Application of the mixed effects model for analysing photovoltaic datasets and interpreting into meaningful insights

Ngoc Thien Le  Widhyakorn Asdornwised  Surachai Chaitusaney  Watit Benjapolakul

Abstract
Energy power studies commonly apply the analysis of variance (ANOVA) method in data analysis context to find the significant difference among many compared groups or to evaluate the impact of an influential factor. However, most of the datasets in these studies are time series data or longitudinal data, which are collected from the same object over some periods. Therefore, violating the independent assumption of ANOVA is the error commonly made. This leads to the misinterpretation of the comparison tests. In this article, this problem is solved by providing the application of mixed effects model (MEM) as a viable alternative to the ANOVA. In detail, two models based on MEM are proposed to analyse the time series data of micro-inverter PV stations located in Concord city, Massachusetts, USA. In the first scenario, the average models are implemented to compare the seasonal variation of monthly yield taking into account the effects of shading conditions and different orientations. In the second scenario, the linear regression model based on MEM is implemented to estimate and compare the decline rate in a 6-year period of PV stations. The analysis results have shown that the models based on MEM performs better than ANOVA in interpreting raw time-series dataset into meaningful insights.

1 INTRODUCTION
A significant portion of data analysis in power energy relates to the longitudinal study, which uses the time series dataset, also called as longitudinal data or panel data [1], for comparing the significant difference among many groups of observed data or for evaluating the impact of an influential factor. These studies commonly apply the analysis of variance (ANOVA) method, developed by Ronald Fisher in 1918, to investigate whether there is a significant difference in the mean, or median values between two or more groups of observed objects [2]. For example, Dusan Kudelas et al. [3], examined the variability of the recorded efficiency of the solar panels and the PV power performance for years of 2012, 2013, and 2015. Authors in [4] and [5] applied ANOVA test for electric power theft detection. In detail, they implemented the state estimation model from the customer metering data to generate normal power consumption. Then, ANOVA is performed to judge whether an individual meter data deviates from normal power consumption. Ni Ding et al. [6], built the Fourier time series model to predict the one-day-ahead and two-day-ahead power consumption. They used ANOVA to choose the most significant frequency components. In cluster analysis study, Jenny Palm et al. [7] proposed the ANOVA test to classify individual activity sequence, i.e. cooking, doing the laundry, or watching television, from the time series data of energy consumption at residential level. However, most of the above studies have ignored the required assumptions in order to apply ANOVA test. The most crucial assumption to apply ANOVA is that the observed data should be independent [8]. Unfortunately, such time series data as in the aforementioned studies clearly violated this assumption. The reason is that time series data come from a single object (for example, PV station, smart meter), hence data taken between two arbitrary intervals are obviously correlated. Therefore, ANOVA test is an inappropriate statistical technique in longitudinal study. In addition, ANOVA is a nonparametric method, which only gives the answer whether or not there is a significant difference between many groups. In the comparison of more than two groups,
ANOVA test has failed to address which group is significantly different from the remains. Moreover, ANOVA cannot give in-depth details about the amount of variation within group or between groups caused by many affecting factors.

In this study, we solve the problem by providing the application of mixed effects model (MEM) as a viable alternative to the ANOVA for longitudinal studies in power energy context. In detail, we propose two models based on MEM to analyze the time-series dataset of micro-inverter PV stations located in Concord city, Massachusetts, USA and show their advantage to ANOVA. Firstly, we implement the average models to compare the seasonal variation of monthly yield taking into account the effects of shading conditions and different orientations. Secondly, we implement the linear regression model based on MEM to estimate and compare the decline rate in a 6-year period of PV stations.

Based on the linear model, MEM has been developed to take into account both the variation in time-series data across individuals (repeated measures context) and the variation across groups (multilevel or hierarchical context). MEM has been widely used in other research fields such as biological, clinical, or medical fields. However, a few scholars have explored the power of MEM for longitudinal studies in power energy [9, 10].

In general, the key contributions of our paper are summarized as follows:

- Increasing the awareness and understanding of the mixed effect model, as a superior method to ANOVA, to interpret the meaningful insights of time-series dataset in power energy;
- Proposing the new average model based on MEM to evaluate the variation of energy yield during a year of PV systems under influences of shading and orientation; and
- Proposing the new linear decline model based on MEM to evaluate the monthly energy yield of PV systems under the influence of shading.

The article is structured as follows. Section 2 reviews the related studies that applied mixed effects model for longitudinal analysis in power energy and other research fields. Section 3 presents the data structure of micro-inverter PV stations, followed by a decomposition method to analyze these data. The average model and linear decline model based on MEM are proposed in Section 4. Section 5 provides the analysis results of the proposed models. Finally, Section 6 summarizes our study and discusses future directions.

