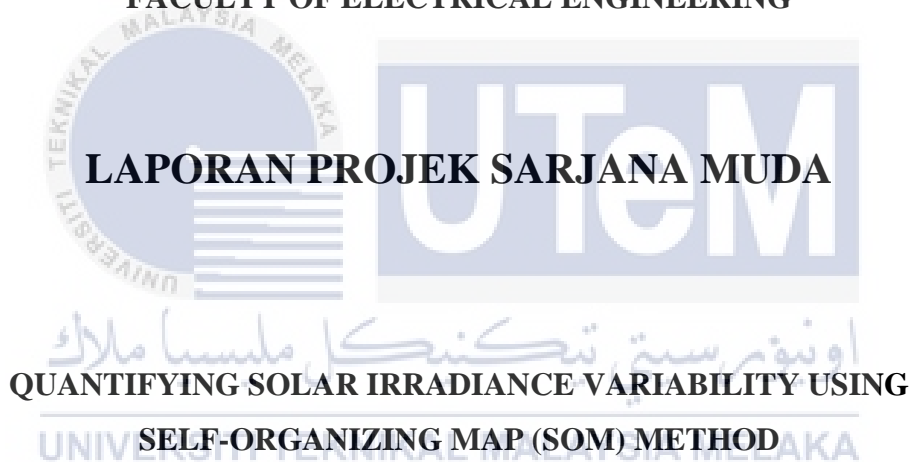




**UNIVERSITI TEKNIKAL MALAYSIA MELAKA  
FACULTY OF ELECTRICAL ENGINEERING**



**LAPORAN PROJEK SARJANA MUDA**

**QUANTIFYING SOLAR IRRADIANCE VARIABILITY USING  
SELF-ORGANIZING MAP (SOM) METHOD**

**By:**

**CHUA SHU YUAN**

**B011410088**

**Bachelor of Electrical Engineering (Power Industry)**

**June 2017**

**Supervised By:**

**Mr. Kyairul Azmi bin Baharin**

“ I hereby declare that I have read through this report entitle “Quantifying Solar Irradiance Variability using Self-Organizing Map (SOM) Method” and found that it has comply the partial fulfilment for awarding the degree of Bachelor of Electrical Engineering (Power Industry) ”

Signature



Supervisor name

.....  
اوتيمور سيتي تيكنيكل مليسيا ملاك

Date

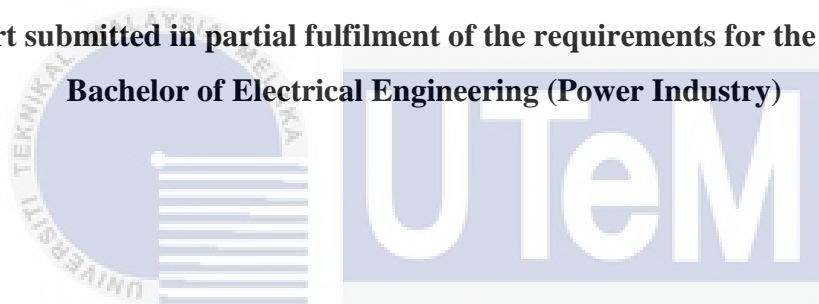
.....  
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

**QUANTIFYING SOLAR IRRADIANCE VARIABILITY USING  
SELF-ORGANIZING MAP (SOM) METHOD**

**CHUA SHU YUAN**

**A report submitted in partial fulfilment of the requirements for the degree of**

**Bachelor of Electrical Engineering (Power Industry)**



**Faculty of Electrical Engineering**  
اونيورسي تيكنيكل ماليسيا ملاك

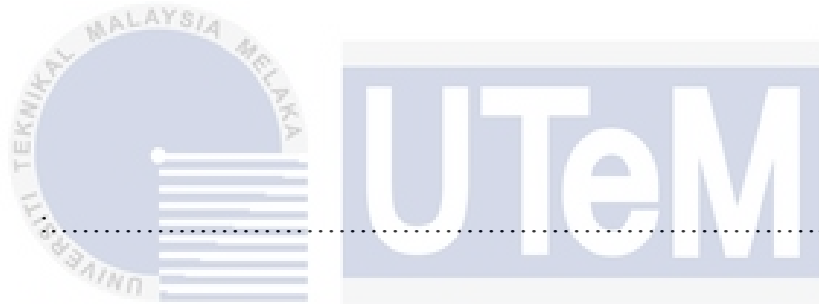
**UNIVERSITI TEKNIKAL MALAYSIA MELAKA**

**UNIVERSITY TEKNIKAL MALAYSIA MELAKA**

**2016/2017**

I declare that this report entitle “Quantifying Solar Irradiance Variability using Self-Organizing Map (SOM) Method” is the result of my own research except as cited in the references. The report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature



Name

اونيورسي تيكنيكل مليسيا ملاك

Date

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

## ACKNOWLEDGEMENT

In preparing this report, I was in contact with many people, researchers, academicians and practitioners. They have contributed towards my understanding and knowledge. I wish to express my sincere appreciation to my project supervisor, En. Kyairul Azmi bin Baharin, for encouragement, guidance critics and friendship. I am also very thankful to my panels En. Azhan bin Ab Rahman, and Mrs Intan Azmira Wan Abdul Razak for their guidance and advice.

My fellow undergraduate students should also be recognized for their support. My sincere appreciation extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips helped me in my research process. Lastly, I am grateful to my family for their continuous love and support.


## ABSTRACT

Among the renewable energy generation in Malaysia, the use of solar energy has increased exponentially throughout the years. The Solar PV generation depends on seasonal variation and clouding which causes variability in the output. The current solution is the discharge of electrical energy from battery storage. However, battery storage is expensive and has a life expectancy. The research here focuses on proposing a classification scheme which can be used for forecasting. Solar irradiance data is clustered into groups for further analysis. The scheme involves the use of unsupervised learning, the self-organizing map (SOM) to cluster the variability profile using solar PV system data from FKE, UTeM. The MATLAB software will be used for SOM simulation. The data used is taken from January to December 2016. The data was tested by varying the map size. After running a few map sizes, the map size 35x35 was chosen. It had a quantization error of 0.844 and error percentage of 10.929%. From the 35x35 map size, a total of 186 days were clustered. Out of the 186 days, 104 days were grouped into 6 categories of variability. The 6 categories were clear sky, high value variability, low value variability, morning to noon variability, noon to evening variability, and overcast. The days clustered were compared to manual clusters and the self-organizing map (SOM) was determined to be highly accurate in clustering high variability irradiance values.

## ABSTRAK

Antara penjana tenaga hijau di Malaysia, penggunaan tenaga suria telah meningkat dengan pesat. Generasi PV Solar bergantung kepada variasi bermusim dan pergerakan awan yang menyebabkan perubahan dalam pengeluaran tenaga suria. Penyelesaian semasa adalah dengan penyaluran tenaga elektrik daripada bateri simpanan. Walau bagaimanapun, bateri simpanan mahal dan mempunyai jangka hayat. Penyelidikan projek ini memberi tumpuan kepada cadangan skema klasifikasi yang boleh digunakan untuk ramalan. Data solar dikelompokkan ke dalam kumpulan untuk analisis selanjutnya. Skim itu melibatkan penggunaan pembelajaran tanpa pengawasan, self-organizing map (SOM) untuk mengumpulkan profil kepelbagaian dengan menggunakan data sistem PV suria dari FKE, UTeM. Perisian MATLAB akan digunakan untuk simulasi SOM. Data yang digunakan akan diambil dari Januari hingga Disember 2016. Saiz peta diubah untuk mencari saiz yang paling sesuai untuk data. Selepas beberapa ujian, saiz peta 35x35 dipilih. Saiz peta tersebut mempunyai ralat penguantuman 0.844 dan peratusan ralat 10.929%. Daripada saiz peta 35x35, sebanyak 186 hari dikelompokkan. Daripada 1866 hari tersebut, 104 hari dipilih untuk dikelompokkan dalam 6 kumpulan. Terdapat 6 kumpulan iaitu langit cerah, nilai tinggi kepelbagaian sepanjang hari, nilai rendah kepelbagaian sepanjang hari, kepelbagaian dari pagi hingga tengah hari, kepelbagaian dari tengah hari hingga petang, dan hari mendung. Hari yang dikelompokkan dibandingkan dengan hari yang dikelompokkan secara manual dan self-organizing map (SOM) dibuktikan bahawa dapat mengelompokkan nilai sinaran tinggi dengan tepat.

## TABLE OF CONTENTS

<b>CHAPTER</b>	<b>TITLE</b>	<b>PAGE</b>
	ACKNOWLEDGEMENT	v
	ABSTRACT	vi
	ABSTRAK	vii
	TABLE OF CONTENTS	viii
	LIST OF TABLE	xi
	LIST OF FIGURE	xii
	LIST OF APPENDICES	xiv
1	<div style="display: flex; align-items: center;">  <div style="margin-left: 20px;">           INTRODUCTION         </div> </div>	1
	1.1 Motivation	1
	1.2 Problem Statement	2
	1.3 Objectives	2
	1.4 Scope	2
	1.5 Outline of the Dissertation	3
2	LITERATURE REVIEW	4
	2.1 Solar Irradiance	4
	2.2 Variability in Solar PV Systems	5
	2.3 Review of Previous Related Work	6
	2.4 Clustering of objects	14
	2.5 Self organizing map (SOM)	14



	2.5.1 Components of Self Organization	15
	2.5.2 Review of past projects with self-organizing map (SOM)	15
	2.6 Groups of days clustered	17
	2.6.1 Clear Sky Day	17
	2.6.2 Morning till noon variability	18
	2.6.3 Noon till evening variability	18
	2.6.4 Overall variability with low irradiance value	19
	2.6.5 Overall variability with high irradiance value	19
	2.6.6 Overcast	20
3	RESEARCH METHODOLOGY	21
	3.1 An overview of methodology	21
	3.2 Flow chart explanation	23
4	RESULT & DISCUSSION	25
	4.1 Clustering of one year using the self-organizing map (SOM)	25
	4.1.1 35x35 map size	26
	4.1.2 50x50 map size	32
	4.1.3 Comparison of 35x35 and 50x50 map size	35
	4.2 Manual Clustering	36
	4.3 Comparison in SOM and manual clustering	37
	4.3.1 Analysis of cluster comparison	38
	4.4 Monthly Clusters	38
	4.4.1 January	38
	4.4.2 February	41
	4.4.3 September	43
	4.4.4 Analysis of month clusters	45
5	CONCLUSION AND RECOMMENDATION	46
	5.1 CONCLUSION	46

5.2 RECOMMENDATION	47
REFERENCE	48
APPENDIX A	51
APPENDIX B	52
APPENDIX C1	53
APPENDIX C2	56



## LIST OF TABLE

TABLE	TITLE	PAGE
2.1	Irradiance and solar system power variability	4
2.2	Review of Previous Studies	9
4.1	Map grid and errors	24
4.2	Days clustered for 35x35	29
4.3	Days clustered for 50x50	32
4.4	Manually clustered days	34
4.5	Comparison of SOM and manual clustering	35
4.6	Different map sizes for January	37
4.7	Comparison of SOM and manual clusters for January	37
4.8	Different map sizes for February	39
4.9	Comparison of SOM and manual clustering for February	40
4.10	Different map sizes for September	41
4.11	Comparison of SOM and manual clustering for February	42
4.12	Comparison of SOM and manual clustering for months	43

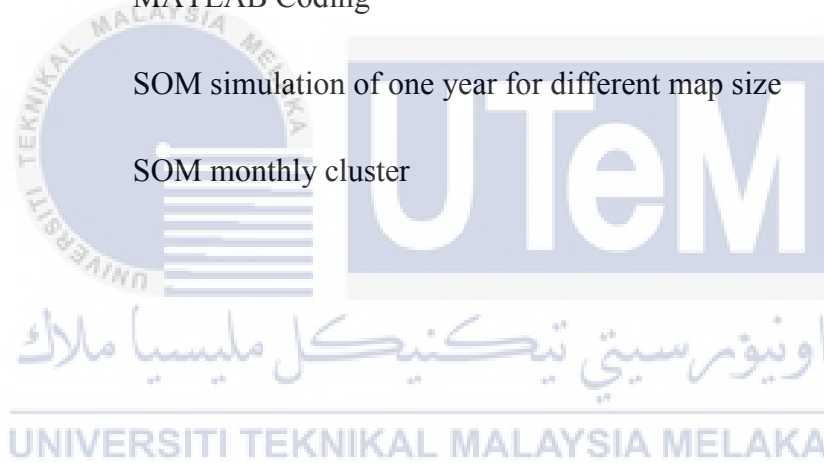
## LIST OF FIGURE

FIGURE	TITLE	PAGE
2.1	Clear sky day	15
2.2	Morning to noon variability	16
2.3	Noon till evening variability	16
2.4	Overall variability with low irradiance value	17
2.5	Overall variability with high irradiance value	17
2.6	Overcast days	18
3.1	Flow chart of entire project flow	20
3.2	Flow chart of SOM cluster method	22
4.1	35x35 SOM	24
4.2	35x35 SOM cluster 1-6	25
4.3	35x35 SOM cluster 7-12	25
4.4	35x35 SOM cluster 13-18	25
4.5	35x35 SOM cluster 19-24	26
4.6	35x35 SOM cluster 25-30	26
4.7	35x35 SOM cluster 31-36	26
4.8	35x35 SOM cluster 37-42	27

4.9	35x35 SOM cluster 43-48	27
4.10	35x35 SOM cluster 49-54	27
4.11	35x35 SOM cluster 55-60	28
4.12	35x35 SOM cluster 61-66	28
4.13	35x35 SOM cluster 67-70	28
4.14	50x50 SOM	30
4.15	50x50 SOM cluster 1-4	30
4.16	50x50 SOM cluster 5-7	31
4.17	January SOM	37
4.18	January SOM cluster	38
4.19	January manual cluster	39
4.20	February SOM	39
4.21	February SOM cluster 1-4	40
4.22	February SOM cluster 5-8	41
4.23	September SOM	41
4.24	September SOM clusters 1-6	42
4.25	September SOM clusters 7-11	43

**LIST OF APPENDICES**

APPENDIX	TITLE	PAGE
A	Gantt Chart	49
B	MATLAB Coding	50
C1	SOM simulation of one year for different map size	51
C2	SOM monthly cluster	54



## CHAPTER 1

### INTRODUCTION

#### 1.1 Motivation

In year 2015, the Asia-Pacific region produced the highest amount of global PV power (59%) with China in the lead. The PV contributed 1.3% to the world's electricity use. Due to its flexibility and adaptability, PV energy is growing at an extremely fast pace[1]. In 2016, 2051MW of PV was installed, and 316GW of solar capacity was produced. The first half of 2016 recorded more than 1000 installations of solar PV every day. As the use of solar increases, the solar prices are slowly dropping in the range of 2-7%. Compared to 5 years ago (2011), solar prices have dropped by 63% [2]. As solar energy becomes more affordable, the number of installations and generation increases.

However, solar power has its disadvantages in terms of variability as it depends on the weather conditions and does not produce energy at night. In this report, we focus on clustering which can be used in forecasting. Clustering has been known as an effective means of grouping objects together based on the similarities of the objects. Many other applications such as data mining, web mining, and voice mining use clustering techniques for further forecast and analysis [3]. Clustering of the solar irradiance can aid in determining the pattern of the solar variability, and know the amount of days with certain characteristics for mitigation strategies. This will allow the use of solar energy to be more efficient.

## 1.2 Problem Statement

When compared to other generation sources, solar PV has the fastest startup time. It only takes seconds to startup. However, the ramp rate is also in seconds which indicates high fluctuations frequency [4]. The current solution for this phenomenon is the use of battery storage which will discharge stored electrical energy when the power generation is low. The battery used is expensive and needs to be changed once it reaches its life expectancy. An alternative solution which is cheaper and long lasting should be used to compensate the PV system's ramp rate. The solar irradiance is clustered to determine the pattern and the variability of the solar PV for mitigation strategies. There have been a few methods used such as Model Tree, Cloud Shadow Model, Artificial Neural Network, and etc. No method is established to characterize irradiance as a large data is involved in the characterization process.

## 1.3 Objectives

1. To use self-organizing map to cluster the variability profile using solar PV system data from FKE, UTeM.
2. To propose a classification scheme from the SOM clustering.
3. To group the clusters into groups of days with different variability.

## 1.4 Scope

This project uses MATLAB software to produce a self-organizing map (SOM) which is an unsupervised learning method to cluster the solar irradiance reading. The data used in this project is obtained from the Photovoltaic and Smart Grid Lab, FKE, UTeM. The data consist of solar irradiance and PV power output from one selected inverter. The duration data is from January to December 2016.



## 1.5 Outline of the Dissertation

Chapter 1 introduces the project overall and the goals to be achieved as well as limitations of the research work. Chapter 2 shows the previous research done on solar clustering and description of the self-organizing map (SOM). Chapter 3 describes the method and steps of the self-organizing map (SOM). Chapter 4 describes the findings and comparison of the self-organizing map (SOM) clusters. Chapter 5 concludes the findings and describes the future research.



## CHAPTER 2

### LITERATURE REVIEW

#### 2.1 Solar Irradiance

Irradiance is a measure of solar power the solar panel receives. The unit is watts per meter squared or  $W/m^2$ . The irradiance received by the panel can vary from  $0 W/m^2$  at night to  $1500W/m^2$  during the day with scattered fluffy clouds. Sunlight travels in a straight line from the sun towards the solar panel while undergoing some atmospheric scattering. Clouds present sometimes reflect extra sunlight towards the solar module, increasing the power produced by the solar system. Irradiance and solar power generated as equally proportional to each other [5]. Due to this factor, the variability of irradiance can determine the variability of solar system output. Based on Yang, Huang, etc, PV output is affected by fluctuate of solar irradiation. For an effective PV output, the solar irradiation has to be forecasted to system reliability and power quality maintenance[6]. Table 2.1 shows the irradiance and solar system power variability. An increase in irradiance directly affects the solar power output.

