Big Data Analysis for PV Applications

Dazhi Yang[1], Idris Lim[2]

[1]Singapore Institute of Manufacturing Technology, Agency of Science, Technology and Research (A*STAR), Singapore

[2]School of Engineering, University of Glasgow

1. Introduction

With increasing photovoltaic (PV) installations, large amounts of time series data from utility-scale PV systems such as meteorological data and string level measurements are collected. Such lengthy time series data from solar systems are highly multi-dimensional and challenging to process, due to the following.

1. Due to fluctuations in irradiance and temperature, PV data is highly stochastic.
2. Spatio-temporal differences with potential time-lagged correlation are also exhibited, due to the wind directions affecting cloud movements.
3. Different types of PV systems in terms of power output and wiring configuration, as well as localised PV effects like partial shading and module mismatches.

2. Objectives

In this work, a data analytics algorithm is applied to mitigate some of the complexities and make sense of the large time series data in PV systems. The time series data is processed to extract features through clustering and identify correspondence between specific measurements and geographical location of the PV systems. This characterisation of the time series data can be used for several PV applications, namely, (1) PV fault identification, (2) PV network design and (3) PV type pre-design for PV installation in locations with different geographical attributes.

3. Methods

3.1. Principle Component Analysis (PCA)

For a centered dataset X, an $n \times p$ matrix, where n is the number of time series (observation, each time series is considered as one observation) and p is the number of time series features (variable), PCA computes the most meaningful basis to re-express X.

If Z is the re-represented data, the above statement can be written as $Z = XA$, where A is an $p \times p$ matrix and its columns are a set of basis vectors for representing of columns of X. PCA assumes all basis vectors are orthonormal.

A normalized direction is selected in p-dimensional space along which the variance in X is maximized; this basis vector is denoted as a_1. In other words, we maximize $V(a_1^T x)$, where x is vector of p random variables. Since the maximum will not be achieved with finite a_1, a normalization constraint is imposed, namely, $a_1^T a_1 = 1$. The subsequent direction is again selected based on the maximum variance criterion, however, due to the orthonormal assumption, the choice is limited to the directions that are perpendicular to a_1. The procedure continues until p directions are selected. Thus $a_k^T x$ is defined as the k-th sample principal components and $z_i k = a_k^T x_i$ is the score for the i-th observation on the k-th PC.

3.2. Biplot

A biplot represents both the observations and variables of a matrix of multivariate data on the same plot. It uses points to represent the scores of the observations on the principal components, and uses vectors to represent the coefficients of the variables on the principal components (Jolliffe, 2002).

At this stage, several interpretations are summarized:

1. Biplots are scatter plots, the points in a biplot can therefore be interpreted the same way: closer points correspond to observations that have similar scores on the PCs. This interpretation is useful for clustering applications.
2. Projection of points onto a vector gives original values of that variable. By examining points along the particular direction of a vector (and its opposite direction as well), samples with anomalous values on what the variable measures can be identified.
3. Angle between two vectors denotes their correlation. Vectors that point in the same direction correspond to variables that have similar response profiles.
4. The apparent length of a vector gives an idea about the variance of that variable. This can be used to conclude the importance of a variable during clustering.

4. Applications

4.1. PV system type identification (generic time series features)

The first, and arguably the simplest, application is PV system type identification. In particular, we are interested in identifying whether a PV system has a fixed orientation or single-axis tracking; these two types of systems are more utilized than dual-axis tracking systems due to their better cost, reliability and energy production trade-off (Mousazadeh et al., 2009; Nann, 1990).

The data used in this application comes from the western wind and solar integration study (WWSIS) conducted by the National Renewable Energy Laboratory (NREL) to explore the operational impact of high renewable penetration into an electricity grid. The full dataset contains approximately 6000 PV plants of different size in western US locations, however, for demonstration purposes, only data from 405 plants in California are used.

Fig. 1 shows some samples of the normalized time series from the WWSIS dataset (five random samples from each type of systems). It can be seen that the power output from those PV systems with trackers (plotted in the column on the right) has a flatter top, due to the DNI gain by the sun-tracking panels. This effect is most apparent in the late morning and early afternoon hours. Although identifying PV systems types by visually inspecting the time series transient is easy, the method is not scalable when thousands or more series need to be identified. The proposed framework can be useful in this application.