2 | RELATED WORKS

The mixed effects model has been widely used over the last couple of decades in biological, clinical, or medical studies [11–16]. This model is suitable to analyse or evaluate the influence of related factors to an observed variable that is repeatedly measured over time such as hourly, daily, or monthly. Such study and repeated measured data are called longitudinal study and longitudinal data or panel data respectively. In recent years, there is a significant number of studies in PubMed and Google Scholar that applied the mixed effects model in many research fields. For example, in hypertension study, Kaneto Mitsumata et al. [17], explored the association of parental hypertension on longitudinal data of blood pressure and metabolism by mixed effects model. In agriculture, Baey Charlotte et al. [18], proposed mixed effects model to describe the interactions among plants and explain the variability observed in the plant’s population. In chemistry, Daniel A. H. et al. [19], applied nonlinear mixed effects model in a longitudinal study of a batch recycle trickle bed reactor system. Furthermore, in order to predict the daily consumption of nature gas, Marek Brabec et al. [20], successfully implemented the nonlinear mixed effects model, which took into account the irregularities of individual customer.

In power engineering, the mixed effects model has recently been attracted and explored by scholars. For example, Cameron Roach [9] used mixed effects model to generate the impact profiles for daily energy demand of the commercial buildings based on data from smart meters. His model was able to quantify the changes in energy caused by building characteristics. In addition, Paulo Fernando et al. [10], utilized the power of mixed effects model combined with the first-order moving average model and first-order autoregressive model to predict the short-term electric power generation from Brazilian distributing utilities. Our previous study in [21] is the first study that found the average degradation rate of micro-inverter PV system of 3% per year by applying the linear version of mixed effects model. To the best of our knowledge, we are among the first research groups exploring the potential applications of mixed effects model to conduct data analytics in power engineering. Such research efforts have opened the new room of applying mixed effects model in many smart grid data analytics in the future.

3 | MONTHLY DATASET OF PV STATIONS

Table 1 illustrates the list of eighteen PV stations that are used in our study. Those PVs all used monocrystalline solar panels and the micro-inverter configuration from Enphase manufacturer [22]. The measurements of generated energy and power are uploaded every five min from the registration day to PV output website [23]. In addition, the technical details of each PV station are the rated PV power; the brand, type of solar panel, total number and power of solar panels; the brand, total number and power of inverters; orientation; shading condition; and tilt degrees of solar panel. The shading situations of PV in Table 1 are classified as follows:

- No shading: PV panel does not suffer from shading;
- Low shading: The shading covers up to 25% of PV panel area;
- Medium shading: The shading covers from 25% to 50% of PV panel area;
- High shading: The shading covers higher than 50% of PV panel area;

These time series data are useful and reliable to perform longitudinal study since all PV stations share the same climate conditions.
TABLE 1 The list of micro-inverter PV stations located in Concord city, Massachusetts, USA (postcode: 01742) in our study. The distance among stations is below 1 km to guarantee the weather condition would vary as little as possible. Shading (e.g. from tree, chimney, or neighbour building) reduces temporarily the amount of radiation reaching the PV panels.

| Station | Installed power (kW) | Orientation | Shading | Tilt degrees |
|---------|----------------------|-------------|---------|--------------|
| 1       | 6.48                 | South       | No      | 45           |
| 2       | 4.9                  | South       |         | 23           |
| 3       | 4.65                 | South West  |         | 33           |
| 4       | 8.2                  | South West  |         | 43           |
| 5       | 10.2                 | South       |         | 1            |
| 6       | 4.16                 | South West  | Low     | 45           |
| 7       | 9.36                 | South       |         | 45           |
| 8       | 6.72                 | South       |         | 45           |
| 9       | 12.48                | South       |         | 45           |
| 10      | 5.76                 | South       |         | 26           |
| 11      | 7.25                 | South       |         | 1            |
| 12      | 14.3                 | South West  |         | 45           |
| 13      | 4.5                  | South       |         | 1            |
| 14      | 4.9                  | South       | Medium  | 45           |
| 15      | 2.22                 | South West  |         | 45           |
| 16      | 6                    | South       |         | 25           |
| 17      | 6.5                  | South       | High    | 22           |
| 18      | 9.9                  | South       |         | 19           |

condition (Concord city) and the measurement data are continuously recorded from 2014 to 2019. The monthly yield $m$ of electric energy is chosen as the variable to investigate the energy performance of a PV station in our study. The calculation of this yield is given in Equation (1),

$$m = \frac{\sum_{j=1}^{M} E_{day,j}}{M P_{pv}} \text{ (kWh/kW)}$$

where $P_{pv}$ is the rated power of PV station, $M$ is the total number of recorded days of the month of interest, and $E_{day,j}$ is the generated energy from PV station on the $j$th day. The resulting monthly yield of PV stations in Table 1 is represented in Figure 1.

