IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 11, No. 4, December 2022, pp. 1333~1343
ISSN: 2252-8938, DOI: 10.11591/ijai.v11.i4.pp1333-1343

Investigation of energy demand correlation during pandemic using self-organizing map algorithm

Mohamad Fani Sulaima[1], Sharizad Saharani[1], Arfah Ahmad[1], Elia Erwani Hassan[1], Zul Hasrizal Bohari[2]
[1]Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[2]Faculty of Electrical and Electronic Engineering Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

Article Info

Article history:

Received Sep 1, 2021
Revised Jun 23, 2022
Accepted Jul 22, 2022

Keywords:

Coronavirus disease 2019 pandemic
Data analysis
Energy demand
Neural network
Self-organizing mapping

ABSTRACT

The world faces a significant impact from the coronavirus disease 2019 (Covid-19) pandemic, which also influences energy consumption. This study investigates the substantial connection of the classified data between power consumption, cooling degree days, average temperature, and covid-19 cases information using mathematical and neural network approaches regression analysis, and self-organizing maps. It is well established that various data mining methods have revamped the classification process of data analytics. Specifically, this study investigates the correlation between the collected variables using regression analysis and selecting the best-matching unit under the normalization method using self-organizing maps. The self-organizing maps become better when the datasets have variations; the result denotes that this method produced high mapping quality based on the map size and normalization method. Furthermore, the data crossing connection is indicated using the regression analysis method. Finally, the classified data results during the movement control order are validated in self-organizing maps to achieve the study objective. By performing these methods, this study established that the correlation between the energy demand towards cooling degree days, average temperature, and covid-19 cases is very weak. The verification has been made where the 'logistic' normalization method has produced the best classification result.

Corresponding Author:

Mohamad Fani Sulaima
Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100, Durian Tunggal, Melaka, Malaysia
Email: fani@utem.edu.my

1. INTRODUCTION

It needs an effective method when dealing with large-scale data for a better future outcome and forecasting [1], mostly when data's growth rate was fast according to the recent trend for the past decades. In 2020, Tenaga Nasional Bhd. Malaysia is now developing a data analytics application for constructing a large solar (LLS) farm in Sepang to improve its operational efficiency [2]. This initiative is for the plan to reduce natural coal usage as a source of electricity generation. The energy demand in Malaysia increased in 2017 due to residential area development projects exceeding 64.4% [3]. This graph-based data classification will help the Malaysian government implement a roadmap to drive efficiency measures to attain progressive energy savings across all sectors. Still, the challenges arise when the information has many factors to be considered [4], [5] to promote energy efficiency in demand side consumers. However, the unsupervised method is required to effectively gain insight into the future to find the correlation between the available variables contributing to the energy demand. Thus, the study analysis supported by regression analysis and

unsupervised self-organizing maps is crucially needed to prove the significant correlation of the independent and dependent data, especially during pandemic Covid-19. Various data mining methods have improved data analytics classification and regression processes. Also, it is needed, especially in dealing with data from the past several years, which is time-consuming and complex to be performed by humans [6].

The regression analysis is a mathematical method for finding relationships between a dependent variable, Y, and an independent variable, X, known as factors affecting the Y. The dependent variable is modeled as being matched by one or more independent variables, known as explanatory variables [7]. The regression method can also investigate the connection strength between 2 or more variables indicated as R^2 [8]. The value of R^2, which is more than 0.75, is considered an accepted model [9]. In the previous study, regression analysis is widely used for forecasting analysis. For example, communication technology has used regression for energy demand forecasting [10]. However, the regression analysis method is considered first layer analysis due to direct calculation settings when dealing with the low connection between data sets or having a low R^2 value. So, it is crucial for any correlation study to get verified by other unsupervised methods such as neural network and self-organizing maps (SOM) algorithm. SOM is an unsupervised clustering method that converts data into low-dimensional visuals in many research areas [11]–[13]. However, a few prior studies related to energy with self-organizing maps exist clustering provides patterns based on similarity for the unlabeled data called clusters [14]. Kohonen stated that self-organizing maps are widely recognized for industry clustering problems and data study [15]. In the power system application, the self-organizing maps are one of the best methods used in the classification process of fair and defective power distribution transformers [16]. The research uses numerical data for self-organizing maps to separate the defect transformer practically.

In past studies, the SOM algorithm was adopted to evaluate the power distribution network to classify the characteristic of buses consisting of 33-bus and 69-bus [17]. It is stated that the SOM can analyze the bus's features by inserting the data information consisting of its power triangle for each of the buses. SOM can be used as a forecaster. Atira *et al.* [18] have applied SOM to forecast the tested data in the medium-term load data by observing the maps and the actual and forecasted data error. The SOM is widely used for isolation analysis between 2 different contradictions of data as well. As presented by [19], the SOM is used to investigate the health of high voltage equipment by measuring its partial discharge level regarding the equipment insulation healthiness. It is found that the SOM can isolate the data between the healthy equipment and having partial discharge. Also, SOM has been successfully used in fault diagnosis [20]. The SOM was able to detect and classify faulty machines based on distance-preserving. The study was performed on the gearbox and bearing in a device by observing its raw vibrations signal coming out from it. The study also found that the SOM can be embedded learning methods into various applications. Using signal processing tools, one can understand the data collected to perform semisupervised learning in various industrial applications. However, the discovered studies are limited by their scope, techniques, and variables. As the positive side of the SOM that has high mapping quality and in contrast and applied competitive learning to error-correction learning [21], [22], the SOM algorithm also differs from other neural networks in the sense that the SOM uses the neighborhood function to sustain the topological properties of the input space [23]. Based on the evidence from the references, the self-organizing maps are applied in another area. Still, there is less research regarding data analytics related to energy consumption patterns [24]. To the best of my knowledge, there is less study on implementing SOM for the data correlation verification. The function of this algorithm would vary and depend on the application required to optimally function.