Due to the simplicity of this application, using only generic time series features would suffice. Hyndman et al. (2015) consolidated a total of 15 generic time series features; these features are listed verbatim in Table 1. After normalization, the series should have zero mean and unit variance. Two features, namely, the mean and variance, can thus be dropped. The 13 time series features are computed for each of the 405 normalized PV power time series. After running PCA, the data points are projected onto the two-dimensional feature space as shown in Fig. 2. After the projection, two linearly separable clusters can immediately be seen. At this stage, any sensible 2-dimensional clustering algorithm could be used to identify the system types. For our choice, k-means clustering with 2 centers and 25 random initialization of centers is used.

Table 1: Fifteen non-seasonal time series features used for PV system type identifications. These generic features are adopted from Hyndman et al. (2015).

Feature	Description
Mean	Mean
Var	Variance
ACF1	First order of autocorrelation
Trend	Strength of trend
Linearity	Strength of linearity
Curvature	Strength of curvature
Entropy	Spectral entropy
Lumpiness	Changing variance
Spikiness	Strength of spikiness
Lshift	Level shift using rolling window
Vchange	Variance change
Fspots	Flat spots using discretization
Cpoints	The number of crossing points
KLscore	Kullback–Leibler score
Change.idx	Index of the maximum KL score

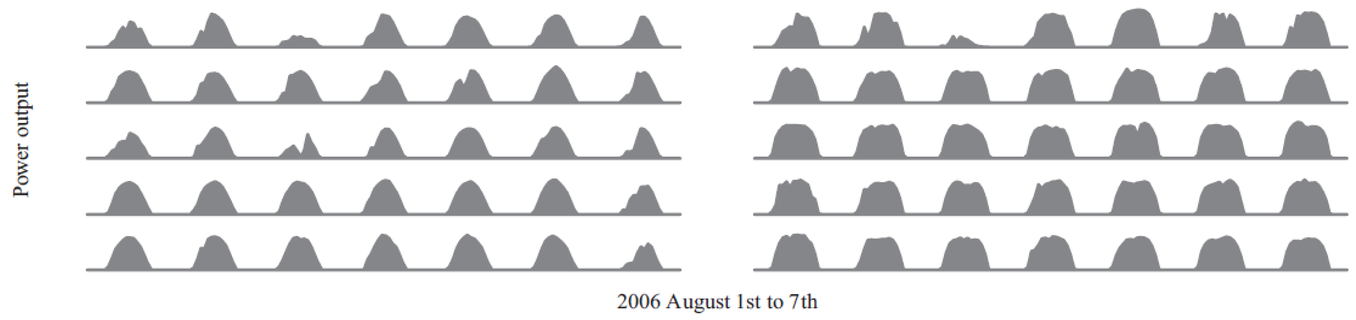

Figure 1: Sample time series from the WWSIS dataset during 2006 August 1–7. Power output from PV systems with trackers (right column) has a flatter top as compared to that from systems with fixed orientations (left column).

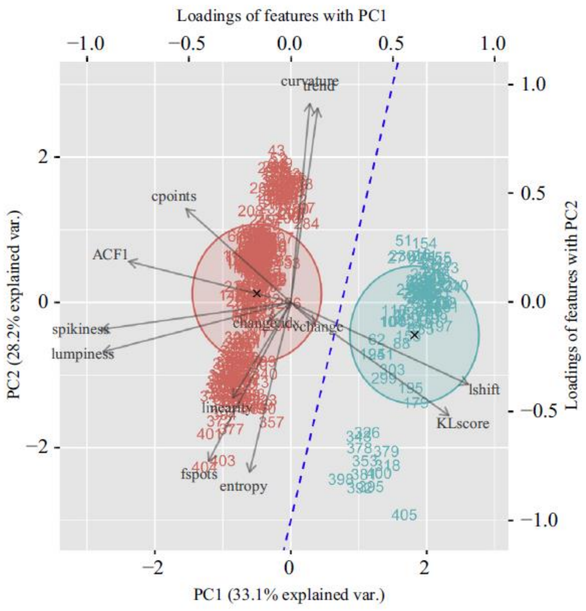

Figure 2: Clustering of PV system types (fixed or tracking) using k-means with principal component analysis. The Indian red cluster shows the PV systems with fixed orientations; the turquoise cluster shows the PV systems with trackers. The cluster centers are indicated with black crosses.