Plotting time series data as in Figure 1 is the important first step in analysing their various components. In the second step, we need a decomposition model for reducing the time series data into three components: trend, seasonal effects, and random errors. From there, we can start to work on evaluation or comparison study based on these components to yield some valuable insights. If we assume an additive decomposition, then we formulate the monthly energy yield $m_i$ of the $i$th PV station as Equation (2),

$$m_i = t_i + s_i + e_i,$$

where $t_i$ is the trend component, $s_i$ is the seasonal component, and $e_i$ is the error (or random) component, all at the cycle of 12 months. The additive decomposition is the most appropriate if the magnitude of the seasonal fluctuations, or the variation around the trend, does not change when the time series value increases. The trend component of a time series is commonly identified using a moving average filter [24, 25]. Since there are 12 months in a year we need to use a moving average of length twelve. This means that at each data point $n$ (where $k$ starts at 7) we average the 6 data points behind and 6 data points in front of each month’s data to calculate the mean value for each month we want to calculate an average value for ($a = 6$). However, since the moving average is an even length we have to take the average of two averages as Equation (3).

$$t_i(n) = \frac{1}{2a} \left[ \sum_{k=-a}^{a} \left( \frac{1}{2a} \right) m_i(n + k) + \sum_{k=-a}^{a} \left( \frac{1}{2a} \right) m_i(n - k) \right]$$

The result of these moving averages yields the overall trend of monthly energy yield from 2014 to 2019 as shown in Figure 2. Note that there are 6 month lag and 6 month lead on each end of these trends. This is due to the moving average which requires 6 months of data both behind and 6 data points in front of the position we want to calculate an average value for ($a = 6$). However, since the moving average is an even length we have to take the average of two averages as Equation (3).

$$e_i = m_i - t_i - \frac{1}{12} \sum_{k=1}^{12} (m_i(k) - t_i(k))$$

The error component, also referred to as the irregular or random component, is composed of all the leftover signals which are not explained by the combination of the trend and
The trend, the seasonal, and the error components of monthly energy yields from 2014 to 2019 in Figure 1 are shown.

To calculate error component, the trend and seasonal components are subtracted from the monthly data as Equation (5).

\[ e_i(n) = m_i(n) - t_i(n) - s_i(n) \]  

The decomposition process of monthly yield of PV stations was implemented using R programming version 3.4.1 [26] and decomposing function in R stats package. Then, the trend component \( t_i(n) \) and seasonal component \( s_i(n) \) of PV stations are used as observed variables in the MEMs in our study.

## 4 THE MIXED EFFECTS MODEL

Mixed effects models are commonly applied when dataset is longitudinal (repeated measures) and classified into many groups of objects. In linear regression aspect, each mixed effects model is combined from two terms: fixed effects and random effects. While the former describes general characteristic of whole dataset or between-groups variability, the latter term reflects the variability among many objects in a group (within-group variability) or the variability in repeated measurements of each object (within-object variability). By calculating the fixed effects terms of grouped objects, any differences among groups can be evaluated easily. Further related theories and method of mixed effects models are referenced to [27–29].

In general, the comparison of many groups using MEM is based on a hypothesis test as below:

- The null hypothesis \( (H_0) \): There is no difference in values of a variable among groups. If the \( P \) value (the probability assuming that the null hypothesis is correct) is larger than 0.05, then \( H_0 \) can be assumed;
- The alternative hypothesis \( (H_1) \): There is real difference in values of a variable among groups \( (P < 0.05) \).

### 4.1 Average model

Figure 3 shows the data structure of the seasonal variation \( r_{ij} \) (kWh/kW) of the \( i \)th PV station and shading group \( j \), collected from the Equation (6) below:

\[ r_{ij} = \max\{t_i(n)\} - \min\{t_i(n)\} \]  

Our objectives are to find out the common seasonal variation of all PV stations, compare the seasonal variations under different shading conditions, and evaluate the effect of orientation to seasonal variation. We propose three average models based on MEM in order to obtain the results as follow.

#### 4.1.1 Average model 1: Finding common seasonal variation

In the first average model, the \( r_{ij} \) of the \( i \)th PV station is treated as the level 1 of model and nested into the level 2 of shading condition as Figure 3. The seasonal variation value of the \( i \)th PV station is formulated as Equation (7),

\[ v_{ij} = V_j + r_{ij} \]  

and \( V_j \) is defined as Equation (8):

\[ V_j = V_0 \]  

where \( V_j \) is assumed as the seasonal variation value of shading group \( j \). \( V_0 \) is the common variation of all PV stations regardless of the shading conditions, \( r_{ij} \) is the residual between the measured variation and estimated variation from model in Equation (7) of the \( i \)th station. The residual term \( r_{ij} \) is assumed to follow a normal distribution with the variance \( \sigma^2 \),

\[ r_{ij} \sim \mathcal{N}(0, \sigma^2) \]  

By substituting Equation (8) to Equation (7), Equation (7) is written as Equation (9) below:

\[ v_{ij} = V_0 + r_{ij} \]  

The term \( V_0 \) in Equation (9) is called the fixed-effect part of model since it represents the common seasonal variation of all PV stations during a 6-year period. The term \( r_{ij} \) is called...
the random-effect since it includes the variant elements in the model. This random-effect explains the variability in \( r_{ij} \) from many PV stations.