Table 2.1: Irradiance and solar system power variability [5]

Irradiance	Solar System Power (%)
2000	200
1750	175
1500	150
1250	125
1000	100
750	75
500	50
250	25
100	10
0	0

Based on the output of a 310Wp solar module, the changes are the most drastic when changes were done in the irradiance levels, operating temperatures, shading effects and other correlated factors. However, among all factors, the change in irradiance affected it the most. The change of irradiance from 1000 W/m<sup>2</sup> to 800 W/m<sup>2</sup> reduced Maximum Power Point (MPP) by 19.83% [7]. (MPP) is the highest value on a power curve in regards with voltage and current.

## 2.2 Variability in Solar PV Systems

PV system operators use every method possible such as forecasting, economic dispatch, scheduling, and reserves to have a reliable and satisfactory output with a low cost. If the variability of the PV system is known earlier, the operators have more options to adjust and maintain the system. The flexibility of a PV generation system is in terms of minimum stable generation, ramp rates, and the time for startup and shut down. Studies are being conducted on the integration of PV systems and the grid to characterize variability and reduce fluctuations and cost. Among other methods, the studies prove that the forecast method is the most efficient and reduces time of dispatch schedules for generation and improvement in flexible generation [8]. Forecasting reduces PV plant management cost by improving the management of variability[9] .

The generation of PV power will vary with time as the sun rises and sets from morning till evening. Based on a single-axis tracking PV plant output, 10-13% of changes can be detected for a time interval of 15 minutes due to changes of the sun. Aside from the sun changes, clouds play a primary role in the solar PV output and forecast. Insolation can be defined as solar energy received over time or irradiance integration. A passing cloud can cause solar insolation to exceed 60% of its peak insolation within a few seconds. The time taken to entirely shade a PV system depends on the system size, cloud speed and height. A 100MW capacity system takes minutes instead of seconds for complete shading. The movement of clouds affects the PV systems output in a non-uniform and uncorrelated way. Clouds may shade a solar plant in half or only partially. Therefore, different changes may occur in one plant and between separate plants[10].

### 2.3 Review of Previous Related Work

Based on [11], wavelet decomposition used with k-clustering improves irradiance forecast of a PV plant. Wavelet functions by breaking the data into the approximation and detailed component which removes the fluctuation from the data to be analysed. Instead of using a single model, historical data is classified into 6 classes and 6 models were simulated with two-layer feed forward network in different conditions. The simulation result had higher accuracy than the single model. However, the limitation in this paper was the data variability causing data with strong variability to be lower than data with weak variability.

Another two papers focuses on Wavelet theory. The first one is by Matthew Lave and Jan Kleissl which use Wavelet Variability Model (WVM) with single irradiance point sensor as the input to simulate a solar PV plant. 4 days with different variability were taken to validate the model. The simulation was tested at a 48MW solar PV plant. The simulated power matched the actual power output with a higher accuracy than the plane of array (POA) point sensor. The simulations also matched the ramp rate (RR) distribution. The simulation proves to be better than POA at short timescales[12]. The second paper by Matthew Lave, Jan Kleissl, and Joshua S. Stein uses the same model but with the additional of spatio-temporal correlations. The research is similar to the first with

additional variability reduction (VR). VR is the ratio of point sensor to PV plant variance. The variability by timescale was accurate when compared with fluctuation power index (fpi). (fpi) is fluctuations of wavelet power content for each timescale. The limitations here are that errors in VR cause errors in fluctuation power index (fpi) on cloudy days or days with long timescale. Cloud movement and GHI sensor location can cause total power output and GHI to be slightly inaccurate in time[13].

According to Clifford W. Hansen, Joshua S. Stein, and A. Ellis, statistical methods can be used to characterize irradiance time series to compare forecast model outputs. Frequency distribution is used to quantify the irradiance that falls within a specific range in that period of time. The distribution of ramps quantifies the change in duration and magnitude for a time period. Lastly, the autocovariance and autocorrelations for time series and ramps in clearness index as quantization distribution and ramp distribution does not correlate the time series values. Piecewise linear function is used to produce a sequence of line segments from the data. This paper suggests the separate simulation of clear sky and cloudy sky models as the bivariate distribution does not have irradiance information of when it changed, therefore similar bivariate distribution might occur for clear day and overcast day[14].

Joshua S. Stein, Matthew J. Reno, and Clifford W. Hansen proposed the idea of using variability index to quantify irradiance and PV output variability. Variability index is the ratio of measured irradiance against time divided by the reference clear sky irradiance. Clear days give a variability index of 1. The higher the variability index, the higher the irradiance variability. Clear or overcast days both have low variability index values. To improve variability index quantization, pair variability index with daily clearness index[15].

The next article focuses on the combination of k- means clustering for classification and Support Vector Machine (SVM) regression for training. The k-means cluster is a vector quantization method and data clustered into 3 clusters according to the daily weather similarity. The SVM regression is a machine learning method and is used for the training of input and output data. The output was separated into two categories: clustered and non-clustered data. Root mean square error (RMSE), mean bias error (MRE), and coefficient of determination ( $R^2$ ) are used to determine the errors detected. For both groups, SVM training had the least errors and better prediction than nonlinear

autoregressive (NAR) and artificial neural network (ANN). Despite the accuracy, the data used for this research is meteorological data. For a real case scenario, forecasted meteorological data will be used and will cause the prediction accuracy to drop due to added errors [16].

Patrick Mathiesen, Daran Rife, and Craig Collier proposed the use of Analog Variability (AnVar) forecast for solar irradiance variability. The currently used numerical weather prediction (NWP) is spatially too coarse for variability prediction. An analog downscale method is created to accurately forecast irradiance variability. The analog technique is a pattern matching algorithm. Historical data is compared to current data to produce an irradiance variability forecast. The AnVar forecast was compared to a 2 km Weather Research and Forecasting (WRF) model. Mean bias error (MBE), mean absolute error (MAE), and root mean squared error (RMSE) were computed based on difference between forecast and observed irradiance. AnVar is more accurate than direct WRF forecast. The advantage of AnVar against the WRF model is that it has less forecast bias as it is trained with observation data [17].

Based on [18], a study on statistical analysis of a solar PV plant was done by measuring the output power ramp rates (kW/min & kW/s) and peak daily power output's maximum dip. Two 10MW solar PV plants provide the data. Low, Medium and High variability days were chosen for comparison of the 5MW and 10MW PV plants in the same and different location. Per minute and second ramp rate percentile comparison (99.99%) shows that variability was low for a larger output PV plant and if the plants are in different locations. One whole month's maximum dip magnitude shows that the relationship between installation size and maximum dips is unclear as clouding happens in the entire PV plant regardless of size. According to [19], geographic smoothing is affected by long timescales and an increment in spatial correlation. The effect geographic smoothing has in reducing variability is studied for a large central PV plant and small distributed PV plant. Variable index is used for the measurement of solar variability. The general model quantifies the solar variability and wavelet decomposition quantifies the fluctuations of solar power. From the study, it is shown that large central plant undergoes higher variability. The coherence spectrum shows that the plant sites are less correlated when taken under 5 minutes. However, when the timescale was extended, the correlation

became stronger. This proves that geographic smoothing decreases with the increase in timescale.

According to [20], a cellular computational network (CCN) method can predict solar irradiance for a PV plant. 1, 2, 3, and 4 cells were planted in different positions in a PV plant. CCN has a group of computational units, and each unit has communication with the next or neighbouring units, forming a network. Each cell predicts irradiance of its own location. The cells are able to function as remote virtual sensors. Mean absolute percentage error (MAPE) is used for accuracy measurement. The results show that 3 cells have the highest accuracy. The advantage of using CCN is the ability to predict irradiance from neighbouring location data and ability to function with insufficient input data.

Matthew Lave and Robert Broderick proposed variability metric to quantify variability and compare the variability of 8 locations in the US. This is due to how different distribution feeder has dissimilar climate region. High frequency data was collected and irradiance ramp rates were computed to be plotted against computed cumulative distributions (cdfs). Las Vegas had the least variability, Oahu Island, Mayaguez, and Lanai had the most variable. Two Albuquerque sites had almost identical values. However, high frequency data collection was inconsistent and the data of Integrated Surface Irradiance Study (ISIS) network was used as replacement. The ISIS data has a 3 minutes interval while the project was done for 30 seconds. The data was still used with validation using high frequency data [21].

Zheng Wan, Irena Koprinska and Mashud Rana did an evaluation on clustering methods to group days based on weather characteristics. Prediction was to be done in the range of half-hourly for the next day. The weather data (temperature, solar irradiance) is used to group days with similar weather. The power data trains individual prediction model. Forecasting models are developed with Neural Networks (NN), k-Nearest Neighbor (k-NN) and Support Vector Regression. Most accurate model is clustering based k-NN. Neural Networks (NN) is the most accurate for non-clustering approach. The performance of algorithms depended on the weather. Solar irradiance works best with clustering based approaches. k-NN, NN and SVR, cluster better than Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing (ES) [22].

Table 2.2: Review of Previous Studies

No	Author	Method	Result/ Contribution	Advantages/ Limitations
1	Shi Su, Yuting Yan, Hai Lu, Zhao Zhen, Fei Wang, Hui Ren, Kangping Li, Zengqiang Mi.(2016)	Wavelet Decomposition with k-means clustering	6 ANN models are used in different conditions. When compare with single model, higher accuracy.	Magnitude of data variability between large values and lower values caused uneven forecast.
2	Matthew Lave and Jan Kleissl (2012)	Wavelet variability model (WVM) with single irradiance point sensor as input to simulate solar PV plant	The WVM simulation matches actual power rating better than plane of array (POA) point sensor. Simulation matches ramp rate (RR) distribution.	Has the same result as POA at long timescales. Better than POA at short timescales.
3	Matthew Lave, Jan Kleissl, and Joshua S. Stein (2013)	Wavelet variability model (WVM) with single irradiance point sensor to simulate solar PV plant output with variability reduction	The WVM simulation matches the actual power output. Variability by timescale was accurately determined when comparisons of fluctuation power index (fpi) was done.	Input requirements only require a single sensor and input data. Errors in VR can cause errors in the fluctuation power index (fpi) on cloudy days/long timescale. Cloud movement and GHI sensor location can cause total power output and GHI to be slightly inaccurate in time.
4	Cliford W. Hansen, Joshua S. Stein, and Abraham Ellis (2010)	Frequency distribution, the distribution of ramps, and the automatic covariance and correlation for the time and ramps in clearness index.	Suggests clear sky and cloudy models to be simulated separately.	The bivariate distribution does not have irradiance information of when it changed, therefore similar bivariate distribution might occur for clear day and overcast day.



5	Joshua S. Stein, Matthew J. Reno, and Clifford W. Hansen (2012)	Variability Index	Low VI values were found for clear and extremely overcast days. High VI values were found when irradiance variability was high. The best VI value is 1, where there is little or no irradiance variability.	For extremely cloudy days, VI also has low values and does not distinguish clear and cloudy days well. The daily clearness index is needed to pair up with VI for better classification.
6	Kuk Yeol Bae, Han Seung Jang, and Dan Jeun Sung (2016)	Combination of k- means clustering for classification and Support Vector Machine (SVM) regression for training.	The result was separated into two categories: clustered and non-clustered data. Root mean square error (RMSE), mean bias error (MRE), and coefficient of determination (R2) are used to determine the errors detected. For both groups, SVM training had the least errors and best prediction than NAR and ANN.	Meteorological data is used in this research. For a real case scenario, forecasted meteorological data will be used and will cause the prediction accuracy to drop due to added errors.
7	Patrick Mathiesen, Daran Rife, and Craig Collier	Analog Variability (AnVar) Forecast	The 10 km numerical weather prediction (NWP) forecast was compared to a 2 km Weather Research and Forecasting (WRF) model. Mean bias error (MBE), mean absolute error (MAE), and root mean squared error (RMSE) were computed based on difference between forecast and observed irradiance. AnVar is more accurate than direct WRF forecast.	The analog model has less forecast bias than the WRF model as it is trained with observation data.

8	Mark Mitchell, Michael Campbell, Kathryn Klement, and Mohammad Sedighy Hatch (2016)	Statistical analysis of the PV power generation variability	Low, Medium and High variability days were chosen for comparison of the 5MW and 10MW PV plants in the same and different location. Per minute and second ramp rate percentile comparison (99.99%) shows that variability was low for a larger output PV plant and if the plants are in different locations.	One whole month's maximum dip magnitude shows that the relationship between installation size and maximum dips is unclear as clouding happens in the entire PV plant regardless of size.
9	Houtan Moaveni, David K. Click, Richard H. Meeker, Jr., Robert M. Reedy and Anthony Pappalardo (2013)	Quantifying Solar Power Variability for a Large Central PV plant and Small Distributed PV Plant using variable index, general model and wavelet decomposition	Large central plant undergoes higher variability. The coherence spectrum shows that the plant sites are less correlated when taken under 5 minutes. However, when the timescale was extended, the correlation became stronger. This shows that geographic smoothing decreases with the increase in timescale.	Correlation for various timescales was done.
10	Iroshani Jayawardene and Ganesh K. Venayamoorthy	Cellular computational network (CCN) prediction method for PV plant	CCN configurations were installed in 2-4 locations. Different CCN configurations are compared. CCN with 3 cells has the best irradiance prediction.	CCN has the ability to predict irradiance from neighbouring location data.

11	Matthew Lave, Robert Broderick (2014)	A variability metric is defined for quantifying and comparing variability of 8 locations in US	High frequency data was collected and irradiance ramp rates were computed to be plotted against computed cumulative distributions (cdfs). Las Vegas had the least variability, Oahu island, Mayaguez, and Lanai had the most variable. Two Albuquerque sites had almost identical values.	High frequency data collection was inconsistent and the data of Integrated Surface Irradiance Study (ISIS) network was used as replacement. The ISIS data has a 3 minutes interval while the project was done for 30 seconds. However, the data was still used with validation using high frequency data.
12	Zheng Wang, Irena Koprinska, Mashud Rana (2016)	Evaluation on clustering methods to group days based on weather characteristics	Weather data (temperature, solar irradiance) groups days with similar weather. Power data trains individual prediction model. Forecasting models are developed with NNs, k-Nearest Neighbor (k-NN) and Support Vector Regression. Most accurate model is clustering based k-NN. NN is the most accurate for non-clustering approach.	Direct forecast was done with PV output. Forecast done simultaneously for all half-hour intervals. The performance of algorithms depended on the weather. Solar irradiance works best with clustering based approaches. k-NN, NN and SVR, cluster better than ARIMA and ES.

## 2.4 Clustering of objects

Clustering is defined as organizing input data into groups that have similar pattern or characteristics. It is essential for data to be clustered into meaningful groups. A meaningful cluster has its similarities maximized in a group while similarities between other groups minimized. The data of the group can be grouped based on a set of measurements or relationship between the group's data. Clustering functions as an exploratory tool to analyse large amount of data and is extensively used for applications such as pattern recognition, machine-learning, and document retrieval [23]. There are a few types of clustering methods, such as hierarchical clustering, grid-based clustering, partitioning clustering and density-based clustering. The self-organizing map (SOM) is also used to develop clustering algorithms [24].

## 2.5 Self organizing map (SOM)

Supervised training is done when the classification is done with a stimulus, and given a corrective feedback. Supervised training reduces the error rate in classification as inputs have a set of rules to follow [25]. Unsupervised training is where networks learn to classify training data without external help. To use this training, we assume the input patterns have common features, and the network identifies and classifies the data with that range of input patterns. The self-organizing map (SOM) is part of the unsupervised system which is based on competitive learning. The output neurons compete among themselves to be energized, and only one is on at any one time. The active neuron is known as winner-takes-all neuron or the winning neuron. Lateral inhibition connection (negative feedback paths) induces such competition among the neurons. With that, the neurons organise themselves in the form of a topological map. SOM transforms the arbitrary dimension of an incoming signal pattern into a one or two dimensional discrete map. The transformation is done adaptively with a topologically ordered method. Neurons are situated at the nodes of the lattice which are one or two dimensional. The neurons are selectively tuned to different input patterns (stimuli) or groups of input patterns in the process of competitive learning. The tuning results in an ordered coordinate system for the input characteristics

are created on the lattice. The topographic map formed can be viewed as a non-linear generalization of principal component analysis (PCA) [26].

### 2.5.1 Components of Self Organization

The four components involved in the SOM:

Initialization: Small random values are used to initialise all the connection weights.

Competition: The neurons of the input data compete based on the discriminant function of their respective values. The specific neuron with the least amount of discrimination function becomes the winner.

Cooperation: The spatial location of excited neurons in a topological map is determined by the winning neuron, which allows for cooperation among nearby neurons.