Therefore, in this study, the SOM algorithm is chosen to investigate the data finding a significant connection between energy consumption in Malaysia, cooling degree days (CDD), numbers of Covid-19 cases, and the average temperature in Malaysia to prove the results of using single multiple regression analysis. The SOM algorithm has validated significant correlation using the best matching unit under the normalization method. This study will help the government and energy providers understand the energy information pattern before and during the movement control order in Malaysia. The significant correlation verification to the several variables is analyzed. Hence, the arrangement of the paper by following this sequence. The research method is in section 2, while section 3 presents results and discussion. The conclusion of the study is explained in section 4 accordingly.

2. RESEARCH METHOD

This section explains the flow of methods on how the regression analysis and SOM algorithm has been implemented. The critical step was to define the variables data related to the energy demand while finding a significant correlation between them. In this study, three related factors were identified. They were CDD, average temperature, and Covid-19 cases. The technique is separated into the following sub-section.

2.1. Variables data formulation

The power consumption demand as a dependent variable is measured in megawatts (MW). The data is gathered from the grid system operator (GSO) within two years periods, which are from 18th March 2019 until 17th March 2020 (before the movement control order (MCO)) and 18th March 2020 until 17th March 2021 (during the Pandemic MCO). This study uses three independent variables collected from various external sources. The first variable is CDD in Malaysia. It is defined that the CDD is the difference between the daily mean and reference temperature [25]. Cooling demand is influenced by factors such as air-conditioning purchasing capacity per capita and operating hours [26], and the population can be a paramount factor. The CDD is generated from the Sepang/KL International Airport weather station, MY (101.70E, 2.72N). The CDD is based on the temperature of Peninsular Malaysia on the equator, and cooling demand is always there and rising during the daytime. CDD is calculated by using (1).

$$CDD=\sum_{i=1}^{n} rd(Ti-Tb) \quad (1)$$

Where n is the number of days in a year, Ti is the daily mean temperature for the day i, Tb is the reference temperature for cooling.

Meanwhile, the second variable is the average temperature, one of the variables generated from (the OGIMET website) and can be calculated using (2). The average temperature data followed energy data as well. The last variable data is the information on the Covid-19 case taken from the Ministry of Health (MOH) website from 18th March 2020, where covid-19 cases were first reported in Malaysia until 17th March 2021. The significant covid circumstances have influenced government decisions on the MCO condition and requirement. Thus, the energy demand affected by this MCO condition has been the early hypothesis for the study.

$$\text{Average Temperature } ^oC=\frac{\text{Min Temp } ^oC}{\text{Max Temp } ^oC} \quad (2)$$

2.2. Regression

The relationships among the energy information-related data are estimated using a statistical process called regression analysis. The performance of the regression analysis depends on the form of the generated process data and how it relates to the regression approach being used. It is widely used for prediction and forecasting with (3), represented by the Y-dependent variable, m-variable coefficient, and C, the trendline intercept. The equation will produce the future estimated value by inserting the importance of specific years. Table 1 shows the selection of dependent and independent variables.

$$Y=mx+C \quad (3)$$

The dependent variables are compared with each independent variable for the R^2 observation. The R^2 can be calculated using (4), where $\hat{y}i$ is a model predicted energy value using a particular point measured from the independent variable, $\bar{y}$ is the mean value of energy values, and yi is the actual observed energy value. This study adopted the single regression and multi-regression analysis from the Microsoft Excel data analysis tools. The significant regression equation and correlation value were powered during a single regression plot. Meanwhile, details verification of the value of the other was collected from multiple regression simulation. The range of the R^2 value is from 0 to 1. Good correlation is defined when the R^2 value is close to one and vice versa. Hence, results and discussion of the significant value of R2 are presented in section 3 accordingly.

$$R^2=\frac{\sum(\hat{y}i-\bar{y})^2}{\sum(yi-\bar{y})^2} \quad (4)$$

2.3. SOM algorithm implementation

In this study, input data collection data consisting of power consumption demand, CDD, average temperature, and covid-19 cases are organized in (m files) to execute the SOM data reader function. This function is offered for compatibility with the input datasets holding a space dimension. Apart from that, SOM normalization is used in the workspace to perform linear and logarithmic scaling. It is essential for measuring distances between vectors using the Euclidean metric. The variables will be scaled linearly so that the data variances are identical. This function is already in the toolbox to apply normalizations to the data input, and the normalization method is always a one-variable operation, as shown in Table 2. Normalization is a scaling technique that shifts the input between 0 and 1. In self-organizing maps, there are four types of normalization to be considered: ‘var’, ‘range’, ‘log’, and ‘logistic’. The variation for these normalization types is needed in

the training process of the data and to get the best mapping for the Malaysia Energy supply information. ‘var’: the variance will be normalized by the data input to unity and the means to zero, ‘range’: the variable values are scaled the input data between zero and one, ‘log’: this type of normalization is using logarithmic transformation, ‘logistic’: all the possible values are scaled between zero and one [17], [18]. Self-organizing map normalization is used in the workspace to perform linear and logarithmic scaling. It is essential for measuring distances between vectors using the Euclidean metric. The variables will be scaled linearly so that the data variances are identical. This function is already in the toolbox to apply normalizations to the data input, and the normalization method is always a one-variable operation.

