1) &= \hat{y}(t|t-1) + \hat{y}(t-1|t-1) + \hat{z}(t-2|t-1) \\
\hat{z}(t|t-1) &= \hat{y}(t|t-1) + \hat{y}(t-1|t-1) + \hat{y}(t-2|t-1) + \hat{z}(t-3|t-1) \\
\hat{z}(t|t-1) &= \hat{y}(t|t-1) + \sum_{k=1}^{\infty} \hat{y}(t-k|t-1) \qquad (9)
\end{aligned}$$

Moreover, we can find the estimated SARIMA model in 10.

$$\begin{aligned}
\hat{I}(t|t-1) &= \hat{z}(t|t-1) + \hat{I}(t-T|t-1) \\
\hat{I}(t|t-1) &= \hat{z}(t|t-1) + \hat{z}(t-T|t-1) + \hat{I}(t-2T|t-1) \\
\hat{I}(t|t-1) &= \hat{z}(t|t-1) + \hat{z}(t-T|t-1) + \hat{z}(t-2T|t-1) + \hat{I}(t-3T|t-1) \\
\hat{I}(t|t-1) &= \hat{z}(t|t-1) + \sum_{k=1}^{\infty} \hat{z}(t-kT|t-1) \qquad (10)
\end{aligned}$$

3.1 Fitting seasonal trend

This section is describes the finding of the seasonal trend and this seasonal trend is used to remove from the data. The aim of this section is making time series data stationary and find seasonal period. From [2], the seasonal trend can represent into fourier series

$$s(t) = \sum_{i=1}^{k} (a_i \cos(\omega_i t) + b_i \sin(\omega_i t)) \tag{11}$$

where $t = 1, 2, \cdots, N$, a_i is the coefficient of co-sinusoidal component of each frequency ω_i and b_i is the coefficient of sinusoidal component of each frequency ω_i.To find the frequency ω_i, the data must be transformed to frequency domain to find the power spectral density. Then we choose the high energy frequency to be selected as ω_i. Before finding the power spectral density, we use Fast Fourier Transform (FFT) to transform to the frequency domain. According to [3], Fast Fourier Transform is an algorithm to find the discrete fourier transform (DFT) which show in (12)

$$S(k) = \sum_{t=0}^{N-1} s(t) e^{\frac{j2\pi kt}{N}} \tag{12}$$

where $\omega_k = \frac{2\pi k}{N}$ and $k = 0, 1, \cdots, N-1$. After using FFT, we find $|S(k)|$ to find ω_k. Only high-energy frequency will given to ω_i in (11).

3.2 Estimation of integrated part

This section is describes a process to find the integrated order d. After removing seasonal trend, ACF maybe slowly decay. From (5) which are already substitute,

$$A(L)z(t) = C(L)e(t)$$

The data $y(t)$ was differentiated by subtracting $y(t-1)$.

$$\Delta A(L)z(t) = A(L)z(t) - A(L)z(t-1) = A(L)(1-L)z(t)$$

We can also differencing 2$^{\text{nd}}$ time

$$\Delta^2 A(L)z(t) = \Delta A(L)z(t) - \Delta A(L)z(t-1) = A(L)(1-L)^2 z(t)$$

If we differencing the data in d time, we will write in equation

$$\Delta^d A(L)z(t) = A(L)(1-L)^d z(t) \tag{13}$$

If an autocorrelation function is cut-off at some lags after differencing d times, we can conclude that this model has an integrated order d.

$$A(L)(1-L)^d z(t) = C(L)v(t) \tag{14}$$

3.3 Estimation of ARMA model

This section is describes the estimation of AutoRegressive Moving Average(ARMA) model. In this report we will use maximum likelihood estimation to find the parameter of AutoRegressive polynomial and Moving Average polynomial in (2).
Maximum likelihood estimation is one of the method to find the parameter by maximizing a cost function which is defined by

$$\mathcal{L}(y|\theta) = f(y(1), y(2), \cdots, y(N)|\theta) \tag{15}$$

where $f(y(1), y(2), \cdots, y(N)|\theta)$ is defined by

$$f(y(1), y(2), \cdots, y(N)|\theta) = \prod_{t=1}^{N} \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big) e^{-\frac{v(t)^2}{2\sigma^2}} \tag{16}$$

In (15), this cost function also called likelihood function where

$$\theta = \begin{bmatrix} A_1 & A_2 & \cdots & A_p & C_1 & C_2 & \cdots & C_q & \sigma^2 \end{bmatrix}^T$$

$f(y|\theta)$ is conditional probabilitiy density function (conditional pdf) of $v(t)$ in (1) and N is a number of data.
From 7 can also write

$$y(t) = A_1y(t-1)+A_2y(t-2)+\cdots+A_py(t-p)+v(t)+C_1v(t-1)+C_2v(t-2)+\cdots+C_qv(t-q) \tag{17}$$

Thus, we can find $v(t)$ in term of $A_1, A_2, \cdots, A_p, C_1, C_2, \cdots, C_q$ from (17)

$$v(t) = y(t)-(A_1y(t-1)+A_2y(t-2)+\cdots+A_py(t-p))-(C_1v(t-1)+C_2v(t-2)+\cdots+C_qv(t-q)) \tag{18}$$

If $v(t)$ has normal distribution which has zero mean and variance σ^2. Then the log-likelihood function according to [1] is

$$\mathcal{L}(\theta) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \sum_{t=1}^{N}\frac{v(t)^2}{2\sigma^2} \tag{19}$$

From [1], if $y(t)$ has real value from 1 to p and $v(t) = 0$ since $t = p, p-1, \cdots, p-q+1$, so that $y(t)$ also has normal distribution. Thus, we start at $t = p+1$. At the same time, the conditional likelihood function is change to (20)

$$\mathcal{L}(y|\theta) = f(y(p+1), y(p+2), \cdots, y(N)|y(1), y(2), \cdots, y(p), \theta) \tag{20}$$

Thus, the likelihood function is

$$\mathcal{L}(y|\theta) = (\frac{1}{2\pi\sigma^2})^{N-p}e^{-\sum_{t=p+1}^{N}\frac{v(t)^2}{2\sigma^2}} \tag{21}$$