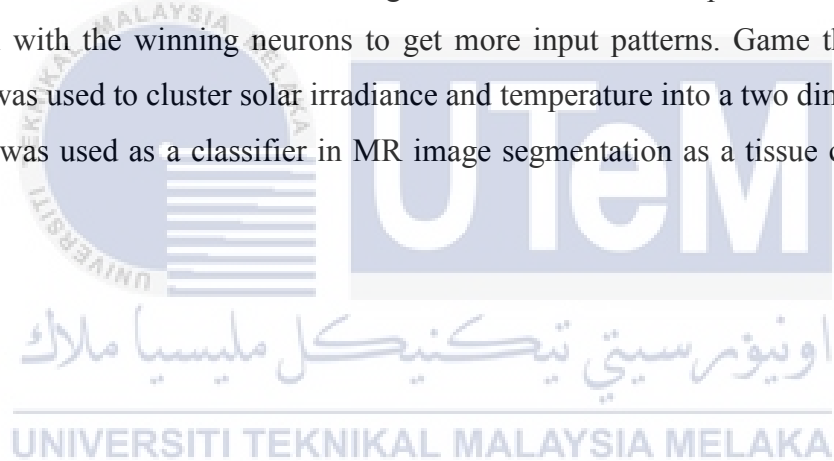
Adaptation: The discriminant function of the excited neurons is decreased by adjusting the associated connection weights so that the winning neuron response to the subsequent application which has a similar pattern is improved.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

### 2.5.2 Review of past projects with self-organizing map (SOM)

Leandro J. Moreira and Leandro A. Silva combined SOM with INN (Informative Nearest Neighbors). SOM was first used to reduce the dataset into a simpler organization by competitive process. The reduced dataset is later utilized by INN which picks the most informative object for classification. The combination resulted in reduction of classification time and statistically equivalent to the (K Nearest Neighbors) classifier [27]. Juha Vesanto and Esa Alhoniemi, proposed a two stage procedure for SOM. First, the large set of prototypes is clustered by SOM or other quantization algorithm. Then, prototypes are combined into actual clusters. This two level clustering reduces computational costs [28].

Based on [29], satellite image analysis and hybrid exponential smoothing state space (ESSS) model use SOM for classification. Cloud cover index is rated based on pixel opaqueness to determine the obstruction of solar irradiance. To determine the exact degree of opaqueness, SOM is used for classification. In another paper, a research to analyse PV power output considering weather type used SOM to classify the PV output by dividing it into 4 categories. The categories were classified based on weather fluctuation characteristics[30]. A weather based hybrid method to forecast PV power output one day ahead used SOM to classify the historical data of solar PV power output. SOM was used to extract the daily features of the data patterns to classifying into different weather types [6]. A hybrid solar radiation forecasting method was done by combining game theoretic concepts. Game theory can be defined as a game where each player has their individual payoff function and strategy. Each player is independent with conflicting interest. The focus of GTSOM is on the non-winning neurons. GTSOM helps them to improve in competition with the winning neurons to get more input patterns. Game theoretic SOM (GTSOM) was used to cluster solar irradiance and temperature into a two dimensional map [31]. SOM was used as a classifier in MR image segmentation as a tissue classifier [32].



## 2.6 Groups of days clustered

The clustered days below uses the Ineichen clear-sky model from [4] as reference. The graphs are based on the manual clusters done and readings taken from FKE's Solar lab.

### 2.6.1 Clear Sky Day

The clear sky day is has very little variations. The solar irradiance value increases steadily and shows a peak in the afternoon approximately at 2pm. After 2pm, it decreases until evening 7pm. The peak of the clear sky has a value of  $1000\text{W/m}^2$ , however the readings taken from the FKE's solar lab as shown in the graph below has readings exceeding  $1000\text{W/m}^2$ .

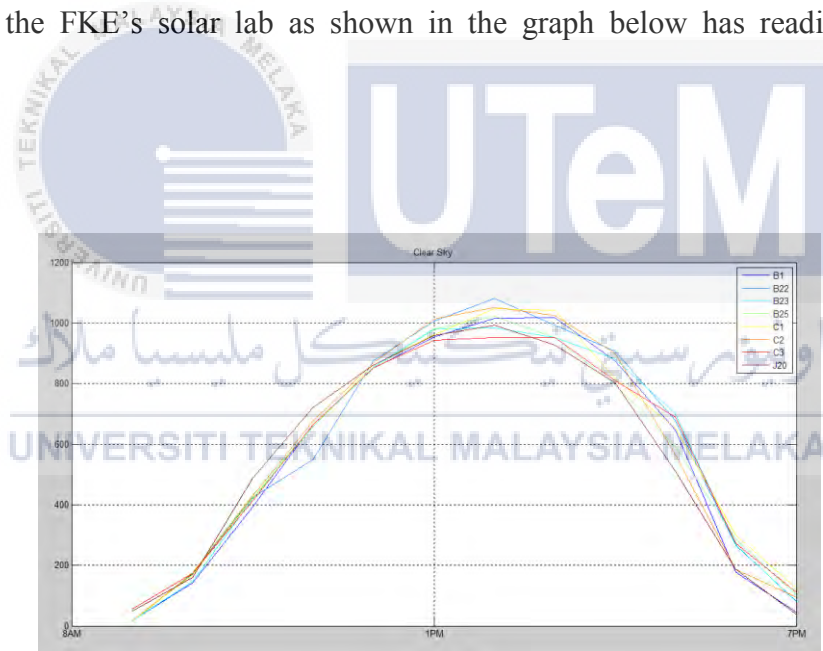


Figure 2.1: Clear sky day

### 2.6.2 Morning till noon variability

The variability is observed from morning till afternoon around 3pm where the variability reduced till evening. The peak ranges from 800 to 1000 W/m<sup>2</sup>.

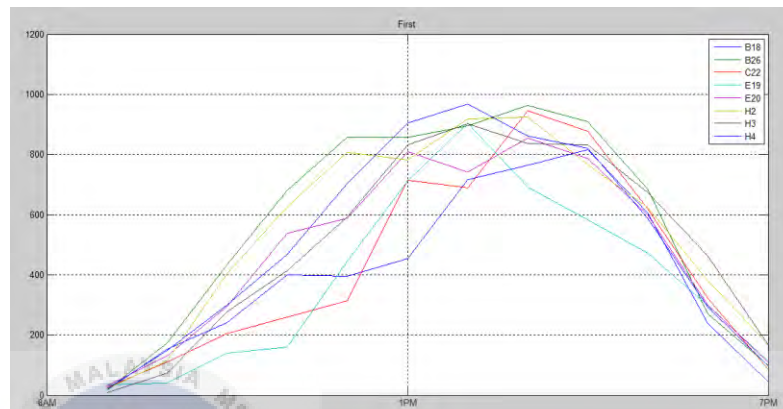


Figure 2.2: Morning to noon variability

### 2.6.3 Noon till evening variability

The graph shows clear sky and high values during the first half of the day and the rest of the day after 12pm shows high variability. The peak ranges from 800 to 1000 W/m<sup>2</sup>.

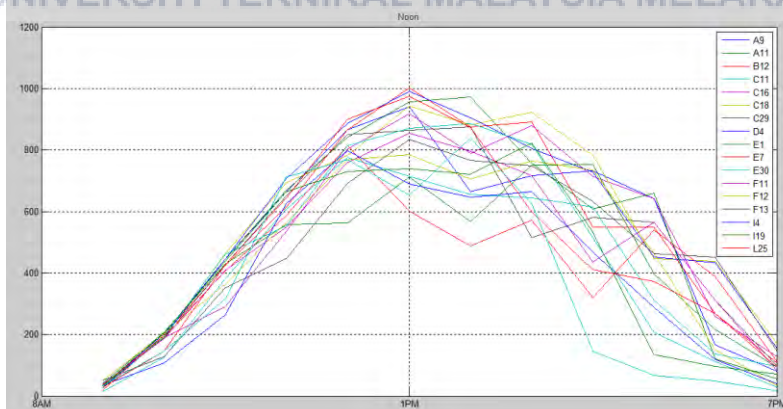


Figure 2.3: Noon till evening variability



### 2.6.4 Overall variability with low irradiance value

There is variability present throughout the day and the peak ranges from 600 to 800 W/m<sup>2</sup>.

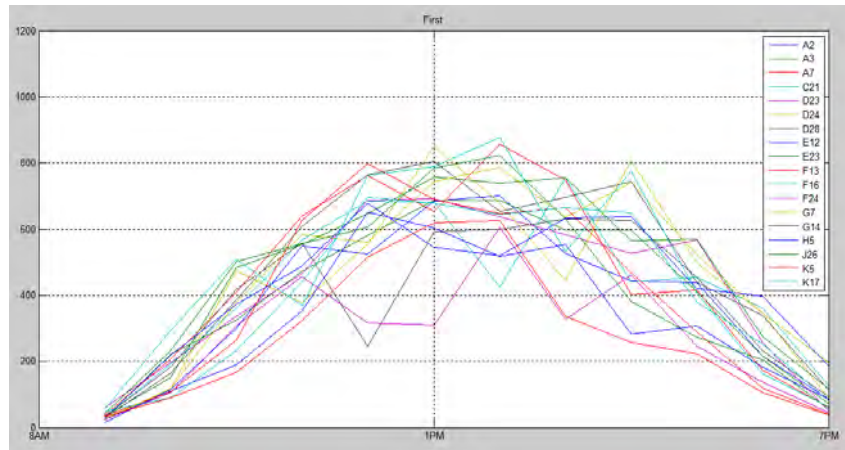


Figure 2.4: Overall variability with low irradiance value

### 2.6.5 Overall variability with high irradiance value

There is variability present throughout the day and the peak ranges from 800 to 1000 W/m<sup>2</sup>.

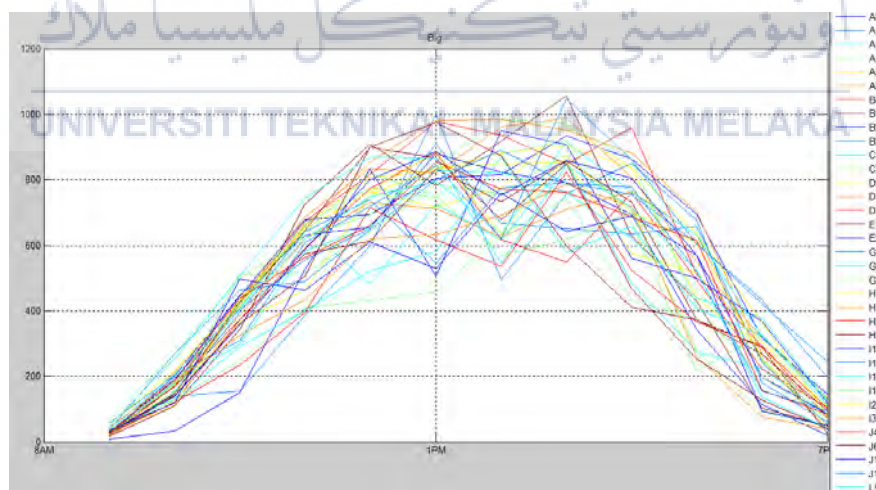


Figure 2.5: Overall variability with high irradiance value

### 2.6.6 Overcast

Overcast has a peak value of below 300W/m<sup>2</sup>.

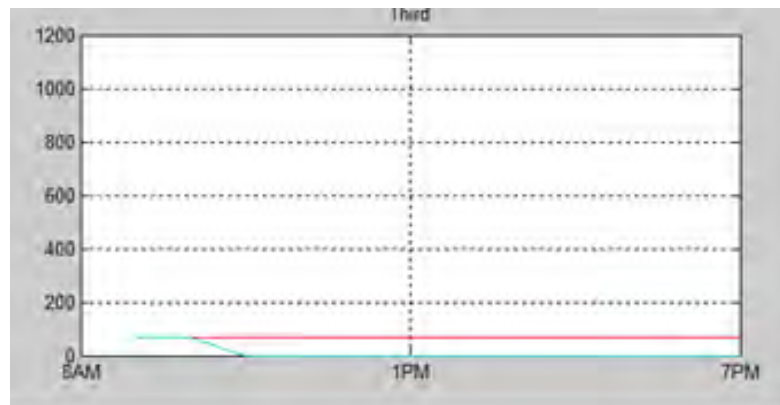


Figure 2.6: Overcast days



## CHAPTER 3

### RESEARCH METHODOLOGY

This chapter presents the flow of the research. The methodology of this research includes the principal of the method that will be carried out to complete the research. Data will be clustered based on the proposed and method and grouped into clusters of variabilities.

#### 3.1 An overview of methodology

A flow chart that summarizes an overall flow of this research work is shown in Figure 3.1. This research started with the findings of literature review. Literature review is an investigation of previous research on the similar clustering. Next, a SOM coding will be developed using Matlab software. There are 3 stages to analyze the irradiance readings. Firstly, for stage 1 it involves the running and getting the shape of the data. Then, for stage 2 is to identify the clusters using SOM and for stage 3 is grouping the clusters into groups.

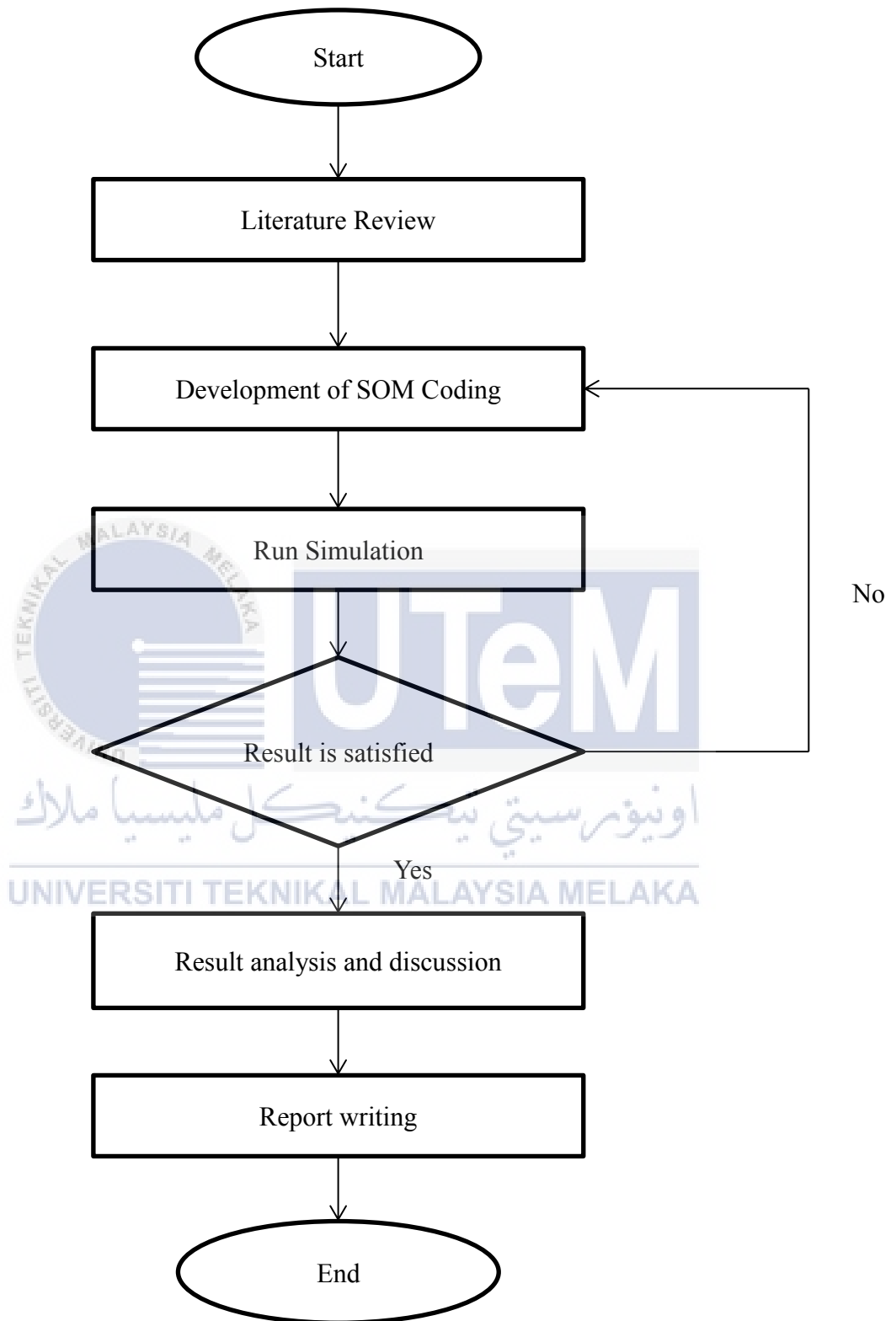
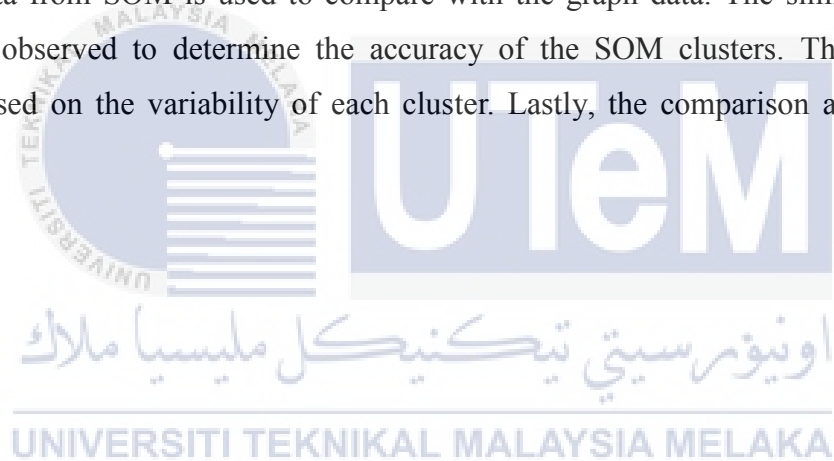


Figure 3.1: Flow chart of entire project flow

### 3.2 Flow chart explanation

Figure 3.2 shows the flowchart of the SOM cluster process. The SOM Toolbox is used here to implement the SOM algorithm. As mentioned in the literature review, the SOM will classify the data based on their similarities and no supervision. The solar PV data read by the SOM will be clustered based the similarities in value. The times of similar PV readings will be clustered. Data is obtained from FKE's solar lab. For SOM data organization, the solar PV data is sorted out and arranged in a notepad for the required time which is 8am to 6pm. The data is run by the training system. The two training phases are rough training and fine tuning phase. The map size is determined. The method for normalization is changed and adjusted to see how it affects data. The days in a month which have the least Euclidean distance and close to each other is grouped separately. The grouped data from SOM is used to compare with the graph data. The similarities of the graphs are observed to determine the accuracy of the SOM clusters. The clusters are grouped based on the variability of each cluster. Lastly, the comparison and results are analysed.



