On the Interdependence and Importance of Meteorological Variables for Photovoltaic Output Power Estimation

Tarek AlSkaif†1, Soumyabrata Dev†2, Lennard Visser1, Murhaf Hossari2, and Wilfried van Sark†1
1 Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands
2 The ADAPT SFI Research Centre, Trinity College Dublin, Dublin, Ireland

Abstract—While the large-scale deployment of Photovoltaic (PV) systems plays an important role in limiting global warming, the variability of PV output power poses challenges in grid management. Typically, the PV output power is dependent on various meteorological parameters at the PV site. In this paper, we analyse the interdependence of different meteorological variables and show their importance for PV output power estimation. Using Principal Component Analysis (PCA), we identify the primary meteorological variables for PV output power estimation. The numerical evaluation is performed using 3 years long of 9 meteorological variables data and PV output power data of 10 distinct rooftop PV systems, located in the city of Utrecht, the Netherlands. Simulation results show the interdependence between the meteorological variables and demonstrate that relative humidity, visibility, temperature and cloud cover are the most important variables for estimating PV output power in Utrecht.

Index Terms—Photovoltaic, Meteorological variables, Cross-correlation, Principal component analysis, Regression.

I. INTRODUCTION

To support reliable grid operation and increasing Photovoltaic (PV) systems integration, it is important to accurately estimate and subsequently forecast the PV output power and understand its variability over time. As a result, the field of PV output power estimation and forecast has received an increasing attention among researchers over the past decade. The different methods of PV power forecast can be generally divided into four classes: statistical, cloud imagery, physical and hybrid methods [1]–[3].

Typically, those forecasting methods integrate various meteorological variables as exogenous input in order to achieve more accurate forecasts since these variables are considered as a source of uncertainty in PV output power [1], [2]. Despite their relevance and for undefined reasons, most of the forecasting methods consider only few meteorological variables as data input. Additionally, these methods are output oriented and do not consider to what extent different meteorological variables are capable of capturing the solar power variability. Subsequently, in literature, less attention was paid to the interdependence of different meteorological variables and their individual importance in the results of different methods of forecast. In [4], an analysis of multiple meteorological variables was proposed but for the purpose of rainfall detection. In [5], weather forecasts were used to predict solar power generation but without analyzing their importance and interdependence.

† Authors contributed equally.

In this paper, we bridge the gap in the literature in the field of solar energy analytics by providing a comprehensive analysis of the input meteorological variables and evaluating their importance for PV output power estimation. We attempt to solve the following research question: how are the different meteorological variables interdependent amongst each other, and how do they impact the PV power output estimation? We perform a dimensionality reduction of the meteorological variables and show how the data can be represented in the reduced subspace using the Principal Component Analysis (PCA) method.

Using 3 years long data holding several meteorological variables and PV output power of 10 rooftop PV systems (i.e., households), we will point out the variables that are most significant to consider when estimating the PV output power. The meteorological variables data are measured by a weather station in the city of Utrecht, the Netherlands and the PV systems are also located in Utrecht. A wide set of meteorological variables are considered in this study, namely temperature (T), dew point temperature (DP), humidity level (RH), visibility (V), air pressure (P), wind speed (WS), cloud cover (CC), wind bearing (WB) and precipitation (R). The study provides insights and future directions for accurate PV output power forecast by helping researchers in the area of solar energy analytics generate forecasting models that use a lower-dimensional subspace of meteorological variables.

The structure of the paper is organized as follows. Section II describes the PCA method. In Section III, we present the interdependence results and analyze the importance of different meteorological variables for PV output power estimation. Finally, we conclude the paper and give pointers for future directions in Section IV.

II. PRINCIPAL COMPONENT ANALYSIS (PCA)

The PCA is an unsupervised linear transformation technique that is prominently used for feature extraction and dimension reduction. It is used to transform the original data onto a new feature space that maintains the most relevant information. In addition to improving the storage space or the computational efficiency, this can also help improving the predictive performance of a learning algorithm [6].