### 4.1.2 Average model 2: Interpreting the effect of shading

In the second average model, the seasonal variation model is formulated in the same way as Equation (7). Since we explore whether there is any difference in seasonal variations among shading groups \( j \), the Equation (8) is rewritten as Equation (10) below:

\[
V_j' = V_0 + V_{0,j} + \lambda_{0j}
\]  

(10)

where \( V_{0,j} \) is assumed as the amount of difference in \( V_j' \) caused by shading group \( j = \{0, 1, 2, 3\} \). Then Equation (7) is written as Equation (11) below:

\[
r_{ij} = (V_0 + V_{0,j} + \lambda_{0j}) + r_{ij}
\]

\[
= (V_0 + V_{0,j} + \lambda_{0j} + r_{ij})
\]

(11)

The fixed-effect part \((V_0 + V_{0,j})\) in Equation (11) represents the seasonal variation of PV stations belonging to group \( j \). \( \lambda_{0j} \) is the residual between the variation of PV stations in each shading group and the common variation \( V_0 \). The residual term \( \lambda_{0j} \) is assumed to follow a normal distribution with the variance \( \sigma_{\lambda}^2 \):

\[
\lambda_{0j} \sim \mathcal{N}(0, \sigma_{\lambda}^2)
\]

### 4.1.3 Average model 3: Interpreting the effect of orientation

In the third average model, the seasonal variation model is the same as Equation (7). However, we assume that there exists an effect of the orientation of PV panel (South (coded as \( u = 0 \)) or South West (coded as \( u = 1 \))) to \( r_{ij} \) and this effect is constant regardless of the shading conditions. Hence, the value of \( V_j' \) is formulated as Equation (12) below:

\[
V_j' = V_0 + V_{0,u} + \lambda_{0j}
\]

(12)

The Equation (7) is then rewritten as Equation (13) below:

\[
r_{ij} = (V_0 + V_{0,u} + \lambda_{0j}) + r_{ij}
\]

\[
= (V_0 + V_{0,u} + \lambda_{0j} + r_{ij})
\]

(13)

The term \((V_0 + V_{0,u})\) in Equation (13) is called the fixed-effect part of model since it represents the seasonal variation plus the effect of orientation on the variation of PV stations. The resulting analysis of three average models are shown in Subsection 5.1.1.

### 4.2 Linear decline model

Figure 4 shows the data structure of the trend values \( t_i \) of the \( i \)th PV station during 6-year period. In this scenario, PV stations are classified into only two groups: non-shading \( (S = 0) \) and shading \( (S = 1) \) since we intend to compare the results to our previous study [21] at yearly scale. As the trend values tend to decrease over time, we propose two linear decline models as follows.

#### 4.2.1 The unconditional model

In the unconditional model, we treat the trend dataset in Figure 2 as only one group regardless of the shading conditions. The trend of the \( i \)th PV station is represented as Equation (14):

\[
t_i = \alpha_i + \beta_i u + \epsilon_i
\]

(14)

where \( t_i \) is the trend values for individual PV station \( i = 1, 2, \ldots, 18 \), measured every month from July 2014 to June 2019 and represented by the index of month \( u = 0, 1, 2, \ldots, 71 \). The \( \alpha_i \) and \( \beta_i \) are named the initial trend (kWh/kW) and decline rate of the \( i \)th PV station, respectively. The meaning of initial trend is the first value of trend that was calculated in July 2014 \( (n = 0) \). \( \epsilon_i \) is the residual between the measured trend and estimated trend from model of the \( i \)th PV station. The residual \( \epsilon_i \) is assumed to follow a normal distribution with zero mean and variance \( \sigma_{\epsilon}^2 \),

\[
\epsilon_i \sim \mathcal{N}(0, \sigma_{\epsilon}^2)
\]

Since all the PV stations are assumed as one group, it is reasonable to imply that they all shared a common initial trend and common decline rate. Therefore, we interpret the parameters \( \alpha_i \) and \( \beta_i \) as Equation (15),

\[
\alpha_i = A + \mu
\]

\[
\beta_i = B + \nu
\]

(15)

where \( A \) and \( B \) are the common initial trend and the common decline rate of all PV stations, respectively. The residual terms
n and v are assumed to follow a normal distribution with their respective variances \( u \sim \mathcal{N}(0, \sigma_u^2) \) and \( v \sim \mathcal{N}(0, \sigma_v^2) \).

By substituting Equation (15) to Equation (14), Equation (14) is written as below:

\[
t_i = (A + u) + (B + v)n + e_i
\]
\[
= (A + Bn) + (u + vn) + e_i
\]  
\[
(16)
\]

The term \((A + Bn)\) in Equation (16) is called the fixed-effect part of model since it represents the general trend of all PV stations during a 6-year period. The random-effect term \((u + vn)\) includes the variant elements in the model. This random-effect explains the variability in trends from many PV stations.

4.2.2 The conditional model

In this model, the trend model is formulated in the same way as Equation (14). However, the PV systems are classified into non-shading and shading group as Figure 4 since we explore whether there is any difference in trends between two groups. Therefore, the Equation (15) is rewritten as Equation (17) below:

\[
\alpha_i = A_0 + A_1s_i + u
\]
\[
\beta_i = B_0 + B_1s_i + v
\]  
\[
(17)
\]

where \(A_0\) and \(B_0\) have the same meaning as \(A\) and \(B\) in Equation (15). The parameters \(A_1\) and \(B_1\) reflect any differences in the initial trend and in the decline rate between two PV groups respectively. The \(s_i\) equals to 0 if PV station belongs to non-shading group and to 1 if it belongs to shading group as shown in Figure 4.