Table 1. Dependent and independent variables selection

Dependent variable	Independent variable
Consumers power consumption demand	Daily cooling degree days value Daily average temperature Daily Covid-19 cases

Table 2. Normalization method

Method	Description
‘var’	Linear operation: variance is normalized to one
‘range’	Linear operation: values are normalized between 0 and 1
‘log’	Logarithmic is applied to the values: x= log(x-m+1) where m=min(x)
‘logistic’	Logistic conversion to scale all possible values between 0 and 1

The following steps explain the SOM process for the data analysis and mapping.

Step 1*:* From the initialization of SOM input data of dependence and independence variables represented by the number of neurons, the number of them placed in the input space is first decided. This step is necessary to allow the network to begin processing the data. The SOM algorithm will generate weights with the same vector dimension as the input data. The input vector will be picked randomly from training datasets at the training and learning phase before the winning neuron is determined until it gets the nearest one to the input vector.

Step 2*:* The weights for neurons and neighborhoods will be improved every iteration until their target is reached. The algorithm measures the distance between the data point, and the neuron weight vector before the best matching unit (BMU) is selected using (5).

$$d= \sum_{i=1}^{n} (qi-pi)^2 \quad (5)$$

Where d is the distance between data nodes, qi is the current input vector, pi is the node’s weight vector.

Step 3*:* The BMU is selected based on the minimum distance value or most similar to the input vector using (6). The weight must be adjusted every iteration to have a more significant change for neighbors closer to the BMU.

$$d= \min\left(\left|\left|\vec{x}-\vec{u}ij\right|\right|\right) \quad (6)$$

Where d is the lowest distance, $\vec{x}$ is the input vector and $\vec{u}ij$ is the updated weights. The modification of node’s weight of the BMU and neighbors is based on (7).

$$uij(t+1)=uij(t)+ai(t)[x(t)-uij(t)] \quad (7)$$

Where, the $uij(t+1)$ is new weigthe ht, $[x(t)-uij(t)]$ is the learning rate and $ai(t)$ Influence rate. The learning rate is decaying for each iteration as the training goes on, the neighborhood gradually shrinks. The distance between a node and the BMU is known assurance rate. It shows how much influence the latter has over the former. From a random arrangement of weights and across several iterations, SOM can arrive at a map of stable zones or reach the target.

Step 4*:* Finally, the U-matrix mapping analysis is accomexaminesssifier results graphically under the four normalization methods. The trained map is labeled using the SOM auto labeling function, in which the map will automatically be labeled based on the data or map. The best matching unit of each vector at the training

phase is found from the vectors in spin density (SD). The actual labels are selected based on the mode determined in this function: 'vote', the label with most instances is the only accepted label by this mode. Next, the U-matrix will be analyzed based on its color concentration. The light hexagonal occurs when the nodes are closed to each other and will be the opposite situation when the hexagonal color is dark [17]. The results of 'log', 'logistic', 'var', and 'range' normalization approach with 500 until 900 neurons with 50 step size that is collected along with their quantization error, topographic error, training time, and map size to get the best classifier.

3. RESULTS AND DISCUSSION

3.1. Data correlation (regression analysis)

In this section, the explanation of the study results and a comprehensive discussion will be presented simultaneously. The single regression analysis is performed, and the data taken is from before the MCO takes place. Figure 1 summarizes the correlation between energy demand and independent variable factors before MCO. Figure 1(a) shows the equation and significant result of the R^2 value between CDD and power demand. Figure 1(b) presents the tabulated data of the average daily CDD and power demand before regression analysis was done. Meanwhile, Figure 1(c) and 1(d) present the regression and correlation analysis between average temperature and energy demand before the MCO. The single value for R^2 is 0.032. It is found that every independent variable in this study; CDD, the average temperature has a low correlation with power demand in Peninsular Malaysia, which is means that the CDD and the average temperature are not the primary influence on the energy demand in Peninsular based on its regression at the first layer of investigation. The point is that the power demand is not affected by any changes that happened throughout the year, signifying that in the regression model, the R^2 must be ~0.5 to 0.75 and above to be accepted in the regression model.

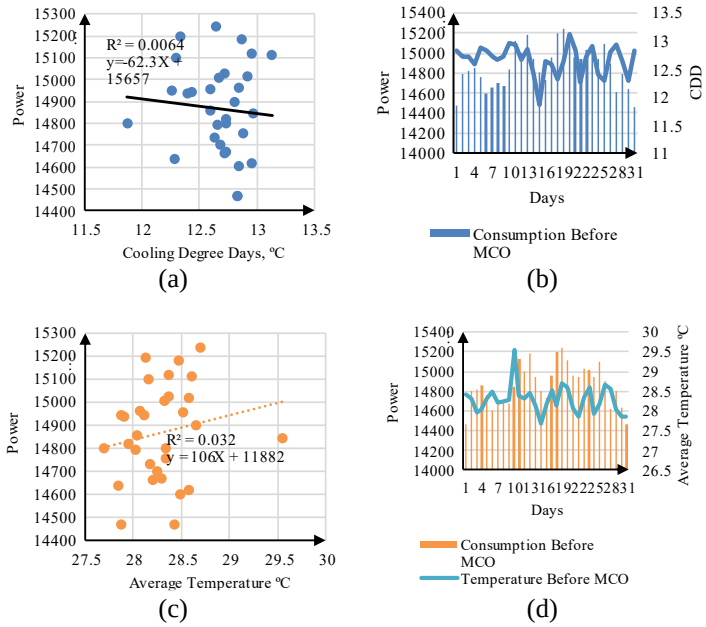


Figure 1. Variables data before pandemic Covid-19 MCO is presented (a) Scatter plot for the CDD and power demand correlation, (b) Power demand and CDD data comparison, (c) Scatter plot for the average temperature and power demand, and (d) Power demand consumption and average temperature comparison