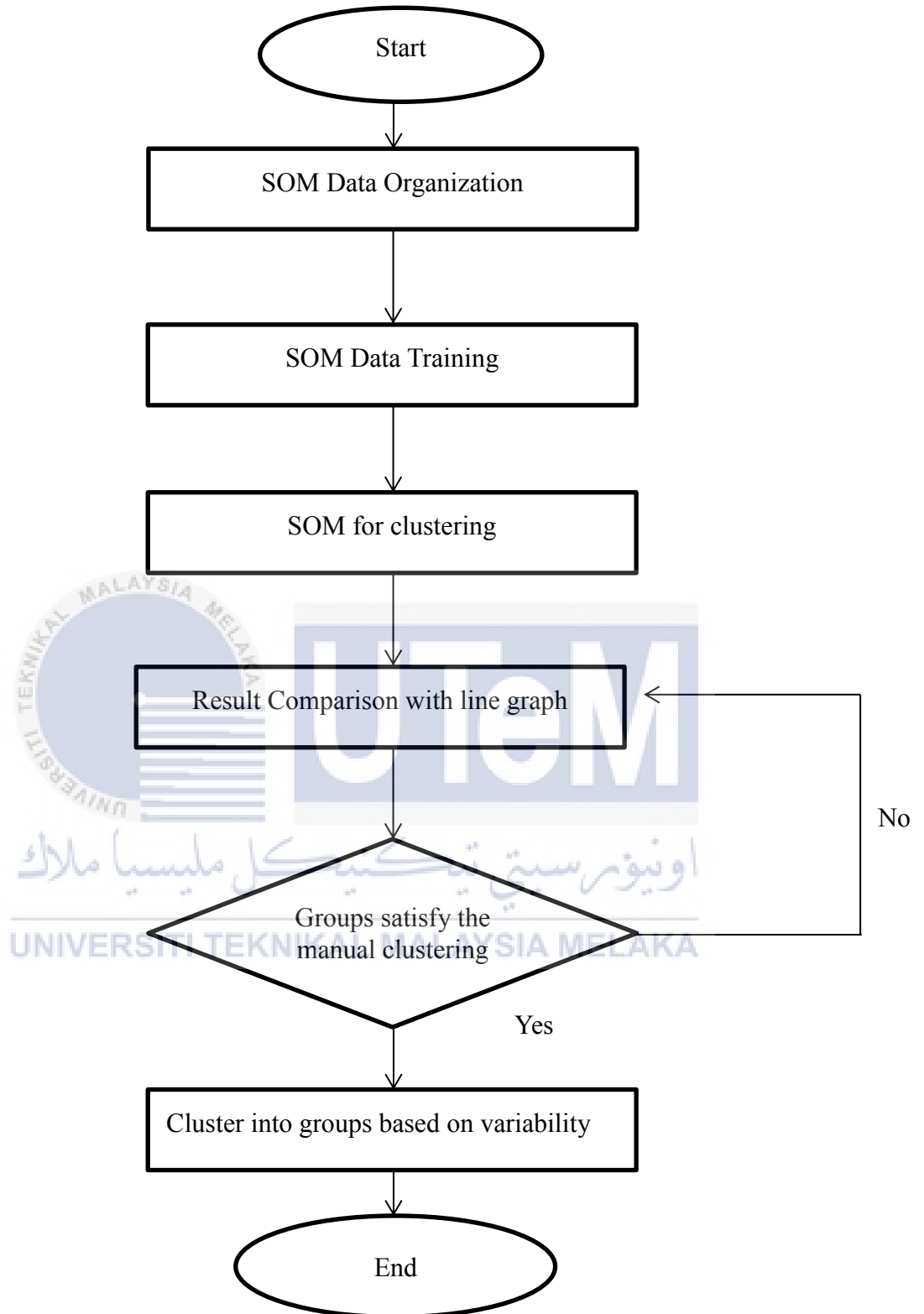


Figure 3.2: Flow chart of SOM cluster method

## CHAPTER 4

### RESULT & DISCUSSION

#### 4.1 Clustering of one year using the self-organizing map (SOM)

Each month was assigned alphabets ranging from A to L following the month January to December 2016. The normalization method used for the coding is 'var'. Var stands for variance in which the data is normalized to one. The map is initialized with 'msize', which changes the map grid size. The SOM clustering for one year was run on a few map sizes. Then smallest map size tested was 30x31 and any map size smaller than that had topographic error. The largest map size tested was 50x60. As the map size became bigger, the number of clusters was reduced. Out of the map sizes run, map size 35x35 was chosen.

The quantization error and topographic error are based on the data input. Both errors affect the quality of the map. The quantization error measures the average distance between each data point and its best matching unit (BMU). The resolution of the map is measured. The topographic error is the proportion of all data vectors when the first and second BMUs are not neighbouring units. The topographic error measures the topology preservation. The error percentage is measured by the error value of the whole map[33]. Based on table 4.1, the increase in map grid size decreases the final quantization error and error percentage. Any map size less than 30x31 will start to have topographic error. The smaller the error percentage and quantization error, the clustering of the self-organizing map (SOM) will be more accurate. However, the lowest final quantization error only produced a few clusters which resulted in the clustering of only a few days. By choosing the map size based on the quantization error will result in inaccurate clustering of days.

As observed in table 4.1, the map grid size smaller than 30x31 has a topographic error of 0.005. As the map size increased, the final quantization error and error percentage decreased. However, at 50x60, there were only 4 clusters and 8 days present from one year. The days clustered were too little. Therefore, map size 35x35 was chosen. In order to compare the clusters, map size 35x35 and 50x50 are elaborated in this report.

Table 4.1: Map Grid Sizes and Errors

Map Grid Size	Final Quantization Error	Final Topographic Error	Error Percentage
50x60	0.220	0.000	1.9126
50x50	0.333	0.000	3.0055
40x40	0.664	0.000	6.0109
35x35	0.844	0.000	10.9290
31x31	0.977	0.000	13.3880
30x31	0.996	0.000	14.2077
30x30	1.027	0.005	16.6667

#### 4.1.1 35x35 map size

Figure 4.1 shows the 35x35 map size which provided 70 clusters. Each cluster was recorded. The clustered days are highlighted on the left and the right shows the Euclidean distance between the particles. As seen, many particles are closely knitted as the Euclidean distance between them ranges from a small value of 0.163 to 1.45.

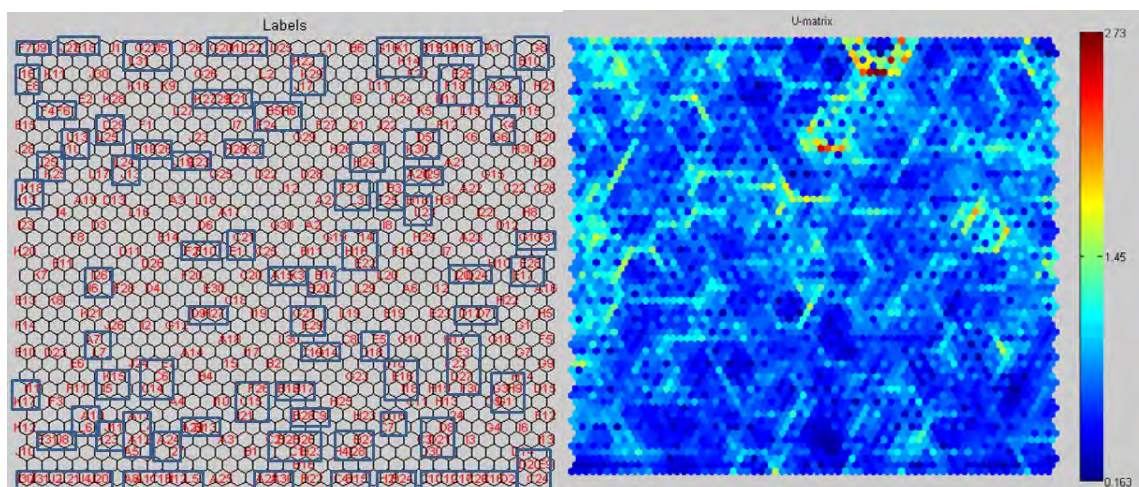


Figure 4.1: 35x35 SOM



The clustered results were plotted in graphs with the B23 data as the clear sky reference. The reference is the blue line in all graphs.