From the maximum likelihood estimation, we can find the estimator from log-likelihood function. Finally, a cost function to find the estimator from the maximum of the cost function in (22)

$$\log\mathcal{L}(y|\theta) = -\frac{N-p}{2}\log(2\pi) - \frac{N-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{N}\frac{v(t)^2}{2\sigma^2} \tag{22}$$

where $\log\mathcal{L}(y|\theta)$ is the cost function of the problem and the last term in (22) can be considerd in 2-norm. Thus we can write into matrix form

$$\begin{bmatrix} v(p+1) \\ v(p+2) \\ \vdots \\ v(N) \end{bmatrix} =$$

$$\begin{bmatrix} y(p+1) \\ y(p+2) \\ \vdots \\ y(N) \end{bmatrix} - \begin{bmatrix} y(p) & y(p-1) & \cdots & y(1) & v(p) & v(p-1) & \cdots & v(p-q+1) \\ y(p+1) & y(p) & \cdots & y(2) & v(p+1) & v(p) & \cdots & v(p-q+2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y(N) & y(N-1) & \cdots & y(p) & v(N) & v(N-1) & \cdots & v(p) \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_p \\ C_1 \\ C_2 \\ \vdots \\ C_q \end{bmatrix}$$

The numerical solution of $\hat{\theta}_{ML}$ can find from many optimization method example steepest-descent, quasi-newton, conjugate-gradient and the other method. However, doing ML in MATLAB and doing ML by hand-out probably not given the same parameter especially the parameter in seasonal moving average polynomial because the simple estimation in 4 show that it has pole lie in unit circle.

3.4 Computation of Forecasting

After finding the model, it will be used to forecast the solar irradiance in the next h-step. From the ARMA model in 7.

$$A(L)y(t) = C(L)v(t)$$

According to [5] and[6], the optimal prediction from the ARMA model by using PEM is

$$\hat{y}(t|t-1) = (1 - C^{-1}(L)A(L))y(t) \tag{23a}$$

$$e(t) = C^{-1}(L)A(L)y(t) \tag{23b}$$

We can find the estimated ARMA(p, q) model

$$\begin{aligned} C(L)\hat{y}(t|t-1) =& (C(L) - A(L))y(t) \\ (C(L) - 1)\hat{y}(t|t-1) + \hat{y}(t|t-1) =& (C(L) - 1)y(t) - (A(L) - 1)y(t) \\ \hat{y}(t|t-1) =& (C(L) - 1)(y(t) - \hat{y}(t|t-1)) - (A(L) - 1)y(t) \\ \hat{y}(t|t-1) =& (C(L) - 1)e(t) - (A(L) - 1)y(t) \\ \hat{y}(t|t-1) =& (C_1L + C_2L^2 + \cdots + C_qL^q)e(t) + \\ & (A_1L + A_2L^2 + \cdots + A_pL^p)y(t) \end{aligned} \tag{24}$$

So that we can compute one-step ahead prediction of ARMA model

$$\begin{aligned} \hat{y}(t+1|t) =& (C(L) - 1)e(t+1) + (1 - A(L))\hat{y}(t+1|t) \\ =& (C_1L + C_2L^2 + \cdots + C_qL^q)e(t+1) + (A_1L + A_2L^2 + \cdots + A_pL^p)\hat{y}(t+1|t) \\ =& C_1e(t) + C_2e(t-1) + C_3e(t-2) + \cdots + C_qe(t-q+1) + \\ & A_1\hat{y}(t|t) + A_2\hat{y}(t-1|t) + \cdots + A_p\hat{y}(t-p+1|t) \end{aligned} \tag{25}$$

And we can compute h-step prediction of ARMA model

$$\begin{aligned} \hat{y}(t+h|t) =& (C(L) - 1)e(t+h) + (1 - A(L))\hat{y}(t+h|t) \\ \hat{y}(t+h|t) =& C_1e(t+h-1) + C_2e(t+h-2) + C_3e(t+h-3) + \cdots + C_qe(t+h-q) + \\ & A_1\hat{y}(t+h-1|t) + A_2\hat{y}(t+h-2|t) + \cdots + A_p\hat{y}(t+h-p|t) \end{aligned} \tag{26}$$

where

$$\hat{y}(t+h|t) = \begin{cases} \hat{y}(t+h|t) & t > 0 \\ y(t+h) & t \leq 0 \end{cases} \tag{27}$$

$$e(t+h|t) = \begin{cases} 0 & t > 0 \\ e(t+h) & t \leq 0 \end{cases} \tag{28}$$

So that we can compute h-step prediction of ARIMA$(p, 1, q)$ model

$$\hat{z}(t+h|t) = \hat{y}(t+h|t) + \sum_{k=0}^{\infty} \hat{y}(t-k+h-1|t) \tag{29}$$

Also we can compute h-step prediction of SARIMA$(p, 1, q)(0, 1, 1)_T$ model

$$\hat{I}(t+h|t) = \hat{z}(t+h|t) + \sum_{k=1}^{\infty} \hat{z}(t-kT+h|t) \tag{30}$$

MATLAB have command to forecast the h-step prediction after estimated the model. There are

1. `Infer`
   Infer gives the residual error and conditional variance from the data which we use. Then the fitted numerical value can find from the different between the data which use in this command and residual error.

2. `Forecast`
   Forecast give the predicted value from the estimated model and the data by using PEM.

3.5 Model Selection

After we specify some properties of the models. we have to consider some criterior score to find the optimal order of SARIMA models. This study uses Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Both AIC and BIC explain a trade-off between a complexity of the model and goodness of fit. Both AIC and BIC are defined by

$$\mathrm{AIC} = -2\mathcal{L} + 2d \tag{31}$$
$$\mathrm{BIC} = -2\mathcal{L} + d\log(N) \tag{32}$$

where $\mathcal{L}$ is a log-likelihood function, N is a number of data and d is a number of parameter in each models.

4 Recursive Prediction Error Method (RPEM)

After finding the optimal prediction, we use recursive identification method to adapt the model with a new data. The following criterion which we consider as cost function is in (33)

$$V_t(\theta) = \frac{1}{2}\sum_{s=1}^{t} \lambda^{t-s}\varepsilon(s,\theta)^T Q\varepsilon(s,\theta) \tag{33}$$

where λ is a forgetting factor which has a value between 0 and 1, Q is a positive definite weight matrix and $\varepsilon(s,\theta)$ is a residual error.

