Multivariate Statistical Process Control for Time-varying Systems: A Case Study Solar Photovoltaic System

Bundit Boonkhao[1,*],Tararat Mothayakul[2], Chanida Yubolsai[2] and Pornpimol Kavansu[2]

[1]Division of Industrial Engineering and Management, Faculty of Engineering,
Nakhon Phanom University, Nakhon Phanom, Thailand
[2]Research & Development Institute, Nakhon Phanom University, Nakhon Phanom, Thailand

Abstract: Time-varying multivariate statistical process control (TMSPC) has been suggested as a method for monitoring processes, detecting faults, and diagnosing issues in systems where variables change over time. This is an adaptation of multivariate statistical process control (MSPC), which is typically used for processes where variables are stable and not influenced by time. However, in certain processes, such as those in solar photovoltaic systems, variables like temperature, voltage, and current fluctuate over time. Thus, TMSPC has been proposed as a monitoring and diagnostic tool for these time-dependent processes. The effectiveness of this technique has been demonstrated using a solar photovoltaic system at Research & Development Institute, Nakhon Phanom University, Thailand (RDI-NPU).

***Index Terms*— Solar photovoltaic system, multivariate statistical process control, principal component analysis, monitoring system, time-variant**

I. Introduction[1]

SOLAR energy has seen widespread adoption in numerous electrical applications, including homes, farms, and industries [1]. This trend is driven by significant reductions in the costs associated with manufacturing, installation, and operation. Despite these cost reductions, solar energy implementation remains confined to regions with high sunlight intensity and is also restricted during periods with low light, such as rainy or winter seasons. Consequently, the successful use of solar energy is heavily influenced by environmental conditions.

Light is a crucial factor in solar energy production. Typically, the photons from light hitting the solar panels vary throughout the day. On bright days, energy production is high, whereas on foggy or cloudy days, it decreases. This variability shows that solar energy production is inconsistent and depends on the surrounding environment. Fluctuations in light lead to fluctuations in energy output, resulting in suboptimal battery charging performance, which poses challenges for energy management. Effective energy management requires continuous monitoring and analysis for future energy planning and operations. Additionally, other environmental factors around the solar cell installation site should also be measured and monitored.

To efficiently monitor and diagnose processes with numerous variables, multivariate statistical process control (MSPC) should be used [2] [3] [4]. This technique employs feature extraction methods to select the variables with the most variance, which are then utilised to create monitoring charts and contribution plots for fault diagnosis. MSPC is typically limited to multivariable processes that operate under normal operational conditions (NOC), where the process set point remains constant over time. However, solar photovoltaic systems are dynamic processes. Thus, implementing MSPC for these systems requires adaptation for time-varying conditions. In this article, time-varying multivariate statistical process control (TMSPC) is proposed and demonstrated using a solar photovoltaic system.

II. Time-varying Multivariate Statistical Process Control

The procedure of TMSPC is composed of data preparation, training process, monitoring charts and fault detection & diagnosis. This section will briefly describe the proposed TMSPC.

[1] The manuscript received May 28, 2024; revised June 23, 2024; accepted June 27, 2024; Date of publication June 30, 2024

[*] Corresponding author: Bundit Boonkhao, division of industrial engineering and management, faculty of engineering, Nakhon Phanom University, Nakhon Phanom, 48000 Thailand, (E-mail: bundit@npu.ac.th)

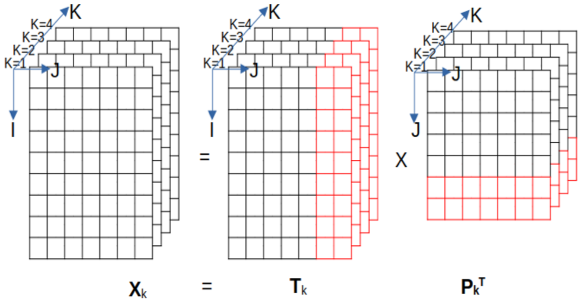

Fig. 1. Block of dataset for time-varying principal component analysis (TPCA) processing.