Figure 2 summarizes the scatter plot and comparison for the significant variables towards power demand during pandemic Covid-19 MCO, respectively. The CDD correlation toward energy demand during MCO is presented in Figure 2(a). Meanwhile, Figure 2(b) illustrates the tabulated data of the daily CDD and power demand concurrently. During the MCO, the R^2 value improved due to decreased power demand. On the other hand, Figure 2(c) and 2(d) show the significant correlation finding for average temperature towards energy demand during MCO. The different topology of average temperature compared to before MCO has increased the value of R^2. During the MCO, the other additional variable was considered, and data was

collected for the number of Covid-19 cases. Thus, Figure 2(e) demonstrates the plotted regression analysis where the R^2 was 0.14. Otherwise, Figure 2 (f) represents the tabulated data between power demand and covid-19 daily cases in Malaysia. Overall R^2 value during the movement control order also has a low correlation but slightly improves, which are 0.05, 0.17, and 0.14 for CDD, average temperature, and Covid-19 daily cases, respectively.

For verifying all the variable's data correlation, the multiple regression analysis has been made accordingly. Table 3 shows the summary output for numerous regressions in this study. With the power demand as the dependent variable and the other set as independent variables, the data correlation indicated by R^2 shows that the datasets have an infirm correlation lower than 0.75. The average temperature is the only variable that increases as the energy demand increases, and the others tend to decrease based on their coefficient. From the results, the regression analysis is performed to observe whether there is a significant relationship between the dependent and independent variables. It also indicates the relative strength of different independent variables influencing the energy demand. The produced equation cannot be used to put independent variables that influence power demand in Malaysia. Thus, Table 4 presents since the values of SE, t-statistic, and P-value do not fulfill the significant correlation condition. The significance t-statistic should be higher than 2, and the P-value should be less than 0.05.

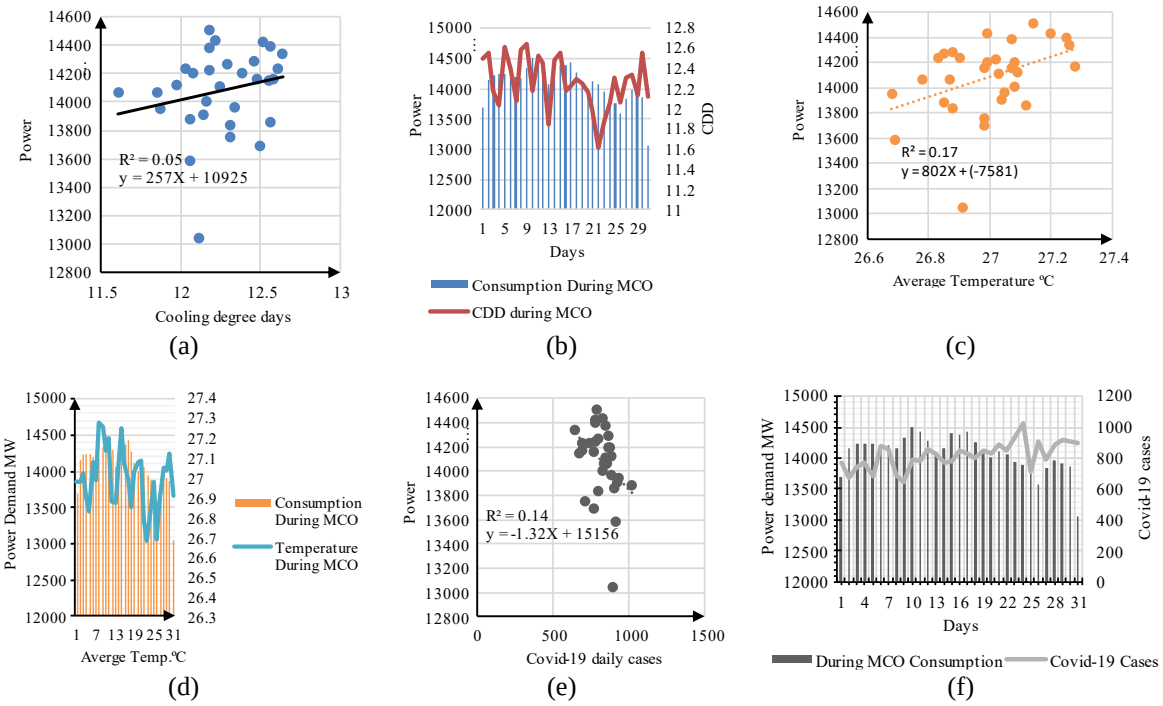


Figure 2. Variables data during pandemic Covid-19 MCO is presented (a) Scatter plot for the CDD and power demand correlation, (b) Power demand and CDD data comparison, (c) Scatter plot for the average temperature and power demand, (d) Power demand consumption and average temperature comparison, (e) Scatter plot for the Covid-19 cases and power demand, (f) Power demand consumption and Covid-19 cases comparison

Table 3. Multiple regression analysis summary output during MCO

Summary Output	
Regression Statistic	
Multiple R	0.493688917
R^2	0.243728747
Adjusted R^2	0.159698607
Standard Error	273.1740641
Observations	31

3.2. Verification using SOM output

In Table 5, the 'logistic' normalization approach produces four neurons that can get zero quantization and topographic error during MCO, which are 500, 550, 600, and 650 neurons. The 500 neurons

are selected for 'logistic' normalization since it has a lower training time than the others. It is enough to make an accurate mapping analysis since it has zero quantization and topographic errors.