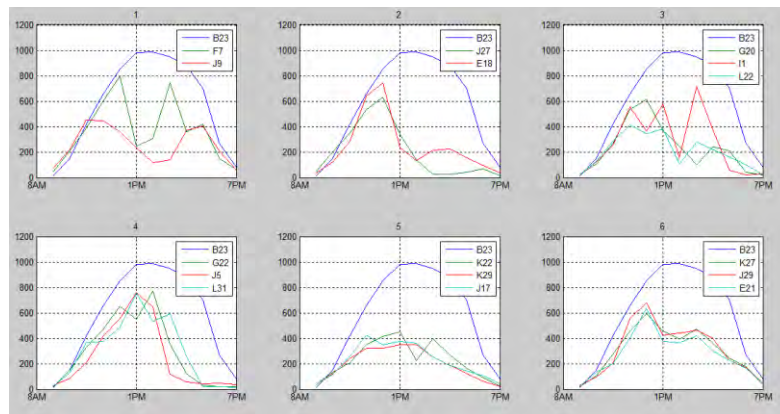


Figure 4.2: 35x35 SOM cluster 1-6

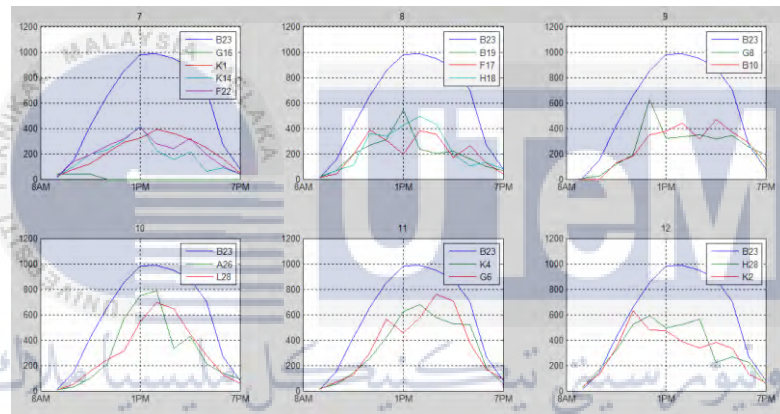


Figure 4.3: 35x35 SOM cluster 7-12

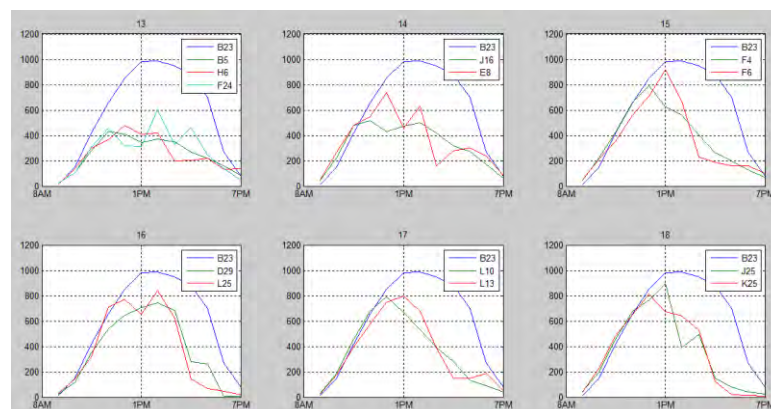


Figure 4.4: 35x35 SOM cluster 13-18

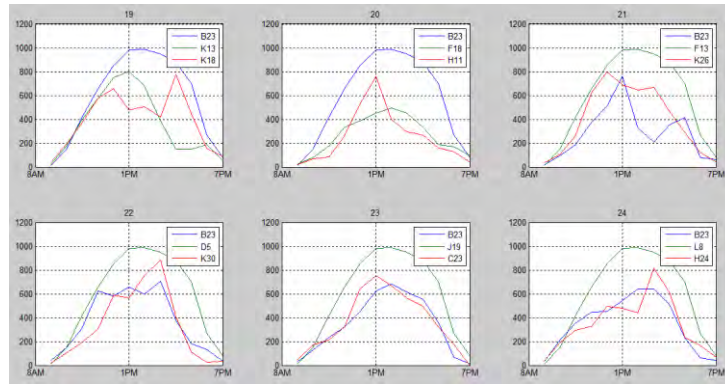


Figure 4.5: 35x35 SOM cluster 19-24

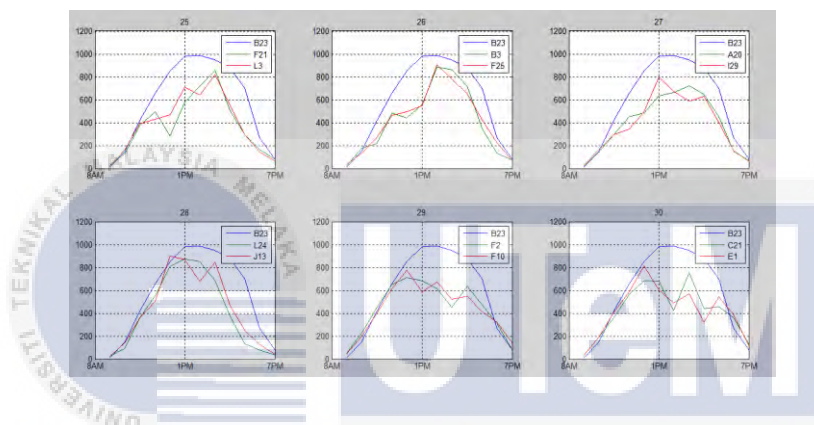


Figure 4.6: 35x35 SOM cluster 25-30

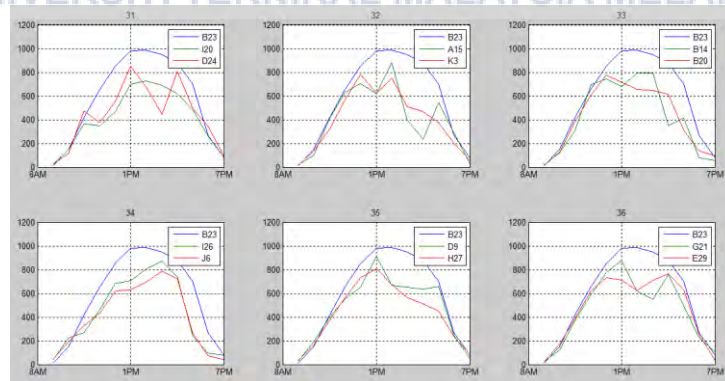


Figure 4.7: 35x35 SOM cluster 31-36

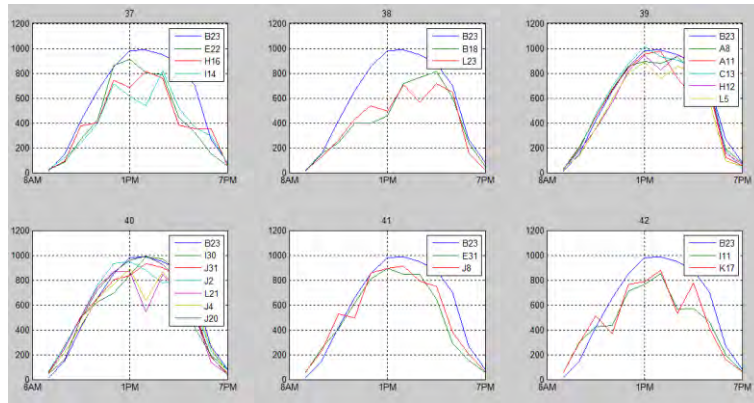


Figure 4.8: 35x35 SOM cluster 37-42

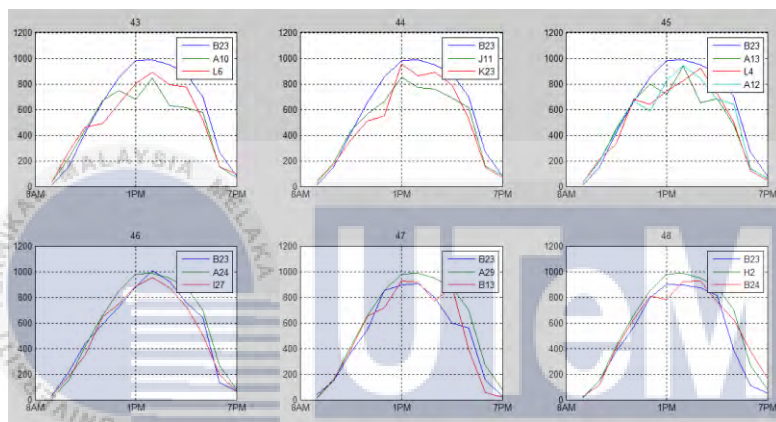


Figure 4.9: 35x35 SOM cluster 43-48

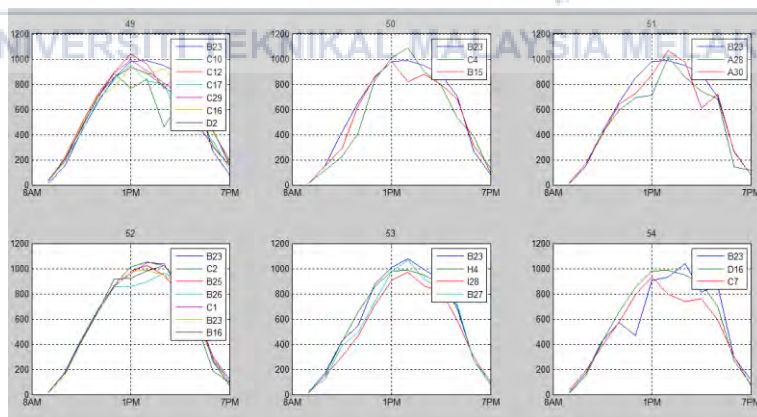


Figure 4.10: 35x35 SOM cluster 49-54



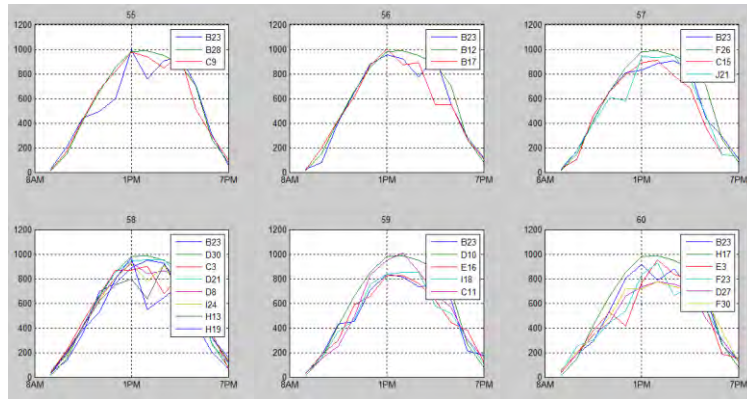


Figure 4.11: 35x35 SOM cluster 55-60

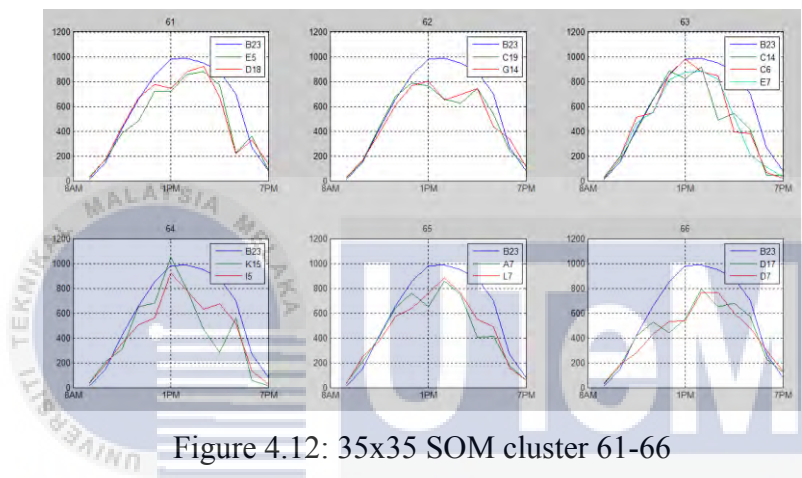


Figure 4.12: 35x35 SOM cluster 61-66

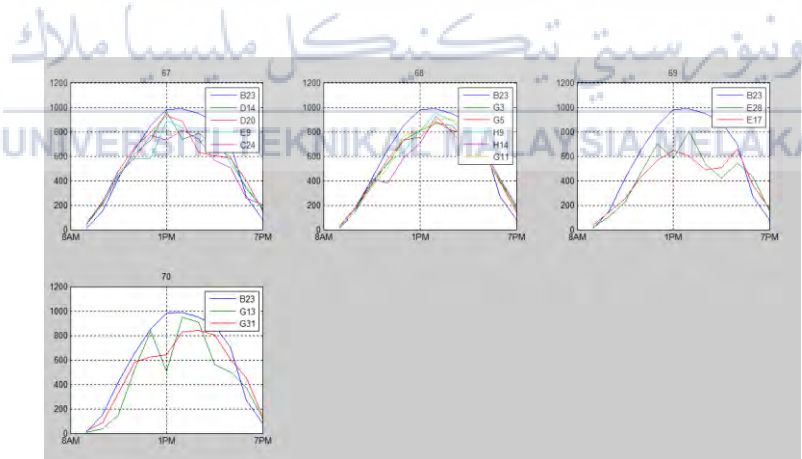


Figure 4.13: 35x35 SOM cluster 67-70

Table 4.2: Days clustered for 35x35

Categories	Days Clustered	Comments
Clear Sky	January(1), February(3), March(3), August(1), September(2)	There are a total of 10 clear sky days in the year 2016. Based on the SOM, month February and March has the most clustered days. A total of 3 clusters made up the clear sky days. However, cluster 52 had an error where two days with high variability was clustered as clear sky days.
Noon to evening variability	March(5), April(3), May(5), June(4), July(1), August(2), October(4), November(3), December(3)	The solar irradiance showed variability after noon for the total of 30 days. The months the variability mostly occurred was March and May with 5 days respectively.
Morning to noon variability	February(1), April(2), June(2), July(3), August(3), September(1), October(1), December(2)	There are a total of 15 days for morning to noon variability. The morning variability occurs mostly in the month of July.
Overall Variability (Low value)	January(1), March(1), July(1), October(2), December(2)	The overall variability which produced a low amount of solar irradiance happened in 7 days of the year.
Overall Variability (High value)	January(5), February(3), March(5), April(4), May(4), June(1), July(3), August(3), September(2), October(3), November(1), December(1)	The overall variability of a high amount of solar irradiance which almost reached clear sky days are 35 days. This happened mostly in the month of January and March.
Overcast	None Clustered	No days with readings below 300 were clustered.

#### 4.1.2 50x50 map size

Figure 4.14 shows the 50x50 self-organizing map (SOM) clusters. As observed, there are 7 clusters. The Euclidean distance is a lot lesser than the 35x35 map size, with many nodes being bright to light blue, having a distance of 0.0176 to 1.93. The euclidean distance for map size 50x50 is smaller than map size 35x35.

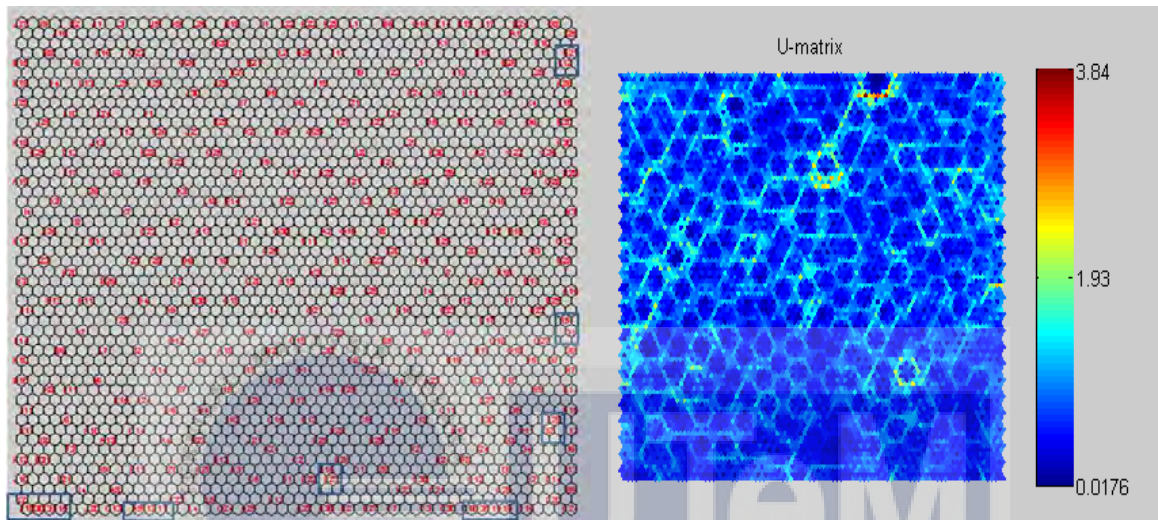


Figure 4.14: 50x50 SOM

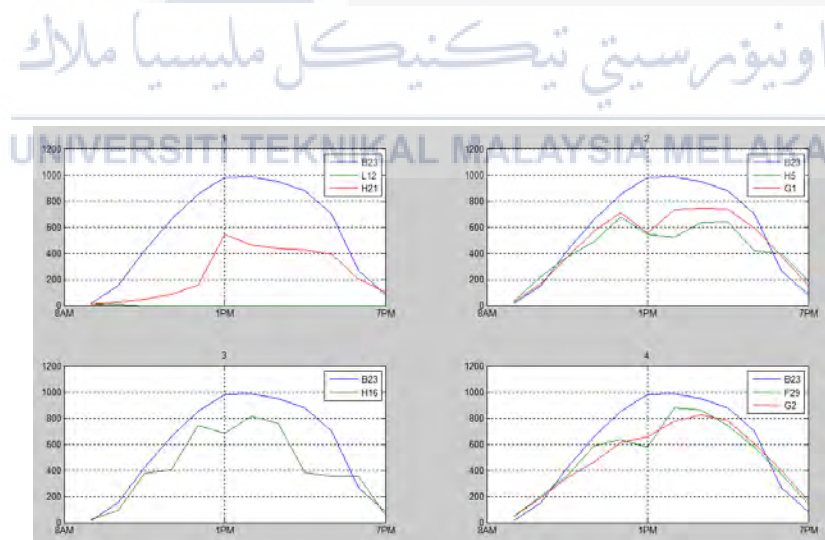


Figure 4.15: 50x50 SOM cluster 1-4

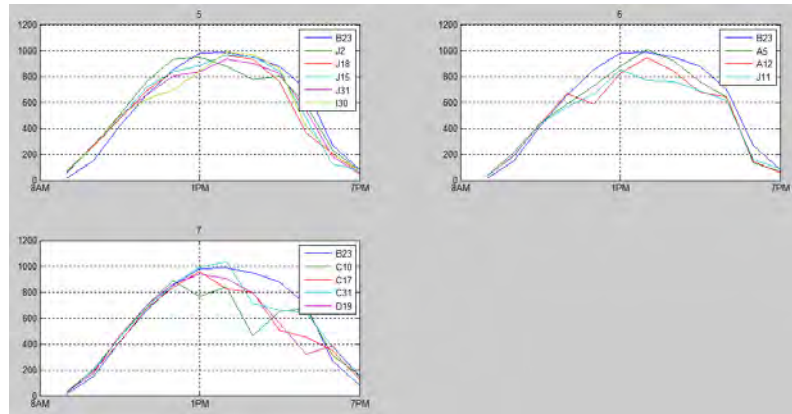


Figure 4.16: 50x50 SOM cluster 5-7



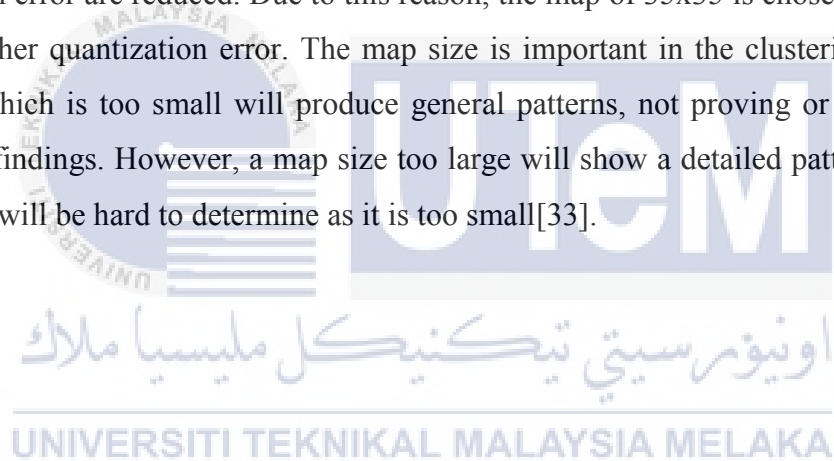
Table 4.3: Days clustered for 50x50

Categories	Days Clustered	Comments
Clear Sky	October(3)	There are a total of 3 clear sky days in the year 2016. The clear sky cluster was obtained from cluster 5 which had a mixture of other non-clear sky readings.
Noon to evening variability	March(3), April(1)	Only a total of 4 days were clustered. The cluster occurs the most in the March month.
Morning to noon variability	June(1), July(1)	There are a total of 2 days for morning to noon variability.
Overall Variability (Low value)	July(1), August(1)	The days clustered based on variability of low value is 2.
Overall Variability (High value)	January(2), October(1)	The overall variability of a high amount of solar irradiance are 3 days. This happened mostly in the month of January.
Overcast	None clustered	Cluster 1 had a reading of overcast days, however the days were not clustered accurately.



#### 4.1.3 Comparison of 35x35 and 50x50 map size

The 35x35 map clustered a total of 186 days into 70 clusters. Out of the 186 days, 104 days were clustered into 6 groups. The other days were uncategorized. The 50x50 map clustered 21 days into 7 clusters. Out of the 21 days, 14 days were chosen. The rest of the days were uncategorized as the pattern could not fit in the groups. Map 50x50 has a lower quantization error compared to 35x35. A lower quantization error produces a better quantization. The quantization error determines the map's performance and is based on the number of weights and neighbourhood size. Quantization error decreases with the decrease of weight numbers[34]. However, in this comparison, even though 50x50 has a lower quantization value, the clusters are few. This is due to overfitting of the data. The increase in map size causes the data to scatter more finely on the map. Therefore, the clusters and quantization error are reduced. Due to this reason, the map of 35x35 is chosen even though it has a higher quantization error. The map size is important in the clustering of data. A map size which is too small will produce general patterns, not proving or detecting any significant findings. However, a map size too large will show a detailed pattern where the differences will be hard to determine as it is too small[33].



## 4.2 Manual Clustering

Table 4.4: Manually clustered days

Categories	Days Clustered	Comments
Clear Sky	February(4), March(3), October(1)	There are a total of 8 clear sky days in a year. The month February has the most clear sky days.
Morning to noon variable	February(2), March(1), May(2), August(3)	There are 8 days where the solar irradiance shows fluctuations in the morning till noon. August has the most compared to the other 3 months.
Noon to evening variable	January(2), February(1), March(4), April(1), May(3), June(3), September(2), December(1)	There are a total of 17 days where the variations in the readings occur from noon till evening with the morning being clear. This condition often happens in March.
Overall Low	January(3), March(1), April(3), May(2), June(3), July(2), August(1), October(1), November(2)	The total for days which have low irradiance values and fluctuations throughout the day is 18 days.
Overall High	January(6), February(4), March(2), April(3), May(2), July(3), August(4), September(6), October(4), December(4)	For overall high, they are a total of 38 days which produce high readings with variations throughout the day. The highest are the months of January and September.
Overcast	February(4), March (2), July(1), December(1)	For overcast days, a total of 8 days were determined. February has the highest overcast days.

### 4.3 Comparison in SOM and manual clustering

Table 4.5: Comparison of SOM and manual clustering

Categories	Days clustered with SOM	Days clustered manually	Comments
Clear Sky	January(1), February(3), March(3), August(1), September(2)	February(4), March(3), October(1)	The total number of days clustered with SOM is 10 days while manually 8 days were clustered. There is a difference of 2 days between the clusters. The months which both have in common are February and March.
Noon to evening variability	March(5), April(3), May(5), June(4), July(1), August(2), October(4), November(3), December(3)	January(2), February(1), March(4), April(1), May(3), June(3), September(2), December(1)	SOM clustered 30 days while manually 17 days were determined. There is a difference of 13 days. The months in common are March, April, May, June, and December.
Morning to noon variability	February(1), April(2), June(2), July(3), August(3), September(1), October(1), December(2)	February(2), March(1), May(2), August(3)	15 days were clustered from SOM while 8 days were clustered from manual observation. There is a difference of 7 days. The months in common are February and August.
Overall Variability (Low value)	January(1), March(1), July(1), October(2), December(2)	January(3), March(1), April(3), May(2), June(3), July(2), August(1), October(1), November(2)	The overall variability for a lower value clustered for SOM is 7 days while manually is 18 days. SOM has 11 days less than manual clusters. The months in common are January, March, July, and October.
Overall Variability (High value)	January(5), February(3), March(5), April(4), May(4), June(1), July(3), August(3), September(2), October(3), November(1), December(1)	January(6), February(4), March(2), April(3), May(2), July(3), August(4), September(6), October(4), December(4)	SOM clustered 35 days while manually 38 days were determined. There is a difference of 3 days between the clusters. The clusters have 10 months in common.
Overcast	None Clustered	February(4), March(2), July(1), December(1)	SOM is unable to cluster overcast days while manually 8 days were clustered.

### 4.3.1 Analysis of cluster comparison

From table 4.5, we can determine that the self-organizing map (SOM) clusters the high value irradiance readings better. Clear sky and high value variability days were clustered more accurately compare to lower value days. The days of the months clustered for high value readings are also similar. As the irradiance value decreases, SOM clusters less days. SOM is unable to predict overcast days for the type of coding used here as it did not cluster any values less than 300W/m<sup>2</sup>.

## 4.4 Monthly Clusters

4 different map sizes were tried on each 12 months individually. Out of the 12 months, 3 months (January, February, and September) were analysed. Out of the 4 map sizes, the map size that had the best clustering was chosen. The results were compared with manual clustering.