A. Data preparation

For the collected variables of solar cell data, the dataset is designed as three-dimensional (3D) data **X**[I×J×K] which I is the date of collection, J is the variable and K is series of time. This can be represented as a block shown in Fig 1. The dataset must be standardised prior to process dimension reduction. Then this dataset will be extracted feature by using principal component analysis (PCA) technique [5] which will be described in the next section.

B. Time-varying principal component analysis

PCA can only handle two-dimensional (2D) datasets, so it is necessary to convert a 3D dataset into a 2D one. Traditional method to transform from multi-dimension to 2D is a multi-way method [6] [7] [8] [9]. The method tries to concatenate the two variables into on variable so the dimension is reduced from 3D to 2D. However, this article suggests reducing dimensions by slicing the dataset at each time interval. Thus, the dataset **X**[I×J×K] is transformed into $\mathbf{X}_K[I\times J]$, where $\mathbf{X}_K$ represents the data with [I×J] dimensions at time K, a process known as time-varying principal component analysis (TPCA). The following step involves applying the PCA procedure.

PCA is a method used to extract features from a multivariable dataset. Given a dataset **X**[I×J], it can be decomposed as a linear combination:

$$\mathbf{X} = \breve{\mathbf{T}}\breve{\mathbf{P}}^T \tag{1}$$

where $\breve{\mathbf{T}}$ is score matrix [I × J], $\breve{\mathbf{P}}$ is loading matrix [J × J] and superscript T is transpose of matrix. By reduction of dimension, number of variables J may be selected only R variables, $R \leq J$. So, the dataset **X**[I × J] can be rewritten as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2}$$

where **T** is score matrix [I × R], **P** is loading matrix [J × R] and **E** is error matrix [I × J]. This is a training process to obtain score and loading matrix with reduction size of variable to R.

The above process is for 2D dataset so for this proposed each time series will be trained and obtained both score and loading matrix at each time K.

$$\mathbf{X}_k = \mathbf{T}_k\mathbf{P}_k^T + \mathbf{E}_k \tag{3}$$

The next process will be using those parameters to construct the monitoring chart and fault analysis.

C. Hotelling's T^2 monitoring chart

Hotelling's T^2 will be constructed as a monitoring chart. It is constructed from score of each variable which can be calculated from

$$\boldsymbol{t}_k = \boldsymbol{x}_k\mathbf{P}_k \tag{4}$$

where $\boldsymbol{x}_k$ vector data [1 × J] at time k and $\boldsymbol{t}_k$ vector score [1 × R] at time k. Monitoring chart can be constructed from the new data, in here, Hotelling's T^2 (T_k^2) has been constructed.

$$T_k^2 = \sum_{r=1}^{R}\frac{t_{k,r}^2}{s_{t_{k,r}}^2} \tag{5}$$

where $t_{k,r}^2$ – square score at index r and $s_{t_{k,r}}^2$ – variant of score at index r at time k.

Likewise other control system, monitoring chart will alarm when T_k^2 is over control limit. Only upper control limit (UCL) is applied for Hotelling's T^2 monitoring chart. In case of trained data (Phase I), the UCL is caculated from

$$T_{UCL}^2 = \frac{(I-1)(I-1)R}{I(I-R)}F_\upsilon(R, I-R) \tag{6}$$

where T_{UCL}^2 is upper control limit of T_k^2, $F_\upsilon(\cdot,\cdot)$ is F – distribution.

The control limit is also used as a criterion for evaluating the trained dataset. If T_k^2 of all training data is less than T_{UCL}^2, then all data used in the training process is acceptable. However, if any data has $T_k^2 > T_{UCL}^2$, it indicates that the data is not acceptable and should be removed from the model. The training process should then be repeated without this data. When the training data passed this criterion, it will be used as the standard model for further new data.