Table 4. SE, t-Stat and P-Value for the simulation of multiple regression

	Coefficient	**Standard Error**	**t-Stat**	P-value
Intercept	-2892.672877	9889.039612	-0.292513024	0.772130541
Temperature	765.5529854	400.3397351	1.91225831	0.066507568
CDD	-1.149254182	0.710932336	-1.616545098	0.117602355
Covid-19	-224.5472443	268.3051038	-0.936910074	0.409989777

Table 5. 'Logistic' normalization simulation results

	Classification Results			
No of Neuron	Map Size	Quantization Error	Topographic Error	Training Time
500	[20:8]	0.000	0.000	4
550	[20:9]	0.000	0.000	5
600	[22:9]	0.000	0.000	7
650	[22:10]	0.000	0.000	8
700	[24,10]	0.000	0.032	10
750	[24:11]	0.000	0.032	12
800	[25:11]	0.000	0.065	15
850	[27:11]	0.000	0.065	18
900	[27:12]	0.000	0.032	22

From Figure 3(a), U-matrix mapping outcome with 500 neurons with (20x8) map size. The label is consisting of day 1 until day 31 or A until E1, respectively. The 'logistic' normalization before the MCO U-matrix shows that only one group is formed, consisting of B, E, H, and I (days 2, 5, 8, and 9). It also shows that the power demand is moving to the right, CDD to the top and the average temperature grouped at the top right. Figure 3(b), U-matrix mapping is arranged with 500 neurons with (20x8) map size. The 'logistic' normalization U-matrix shows that there are two groups are formed during MCO. The red-dotted box consists of days (6, 7, 12, 13, 19, 20, 21, 28, 29) labeled as (F, G, L, M, S, T, U, B1, C1), and the second group of data is in the green-dotted box consisting of the day (8, 9, 10, 11) labeled as (H, I, J, K) are placed in a single, separated group, which means that the average data from these days have a connection or dataset is the same but separated with the others. U-matrix shows that every variable in the 'logistic' approach has a different type of topology grouped in black boxes.

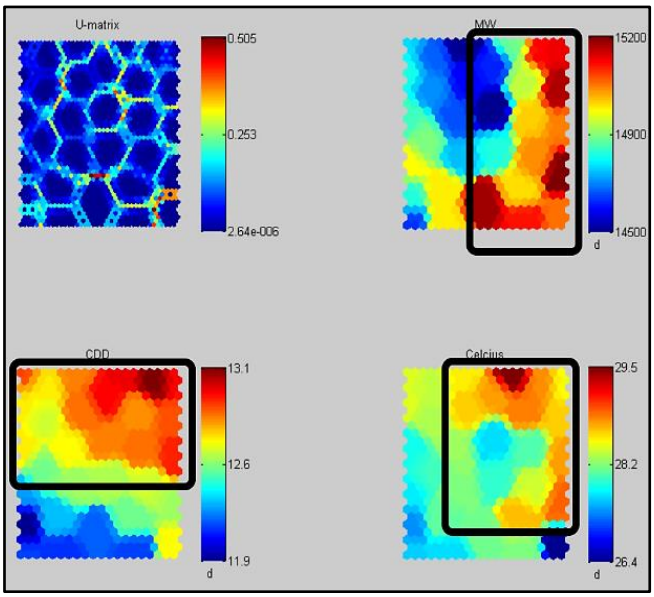
3.2.1. The best U-matrix discussion

The hexagonal topology was chosen to achieve a high mapping quality. The U-matrix result among all normalization methods with zero topographic error and quantization error are represented by 'logistic' and 'range' with 500 neurons and lower training time, which is 4 seconds was selected. 'log' approach is not able to isolate the datasets with minor differences where neurons I and E1 is the only separated datasets contributed by a considerable value in different, which is low energy demand and low covid cases according to its mapping component refer to Table 6. It is proven that the clustering method objective is achieved by observing the U-matrix separated with brighter colors among the datasets, which means that the similarities between data are more minor. The lower distances or high similarities are placed inside, the darker color.

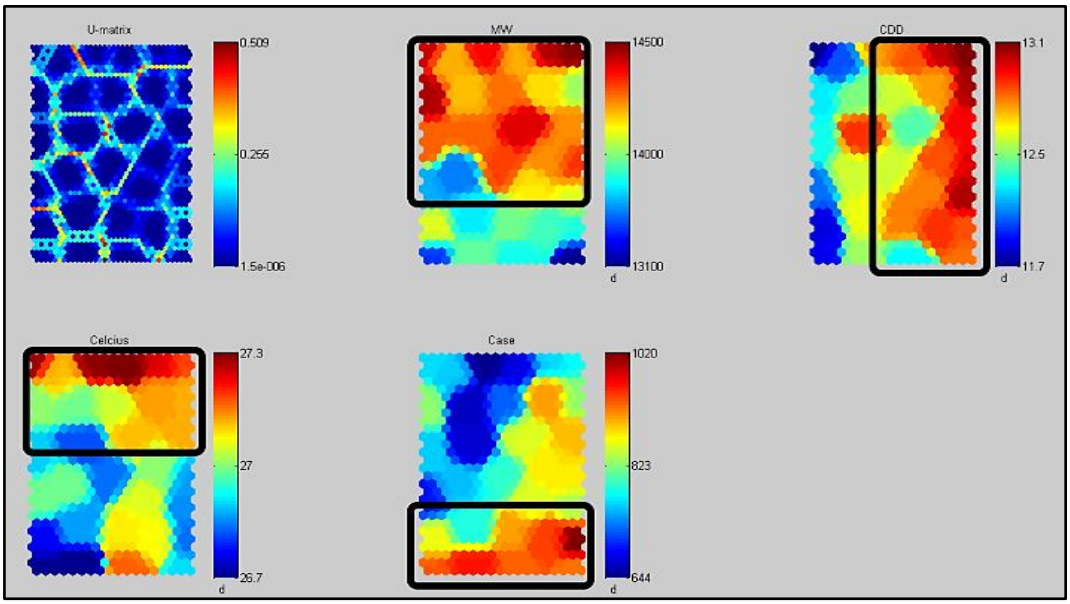
The 'Logistic' U-matrix is chosen, as shown in Figures 4(a) and 4(b), since it has a significant result showing that the connection between datasets is low. In addition, the SOM is easy to analyze since it converts high dimension input to low dimension map [18]. The classification summary is shown in Table 7, where the logistic normalization approach is the best classifier among the others. It is proved that most of the datasets do not correlate as stated in their R2, which is 0.12 and 0.24 for before MCO and during MCO, respectively. However, the U-matrix analysis shows that small groups contributed to the small value of R2 where two groups are formed in both periods. Before the MCO, group 1 (F,G,P,B1,C1) representing day (6,7,16,28,29) and group 2 (B,E,H,I) representing day (2,5,8,9). It is the same with during the MCO, consisting of group 1 (F,G,L,M,S,T,U,B1,C1) representing day (6,7,12,13,19,20,21,28,29) and group 2 (H,I,J,K). The datasets inside its group are correlated but separated from other datasets.