### 4.4.1 January

Table 4.6 shows the number of map sizes tested. The map size of 9x8 was chosen as it had a sufficient number of clusters compared to the other map sizes. 8x9 had almost the same number of clusters; however it had a higher total percentage of error. The 10x8 reading gave results whose days could not be clustered. The 8x8 had too many days clustered together and made the separation of clusters difficult. Based on figure 4.23, the 9x8 map size produced 5 clusters. Manual clustering gives 6 clusters.

Table 4.6: Different map sizes for January

Map Size	Final Quantization Error	Final Topographic Error	Total Percentage of Error
10x8	1.199	0.000	6.4516
9x8	1.367	0.000	12.9032
8x8	1.352	0.000	6.4516
8x9	1.298	0.000	16.1290

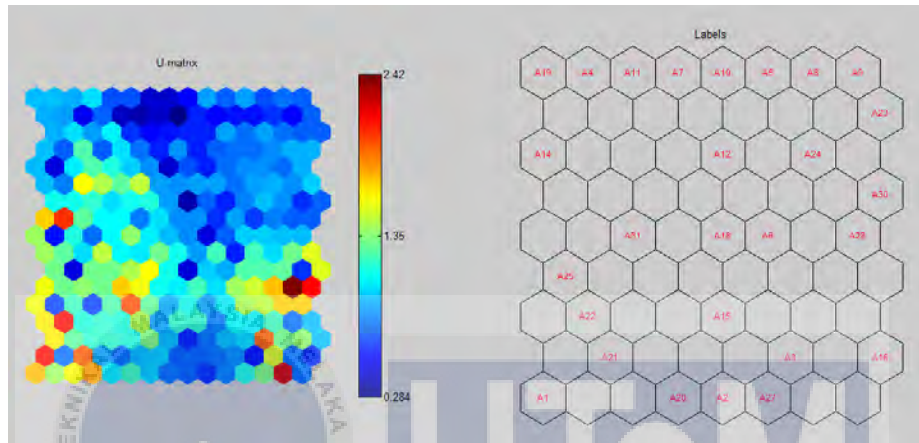


Figure 4.17: SOM January

Based on table 4.7, we observe that SOM has clustered 20 days into 5 groups while manually 26 days were clustered into 6 groups.

Table 4.7: Comparison of SOM and manual clusters for January

No	SOM clusters	Manual Clustering
1	A19, A4, A11, A7, A10, A5, A8, A9, A23	A19, A4, A14
2	A28, A30	A11, A5, A12
3	A18, A6	A8, A9, A24
4	A25, A22, A21	A25, A31
5	A20, A2, A27	A18, A30, A28, A6
6		A20, A2, A3

For SOM clustered days as seen in figure 4.18, the first group has a peak value of between 800 to 1000W/m<sup>2</sup>. Only one day has a sudden drop in the value reaching to zero and increasing again after the drop. The second group shows a similar shape with peak ranging from 1000 W/m<sup>2</sup> and above that. The third group shows similarities with peak between 800 to 1000W/m<sup>2</sup>. The fourth cluster shows inaccurate clustering. Cluster 5 has a peak between 600 to 800W/m<sup>2</sup> and has days of different shape. Our of the five clusters, we can group cluster 1, 2, and 3 into groups of high value with variability throughout the day which equals to 13 days and cluster 5 as low value with variability throughout the day which is 3 days. Based on figure 4.19, the first 5 clusters are grouped into days with high value with variability with peak between 800 to 1000W/m<sup>2</sup> which are 16 days. The last cluster is grouped into low value with variability which is 3 days. When compared to manual cluster, SOM is able to cluster accurately.

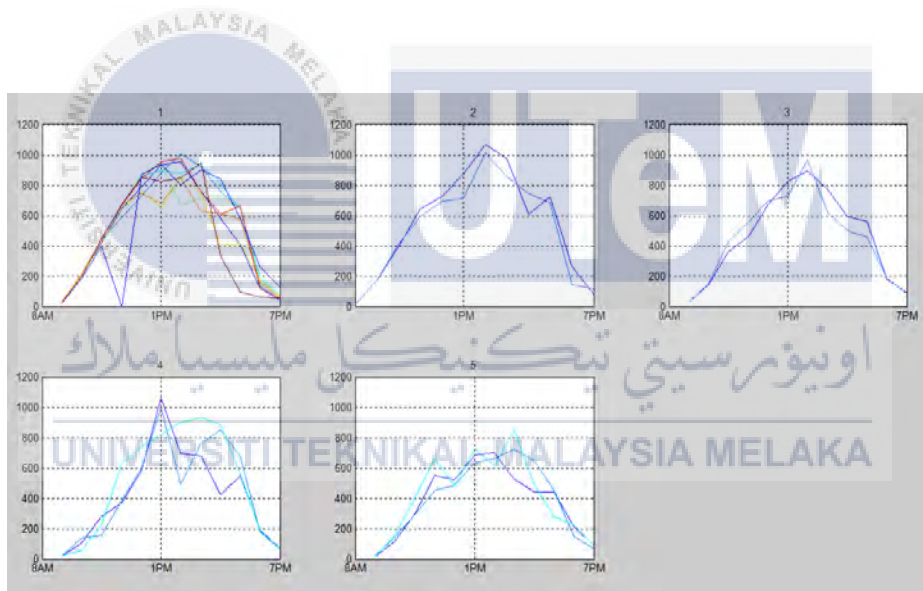


Figure 4.18: January SOM clusters

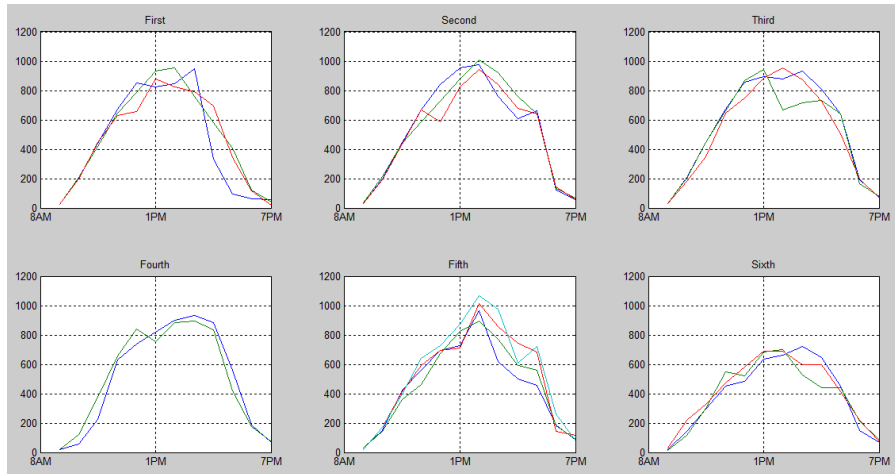


Figure 4.19: January manual clusters

### 4.4.2 February

Table 4.8 shows the number of map sizes tested. The map size of 8x9 was chosen as it had sufficient clusters even though the total percentage error was high. The other map sizes produced a clustered which could not be analysed as they were all connected. The 8x9 map size produced 4 clusters. Manual clustering found 4 clusters.

Table 4.8: Different map sizes for February

Map Size	Final Quantization Error	Final Topographic Error	Total Percentage of Error
10x8	0.587	0.000	20.6897
9x8	0.677	0.000	17.2414
8x9	0.642	0.000	24.1379
8x10	0.611	0.000	17.2414

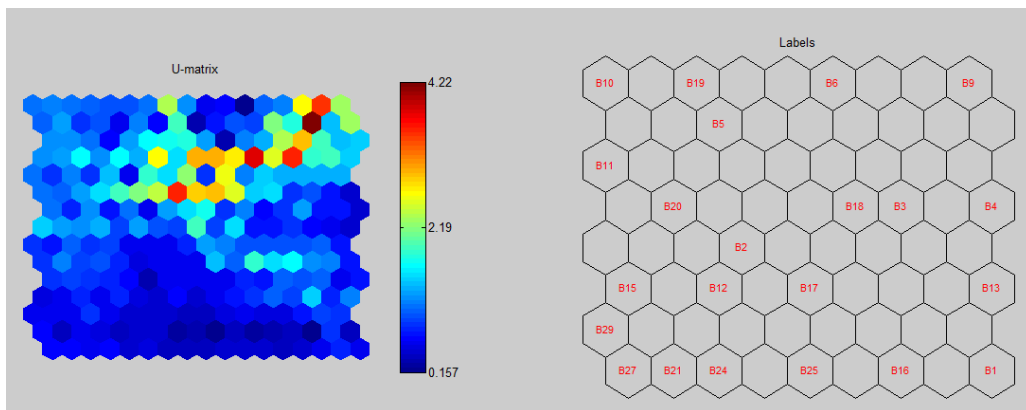


Figure 4.20: February SOM



Table 4.9: Comparison of SOM and manual clusters for February

SOM clusters	Manual Clusters
B19, B5	B10, B19, B5
B18, B3	B20, B18, B3, B4
B2, B12	B6, B7, B8, B9
B15, B29, B27, B21, B24	B15, B29, B24, B17, B25, B16, B13, B1

Based on figure 4.21, the first cluster has grouped 2 days and they have peak values of between 400 and 600W/m<sup>2</sup>. The second cluster has 2 days grouped and low values in the morning till late noon and higher values at approximately after 2pm. The third cluster has 2 days grouped and shows variability after 1pm. The last cluster has variability throughout the day with a peak between 800 to slightly above 1000W/m<sup>2</sup>. A total of 11 days were clustered. Cluster 1 is under the low value variability group. Cluster 2 (2days) is grouped as morning to noon variability. Cluster 3 is grouped in afternoon to evening variability. Cluster 4 is grouped as high value variability. Based on figure 4.22, the first cluster has low values below 600W/m<sup>2</sup> and is grouped in low value variability. There are 3 days in this group. The second cluster has morning to noon variability till approximately 3pm. 4 days are grouped in it. The 3<sup>rd</sup> cluster has 4 days as overcast as the values are below 300W/m<sup>2</sup>. The 4<sup>th</sup> cluster has 8 high value variability days. Based on the comparison, SOM is unable to cluster overcast days.

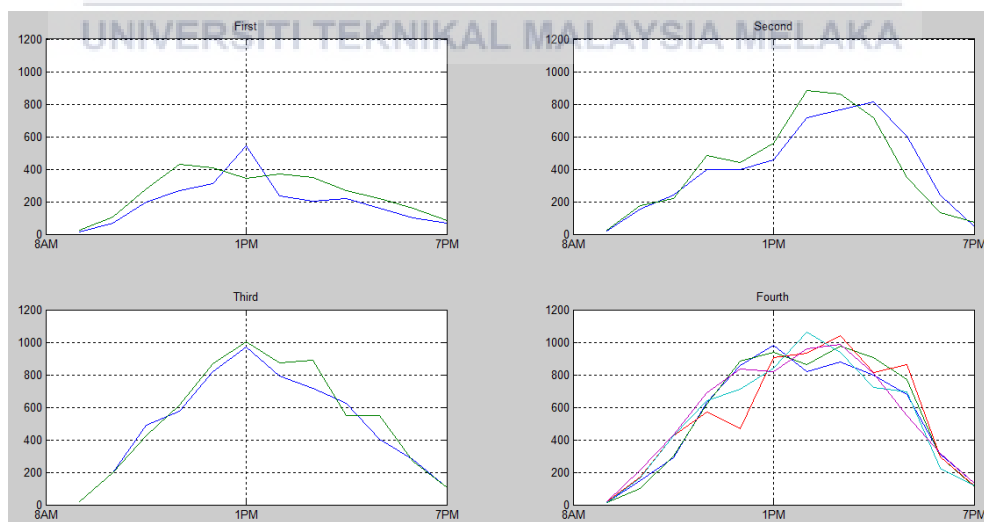


Figure 4.21: February SOM clusters 1-4



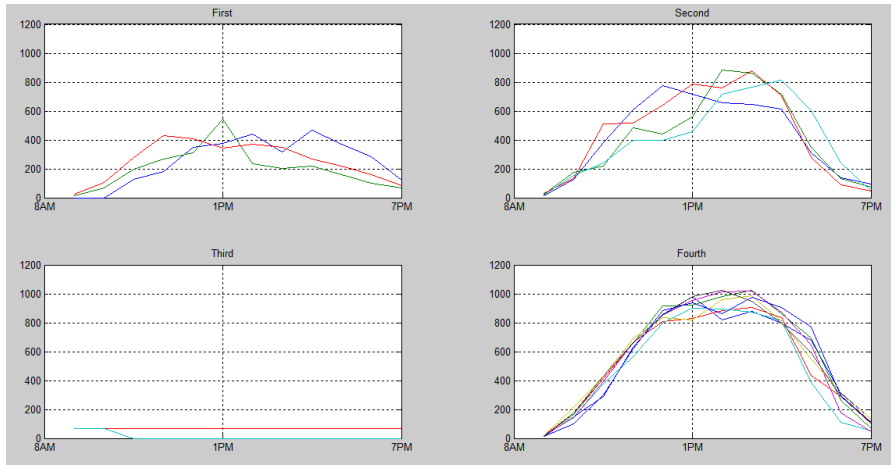


Figure 4.22: February SOM clusters 5-8

4.4.3 September

Table 4.10 shows the number of map sizes tested. The map size of 10x8 was chosen. It produced 6 clusters. Manual clustering gives 5 clusters.

Table 4.10: Different map sizes for September

Map Size	Final Quantization Error	Final Topographic Error	Total Percentage of Error
10x8	1.164	0.000	16.6667
8x10	1.177	0.000	10.0000
8x9	1.268	0.000	10.0000
9x8	1.307	0.000	10.0000

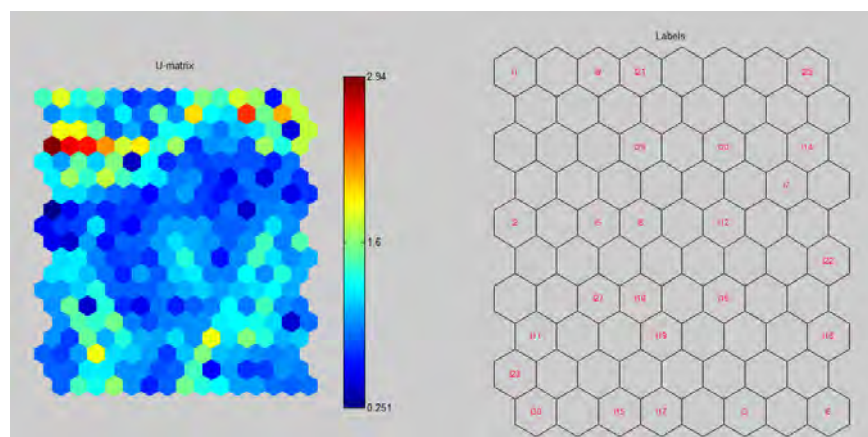


Figure 4.23: September SOM

Table 4.11: Comparison of SOM and manual clusters for September

No	SOM clusters	Manual Clusters
1	I9, I21	I14, I17, I20, I29, I12
2	I14, I17	I2, I5, I8
3	I5, I8	I27, I10, I19, I15, I7
4	I27, I10, I19	I3, I6
5	I11, I23, I30	I22, I18
6	I15, I7	

Based on figure 4.24, the first, 5<sup>th</sup>, and 6<sup>th</sup> cluster is inaccurately clustered. The second cluster is both have peaks which reach slightly above 800W/m<sup>2</sup>. The third and fourth cluster has peaks between 800 to 1000W/m<sup>2</sup>. A total of 7 days are clustered. The three clusters can be grouped under high value variability. Based on figure 4.25, the first cluster has a peak between 600 to 800W/m<sup>2</sup> and is clustered under low value variability. Clusters 2, 4, and 5 have peaks between 800 to 1000W/m<sup>2</sup> and are clustered under high value variability. Cluster 3 shows variability from noon to evening. A total of 17 days were clustered.