When obtaining new data at time k, the score of the new data can be obtained by multiplying the loading matrix, which obtained from TPCA, following Eq. (4). Then Hotelling's T^2 monitoring chart and control limit for new data will be constructed. However, for new data (Phase II), the UCL will be calculated from

$$T_{UCL}^2 = \frac{(I-1)(I+1)R}{I(I-R)}F_\upsilon(R, I-R) \tag{7}$$

Next section, when the fault occurs, fault detecting and diagnosing will be described.

D. Fault detection & diagnosis

If $T_k^2 > T_{UCL}^2$, it signifies that a fault has been detected or the system is out of control. To identify which variables caused the fault, a contribution plot can be used [10] [11].

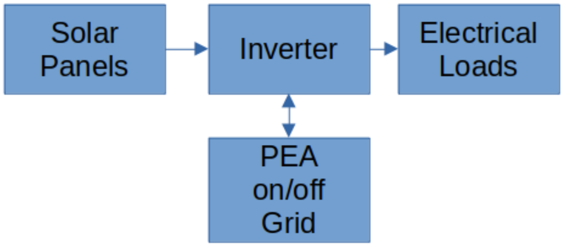

Fig. 2. Diagram of solar photovoltaic system install at RDI-NPU.

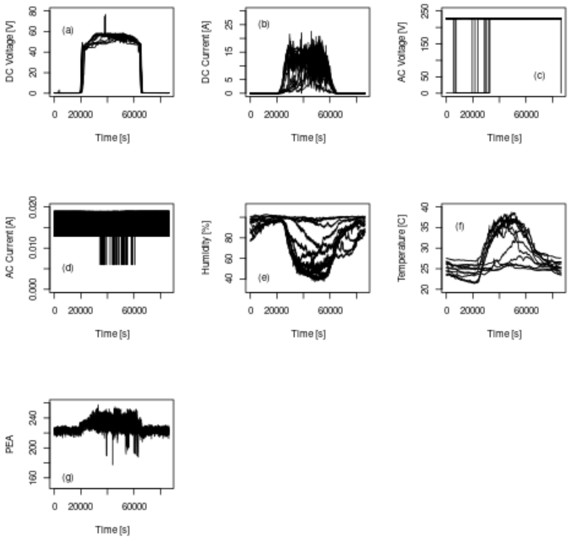

Fig. 3 Dataset for training which composed of 12 days and 7 variables: (a) DC voltage, (b) DC current, (c) AC voltage, (d) AC current, (e) humidity, (f) temperature and (g) voltage from PEA.

Referring back to Eq. (4), the contribution plot shows the projection of individual variables onto the loading matrix to generate their scores. The variable with the highest score is identified as the root cause of the process fault because it is the highest variance on that time. For better visual understanding, the contribution plot can be displayed as a bar chart, showing the scores of the individual variables.

This section is the proposed method of TMSPC which composes of data collection, transformation and feature extraction, then monitoring chart, control limit and fault dection and diagnosis are established. The next section will be a demonstration of applying TMSPC to solar photovoltatic system.

III. CASE STUDY

A. *Solar Photovoltaic System*

To illustrate the use of TMSPC, a solar photovoltaic system at Research & Development Institute, Nakhon Phanom University, Thailand (RDI-NPU) is used as a case study. Fig. 2 depicts the diagram of the solar photovoltaic system installed at RDI-NPU. The system consists of 30 crystalline solar panels (SPOT model S-170-24), with 6 panels connected in series and 5 of these series-connected sets arranged in parallel. The panels are linked to an inverter (LEONICS model G-304), which converts the electricity for distribution to the RDI-NPU building. Additionally, the building draws electricity from the Provincial Electricity Authority (PEA). During operation, seven variables are recorded every second: DC voltage [V], DC current [A], AC voltage [V], AC current [A], humidity [%], temperature [°C], and electricity from PEA [V]. The data is logged in a file on the data logger (DX2000-Wisco Industrial Instruments) and can be analysed using universal viewer software (SMARTDAC+ STANDARD).