The resulting clusters are displayed in Fig. 2. It can be seen from the biplot that some features, namely, vchange and change.idx, are less variable than others across all time series; they contribute less in terms of separating the systems. The figure also reveals some opposing features, e.g., curvature and linearity; lshift and cpoints. Time series with high linearity in general expects a low curvature. It is worth to mention that the separation of the two clusters is along the direction of lshift, the level shift using rolling window. As this feature computes the maximum absolute difference between consecutive mean values from a rolling window, the tracker systems with more rapid power changes have higher lshift values than those systems with fixed orientations.

4.2. Irradiance monitoring network design (application-specific features)

4.2.1. SUNY data

The State University of New York (SUNY) gridded satellite derived irradiance data is used in this application. The full dataset contains hourly estimates (using the Perez et al., 2002 model) of global, diffuse and direct irradiance over a 10 km (about 0:1 latitude and longitude) grid for all states in the United States, except for Alaska where satellite cannot resolve cloud cover information, for 1998–2005.

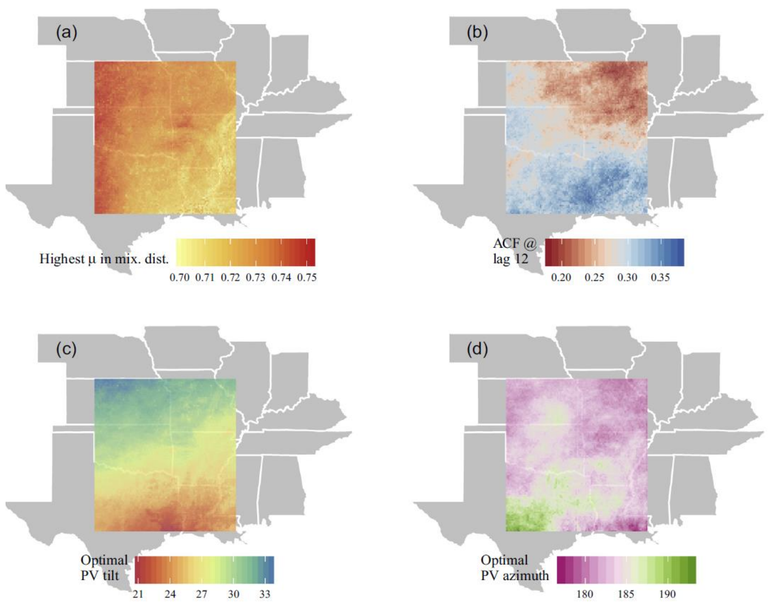

Figure 3: Maps of four PCA features extracted from k_t time series. (a) The highest mean value (location parameter) from the mixture distribution fitting; (b) the autocorrelation at lag 12; (c, d) optimal PV tilt and azimuth angles that maximize the annual yield of a flat surface collector.

Table 2: Six application-specific time series features used for irradiance monitoring network design.

Feature	Description
Lat	Latitude
Lon	Longitude
ACF12	Lag 12 autocorrelation
μ_1	The highest location in the fitted skew-normal mixture distribution
Opt.Tilt	Optimal PV tilt angle that maximizes annual yield
Opt.Azimuth	Optimal PV azimuth angle that maximizes annual yield

4.2.2. Choice of features

We are interested in constructing a proximity so that the similarity in irradiance at different locations can be quantified. The most intuitive features under this consideration are perhaps the geographical locations, namely, the latitude and longitude of each pixel. Besides the geographical locations, the next best approach to generate characteristics is using statistics. Watanabe et al. (2016) used sample mean, variance and entropy to evaluate the variation in solar irradiance.