By substituting Equation (17) to Equation (14), then Equation (14) is written as Equation (18) below:

\[
t_i = (A_0 + A_1s_i + u) + (B_0 + B_1s_i + v)n + e_i
\]
\[
= [(A_0 + A_1S_i) + (B_0 + B_1S_i)n] + (u + vn) + e_i
\]  
\[
(18)
\]

In this case, the fix-effect term is \([(A_0 + A_1S_i) + (B_0 + B_1S_i)n]\). Unlike the fix-effect term in unconditional model, it now separately represents the decreasing trends of non-shading PV stations and shading PV stations. In addition, if there exists any significant difference between two groups, then \(A_1\) and \(B_1\) identify how large the difference is. The resulting analysis of both unconditional model and conditional model are shown in Section 5.2.2.

5 RESULTS AND DISCUSSION

All algorithms and models in Section 4 were implemented using R programming version 3.4.1 [26] and mixed effects analysis from nlme package [30] or lme4 package [31]. The random process used the same number of generators to ensure the reproducibility.

5.1 Seasonal variation analysis

5.1.1 Average models based on MEM

Table 2 shows the fix-effect results of three average models, respectively. From these results, we conclude the findings about the seasonal variation of PV stations as follows:

- The common seasonal variation for the entire PV stations regardless of shading conditions and orientation is 2.86 (kWh/kW). Furthermore, this result is significant since the corresponding P value is much smaller than 0.05; (average model 1)
- Taking into account the shading conditions, the PV stations which suffer from high shading have the largest seasonal variation of 3.38 (= 2.44 + 0.94) (kWh/kW). On the contrary, the non-shading PV station have the smallest seasonal variation of 2.44 (kWh/kW) and the seasonal variation tends to increase whenever the shading level (low, medium, high) increases; (average model 2)
- The seasonal variation of high shading group is significantly different from the non-shading group since the corresponding P value is smaller than 0.05 (<0.0001). In addition, the amount of this difference is about +0.94 (kWh/kW). Meanwhile, there are not any significant difference in the seasonal variation of low shading and medium shading groups since its corresponding P values (0.16 and 0.66, respectively) are larger than 0.05; (average model 2)
- The impact of orientation to seasonal variation is not significantly different as we expected. Since the P value of South West group is 0.58, which is larger than 0.05. Therefore, there is no significant difference in seasonal variation between two orientations of South and South West. (average model 3)

The mean values and its corresponding 95% confidence intervals (CI) in Figure 5 confirms again our conclusions. The overlapping ranges of 95% CI indicate that any differences among PV groups are not significant, except the non-overlapping ranges between high shading group and non-shading group.

Table 3 shows the random-effect results of three average models and the corresponding R-squared. This score, calculated as \(\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\epsilon^2)\), represents the proportion of variance in the seasonal variation that is explained by the model. We interpret the meaningful insights about the effectiveness of models as follows:

- The R-squared of the average model 2 is higher than that of the average model 3 indicates that the average model 2 is more fit to the dataset than the average model 3.
- Based on both the variance and R-squared results, the impact of shading to the seasonal variation is higher than the impact of orientation.
TABLE 2  The fixed-effect results of average models

| Model                  | Mean square error (MSE) | Variation | Meaning                                                   | Value (kWh/kW) | P value |
|------------------------|-------------------------|-----------|-----------------------------------------------------------|----------------|---------|
| Average model 1        | 0.295                   | $V_0$     | Common seasonal variation                                 | 2.86           | 0.0001  |
| Average model 2        | 0.195                   | $V_0$     | Seasonal variation of non-shading group                   | 2.44           | 0.02    |
|                        |                         | $V_{01}$  | Difference in seasonal variation between low shading      | +0.42          | 0.16    |
|                        |                         |           | group and non-shading group                               |                |         |
|                        |                         | $V_{02}$  | Difference in seasonal variation between medium shading   | +0.76          | 0.66    |
|                        |                         |           | group and non-shading group                               |                |         |
|                        |                         | $V_{03}$  | Difference in seasonal variation between high shading     | +0.94          | <0.0001 |
|                        |                         |           | group and non-shading group                               |                |         |
| Average model 3        | 0.298                   | $V_0$     | Seasonal variation of South group                         | 2.88           | 0.0001  |
|                        |                         | $V_{01}$  | Difference in seasonal variation between South West group | +0.15          | 0.58    |
|                        |                         |           | and South group                                           |                |         |

FIGURE 5  The mean values of seasonal variation and its 95% confidence intervals from average model 2 and average model 3 in Table 2

Finally, the Quantile–Quantile (Q–Q) plot in Figure 6 confirms that the assumption of normal distribution of residual terms is rational since if the two sets have the same distribution, the points should fall approximately along a reference line.