From an offline estimation, $\hat{\theta}(t)$ cannot be found analytically except in recursive least-square. From [6], assume that

- $\hat{\theta}(t-1)$ minimize $V_t(\theta)$
- minimum point of cost function $V_t(\theta)$ is close to $\hat{\theta}(t-1)$

Then approximated the cost function $V_t(\theta)$ by the second order taylor series expansion around $\hat{\theta}(t-1)$

$$V_t(\theta) \approx V_t(\hat{\theta}(t-1)) + \nabla V_t(\theta(t-1))^T(\theta - \hat{\theta}(t-1)) + \frac{1}{2}(\theta - \hat{\theta}(t-1))^T\nabla^2 V_t(\theta(t-1))(\theta - \hat{\theta}(t-1)) \tag{34}$$

From assumption, we get the parameter $\hat{\theta}(t)$ by given $\theta = \hat{\theta}(t)$. Thus, we will get an updated formula in (35)

$$\hat{\theta}(t) = \hat{\theta}(t-1) - 2\nabla^2 V_t(\theta(t-1))\nabla V_t(\theta(t-1)) \tag{35}$$

From (35), it shows that the updated formula of estimator is similar to Newton-Raphson step. In this study we might skip to the final step of an update formula of this method. According to [7], the final step of an update formula which give forgetting factor is equal to 1 and Q is identity matrix are shown in

$$\hat{\theta}(t) = \hat{\theta}(t-1) + P(t)K(t)\varepsilon(t) \tag{36}$$
$$g(t) = P(t)K(t)[1 + K^T P(t-1)K(t)]^{-1} \tag{37}$$
$$P(t) = P(t-1) - P(t-1)K^T(t)g(t) \tag{38}$$

where $\varepsilon(t) = y(t) - \hat{y}(t|t-1)$, $K(t) = -\nabla\varepsilon(t)$

5 Preliminary results

This section is describes the estimation of SARIMA model with solar irradiance data at Chulalongkorn University. According to [4], because the data come from different source, the estimated model is different in term of order of each polynomial and integrated order but uses SARIMA model to forecast solar irradiance. In this study we select select the optimal model by using AIC and BIC. The data used for training is GHI from 1$^{\text{st}}$ January to 31$^{\text{st}}$ May 2017 at 6.00 AM to 6.00 PM.

5.1 Fitting seasonal trend

The aim of finding seasonal trend is finding the seasonal period T and removing the seasonal trend to make the time series data have stationary process.

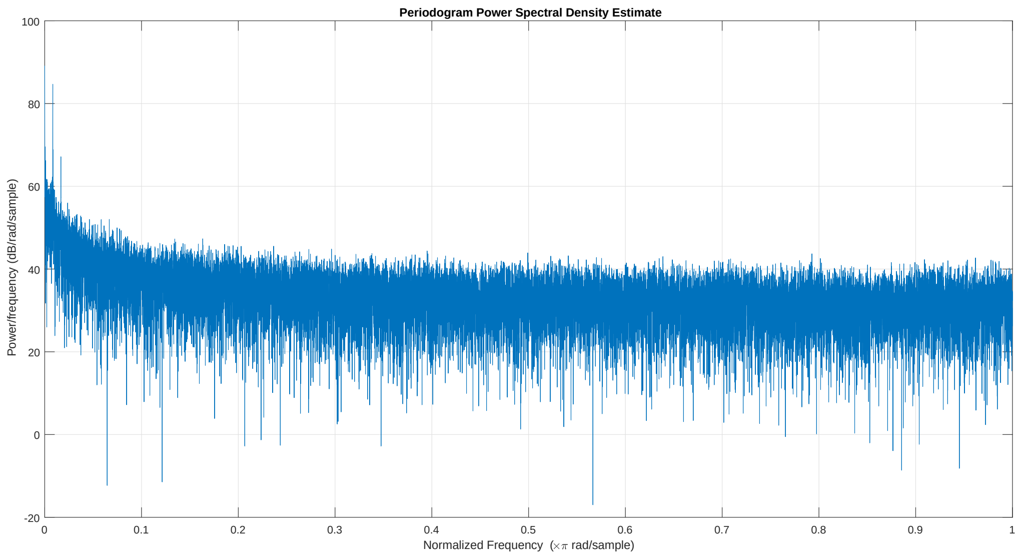


Figure 3: Power Spectral Density by using Fast Fourier Transform

First, we use FFT finding the power spectral density to find the power spectral density. Then we choose only high-energy frequency and label them ω_i. Figure 3 show that there are two peaks at the frequencies $\omega_1 = 0.0083\pi$ and $\omega_2 = 0.0166\pi$

After choosing frequency ω_i, we can find each a_i, b_i and α in 3 and 11 from the least square method. Then we will get the data with removing seasonal trend in Figure 4. This set of data are used for finding integrated order.

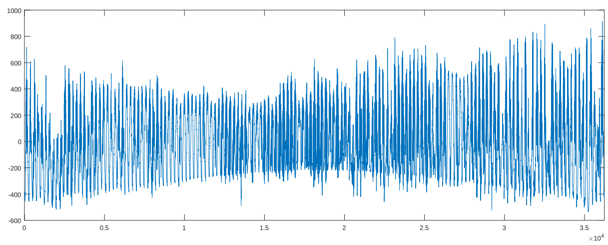
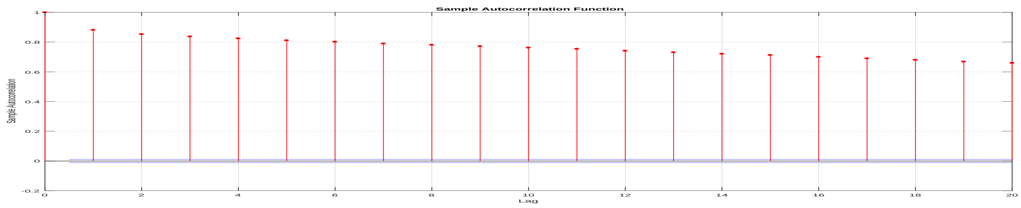


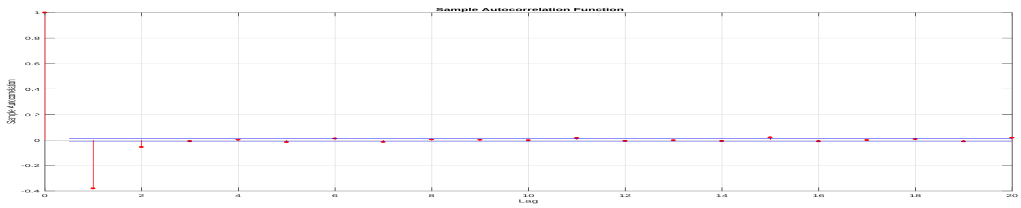
Figure 4: Component with removing seasonal sompontent