Based on the overall results, all normalization methods provide data crossing visualization and characteristics. It was proved that the normalization style does the training time. Still, lattice size does, and the mapping quality, the more extensive the lattice size, will have a good resolution. The selection is based on its topographic error and the quantization, the quantization error,y refers to the U-matrices not affecting, the dataset and has a low connection to verify that the pandemic Covid-19 situation in terms of CDD, average temperature, and the number of Covid cases is not significant. The U-matrix produced shows that most of the data do not

correlate. The hexagonal light color separates most datasets, and each variable's mapping component has different topology. However, as shown in Figure 4, small groups explicitly connected by observing the maps have indicated lower correlation to reflect regression analysis output at first layer analysis.



(a)



(b)

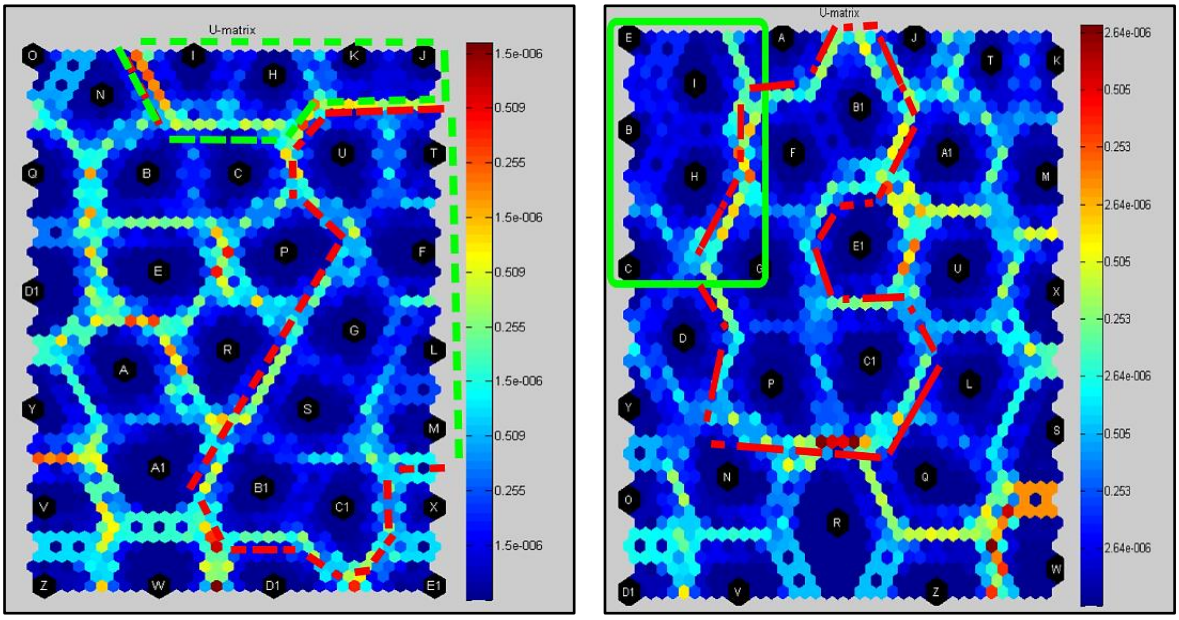
Figure 3. Comparing simulation results of the 'logistic' normalization method of (a) before MCO and (b) during MCO

Table 6. Summary analysis of the clustering for the U-matrix

Normalization Method	No of Neurons	Map size	Quantization Error	Topographic Error	Training Time (s)	Analysis
Logistic	500	(20x8)	0	0	4	Excellent classification with significant separator
Range	500	(20x8)	0	0	4	Well define data classification
Var	600	(22x9)	0	0	9	Good mapping quality but longer training time
Log	700	(25x11)	0	0	10	Poor data severance

Table 7. Summary analysis of the clustering for the U-matrix

Normalization	Period	Group 1	Group 2	Regression, R^2
Logistic	Before MCO	F, G, P, B1, C1 Day (6,7,16,28,29)	B, E, H, I Day (2,5,8,9)	0.12
	During MCO	F, G, L, M, S, T, U, B1, C1 Day (6, 7, 12, 13, 19, 20, 21, 28, 29)	H, I, J, K Day (8,9,10,11)	0.24



(a) (b)