Generally speaking, a random variable can be well characterized by its distribution function (Kobayashi et al., 2011). There are various ways to describe a distribution function. For instance, we can use quantiles to describe an empirical distribution function, and use parameters to describe a parametric distribution function. In the monitoring network design problem, it is found that some descriptive statistics are more useful than others. Two features, namely, the lag 12 autocorrelation of the clearness index time series and the highest location parameter in the fitted skew normal mixture distribution, are most informative. Besides statistical features, features with physical and engineering implications can also be considered.

For example, it is well known that the optimal orientation of a PV system, in terms of maximizing its annual yield, is subjected to not only the Sun path, but also the intricate geographical and climatic conditions (Smith et al., 2016; Lave et al., 2015; Khoo et al., 2014). Suppose one of the tasks of the designed monitoring network is to help monitor the PV performance in its proximity, features such as the optimal tilt and azimuth angles are useful. These two features, together with the earlier features, are arranged in Table 2; four feature maps are plotted in Fig. 3. It is observed that all four features contain strong spatial structure (see discussion below), which is in favor of the linearity assumption in PCA.

4.2.3. Clustering results

Once the data matrix is prepared, PCA is performed and the corresponding biplot is shown in the left panel of Fig. 4. Unlike the previous application, it is observed that the clusters are not linearly separable in this case. It is therefore necessary to choose a "k" during clustering. As mentioned earlier, Zagouras et al. (2013) computed two indices, namely, the DB index (Davies and Bouldin, 1979) and CH index (Calin´ ski and Harabasz, 1974). As both indices decrease with the number of clusters, the elbow method (Thorndike, 1953) was then used to identify the optimal value of k. Besides these two indices, the elbow method can be applied to many other evaluation metrics, such as the percentage of variance explained (Goutte et al., 1999) and Silhouette index (Rousseeuw, 1987). It is not our immediate interest to advise on the "most appropriate" validation index for irradiance monitoring network design in this work. Instead, a fixed number of clusters, 10, is adopted. This fixed number can be thought of in a practical context as the number of sensors that an installation budget allows. Based on the setting of k = 10, the final cluster map is shown on the right panel of Fig. 4.

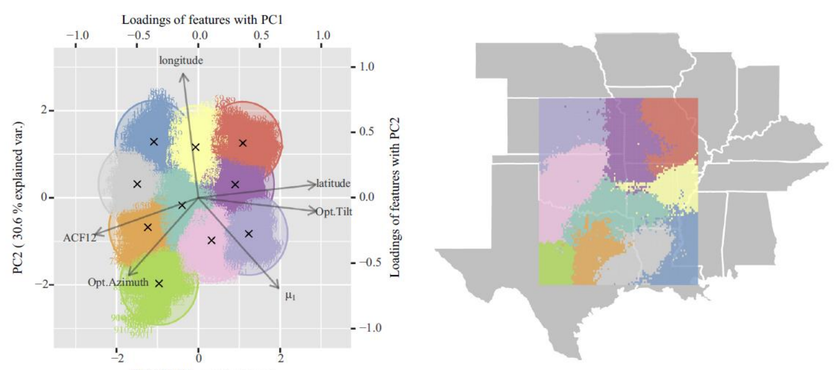

Figure 4: (Left) Biplot of the satellite-derived irradiance data; a total of 6 features are considered during PCA. (Right) The k-means clustering results. Pixels with same colors have similar properties, and thus can be approximated by a single sensor.

The clustering results show a series of important findings. Firstly, latitude and longitude appear to be the major features defining the first and second principal components, respectively. This indicates that the geographical location contributes the most to the overall variation in the time series feature space, i.e., a major deciding factor in our network design setup. Secondly, the small angle between latitude and the optimal PV tilt suggests that the optimal PV tilt depends largely on the site's latitude (the two variables are highly correlated), but not on longitude; this knowledge is known a priori. Another important observation is that the clustering results shown in Fig. 4 agree well with the features maps. For example, the chartreuse cluster along the vector representing optimal PV azimuth can be related to the observations shown in the bottom left corner of Fig. 3(d), namely, the lime green patch; the pink and light steel blue clusters along the μ_1 vector can be linked to the high μ_1 values depicted in the left side of Fig. 3(a). We conclude that the geographical shapes of the final clusters are influenced by the selected features. This linkage between the feature maps and clustering results gives flexibility and practical advantages to our network design approach. The design framework can be applied to any features that are thought appropriate, and the results are readily interpretable.