5.2 Finding Integrated Order

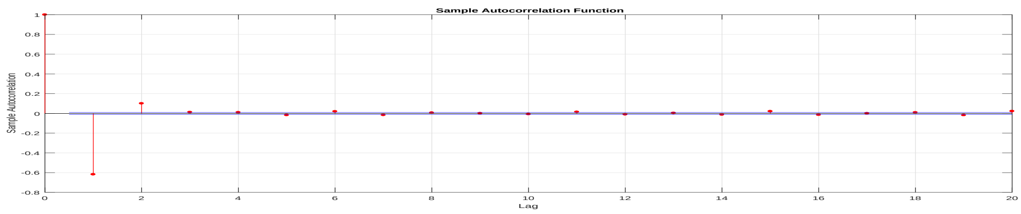
This aim of differencing is finding the integrated order d and also making time series stationary. After removing the seasonal trend, the autocorrelation function(ACF) of the residual error is shown in figure 5(a). Figure 5(a) show that there is non-stationary because autocorrelation function is slowly decay. So that we must differentiate the data before using maximum likelihood estimation to find the ARMA model. The autocorrelation function of one time differencing data and two time differencing data are shown in figure 5. From Figure 5, we can conclude that $d = 1$ is the best choice because it has more similar to white noise than $d = 2$.



(a) without differencing



(b) 1 time differencing

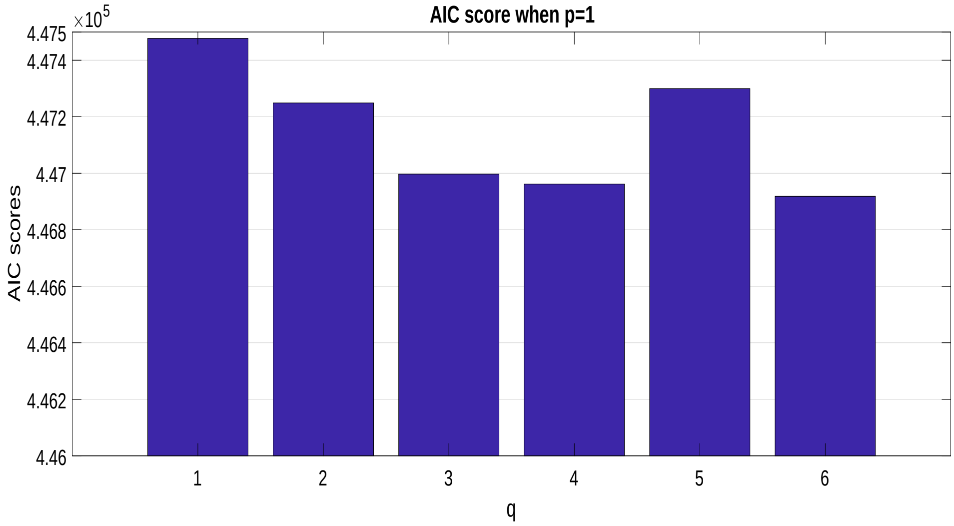


(c) 2 times differencing

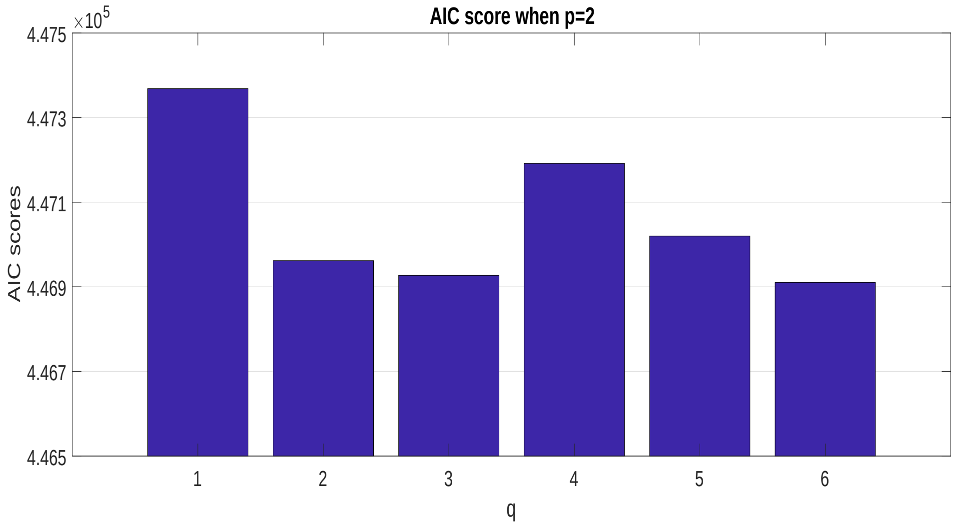
Figure 5: ACF of both one and two time differencing

5.3 Maximum Likelihood Estimation and Model Validation

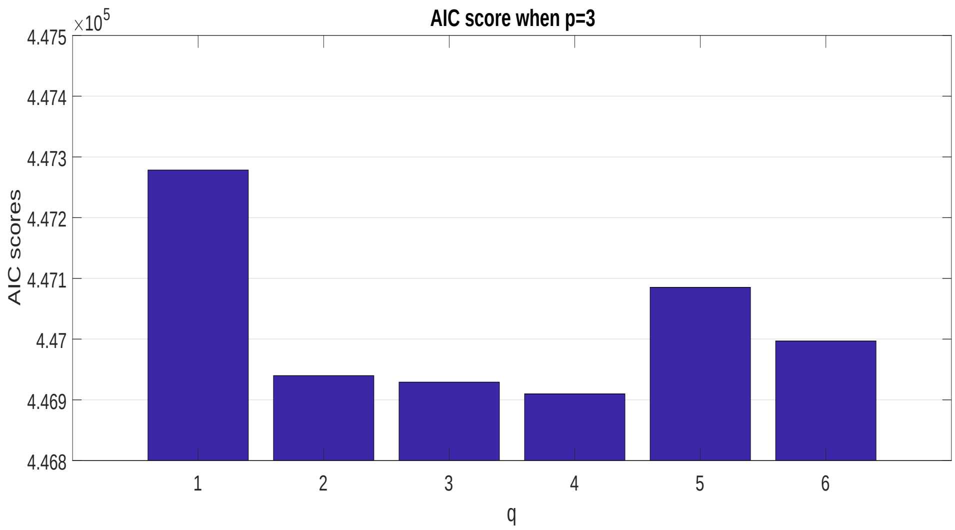
The aim of maximum likelihood estimation is finding the parameter in the ARMA model and the aim of model validation is finding the order p and q. After data are stationarized, this set of data are used for maximum likelihood estimation to estimate the parameter in the ARMA model. Then use each model to find AIC and BIC scores for finding the optimal model and to find mean absolute error and root mean square error on validation data set. Let p and q be in the range 1 to 6. The result of both AIC and BIC scores are shown in Figure 6, 7, 8, 9 respectively.