TABLE 3  The random-effect results of average models

| Model                  | Parameter | Source of variance | Variance | R-squared (%) |
|------------------------|-----------|--------------------|----------|---------------|
| Average model 1        | $\sigma_r^2$ | Residual of PV     | 0.31     | No applicable |
|                        |           | stations           |          |               |
| Average model 2        | $\sigma_r^2$ | Residual between  | 0.39     | 61%           |
|                        |           | groups             |          |               |
|                        | $\sigma_r^2$ | Residual of PV     | 0.25     |               |
|                        |           | stations           |          |               |
| Average model 3        | $\sigma_r^2$ | Residual between  | 0.10     | 28%           |
|                        |           | groups             |          |               |
|                        | $\sigma_r^2$ | Residual of PV     | 0.26     |               |
|                        |           | stations           |          |               |

5.1.2  Comparison with ANOVA

Table 4 shows the analysis results from ANOVA to investigate the effects of shading and orientation. In both cases, ANOVA reports that there are no group differences since the $P$ values are larger than 0.05 (0.11 and 0.82). This result confirms the disadvantage of ANOVA in multiple groups comparison. Even if the $P$ value is smaller than 0.05, we have to take multiple pairwise comparisons by using post-hoc test such as Bonferroni test [32] or Tukey’s Test [33] to determine the difference between groups.

5.2  Decline trend analysis

5.2.1  Linear decline model based on MEM

Tables 5 and 6 show the resulting parameters of the fix-effect and random-effect terms of the decline trend modeling, respectively. From these results, the trend models of each PV group are rewritten as follow:
**TABLE 4** The results of ANOVA tests to find any difference in seasonal variation of PV groups in Table 1. The F-test value is the ratio of between-group variation and within-group variation.

| ANOVA test         | Source of variation | Degrees of freedom | Sum of square | Mean square | F-test value | P value |
|--------------------|---------------------|--------------------|---------------|-------------|--------------|---------|
| Effect of shading  | Between-groups      | 3                  | 1.79          | 0.59        | 2.38         | 0.11    |
|                    | Within-group        | 14                 | 3.51          | 0.25        |              |         |
| Effect of orientation | Between-groups   | 1                  | 0.02          | 0.02        | 0.06         | 0.82    |
|                    | Within-group        | 16                 | 5.28          | 0.33        |              |         |

The trend model for all PV stations is:

$$t_i = 2.98 - (5.5 \times 10^{-3})n$$  \hspace{1cm} (19)

The trend model for non-shading PV group ($S = 0$):

$$t_i = (A_0 + A_1S_i) + (B_0 + B_1S_i)n$$

$$= (3.10 - 0.16)0 + (-8.6 \times 10^{-3} + 4.3 \times 10^{-3})0)n$$

$$= 3.10 - (8.6 \times 10^{-3})n$$  \hspace{1cm} (20)

The trend model for shading PV group ($S = 1$):

$$t_i = (A_0 + A_1S_i) + (B_0 + B_1S_i)n$$

$$= (3.10 + (-0.16)1 + (-8.6 \times 10^{-3} + 4.3 \times 10^{-3})1)n$$

$$= 2.94 - (4.3 \times 10^{-3})n$$  \hspace{1cm} (21)

However, there is no significant difference between trend models in Equation (20) and Equation (21) since the obtained $P$ values of $A_1$ and $B_1$ are 0.35 and 0.06 which are both greater than 0.05. In addition, the 95% CI. ranges of $A_1$ and $B_1$ in Table 5.

**TABLE 5** The fixed-effect results of both models.

| Model            | Parameter | Meaning                          | Value (kWh/kW) | 95% CI. (kWh/kW) | $P$ value |
|------------------|-----------|----------------------------------|----------------|------------------|-----------|
| Unconditional model | $A$     | Initial trend                    | 2.98           | 2.83 – 3.13      | $<0.0001$ |
|                  | $B$     | Decline rate                     | $-5.5 \times 10^{-3}$ | $(-7.6 \times 10^{-3} - (-3.4) \times 10^{-3})$ | $<0.0001$ |
| Conditional model | $A_0$   | Initial trend                    | 3.10           | 2.81 – 3.39      | $<0.0001$ |
|                  | $B_0$   | Decline rate                     | $-8.6 \times 10^{-3}$ | $(-12.3) \times 10^{-3} - (-4.8) \times 10^{-3})$ | $<0.0001$ |
|                  | $A_1$   | Difference in initial trend between two groups | $-0.16$ | $(-0.53) - 0.2$ | 0.35      |
|                  | $B_1$   | Difference in decline rate between two groups | $4.3 \times 10^{-3}$ | $(-0.14) \times 10^{-3} - (8.7) \times 10^{-3})$ | 0.06      |

**TABLE 6** The random-effect results of both models.

| Model            | Parameter | Source of variance | Variance | R-squared (%) |
|------------------|-----------|--------------------|----------|---------------|
| Unconditional model | $\sigma_u^2$ | Initial trend | 0.32      | 74.7%         |
|                  | $\sigma_v^2$ | Decline rate | $4.52 \times 10^{-3}$ |               |
|                  | $\sigma_e^2$ | Residuals        | 0.11      |               |
| Conditional model | $\sigma_u^2$ | Initial trend | 0.32      | 74.6%         |
|                  | $\sigma_v^2$ | Decline rate | $4.19 \times 10^{-3}$ |               |
|                  | $\sigma_e^2$ | Residuals        | 0.11      |               |
also cross the zero value. Therefore, we conclude that there is no significant difference in trend of monthly energy yield of micro-inverter PV systems under shading and non-shading conditions. This means that the common model in Equation (19) can be used to represent the decline trend of two PV groups. This result has confirmed the ability of the micro-inverter PV system to eliminate the effect of shading.