5. Concluding Remarks

An analytics method for handling big time series data is discussed. In this method, each time series is reduced to a set of features, either generic or application-specific. The reduced feature space facilitates visualization and analyses. Two solar engineering applications are considered to demonstrate the idea. Traditional approaches to these applications consider data points as individuals; the present approach that considers time series as entities is thus novel in terms of data handling. Principal component analysis and biplot are the main tools in all three applications. Biplots make the results of PCA interpretable. By examining the biplots, geometrical relationships among the features and original time series can be established, which leads to insights that are otherwise unobservable using traditional methods. The analytics method is flexible in terms of feature design and can be applied to a variety of other applications. However, common to all dimension reduction strategies, extracting time series features may result in information loss. One should be cautious when replacing the traditional methods with the present method.

6. References

Jolliffe, I., 2002. Principal Component Analysis. Springer Verlag, New York.

Mousazadeh, H., Keyhani, A., Javadi, A., Mobli, H., Abrinia, K., Sharifi, A., 2009. A review of principle and sun-tracking methods for maximizing solar systems output. Renew. Sust. Energy Rev. 13, 1800–1818. http://dx.doi.org/10.1016/jrser.2009.01.022.

Nann, S., 1990. Potentials for tracking photovoltaic systems and V-troughs in moderate climates. Solar Energy 45, 385–393. Hyndman, R.J., Wang, E., Laptev, N., 2015. Large-scale unusual time series detection, in: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp.1616–1619. doi:http://dx.doi.org/10.1109/ICDMW.2015.104.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58, 453–467. http://dx.doi.org/ 10.1093/biomet/58.3.453.

Perez, R., Ineichen, P., Moore, K., Kmiecik, M., Chain, C., George, R., Vignola, F., 2002. A new operational model for satellite-derived irradiances: description and validation. Solar Energy 73, 307–317. http://dx.doi.org/10.1016/S0038-092X (02)00122-6.

Watanabe, T., Takamatsu, T., Nakajima, T.Y., 2016. Evaluation of variation in surface solar irradiance and clustering of observation stations in Japan. J. Appl.Meteorol. Climatol. 55, 2165–2180. http://dx.doi.org/10.1175/JAMC-D-15-0227.1.

Kobayashi, H., Mark, B.L., Turin, W., 2011. Probability, Random Processes, and Statistical Analysis: Applications to Communications, Signal Processing, Queueing Theory and Mathematical Finance. Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511977770.

Smith, C.J., Forster, P.M., Crook, R., 2016. An all-sky radiative transfer method to predict optimal tilt and azimuth angle of a solar collector. Solar Energy 123, 88– 101. http://dx.doi.org/10.1016/j.solener.2015.11.013.

Lave, M., Hayes, W., Pohl, A., Hansen, C.W., 2015. Evaluation of global horizontal irradiance to plane-of-array irradiance models at locations across the United States. IEEE J. Photovolt. 5, 597–606. http://dx.doi.org/10.1109/JPHOTOV. 2015.2392938.

Khoo, Y.S., Nobre, A., Malhotra, R., Yang, D., Rüther, R., Reindl, T., Aberle, A.G., 2014. Optimal orientation and tilt angle for maximizing in-plane solar irradiation for PV applications in Singapore. IEEE J. Photovolt. 4, 647–653. http://dx.doi.org/ 10.1109/JPHOTOV.2013.2292743.

Zagouras, A., Kazantzidis, A., Nikitidou, E., Argiriou, A., 2013. Determination of measuring sites for solar irradiance, based on cluster analysis of satellitederived cloud estimations. Solar Energy 97, 1–11.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1, 224–227. http://dx.doi.org/10.1109/ TPAMI.1979.4766909.

Calin´ ski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. 3, 1–27. http://dx.doi.org/10.1080/03610927408827101.

Thorndike, R.L., 1953. Who Belongs in the Family? Psychometrika 18, 267–276. doi: http://dx.doi.org/10.1007/BF02289263.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F.Å., Hansen, L.K., 1999. On clustering MRI time series. NeuroImage 9, 298–310. http://dx.doi.org/10.1006/nimg.1998. 0391.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7.