(a) AIC score when p=1

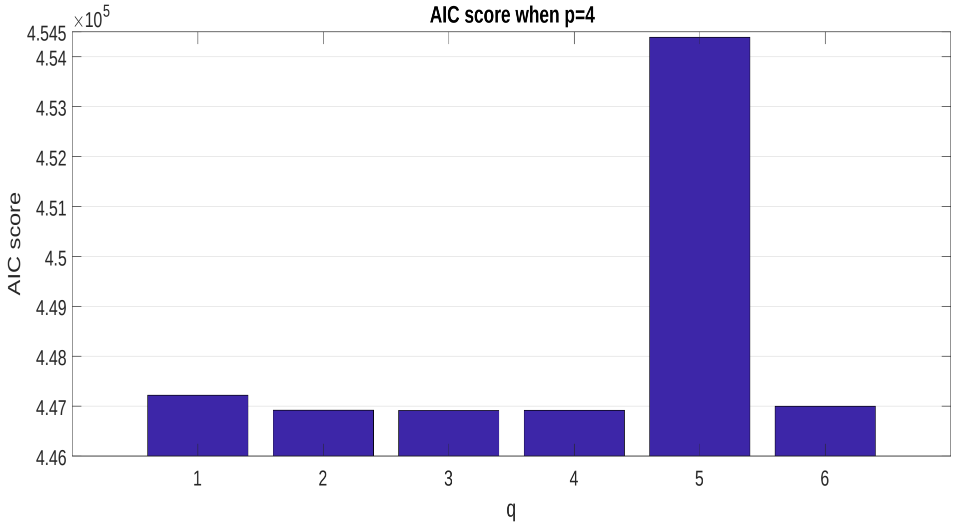


(b) AIC score when p=2

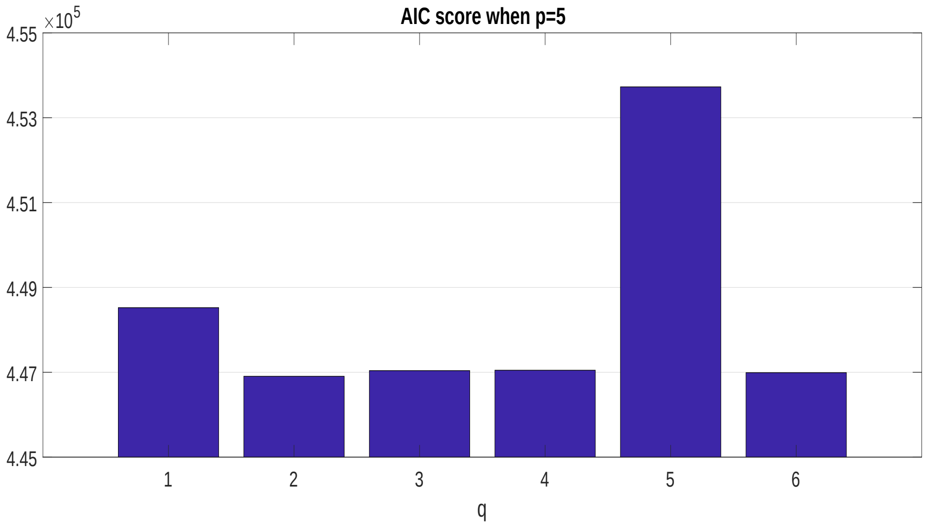


(c) AIC score when p=3

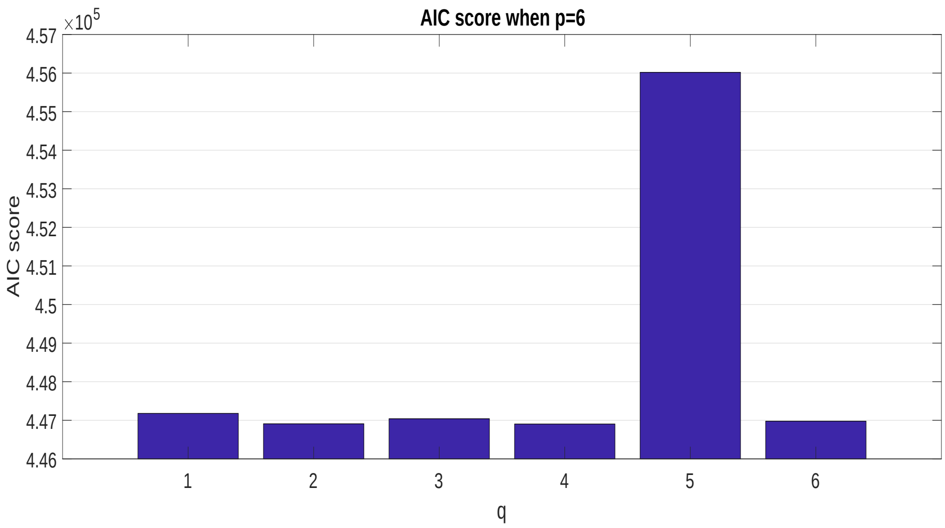
Figure 6: AIC score when p=1-3



(a) AIC score when p=4

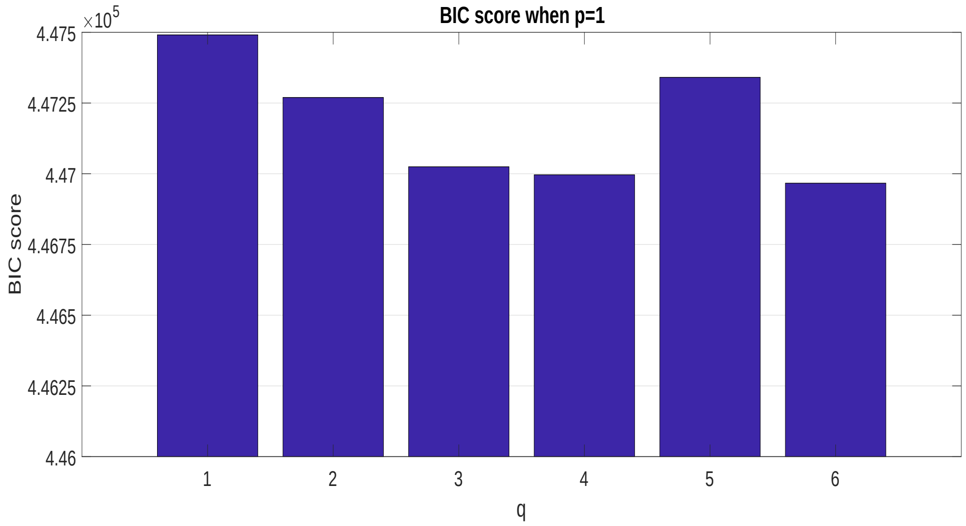


(b) AIC score when p=5

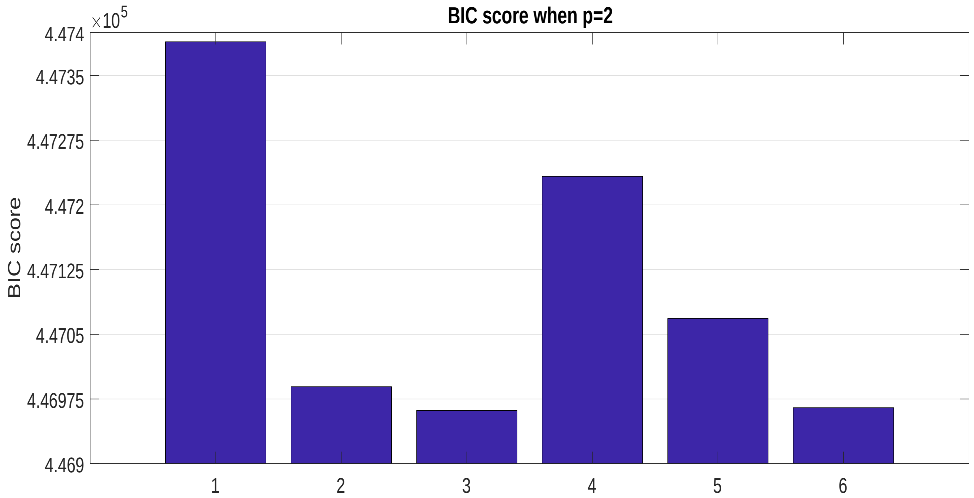