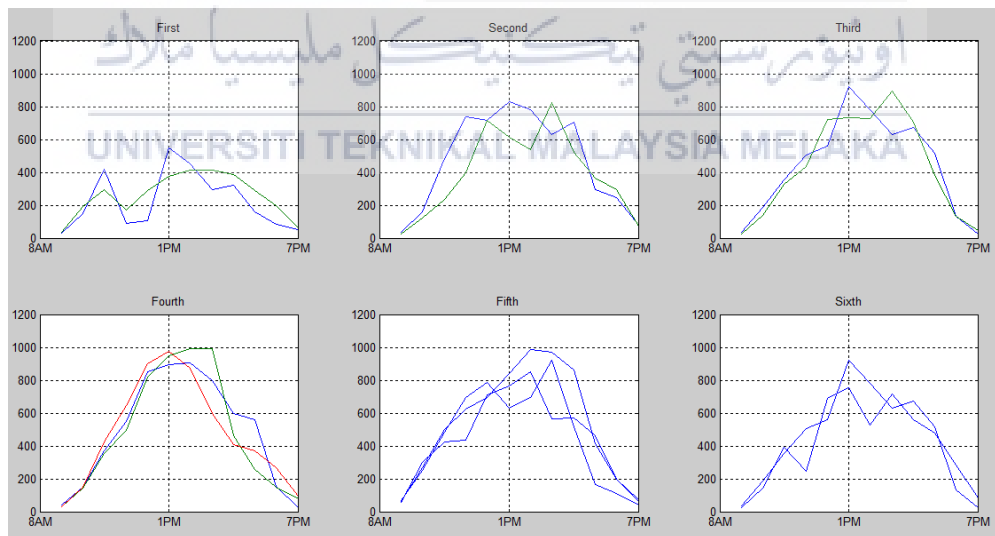


Figure 4.24: September SOM clusters 1-6

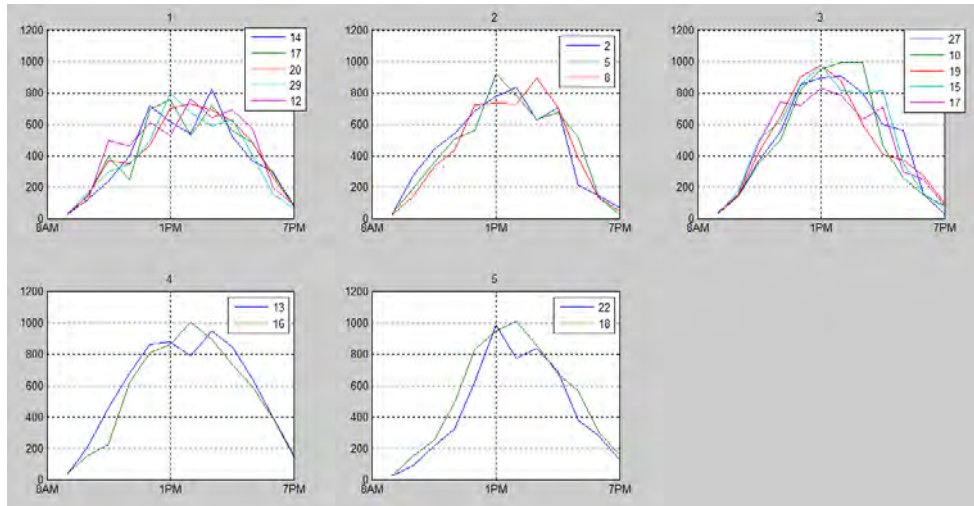


Figure 4.25: September SOM clusters 7-11

#### 4.4.4 Analysis of month clusters

As displayed in table 4.12, the self-organizing map is able to cluster the high value days better than low value days and does not cluster overcast days. The clustering of the days in a month is similar to the clustering of the year.

It is observed that monthly clustering produces more days off the month than clustering of days in a year. From the days of the year clustered, the results showed 7 days from January, 9 days from February, and 5 days from September. From the days clustered from the individual months, 16 days were clustered from January, 11 days from February, and 7 days were clustered from September. It is harder to determine the similar clusters of days in a year than in a month.

Table 4.12: Comparison of SOM and manual clusters for months

Categories	January		February		September	
	SOM	Manual	SOM	Manual	SOM	Manual
High value variability	13	16	5	8	7	7
Low value variability	3	3	2	3	0	5
Morning to noon variability			2	4		
Noon to evening variability			2	0	0	5
Overcast			0	4		

## CHAPTER 5

### CONCLUSION AND RECOMMENDATION

#### 5.1 CONCLUSION

Solar irradiance data of a year was clustered using Self-organizing map (SOM) based on the variability of the days. The 35x35 map size was chosen to cluster the year. From the topology map, 186 days were clustered into 70 clusters. However, only 104 days were chosen to be clustered into 6 groups. The days were clustered into 6 groups such as clear sky, overcast, high value variability, low value variability, morning to noon variability, and noon to evening variability. Based on the findings of the cluster, the self-organizing map (SOM) has the ability to accurately cluster days with high value irradiance days into groups. The clustering efficiency decreases for days with lower irradiance readings. The group with the most number of days were the high value variability group, which had high irradiance value between 800 to 1000W/m<sup>2</sup>, but it does not fit the clear sky criteria. The months with the most clear sky days were February and March. 10 months had mostly high value irradiance with overall variability. February had the most number of overcast days. Days in a month were clustered to check the clustering pattern and had the same results as a year where days with high irradiance were clustered better. The limitation of SOM is its inability to cluster overcast days, which are values less than 300W/m<sup>2</sup>. Another limitation is its inability to cluster certain days with patterns out of the group range. This limitation causes the prediction of the irradiance on cloudy days or days with low irradiance to be inaccurate.

## 5.2 RECOMMENDATION

To increase the efficiency of the clustering, the use of supplementary tools which can improve the classification of SOM can be implemented. A few researches have been done to improve the use of self-organizing map (SOM), such as Game Theory SOM. It is a hybrid forecasting method with the use of Neural gas (NG) and competitive Hebbian Learning (CHL) to improve the map quality. Non-winning neurons have their competition with winning neurons increased for more input patterns[31]. Another method is the combination of SOM and other methods. Combination of methods such as learning vector quantization (LVQ)[6] or Jaccard New Measure[35] have been researched and proved successful. Other approaches are done by improving the weight adjustments of the neurons by Frequency Sensitive Competitive Learning (FSCL) algorithm for better neuron order and less quantization and topography error [36].



## REFERENCE

- [1] G. Masson, "A Snapshot of Global PV(1992-2015)," 2016.
- [2] T. K. (Solar E. I. Association), "Solar Industry Experiences Record-Breaking Growth," *Ecowatch*, 2016. [Online]. Available: <http://www.ecowatch.com/solar-energy-record-growth-2003130851.html>. [Accessed: 15-Dec-2016].
- [3] L. V Bijuraj, "Clustering and its Applications," pp. 169–172, 2013.
- [4] C. Trueblood *et al.*, "PV Measures Up for Fleet Duty," *IEEE Power Energy Mag.*, no. March/April, pp. 33–44, 2013.
- [5] S. Magee, "Solar & Weather," in *Solar Irradiance & Isolation for Power System*, USA: CreateSpace Independent Publishing Platform, 2910, p. 7.
- [6] H. Yang, S. Member, C. Huang, Y. Huang, and Y. Pai, "A Weather-Based Hybrid Method for 1-Day Ahead Hourly Forecasting of PV Power Output," vol. 5, no. 3, pp. 917–926, 2014.
- [7] A. Singla, K. Singh, and V. K. Yadav, "Environmental Effects on Performance of Solar Photovoltaic Module," 2016.
- [8] B. Kirby and M. Milligan, "Facilitating Wind Development : The Importance of Electric Industry Structure Facilitating Wind Development : The Importance of Electric Industry Structure," no. May, 2008.
- [9] J. C. Smith, M. R. Milligan, E. A. DeMeo, and B. Parsons, "Utility wind integration and operating impact state of the art," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 900–908, 2007.
- [10] A. Mills *et al.*, "Dark shadows: Understanding variability and uncertainty of photovoltaics for integration with the electric power system," *IEEE Power Energy Mag.*, vol. 9, no. 3, pp. 33–41, 2011.
- [11] S. Su, Y. Yan, and H. Lu, "A Classified Irradiance Forecast Approach for Solar PV Prediction Based on Wavelet Decomposition," pp. 1–5, 2016.
- [12] M. Lave and J. Kleissl, "Testing a Wavelet-based Variability Model(WVM) for Solar PV Power Plants," *Summited to IEEE J.*, pp. 1–6, 2012.

- [13] M. Lave, J. Kleissl, and J. S. Stein, "A Wavelet-Based Variability Model ( WVM ) for Solar PV Power Plants," vol. 4, no. 2, pp. 501–509, 2013.
- [14] C. Hansen, J. Stein, and A. Ellis, "Statistical criteria for characterizing irradiance time series," *Sandia Natl. Lab. ...*, no. October, 2010.
- [15] J. Stein, C. Hansen, and M. Reno, "The variability index: A new and novel metric for quantifying irradiance and PV output variability," *World Renew. Energy ...*, no. May, pp. 1–7, 2012.
- [16] K. Y. Bae, H. S. Jang, and S. Member, "Hourly Solar Irradiance Prediction Based on Support Vector Machine and Its Error Analysis," vol. 8950, no. c, pp. 1–11, 2016.
- [17] P. Mathiesen, D. Rife, and C. Collier, "Forecasting Solar Irradiance Variability Using the Analog Method," *43rd IEEE Photovolt. Spec. Conf.*, pp. 1207–1211, 2016.
- [18] M. Mitchell, M. Campbell, K. Klement, and M. Sedighy, "Power Variability Analysis of Megawatt-Scale Solar Photovoltaic Installations," pp. 1–4, 2016.
- [19] H. Moaveni, D. K. Click, R. H. Meeker, R. M. Reedy, and A. Pappalardo, "Quantifying solar power variability for a large central PV plant and small distributed PV plant," *Conf. Rec. IEEE Photovolt. Spec. Conf.*, vol. 2, pp. 969–972, 2013.
- [20] I. Jayawardene, S. Member, G. K. Venayagamoorthy, and S. Member, "Spatial Predictions of Solar Irradiance for Photovoltaic Plants," no. October 2011, pp. 267–272, 2016.
- [21] M. Lave and R. Broderick, "Characterizing Local High - frequency Solar Variability for use in Distribution Studies," pp. 3454–3456, 2014.
- [22] Z. Wang, I. Koprinska, and M. Rana, "Clustering Based Methods for Solar Power Forecasting," pp. 1487–1494, 2016.
- [23] Z. M. Zin, "Cluster and Visualize Data Using 3D Self-Organizing Maps," no. Urai, pp. 163–168, 2014.
- [24] S. Wu and T. W. S. Chow, "Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density," vol. 37, pp. 175–188, 2004.
- [25] B. C. Love, "Comparing supervised and unsupervised category learning," vol. 9, no. 4, pp. 829–835, 2002.
- [26] M. Kubat, "Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7.," *The Knowledge Engineering Review*, vol.



- 13, no. 4. pp. 409–412, 1999.
- [27] L. J. Moreira, S. Paulo, L. A. Silva, and S. Paulo, “Maps and Informative Nearest Neighbor,” pp. 706–713, 2016.
- [28] J. Vesanto and E. Alhoniemi, “Clustering of the self-organizing map,” *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 586–600, 2000.
- [29] Z. Dong, D. Yang, T. Reindl, and W. M. Walsh, “Satellite image analysis and a hybrid ESSS/ANN model to forecast solar irradiance in the tropics,” *Energy Convers. Manag.*, vol. 79, pp. 66–73, 2014.
- [30] C. Li, C. Liu, Y. Huang, D. Liu, and J. Zhang, “Research on Analysis Method for Photovoltaic Power Considering Weather Type,” no. 2.
- [31] M. Ghayekhloo, M. Ghofrani, M. B. Menhaj, and R. Azimi, “A novel clustering approach for short-term solar radiation forecasting,” *Sol. Energy*, vol. 122, pp. 1371–1383, 2015.
- [32] P. K. Prajapati and M. Dixit, “Un-Supervised MRI Segmentation Using Self Organised Maps,” *2015 Int. Conf. Comput. Intell. Commun. Networks*, no. 2, pp. 471–474, 2015.
- [33] M. Chattopadhyay, P. K. Dan, and S. Majumdar, “Application of Visual Clustering Properties of Self Organizing Map in Machine-part Cell Formation,” pp. 1–33.
- [34] Y. Sun, “On quantization error of self-organizing map network,” vol. 34, no. April, pp. 169–193, 2000.
- [35] A. Ahmed, “An Improved Self-Organizing Map for Bugs Data Clustering,” no. October, pp. 135–140, 2016.
- [36] V. Aggarwal, A. K. Ahlawat, and B. N. Pandey, “A Weight Initialization Approach for Training Self Organising Maps for Clustering Applications,” pp. 1000–1005, 2012.





## APPENDIX B

### MATLAB CODING

```

clear
clf reset
f0 = gcf;
clc

sD=som_read_data('1yearXX.txt','X');
sD = som_normalize(sD,'var');
sM = som_make(sD,'msize',[10 10]);
sM = som_autolabel(sM,sD,'vote');

figure(1);
som_show(sM,'umat','all');

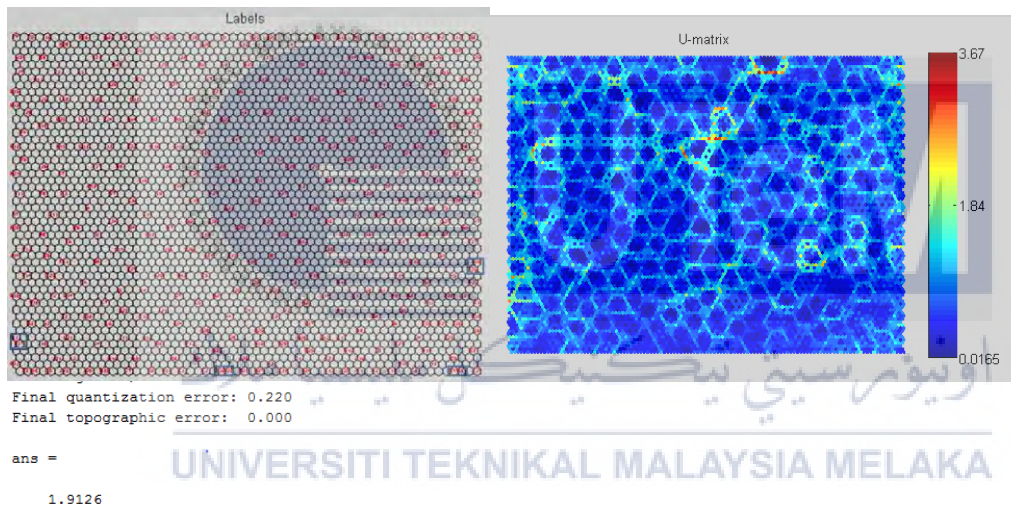
figure(2);
som_show(sM,'umat','all','empty','Labels');
som_show_add('label',sM.labels,'TextSize',8,'TextColor','r')
sD2 = som_label(sD,'clear','all');
sD2 = som_autolabel(sD2,sM);           % classification
ok = strcmp(sD2.labels,sD.labels);    % errors
100*(1-sum(ok)/length(ok))          % error percentage (%)