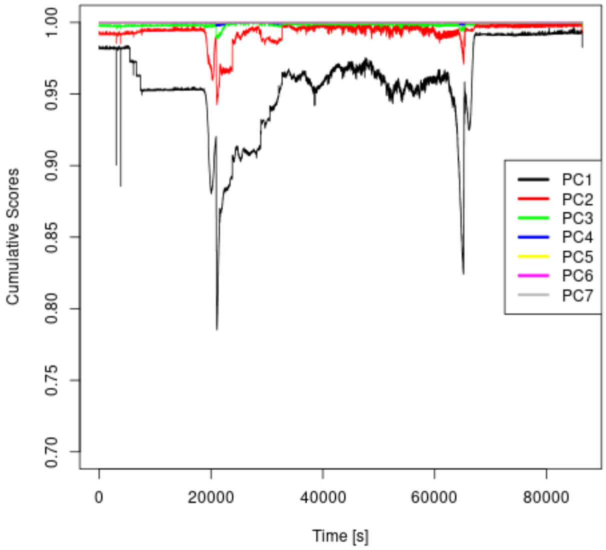

Fig. 4 Score of individual principal component along the time processed by TPCA.

B. *TMSPC for solar system*

In this demonstration, the collection of data for 17 days were use as dataset therefore the dimension of data should be $\mathbf{X}[17 \times 7 \times 86400]$, which means 17 days, 7 variables and 86,400 sec. However, the data were divided for training and verifying as 70:30 or 12 days for training dataset and the rest for verifying. Hence, the training dataset were $\mathbf{X}[12 \times 7 \times 86400]$. Fig. 3 shows the plots of individual variable of training dataset (12 days) in 86,400 sec. Note that the 12 days were selected from the best normal operational condition. For TMSPC, the dataset was transformed to $\mathbf{X}_{86400}[12 \times 7]$. In the pre-processing, data $\mathbf{X}_{86400}[12 \times 7]$ was standardised by subtracting with mean and dividing with standard deviation of individual variable. Then, the TPCA was further processed to determine score and loading matrix of the training dataset at each time.

The score and loading matrices for each time point in seconds can be calculated using TPCA as shown in Eq. (3). Fig. 4 displays the cumulative scores of the principal component (PC) over time. It is evident that the first principal component (PC1) captures at least 77.27% of the information from the trained data. Consequently, using just one PC is

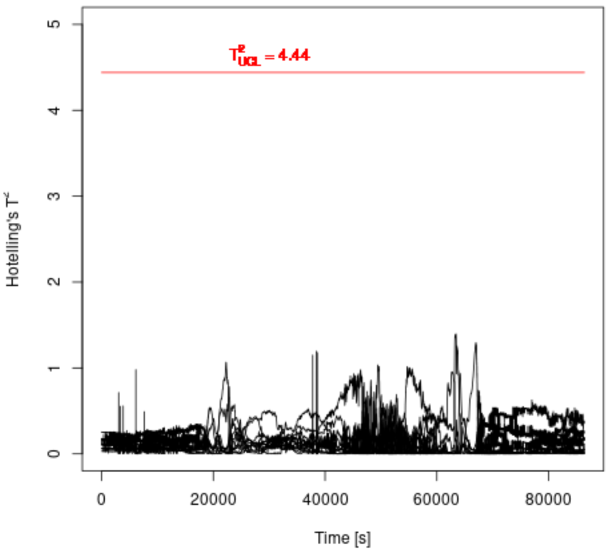

Fig. 5 Transformation of the training dataset from 7 variables to only one principal component (PC1) and all variables are lower than T^2_{UCL}.