(c) AIC score when p=6

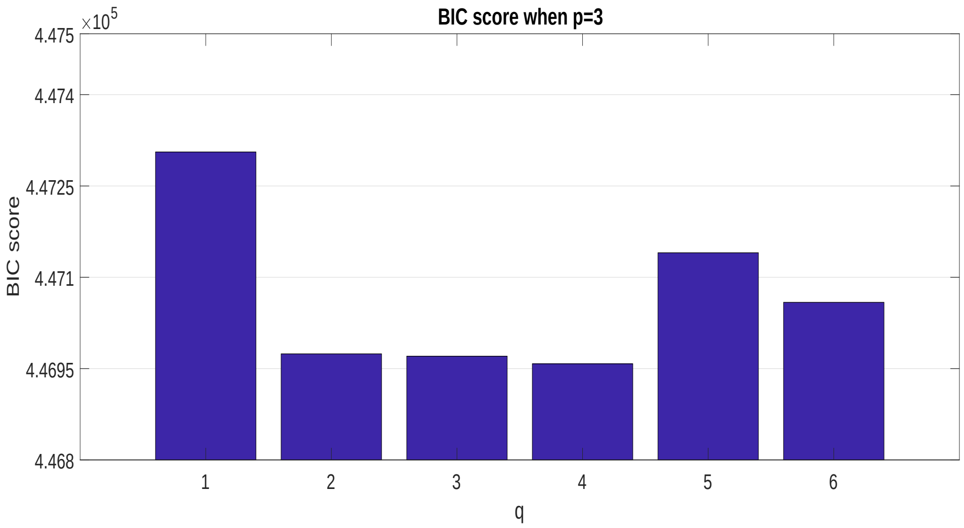
Figure 7: AIC score when p=4-6



(a) BIC score when p=1

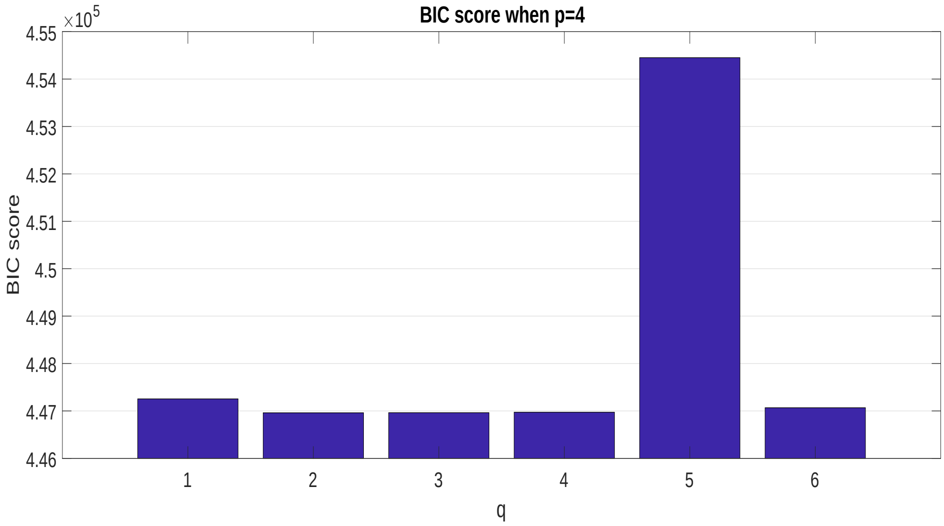


(b) BIC score when p=2

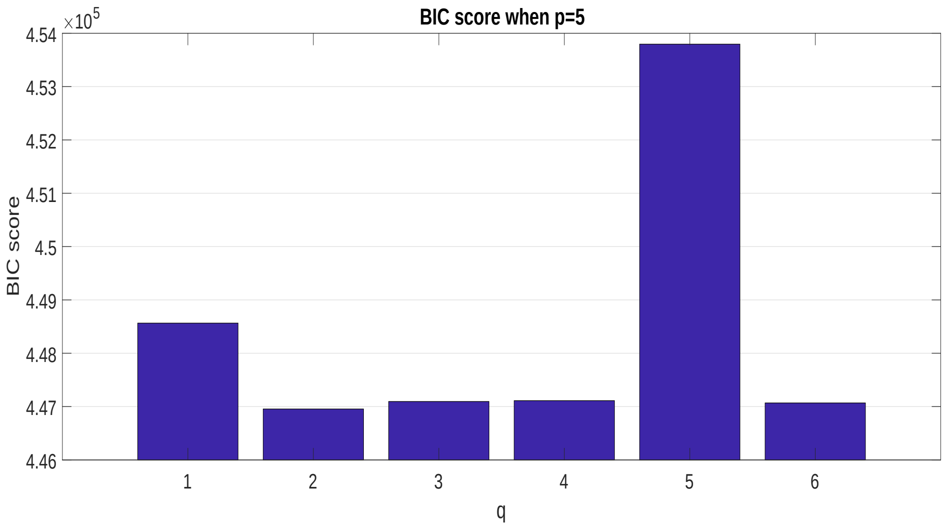


(c) BIC score when p=3

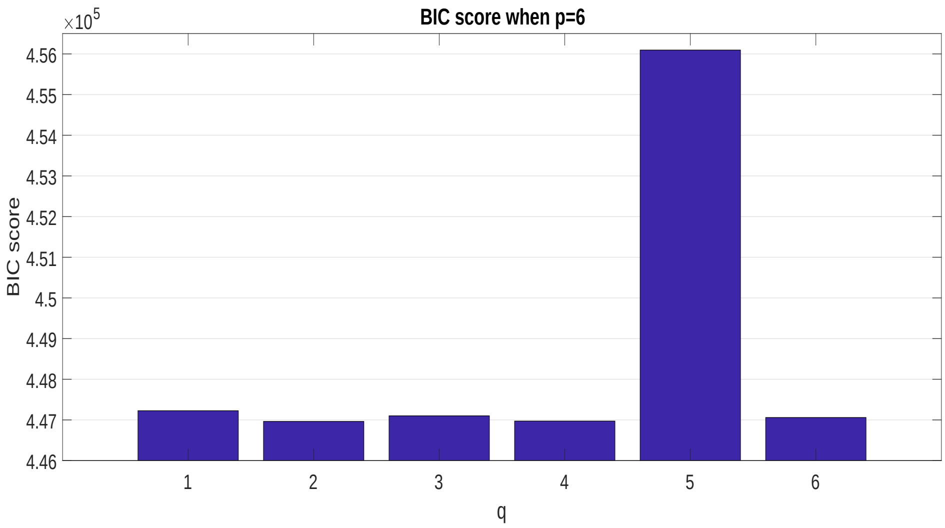
Figure 8: BIC score when p=1-3