We use the PCA method to understand the interdependence of the meteorological variables. We consider different vectors of meteorological variables, indicated by \( v_1, v_2, \ldots, v_n \), where \( n \) indicates the number of meteorological variables.
Each vector has $m$ number of observations (i.e., weather recordings timestamps in the assessment period). We construct the variable matrix $\mathbf{X}$, of dimension $m \times n$, by stacking the $n$ vectors of meteorological variables, as:

$$\mathbf{X} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n].$$  \hspace{1cm} (1)

To find the principal components, we first need to standardize the variable matrix $\mathbf{X}$ across each of the meteorological variables, with respect to the mean and standard deviation of each vector, as:

$$\mathbf{X}_{std} = \left[ \frac{\mathbf{v}_1 - \bar{\mathbf{v}}_1}{\sigma_{v_1}}, \frac{\mathbf{v}_2 - \bar{\mathbf{v}}_2}{\sigma_{v_2}}, \ldots, \frac{\mathbf{v}_n - \bar{\mathbf{v}}_n}{\sigma_{v_n}} \right],$$  \hspace{1cm} (2)

where $\mathbf{X}_{std}$ is the standardized matrix of $\mathbf{X}$.

The next step is to construct the covariance matrix of $\mathbf{X}_{std}$ which stores the pairwise covariances between the different meteorological variables. After that, we perform an eigenvalue and eigenvectors decomposition of the covariance matrix. We sort the eigenvalues by decreasing order and rank the corresponding eigenvectors. The resulted eigenvectors define the new orthogonal components, also called ‘the principal components’. These principal components are mutually uncorrelated to each other even if the input meteorological variables are correlated due to the orthogonality of the decomposed eigenvectors.

### III. Results and Discussions

#### A. Case Study

In this case study, 3 years of hourly meteorological variables data from Feb, 2014 until Feb, 2017 are used to generate the results. The data are measured by a weather station at De Bilt, Utrecht, the Netherlands with coordinates 52.0907° N, 5.1214° E. This weather station belongs to the Royal Netherlands Meteorological Institute (KNMI) and the data is made available online via their website [7]. Rooftop PV output power data, of the same time resolution, of 10 households located also in Utrecht are used in the regression model in order to determine the most important meteorological variables. The surface azimuth angle of the installed solar panels in all considered households is 180 (i.e., south oriented). The considered households have different PV system capacities. Before training the different models, we normalize the PV power output of each system according to their size. Matlab R2016b environment is used to perform the different analyses in this case study.

#### B. Interdependence of Meteorological Variables

In this section, we systematically analyze $n = 9$ meteorological variables, namely temperature (T), dew point temperature (DP), humidity level (RH), visibility (V), air pressure (P), wind speed (WS), cloud cover (CC), wind bearing (WB) and precipitation (R), and check their interdependence amongst each other. We compute the cross-correlation values between each pair of the $n = 9$ meteorological variables by calculating the correlation coefficients of the variables matrix $\mathbf{X}$. The correlation coefficient between a pair of meteorological variables indicates the degree of correlation between them (i.e., a measure of their linear dependence) [8]. For instance, the correlation coefficient of the first two vectors $v_1$ and $v_2$ in matrix $\mathbf{X}$ can be calculated using the Pearson correlation coefficient formula as follows:

$$\rho(v_1, v_2) = \frac{1}{m - 1} \sum_{i=1}^{m} \left( \frac{v_{1i} - \mu_{v1}}{\sigma_{v1}} \right) \left( \frac{v_{2i} - \mu_{v2}}{\sigma_{v2}} \right)$$  \hspace{1cm} (3)

where $m$ is the number of observations in each vector, $\mu_{v1}$ and $\sigma_{v1}$ are the mean and standard deviation of $v_1$, respectively, and $\mu_{v2}$ and $\sigma_{v2}$ are the mean and standard deviation of $v_2$. Consequently, the correlation coefficient matrix of all vectors in matrix $\mathbf{X}$ is a new matrix $\mathbf{R}$ of correlation coefficients for each pairwise variable combination:

$$\mathbf{R} = \begin{pmatrix} \rho(v_1, v_1) & \cdots & \rho(v_1, v_n) \\ \vdots & \ddots & \vdots \\ \rho(v_n, v_1) & \cdots & \rho(v_n, v_n) \end{pmatrix}.$$  \hspace{1cm} (4)

![Fig. 1. Interdependence of the various meteorological variables amongst each other: Temperature (T), dew point temperature (DP), humidity level (RH), visibility (V), air pressure (P), wind speed (WS), cloud cover (CC), wind bearing (WB) and precipitation (R) (best viewed in color).](image)