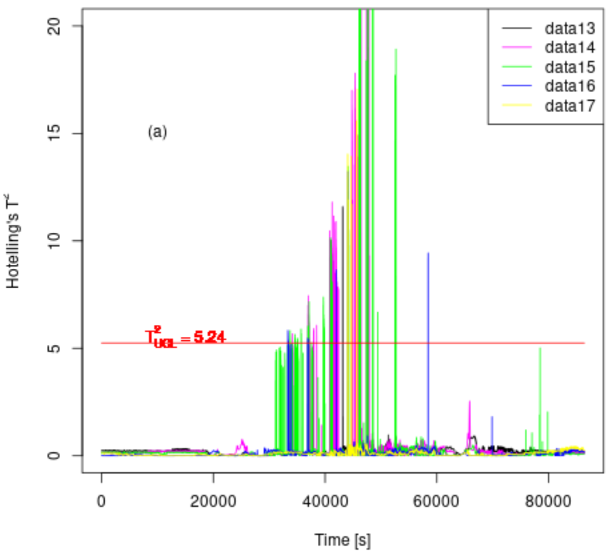

Fig. 6 Hotelling's T^2_k of test 5 new data. There are many faults occurred.

sufficient for creating monitoring charts, meaning that R can be selected of 1. This implies that instead of monitoring seven variables, monitoring only the first PC can account for at least 77.27% of the information.

Fig. 5 displays T^2_k monitoring charts for the trained data. The original seven variables were reduced to a single principal component (PC1) for constructing the chart. Using an upper control limit T^2_{UCL} for Phase I (Eq. (6)) and set at a 95% confidence level, the variables remain below this threshold. This indicates that the trained data is of sufficient quality for future predictions.

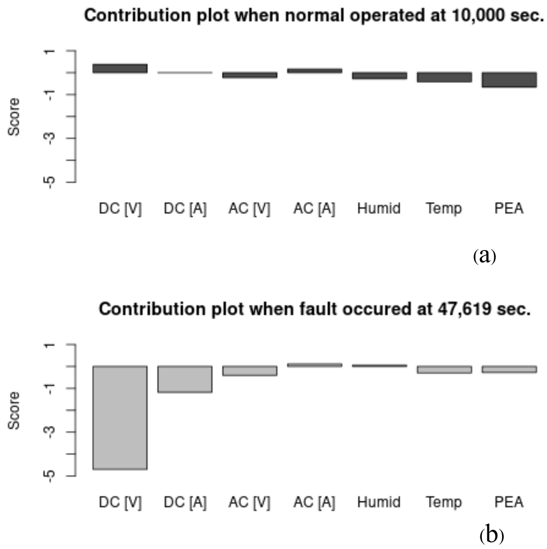

Fig. 7 Contribution plot of all variables; (a) when Hotelling's T^2_k less than T^2_{UCL}, (b) when Hotelling's T^2_k greater than T^2_{UCL}.

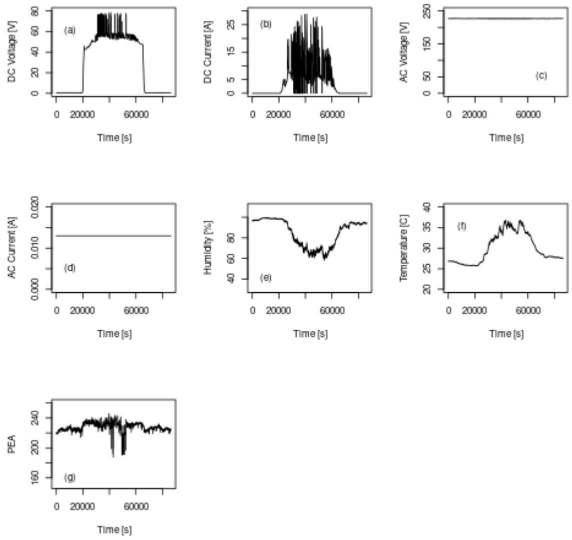

Fig. 8 Plot of all variable of data15 at time 47,619 sec.

C. *Fault detection & diagnosis*

The remaining data was used for evaluating the performance of TMSPC. Firstly, fault detection was evaluated. When the new data obtained, seven variables at individual time k were standardised and the scores of the new data were calculated using Eq. (4) and constructing T^2_k from Eq. (5). Fig. 6 represents the T^2_k of all 5 new data. As can be seen in the figure that, there were some events that T^2_k higher than T^2_{UCL}. This means the faults were detected.