Figure 4. Selected 'logistic' U-Matrix: (a) Before MCO, (b) During MCO

4. CONCLUSION

This study successfully investigates the significant correlation between power demand in Peninsular Malaysia and independent variables such as cooling degree days, average temperature, and Covid-19 cases. The demonstrated method has applied regression analysis and self-organizing maps concurrently. The methods used are qualified to perform time-consuming and complex research by the conventional verification method. The decision is made by observing the R^2 produced by regression and the U-matrix mapping analysis in self-organizing maps. The R^2 value is low before MCO and during MCO, respectively. It indicates no correlation or weak connection between the variables (CDD, average temperature, Covid-19), which does not influence the power demand in Peninsular Malaysia. The correlation status has been rechecked using SOM. The algorithm performs four types of normalization to validate the regression analysis output: 'log', 'logistic', 'range', and 'var'. The U-matrix produced shows that most of the data do not correlate. The hexagonal light color separates most datasets, and each variable's mapping component has different topology. For future research, more datasets and their variation must be considered to achieve the optimum capability of the self-organizing maps. The methodology used in this study can make forecasting decisions if the datasets have a strong correlation.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Teknikal Malaysia Melaka for all the support. The study is funding by the Ministry of Higher Education (MOHE) of Malaysia through the Fundamental Research Grant Scheme (FRGS), No: FRGS/1/2021/FKE/F00465.

REFERENCES

[1] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, "Big data analytics for dynamic energy management in smart grids," *Big Data Research*, vol. 2, no. 3, pp. 94–101, Sep. 2015, doi: 10.1016/j.bdr.2015.03.003.

[2] F. Adilla, "TNB to deploy big data analytics in Sepang LSS to boost efficiency," *New Straits Times*, May 06, 2020. https://www.nst.com.my/business/2020/05/590316/tnb-deploy-big-data-analytics-sepang-lss-boost-efficiency (accessed Jun. 04, 2021).

[3] M. F. Sulaima, N. Y. Dahlan, Z. M. Yasin, M. M. Rosli, Z. Omar, and M. Y. Hassan, "A review of electricity pricing in

peninsular Malaysia: Empirical investigation about the appropriateness of enhanced time of use (ETOU)electricity tariff," *Renewable and Sustainable Energy Reviews*, vol. 110, pp. 348–367, 2019, doi: 10.1016/j.rser.2019.04.075.
[4] D. Fawzy, S. Moussa, and N. Badr, "The evolution of data mining techniques to big data analytics: an extensive study with application to renewable energy data analytics," *Asian Journal of Applied Sciences*, vol. 4, no. 3, pp. 756–766, Jun. 2016.
[5] K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: from big data to big insights," *Renewable and Sustainable Energy Reviews*, vol. 56, no. 2016, pp. 215–225, Apr. 2016, doi: 10.1016/j.rser.2015.11.050.
[6] D. Miljkovic, "Brief review of self-organizing maps," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2017, pp. 1061–1066, doi: 10.23919/mipro.2017.7973581.
[7] S. Bressi, A. Dumont, A. Carter, and N. Bueche, "A multiple regression model for developing a RAP binder blending chart for stiffness prediction," in *International Conference Bituminous Mixtures and Pavements*, 2015.
[8] Y. Yang, "Prediction and analysis of aero-material consumption based on multivariate linear regression model," in *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Apr. 2018, pp. 628–632, doi: 10.1109/icccbda.2018.8386591.
[9] L. Stetz Consulting, "Regression for M & V : reference guide," May, 2012.
[10] N. Regmi and S. B. Pandey, "A regression analysis into Nepali ICT's energy consumption and its implications," in *2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Dec. 2015, pp. 1–8, doi: 10.1109/skima.2015.7400034.
[11] K. Cebrat and Ł. Nowak, "Revealing the relationships between the energy parameters of single-family buildings with the use of self-organizing maps," *Energy and Buildings*, vol. 178, pp. 61–70, Nov. 2018, doi: 10.1016/j.enbuild.2018.08.028.
[12] D. Sarchiz and I. S. Vasarhely, "Reactive power compensation assessment using self-organizing feature map," *IFAC Proceedings Volumes*, vol. 40, no. 8, pp. 147–152, 2007, doi: 10.3182/20070709-3-ro-4910.00023.
[13] K. Chakraborty, A. De, and A. Chakrabarti, "Voltage stability assessment in power network using self organizing feature map and radial basis function," *Computers & Electrical Engineering*, vol. 38, no. 4, pp. 819–826, Jul. 2012, doi: 10.1016/j.compeleceng.2012.03.012.
[14] Z. Kang *et al.*, "Structured graph learning for clustering and semi-supervised classification," *Pattern Recognition*, vol. 110, p. 107627, Feb. 2021, doi: 10.1016/j.patcog.2020.107627.
[15] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52–65, Jan. 2013, doi: 10.1016/j.neunet.2012.09.018.
[16] M. Domínguez, J. J. Fuertes, I. Díaz, A. A. Cuadrado, S. Alonso, and A. Morán, "Analysis of electric power consumption using self-organizing maps.," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 12213–12218, Jan. 2011, doi: 10.3182/20110828-6-it-1002.02092.
[17] M. F. Sulaima, M. H. Jali, Z. H. Bohari, F. Baharom, and N. Baharin, "Evaluation of power distribution network system by using self-organizing map (SOM) as the best practice for buses characteristic classification," *Journal of Theoretical and Applied Information Technology*, vol. 86, no. 2, pp. 299–305, Apr. 2016.
[18] N. N. Atira, I. Azmira, Z. H. Bohari, N. A. Zuhari, and N. F. M. Ghazali, "Medium term load forecasting using statistical feature self organizing maps (SOM)," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 11, no. 2, pp. 25–29, 2019.
[19] Z. H. Bohari, M. Isa, A. Z. Abdullah, A. A. Rahman, and M. F. Sulaima, "Intelligent PD classification via SOM with optimized correlation," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12S2, pp. 53–60, Dec. 2019, doi: 10.35940/ijitee.l1011.10812s219.
[20] W. Li, S. Zhang, and G. He, "Semisupervised distance-preserving self-organizing map for machine-defect detection and classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 5, pp. 869–879, May 2013, doi: 10.1109/tim.2013.2245180.
[21] H. Yin, "The self-organizing maps: background, theories, extensions and applications," in *Computational intelligence: a compendium*, Springer, 2008, pp. 715–762.
[22] Y. Dong, X. Wang, J. Jin, Y. Qiao, and L. Shi, "Effects of eco-innovation typology on its performance: Empirical evidence from Chinese enterprises," *Journal of Engineering and Technology Management*, vol. 34, pp. 78–98, 2014, doi: 10.1016/j.jengtecman.2013.11.001.
[23] S. Clark, S. A. Sisson, and A. Sharma, "Tools for enhancing the application of self-organizing maps in water resources research and engineering," *Advances in Water Resources*, vol. 143, p. 103676, Sep. 2020, doi: 10.1016/j.advwatres.2020.103676.
[24] V. Marinakis, "Big data for energy management and energy-efficient buildings," *Energies*, vol. 13, no. 7, p. 1555, Mar. 2020, doi: 10.3390/en13071555.
[25] Y. Shi, D.-F. Zhang, Y. Xu, and B.-T. Zhou, "Changes of heating and cooling degree days over China in response to global warming of 1.5 °C, 2 °C, 3 °C and 4 °C," *Advances in Climate Change Research*, vol. 9, no. 3, pp. 192–200, Sep. 2018, doi: 10.1016/j.accre.2018.06.003.
[26] M. Isaac and D. P. van Vuuren, "Modeling global residential sector energy demand for heating and air conditioning in the context of climate change," *Energy Policy*, vol. 37, no. 2, pp. 507–521, Feb. 2009, doi: 10.1016/j.enpol.2008.09.051.