(a) BIC score when p=4



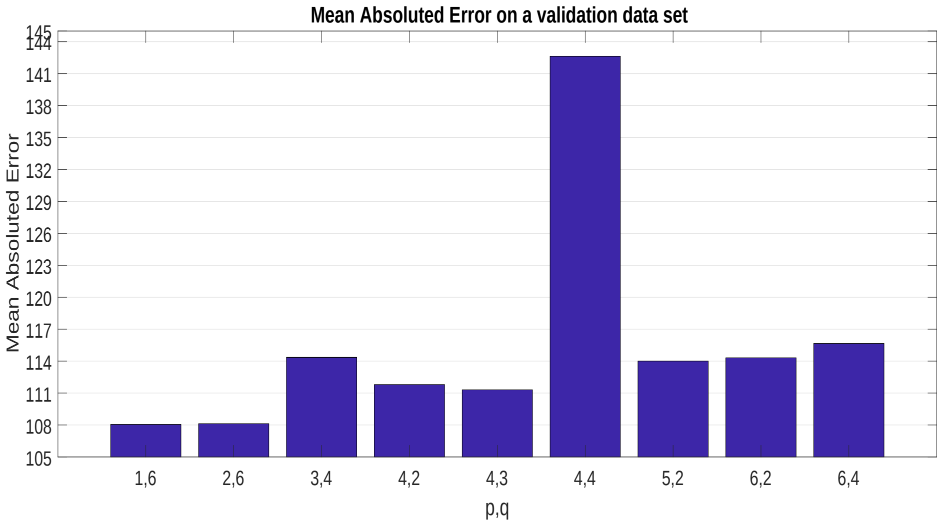
(b) BIC score when p=5



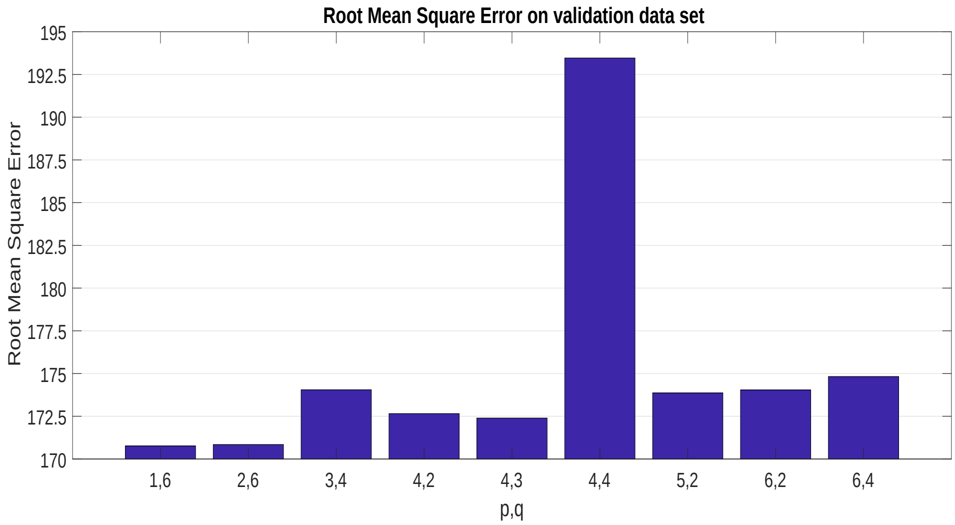
(c) BIC score when p=6

Figure 9: BIC score when p=4-6

Figure 6, 7, 8, 9 show that many models have small AIC score and BIC score. Therefore, we determine RMSE and MAE of these models to select a proper one.