When the fault was detected, fault diagnosing is consequent process for evaluation. By considering individual dataset, as can be seen in Fig. 6 for only **data15**, there were faults detected along the daytime. To diagnose the root cause of the fault, i.e., which variable caused the fault, this can be determined by using contribution plot. By picking the time that $T_k^2 > T_{UCL}^2$, for example at time 47,619 sec, $T_{47,619}^2$ = 46.65 > T_{UCL}^2= 5.24, then calculating contribution plot following from Eq. (4).

Fig. 7 illustrates the contribution plot for the selected time intervals. In Fig. 7(a), the contribution plot during the NOC shows that the score for each variable is very small, and the sum of the squares of the scores is less than T_k^2.

Contribution plot while out of NOC, score of the root cause will be large. At time k = 47,619 sec, the fault was detected and, if considering Fig. 7(b), the variable DC voltage contributed the highest score. This means DC voltage variable was the root cause of the fault at time 47,619 sec. This can be done similarly for all other faults.

Returning to the original variable **data15** at time 47,619 sec or $\mathbf{X}_{47619}$, Fig. 8 shows graphical presentation of all seven variables versus time. As can be seen in Fig. 7(a), DC voltage at time 47,619 sec was fluctuated therefore this should be the reason of fault occurred. This is agreed with the diagnosing using contribution plot (Fig. 8). Although, DC current variable also fluctuated, the cause was less effect than DC voltage and it was the second most effected. The knowledge for this technique may be used as further maintenance or energy management.

IV. Conclusion

TMSPC has been introduced as a method for monitoring and diagnosing faults in solar photovoltaic systems, which are inherently time-varying. This tool uses TPCA to reduce the dimensionality of seven photovoltaic variables down to a single principal component. Subsequently, Hotelling's T_k^2 statistic and a contribution plot are employed as a monitoring chart and diagnostic tool, respectively. Both tools can detect and diagnose faults occurring in the system at specific times. This technique can also be applied to other time-varying processes, such as chemical batch processing.

Acknowledgment

This research was financially supported by National Research Council Thailand (NRCT) under the Community Renewable Energy Center (CREC) project grant number 2562A13402012**.**

References

[1] T. Mothayakul, P. Sripadungtham and U. Boonbumroong, "Solar home with energy management by frugal discharge," *Journal of Thai Interdisciplinary Research,* vol. 13, no. 3, pp. 64 - 68, 2018.

[2] Z. Ge and Z. Song, Multivariate Statistical Process Control: Process Monitoring Methods and Applications, London: Springer, 2013.

[3] R. Mason and J. Young, Multivariate Statistical Process Control with Industrial Applications, Pennsylvania: Society of Industrial and Applied Mathematics, 2002.

[4] E. Santos-Fernandez, Multivariate Statistical Quality Control Using R, New York: Springer, 2012.

[5] I. Jolliffe, Principal Component Analysis, New York: Springer, 2002.

[6] P. Kroonenberg, Applied Multiway Data Analysis, Hoboken: Wiley, 2008.

[7] R. Henrion, "N-way principal component analysis theory, algorithm and applications," *Chemometrics and Intelligent Laboratory Systems,* vol. 25, pp. 1 - 23, 1994.

[8] S. Wold, P. Geladi, K. Esbensen and J. Ohman, "Multi-way principal component-and PLS-analysis," *Journal of Chemometrics,* vol. 1, pp. 41 - 56, 1987.

[9] P. Nomikos and J. MacGregor, "Monitoring batch processes using multiway principal component analysis," *AIChE Journal,* vol. 40, pp. 1361 - 1375, 1994.

[10] P. Miller, R. Swanson and C. Heckler, "Contribution plots: a missing link in multivariate quality control," *Applied Mathematics and Computer Science,* vol. 8, no. 4, pp. 775 - 792, 1998.

[11] J. Westerhuis, S. Gurden and A. Smilde, "Generalized contribution plots in multivariate statistical process monitoring," *Chemometrics and Intelligent Laboratory Systems,* vol. 51, pp. 95 - 114, 2000.