Another result is the common decline rate, counted in a year in percentage (DRPV) of micro-inverter PV systems calculated as Equation (22)

$$\text{DRPV} = -\frac{B}{A} \left( \frac{12 \text{ months}}{\text{year}} \right) (100)$$

(22)

and the 95% CI. range is calculated as Equation (23)

$$95\% \text{ CI.} = -\frac{(-3.4 \times 10^{-3})}{2.98} (12)(100) \approx 1.4\% \text{ to } 3.1\%$$

(23)

Comparing to the 95% CI. range of DRPV from 2% to 4.3% in our previous study [21] at yearly scale, the 95% CI. range of DRPV in this study is more accurate since its 95% CI. range is smaller, that is 1.7% (1.4% to 3.1%) as in Equation (23) compares to 2.3% (2% to 4.3%) in [21]. Finally, the Q–Q norm plot in Figure 7 confirms that the assumption of normal distribution of residual terms in linear decline model is rational.

5.2.2 Comparison with linear regression model

As mentioned in Section 1, using ANOVA for the time series datasets violates the independent variable assumption. Therefore, we applied the traditional linear regression model to compare the results with our proposed models based on MEM. The models of single linear regression and multiple variable linear regression are shown in Equation (24) and Equation (25) respectively.

$$t = \alpha + \beta n + \epsilon$$

(24)

$$t = \alpha + \beta n + \gamma s + \epsilon$$

(25)

The outcome $t$ is the decline trend of PV stations. $\alpha, \beta$, and $\gamma$ are the coefficients of models. The variable $s$ is the index of month. The binary variable $s$ represents shading condition, $s = 0$ for non-shading group and $s = 1$ for shading group.

The result analysis in Table 7 shows the same parameters found by the models based on MEM in Table 5. With the P value equals 0.1 (which is larger than 0.05), the multiple linear regression also indicates that there is no significant difference in the decline trend between two groups.

Furthermore, Table 8 compares the goodness-of-fit of models for two approaches. Besides the MSE and R-squared score, we also applied the Akaike information criterion (AIC). AIC is a method for evaluating how well a model fits the dataset. AIC can be generated from [34, 35]. The lower AIC score, the better the model fits the dataset. Our evaluation are shown as follows:

- The MSE scores of two methods are about the same since the MEM is actually extended from the linear regression method;
- The R-squared and AIC in Table 8 indicate that the models based on MEM are better fit to the datasets than the traditional linear regression models.

6 CONCLUSIONS

The mixed effects model has arisen as a suitable approach to handle the influence analysis of longitudinal study instead of using the conventional ANOVA method. In this study, we illustrated the applications of this model on the time-serial data of micro-inverter PV stations to evaluate the effect of shading.
and orientation to the seasonal variation and decline trend of monthly energy yield. By decomposing the observed variables into fixed-effect and random-effect terms, the analysis results from our proposed models provide more details about the impact of influential factors than ANOVA method without violating the assumption of independent variable. The proposed method via mixed effect model in our study could further be applied to other types of datasets in Smart Grid domain such as power consumption from smart meters or energy production from wind turbines.

In further study, we plan to use the mixed effects model to compare the long-term energy efficiency between micro-inverter PV and string inverter PV system, taking into account various conditions and regions. We will also integrate the mixed effects model into ensemble learning or machine learning models and evaluate the improvement for the forecasting tasks.

ACKNOWLEDGMENT
This research is supported by Ratchadapisek Somphot Fund for Artificial Intelligence, Machine Learning, and Smart Grid Technology (Year 2) Research Unit (RU) and for Postdoctoral Fellowship, Chulalongkorn University.

ORCID
Ngoc Thien Le https://orcid.org/0000-0002-1113-8756
Watit Benjapolakul https://orcid.org/0000-0001-6134-7268