(a) MAE



(b) RMSE

Figure 10: Error on validation data set

The result show that SARIMA$(1, 1, 6)(0, 1, 1)_{250}$ and SARIMA$(2, 1, 6)(0, 1, 1)_{250}$ have less RMSE and MAE than others model. Thus, we select SARIMA$(1, 1, 6)(0, 1, 1)_{250}$ because this model has less complexity than SARIMA$(2, 1, 6)(0, 1, 1)_{250}$. SARIMA$(1, 1, 6)(0, 1, 1)_{250}$ is described in 39.

$$\begin{aligned}(1 - L^{250})(1 - L)(1 - 0.97L)I(t) = \\ (1 - 1.55L + 0.47L^2 + 0.065L^3 + 0.01L^4 - 0.02L^5 + 0.03L^6)(1 - 0.92L^{250})e(t)\end{aligned} \tag{39}$$

Thus, we described the forecasting solar irradiance model which is $\hat{I}(t + h|t)$ are shown in 40

$$\begin{aligned}\hat{y}(t + h|t) =& (-1.55L + 0.47L^2 + 0.065L^3 + 0.01L^4 - 0.02L^5 + 0.03L^6)e(t + h) + (0.97L)\hat{y}(t + h|t) \\ \hat{z}(t + h|t) =& (-1.55L + 0.47L^2 + 0.065L^3 + 0.01L^4 - 0.02L^5 + 0.03L^6)e(t + h) + (0.97L)\hat{y}(t + h|t) \\ & + \sum_{k=0}^{\infty} \hat{y}(t - k + h - 1|t) \\ \hat{I}(t + h|t) =& \hat{z}(t + h|t) + \sum_{k=1}^{\infty} \hat{z}(t - 250k + h|t)\end{aligned} \tag{40}$$

6 Project overview

The mind-mapping of this project is shown in Figure 11

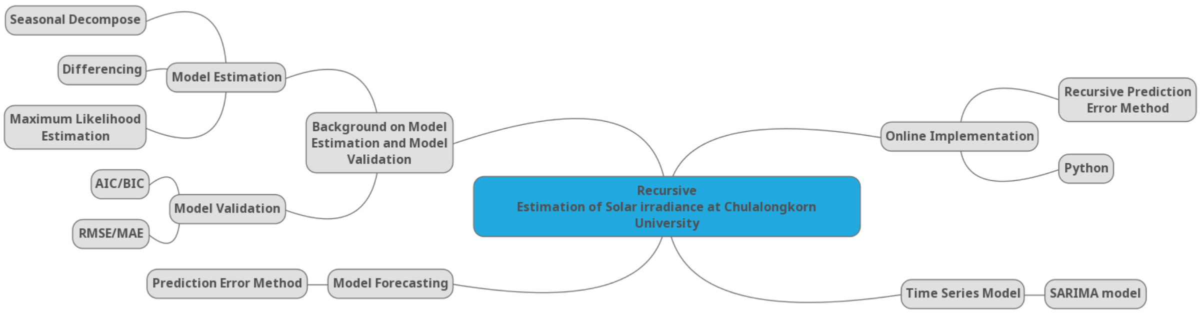


Figure 11: Project Mindmap

6.1 Scope of work

The scopes of this study are

1. We use recursive identification method to the SARIMA model
2. We use data measured at Chulalongkorn University
3. Output solar forecasting is shown on server at Chulalongkorn University

6.2 Expected outcomes

Implementation of recursive estimation adapt with forecasting model to forecasting the solar irradiance data at real-time on the server at Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University.

6.3 Project plans

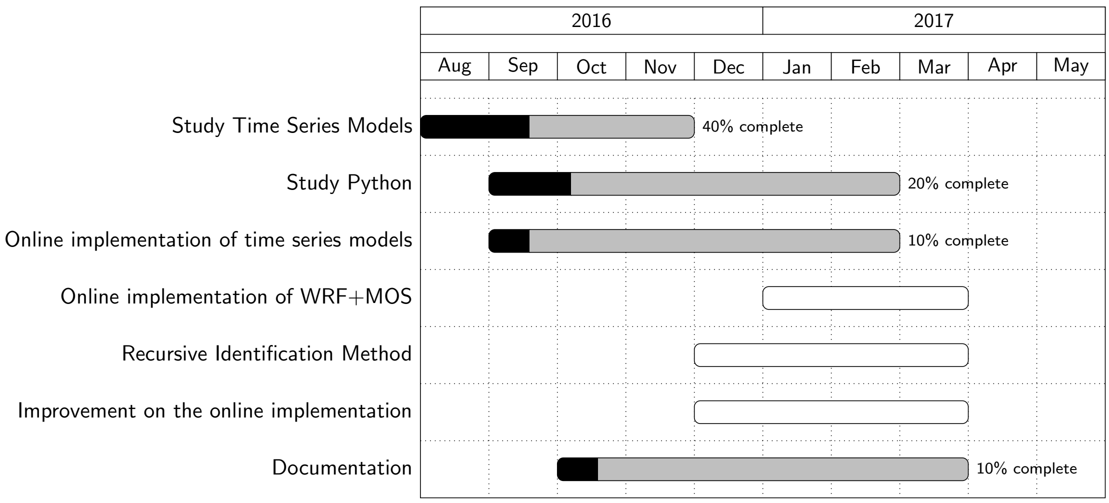


Figure 12: Gantt chart of the project

References

[1] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[2] Nikolaos Kourentzes Juan R. Trapero and A. Martin. Short-term solar irradiation forecasting based on dynamic harmonic regression. *Energy*, 84:289–295, 2015.

[3] W. Khovidhungij. *Signal, Systems and Control*. Chulalongkorn University Press, 2016.

[4] V. Layanun and J. Songsiri. Solar irradiance forecasting for chulalongkorn university location using time series models. http://jitkomut.eng.chula.ac.th/group/vichaya_report.pdf, 2016.

[5] L. Ljung. *System Identification:Theory for user*. Prentice-Hall, 1987.

[6] T. Söderström and P. Stoica. *System Identification*. Prentice Hall International, 1989.

[7] Peter C. Young. *Recursive Estimation and Time Series Analysis*. Springer, 2011.

7 Appendices

7.1 MATLAB code

Finding the estimated model and calculating AIC score, BIC score, RMSE, and MAE for all p,q

6pt

```
% p and q are the highest order of AR and MA polynomial
SpecMdl = arima('Constant',0,'ARLags',1:p,'D',1,'MALags',1:q,'
    Seasonality',240,'SMALags',240);
% Fit the model by using ML
[EstMdl,EstParamCov,logL,info] = estimate(SpecMdl,dataTra);
% Find AIC and BIC
AIC(p,q) = -2*logL+2*(p+q);
BIC(p,q) = -2*logL+(p+q)*log(length(dataVad));
% Find residual error
[Et,Vt] = infer(EstMdl,dataTra);
MAEt(p,q)=sum(abs(Et))/length(dataTra);
RMSEt(p,q)=sqrt(sum(Et.^2)/length(dataTra));
% Find validation error
[Ev,Vv] = infer(EstMdl,dataVad);
MAEv(p,q)=sum(abs(Ev))/length(dataVad);
RMSEv(p,q)=sqrt(sum(Ev.^2)/length(dataVad));
```