BIOGRAPHIES OF AUTHORS

Mohamad Fani Sulaima is serving as Senior Lecturer in the Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka (UTeM). Upon joining UTeM, he served as a Coordinator and Head for the Energy Management Division in the Centre for Sustainability and Environment before being appointed as the first internal University Energy Manager in 2015. He received his bachelor's degree from Tokai University, Japan, in 2010 and a Master's degree from the University of Malaya. He received Ph.D. in Electrical Engineering with a specialization in Energy Demand Side Management from Universiti Teknologi Mara (UiTM), Malaysia, in 2020. His research interests include power system, demand-side management, demand response, energy efficiency, measurement & verification, and artificial intelligence. As a result of his research interest, he has published more than 90 articles, journals, and academic papers. He can be contacted at email: fani@utem.edu.my.

Sharizad Saharani is serving as Biomedical Engineering Technician at Sedafiat Sdn. Bhd. located in Semporna, Sabah. He received his Bachelor's degree in Electrical Engineering with Honour from Universiti Teknikal Malaysia, Melaka in 2022 and a Diploma in Electric and Electronic from Politeknik Ibrahim Sultan, Pasir Gudang, Johor in 2017. He is interested in data analytics and power system. He can be contacted at email: sharizad96@icloud.com.

Arfah Ahmad is a lecturer at Faculty Electrical Engineering, Universiti Teknikal Malaysia Melaka, Malaysia. She had served the university since 2010; previously she was with the Universiti Teknologi Mara (UiTM), Johor for a year. She received her Bachelor's degree and Master's degree in Statistics from Universiti Kebangsaan Malaysia (UKM) in 2008. She is the Graduate Research Assistant with Solar Energy Research Institute (SERI) at UKM while pursuing her master degree. In 2020, she received her Ph.D in data analysis from the School of Electrical, Electronic and Computer Engineering at The University of Western Australia (UWA). Her research interests include data compression, application of mathematics and statistics in power systems engineering, probability, statistical modeling, computing, and analysis. She can be contacted at email: arfah@utem.edu.my.

Elia Erwani Hassan is a Senior Lecturer at the Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka. Begin in year 1995, using a Bachelor of Electrical Engineering qualification she started teaching experience as a lecturer in Universiti Teknologi Mara (UiTM), Shah Alam. She then completed her study in Universiti Teknologi Malaysia (UTM) in Master of Engineering Electrical -Mechatronics and Automatic Control. She did a Ph.D in Environmentally Constraint Economic Dispatch and Reactive Power Planning for Ensuring Secure Operation in Power System. Her research area is interested in Power System and Optimization. Ir. Dr. Elia Erwani Hassan also as a member of Board of Engineer Malaysia (BEM). She can be contacted at email: erwani@utem.edu.my.

Zul Hasrizal Bohari is a senior lecturer at the Faculty of Electrical and Electronic Engineering Technology, Universiti Teknikal Malaysia Melaka. He was graduated with Doctor of Philosophy in Electrical Engineering from Universiti Malaysia Perlis (UniMAP) in 2022 and Master of Electrical Engineering from UNITEN back in 2013 and Bachelor Degree in Electrical and Electronic Engineering from UKM back in 2007. He has published more than 40 journals in various publisher and started to do research works related to high voltage engineering, electrical engineering and artificial intelligence (AI) since 2011. Among the achievement was being awarded with few international innovation awards on new innovation related to renewable energy, HV engineering and artificial intelligence (AI). He can be contacted at email: zulhasrizal@utem.edu.my.