The resulted interdependence of the considered meteorological variables in the 3-years period is illustrated in Fig. 1. The off-diagonal elements show the degree of dependency of each pair of variables. We observe that for an oceanic climate like the one in the Utrecht area, the temperature (T) has a strong positive correlation with the dew point temperature (DP) indicating that when T increases, the DP also increases. The visibility (V) has a strong negative correlation with the relative humidity (RH) indicating that when the horizontal view during observation is high, the RH is low, and vice versa. Furthermore, V is positively correlated with T (e.g., during sunny days in spring and summer seasons) and with wind speed (WS) (i.e., when WS is high, V is usually high). Similarly, cloud cover (CC) has a degree of positive correlation...
with RH. In the Netherlands, it is likely that the humidity is high when it is cloudy but this can also happen during less cloudy days due to the relatively large water surface area.

C. Principal Component Analysis

Fig. 2 shows the results of the principal component analysis. It illustrates the percentage of variance captured by each of the principal components. It is noticed that most of the variance is captured by the first primary principal components, indicating that there is a high degree of correlation amongst the input variables. We observe from Fig. 2, that cumulatively, around 80% of the input variance is captured by the first 5 principal components.

The biplot representation of the input meteorological variables is depicted in Fig. 3. The figure illustrates the contribution of the input variables into the first two primary principal components. The principal components are orthogonal to each other and they provide insights on the meteorological variables correlation. It can be noticed that the data is more spread along the x-axis (i.e., 1st principal component) than the y-axis (i.e., 2nd principal component), which is consistent with the results of Fig. 2 (i.e., more variance is captured by the 1st principal component than the 2nd).

It can be observed that temperature (T), dew point temperature (DP) and visibility (V) are correlated and have high individual contribution to the first principal component. In addition, the precipitation (R) and cloud cover (CC) are correlated with each other and have high individual contributions to the second principal component. Similarly, visibility (V) and relative humidity (RH) are negatively correlated with each other. Wind speed (WS) and wind bearing (WB) are positively correlated but have low contribution to the first two principal components.

D. Impact of Meteorological Variables

In this section, we analyze the importance of each of the input meteorological variables, \( v_1, v_2, \ldots, v_9 \) in \( X \), to estimate the PV output power, \( p \). We train a regression ensemble model, the least squares boosting (LSBoost) algorithm, using the response vector \( p \), and the predictor variables vectors \( v_i \). The MATLAB function \texttt{predictorImportance} is used to estimate the importance of each of these predictor vectors in regressing the PV output power \( p \). It estimates the importance by accumulating the estimates over all the weak learners in the ensemble – a higher value indicates that the particular variable has more importance in regressing the PV output power.

The same \( n = 9 \) meteorological variables across the 3 years data are used to generate the results. Fig. 4 shows the predictor importance for all the input meteorological variables in PV output power estimation. We observe that variables such as
relative humidity (RH), visibility (V), temperature (T) and cloud cover (CC) have the highest importance when estimating the PV output power in Utrecht. Therefore, this reduced space of variables can be used when forecasting the PV output power instead of considering the whole original space of variables (i.e., 9 variables).

It is important to note that the importance and ranking of the meteorological variables is dependent on the climate of the considered area of study. We confirm that the observed ranking is applicable to any region that has a climate similar to the one of Utrecht, the Netherlands (i.e., oceanic climate). Furthermore, the methods are generic and can be used to perform similar analysis for any climate zone.

IV. CONCLUSIONS

In this paper, we analyzed the interdependence of different meteorological variables and assessed their importance for PV output power estimation. A complete 3 years of meteorological data and PV output power data from the city of Utrecht, the Netherlands, were used to establish the relationship between the PV output power and the input meteorological variables. The paper provides a method to reduce the space of input meteorological variables using Principal Component Analysis (PCA). The obtained observations help in selecting and ranking the input meteorological variables for classification and regression tasks in the field of solar energy analytics and forecasting.

Our future work will extend the analysis of the interdependence and importance of meteorological variables to other locations with different climate zones. In addition, we aim to assess and compare the estimation performance of different machine learning based regression models when using the complete space versus the lower dimension subspace of meteorological variables.

V. ACKNOWLEDGEMENTS

This work is partially funded by the Joint Programming Initiative (JPI) Urban Europe project “PARTicipatory platform for sustainable ENergy managemenT (PARENT)” and by eNErgy intrAnetS (NEAT), with support from the Netherlands Science Foundation (NWO).

REFERENCES