```

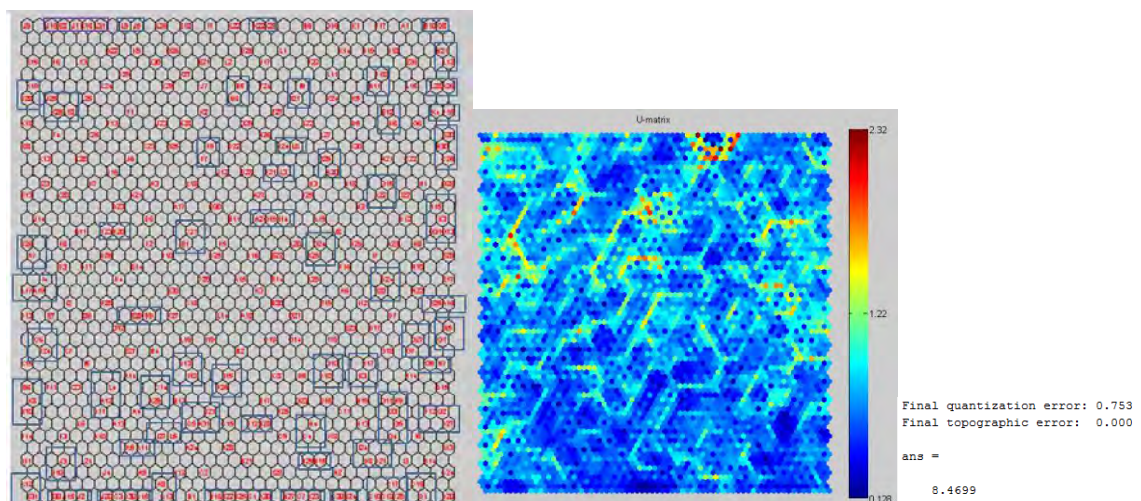
## APPENDIX C1

## SOM SIMULATION OF ONE YEAR FOR DIFFERENT MAP SIZE

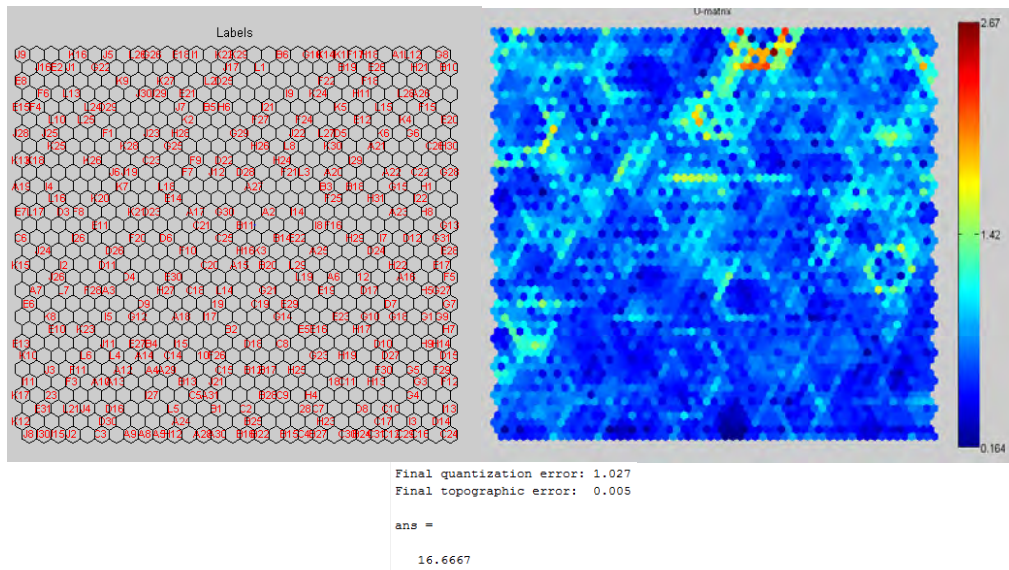
50x60



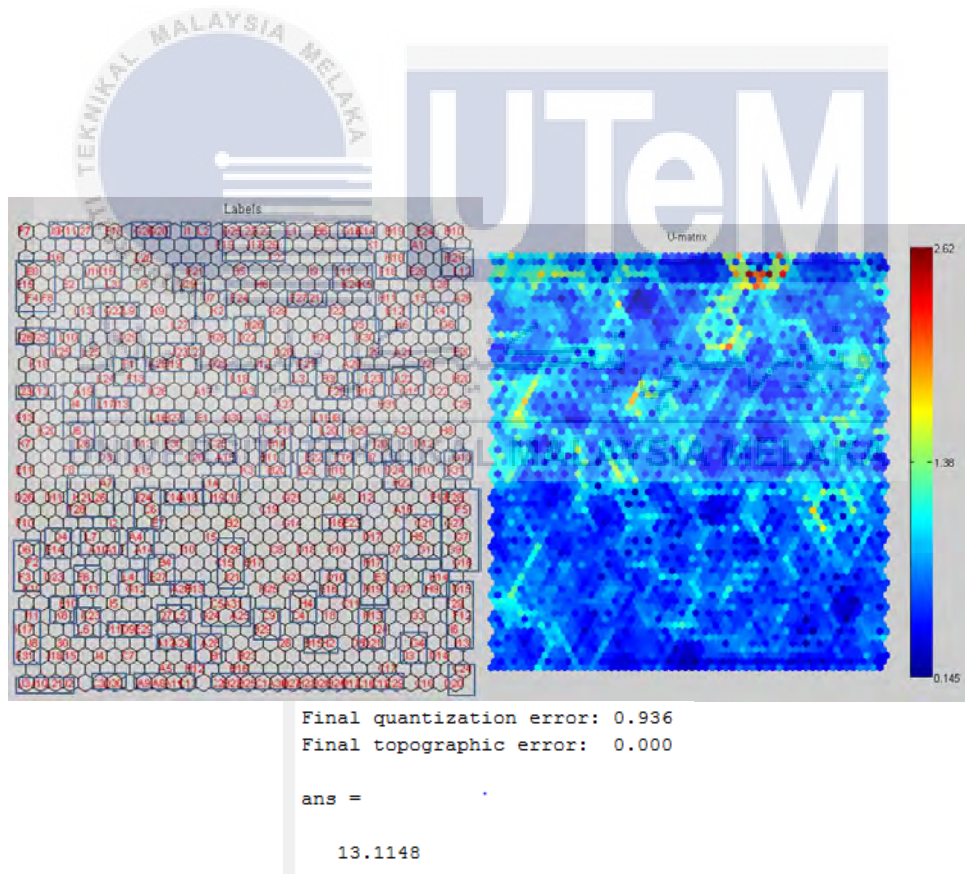
40x35



30x30

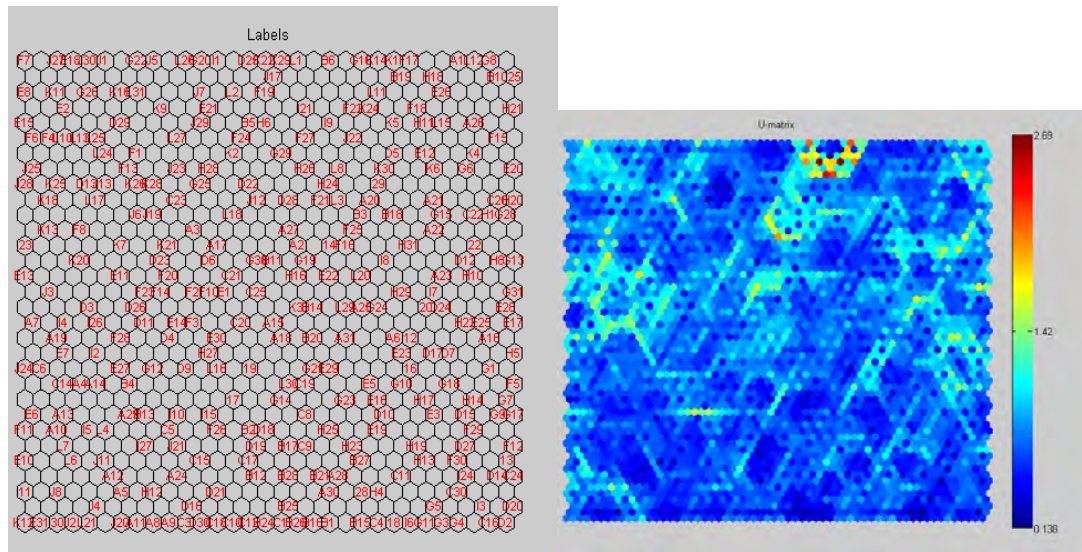


35x30

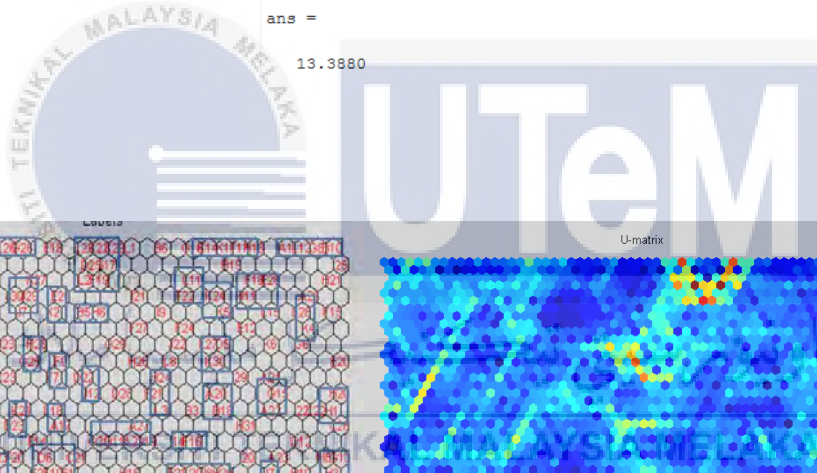
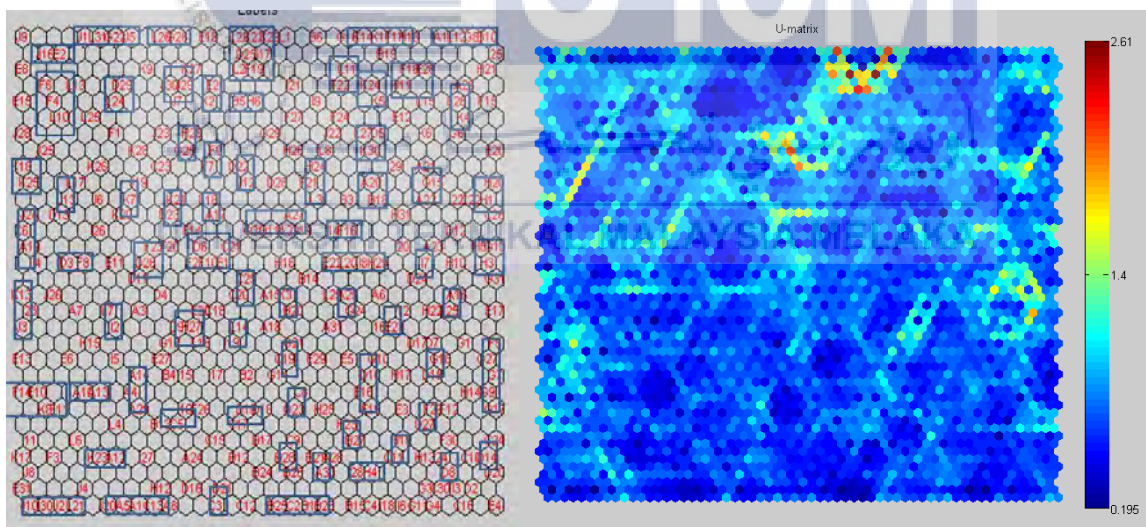




31x31



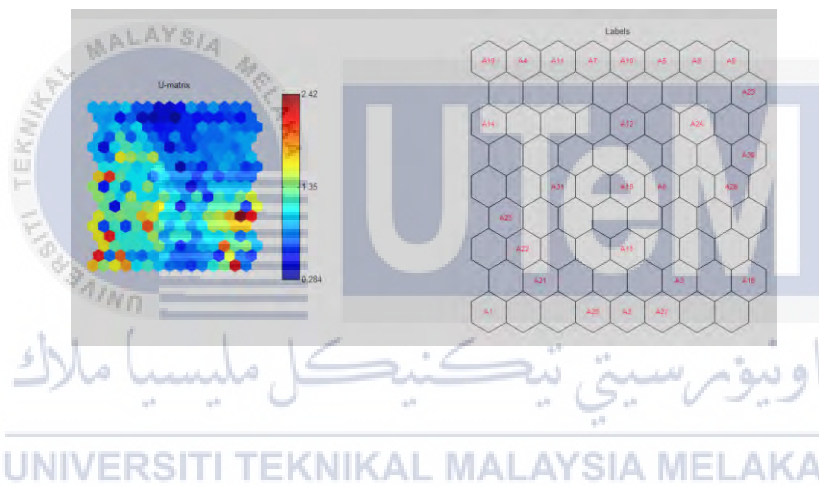
30x31



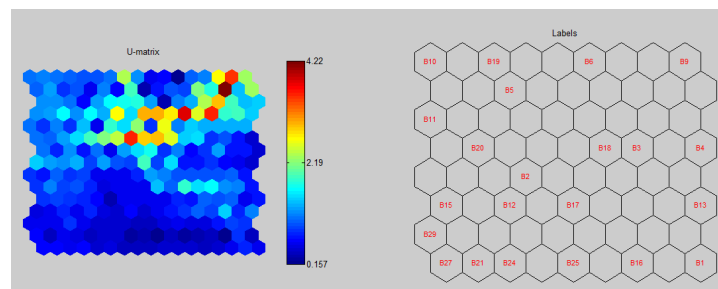
## APPENDIX C2

## SOM MONTHLY CLUSTER

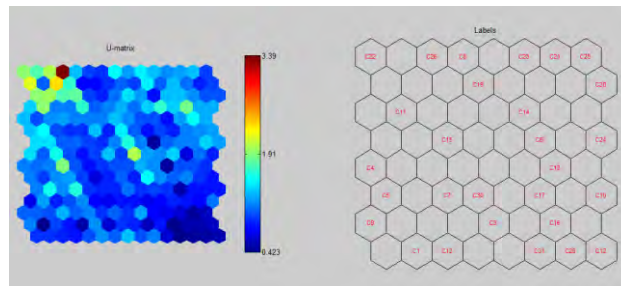
January



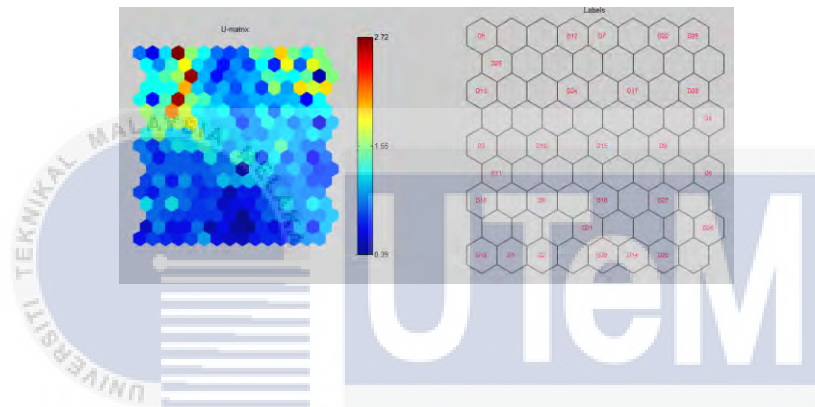
February



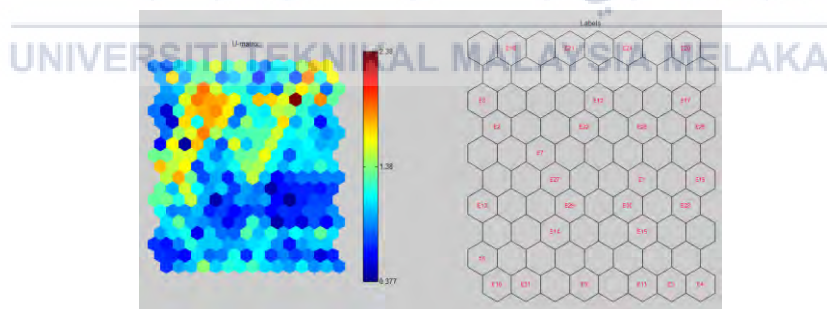
March



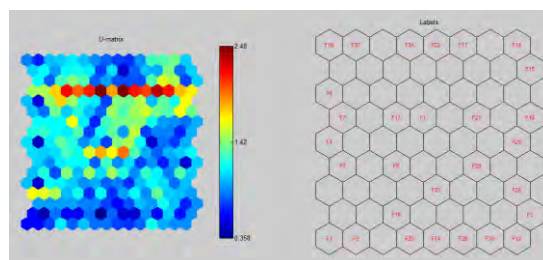
April



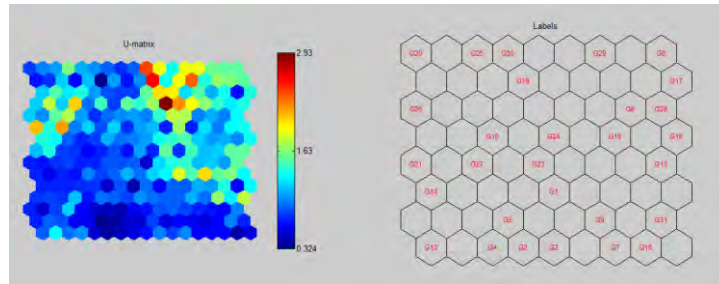
May



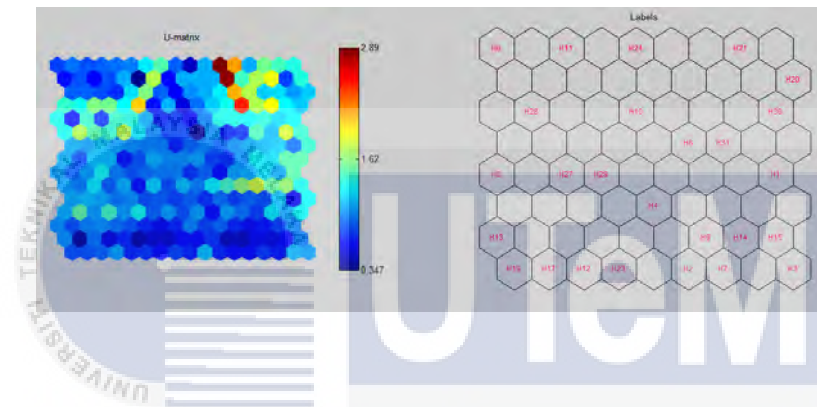
June



July

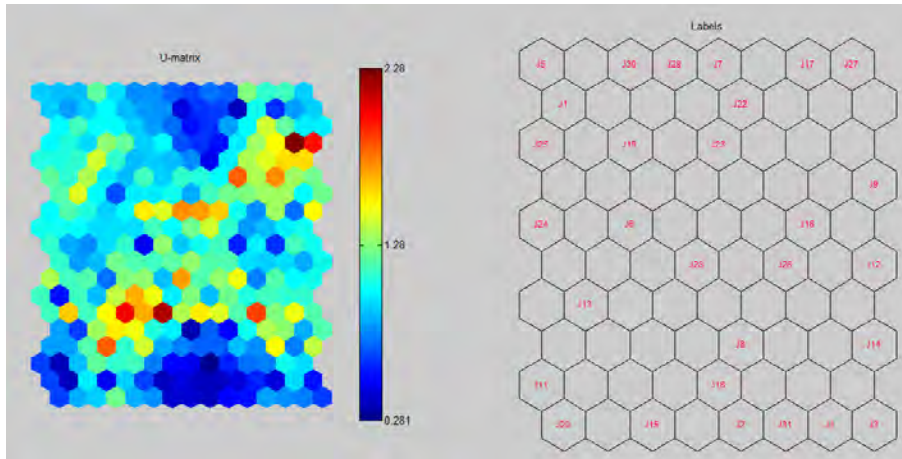


August

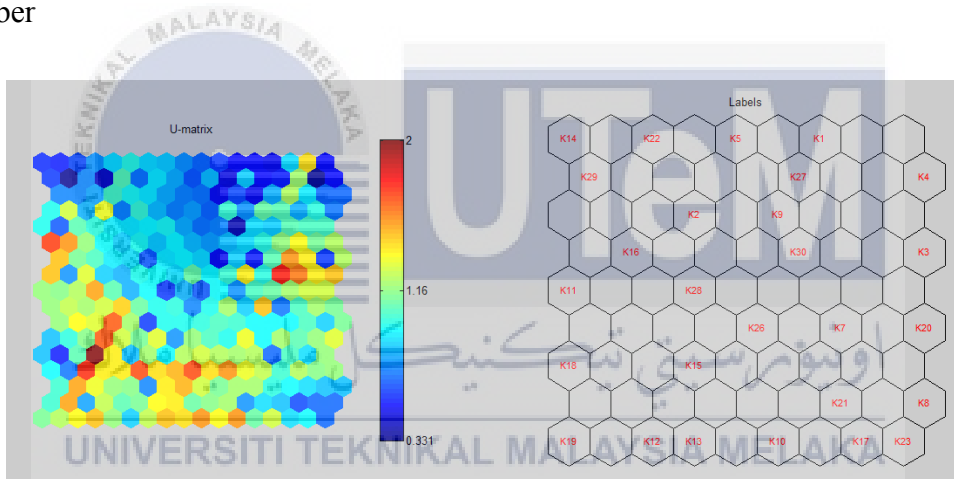




October



November



Dec