REFERENCES
1. Wang, E., Cook, D., Hyndman, R.J.: A new tidy data structure to support exploration and modeling of temporal data. Journal of Computational and Graphical Statistics 29(3), 466–478 (2020)
2. Rizopoulos, D.: ANOVA method for comparison models. https://www.rdocumentation.org/packages/ltm/versions/1.1-1/topics/anova. Accessed 12 June 2020
3. Kudelas, D., et al.: Investigation of operating parameters and degradation of photovoltaic panels in a photovoltaic power plant. Energies 12(19), 4640 (2019)
4. Huang, S., Lo, Y., Lu, C.: Non-technical loss detection using state estimation and analysis of variance. IEEE Trans. Power Syst. 28(3), 2959–2966 (2013)
5. Su, C., Lee, W., Wen, C.: Electricity theft detection in low voltage networks with smart meters using state estimation. In: 2016 IEEE International Conference on Industrial Technology, pp. 493–498. IEEE, Piscataway, NJ (2016)
6. Ding, N., et al.: Time series method for short-term load forecasting using smart metering in distribution systems. In: 2011 IEEE Trondheim PowerTech, pp. 1–6. IEEE, Piscataway, NJ (2011)
7. Palm, J., Ellegård, K., Hellgren, M.: A cluster analysis of energy-consuming activities in everyday life. Building Research & Information 46(1), 99–113 (2018)
8. Aarts, E., et al.: A solution to dependency: using multilevel analysis to accommodate nested data. Nat. Neurosci. 17(4), 491–496 (2014)
9. Roach, C.: Estimating electricity impact profiles for building characteristics using smart meter data and mixed models. Energy Build. 211, 109686 (2020)
10. Simões, P.F.M., et al.: Analysis and short-term predictions of non-technical loss of electric power based on mixed effects models. Socio-Economic Planning Sciences 71, 100804 (2020)
11. Feder, P.L.: Some differences between fixed, mixed, and random effects analysis of variance models. Journal of Quality Technology 6(2), 98–106 (1974)
12. Molenberghs, G., Verbeke, G.: A review on linear mixed models for longitudinal data, possibly subject to dropout. Statistical Modelling 1(4), 235–269 (2001)
13. Goldstein, H.: Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika 73(1), 43–56 (1986)
14. Lindstrom, M.J., Bates, D.M.: Nonlinear mixed effects models for repeated observations using mixed effects model. Biometrika 63(3), 673–687 (1990)
15. Beiter, P.J., Landis, J.R.: A mixed-effects model for categorical data. Biometrics 41(4), 991–1000 (1985)
16. Roy, A.: Estimating correlation coefficient between two variables with repeated observations using mixed effects model. Biometrical Journal 48(2), 286–301 (2006)
17. Mitsumata, K., et al.: Effects of parental hypertension on longitudinal trends in blood pressure and plasma metabolic profile. Hypertension 60(5), 1124–1130 (2012)
18. Baey, C., et al.: Mixed-effects estimation in dynamic models of plant growth for the assessment of inter-individual variability. Journal of Agricultural, Biological and Environmental Statistics 23(2), 208–232 (2018)
19. Hickman, D.A., et al.: Nonlinear mixed-effects models for kinetic parameter estimation with batch reactor data. Chem. Eng. J. 377, 258–269 (2013)
20. Brabec, M., et al.: A nonlinear mixed effects model for the prediction of natural gas consumption by individual customers. International Journal of Forecasting 24(4), 659–678 (2008)
21. Le, N.T., Benjapolakul, W.: Comparative electrical energy yield performance of micro-inverter pv systems using a machine learning approach based on a mixed-effect model of real datasets. IEEE Access 7, 175126–175134 (2019)
22. Enphase: Enphase Microinverter. https://enphase.com/en-us/products-and-services/microinverters. Accessed 14 February 2020
23. PVOutput dataset. https://pvoutput.org/. Accessed 14 February 2020
24. Soliman, S.A., Al-Kandari, A.M.: Chapter 9-electric load modeling for long-term forecasting. In: Soliman, S.A., Al-Kandari, A.M. (eds.) Electrical Load Forecasting, pp. 353–406. Butterworth-Heinemann, Boston (2010)
25. Ma, Y., Feng, S.: Prediction and forecast of a wind power. In: Wang, N., Kang, C., Ren, D. (eds.) Large-Scale Wind Power Grid Integration, pp. 185–209. Academic Press, Oxford (2016)
26. R.: The R-project for statistical computing. (2020). https://cran.r-project.org/bin/windows/base/. Accessed 14 February 2020
27. Verbeke, G., Molenberghs, G.: In: Linear Mixed Models for Longitudinal Data. Springer, New York, NY (2000)
28. Pinheiro, J.C., Bates, D.M.: Mixed-Effects Models in s and s-plus. Springer, New York, NY (2000)
29. Jiang, J.: Linear and Generalized Linear Mixed Models and Their Applications. Springer, New York, NY (2007)
30. Pinheiro, J., et al.: nlme Package: Linear and nonlinear mixed effects models. https://rdrr.io/cran/lme4/man/lmer.html. Accessed 14 February 2020
31. Bates, D., et al.: Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1), 1–48 (2015)
32. Hochberg, Y.: A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75(4), 800–802 (1988)
33. Ghosh, M.N., Sharma, D.: Power of tukey’s test for non-additivity. J. R. Stat. Soc.: Series B (Methodological) 25(1), 213–219 (1963)
34. Hamparsum, B.: Model selection and akaike’s information criterion (aic): The general theory and its analytical extensions. Psychometrika 52(3), 345–370 (1987)
35. Mutua, F.M.: The use of the akaike information criterion in the identification of an optimum flood frequency model. Hydrol. Sci. J. 39(3), 235–244 (1994)