Yang and Reindl *Renewables: Wind, Water, and Solar* (2015) 2:1
DOI 10.1186/s40807-014-0001-x

Renewables: Wind, Water, and Solar
a SpringerOpen Journal

ORIGINAL RESEARCH Open Access

# Solar irradiance monitoring network design using the variance quadtree algorithm

Dazhi Yang[1,2*] and Thomas Reindl[1]

**Abstract**

Our aim is to determine the optimal placement of solar irradiance monitoring stations for renewable energy integration into electricity grids. Hourly SUNY satellite-derived irradiance over a rectangular grid of 34° to 44° N, 100° to 110° W with a 0.1° resolution are used in this work. The variance quadtree algorithm is used to identify the regions with high spatio-temporal variations. The densities of monitoring stations over different regions therefore follow the empirical variation. The network design is compared to the results from the so-called "L-method". A discussion based on the network's predictive performance is also presented. We show that the unique design solution obtained using the L-method cannot capture the spatio-temporal variations embedded in irradiance random fields. A robust design should consider both the design requirements and functionalities of the monitoring network.

**Keywords:** Variance quadtree; Kriging; Ground-based monitoring network

## Background

With an increasing penetration of renewable energy into the electricity grid, monitoring becomes more and more important in resource assessment, system design, energy planning, and grid management. A major aim of setting up monitoring networks is to predict values of an attribute of interest, such as solar or wind resources, at unobserved locations using the observed data at known locations (Yang et al. 2013). To sample complex wind or irradiance spatial distributions, for example, ideally one seeks to deploy as many sensors as possible. However, to minimize costs, an optimal number and placement of monitoring equipment is critical.

The simplest monitoring network has a regular grid. The network therefore has only one design parameter, the inter-station spacing. Kriging (a geostatistical interpolation method) is used to determine the optimal spacing (McBratney et al. 1981; van Groenigen et al. 1999). As the optimal interpolator, kriging has been used in solar energy applications, mostly to estimate and plot the insolation maps of an area (McKenney et al. 2008; Moreno et al. 2011; Righini et al. 2005). The core idea of kriging is to estimate the process value $z(s_0)$ at an unknown location $s_0$ based on a linear combination of weighted values at other observed locations:

$$\hat{z}(s_0) - \mu(s_0) = \sum_{i=1}^{n} w_i \left[z(s_i) - \mu(s_i)\right] \tag{1}$$

where function $\mu(s_0)$ denotes the spatial trend at $s_0$, and symbol ˆ denotes the estimated value. The weights $w_i$ can be calculated by minimizing the variance of the estimator: $\mathrm{var}\left[\hat{z}(s_0) - z(s_0)\right]$. In this application, we consider the measurements to be accurate. Rigorously, in hierarchical models, measurement uncertainties must be considered, see (Cressie and Wikle 2011) for details.

The semivariogram, kriging variance, and standard errors provide valuable information about the predictability of the designed network. However, these techniques are often used in purely spatial sampling problems such as designing a network for groundwater monitoring (Yang et al. 2008).

For spatio-temporal data such as solar irradiance random fields, the temporal evolution brings an additional dimension into network design. In other words, at each spatial location, a time series of the attribute of interest can be observed. Observations sharing neighboring spatial locations with similar temporal characteristics should

*Correspondence: yangdazhi.nus@gmail.com
[1] Solar Energy Research Institute of Singapore (SERIS), National University of Singapore, 7 Engineering Drive 1, Block E3A, #06-01, Singapore 117574, Singapore
[2] Department of Electrical and Computer Engineering, National University of Singapore, 4 Engineering Drive 3, Block E4, #05-45, Singapore 117583, Singapore

be grouped together so that a single sensor can represent the group. It is therefore intuitive to use clustering as the design tool. The temporal observations are used as features of the clustering.

### Fundamentals and limitations of the classic k-means algorithm

A common goal of clustering is finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups. The monitoring will thus take place in each cluster. Many approaches have been used in the literature for robust clustering, and the k-means algorithm (MacQueen 1967) can be considered as one of the most foundational methods. The classic k-means algorithm is outlined as follows:

1. Select $k$ points as the initial centroids.
2. *Repeat*: Assign all points to the closest centroid to form $k$ clusters. Recompute the centroid of each cluster.
3. *Until*: The centroids are stable.

Among many strengths of the algorithm, k-means has many known limitations. The algorithm is very sensitive to initial centroids selection and the outliers. It cannot handle the data with non-spherical shaped clusters. Most importantly, it is incapable of clustering data with ‘true clusters’ that are of different sizes and/or densities. If we aim to cluster the geographical areas using the spatio-temporal solar irradiance random fields so that the monitoring stations can be placed within each clustered region; these limitations are applicable with no exception. Two recent studies (Zagouras et al. 2013, 2014) demonstrate applications of k-means algorithm in solar irradiance monitoring network design. The technique (see below) shown in these works requires long-enough datasets, which are not universally available, especially for countries and regions that need monitoring network design. Therefore, a design algorithm that uses minimal data is needed. The k-means-based algorithm presented in (Zagouras et al. 2013) does not consider geographical structure of the data; instead, only the geometrical structure is considered. This may result in a geographically scattered network; it is therefore difficult to discern the clusters (Minasny et al. 2007). In addition, a monitoring network design algorithm which considers the local variations is desired. The density of the monitoring stations should be higher within the regions/areas with higher spatio-temporal variations. Lastly, the predictive performance of the design network is not tested in (Zagouras et al. 2013).

We therefore seek a network design tool with the following attributes: i) small data requirement, ii) considers both geographical and geometrical structures of the data, iii) considers local variation, and iv) good predictive performance. We introduce the variance quadtree algorithm as a tool for solar irradiance monitoring network design.

## Data

The State University of New York (SUNY) gridded satellite-derived data is used in this work. The data is freely available at ftp://ftp.ncdc.noaa.gov/pub/data/nsrdb-solar. The dataset covers hourly estimates of global, diffuse, and direct irradiance over a 10 km (about 0.1° in latitude and longitude) grid for all states in the United States except for Alaska where satellite cannot resolve cloud cover information for 1998 to 2005. The SUNY dataset was created using the model developed by Perez et al. (2002) through Geostationary Operational Environmental Satellites (GOES) imagery. It has been used and verified numerous times in the literature (Vignola et al. 2007), and its accuracy can be considered as sufficient for our study here. As the dataset carries a large amount of information and is in a continental scale, we only select a subset of data, namely, a 10° by 10° square over the state of Colorado and part of its surrounding states, from years 2004 and 2005. Colorado is notable for its diverse geography, thus it is a suitable area for our study. Figure 1 shows the area of study.

There are several points to take note before the dataset can be used. The selected region crosses two different timezones in the US, namely, the Mountain timezone and the Central timezone. As the data is recorded at local time, it is important to make the time synchronizations among all the pixels. To achieve this, a shape (.shp) file of the world timezones is used to identify each pixel for its respective timezone. The shape file is downloadable at http://efele.net/maps/tz/world/. The SUNY data is derived from two satellites, namely, GOES-East and GOES-West. The image capturing of these two satellites take place at 15 mins past an hour and on the hour, respectively. Although the SUNY gridded data is shifted in time (using interpolation) for consistency, we find the hourly readings from two adjacent pixels (within the same timezone but obtained using different satellites) can be very different. To further stabilize the variance in the temporal data, the daily global horizontal insolation (Wh/m$^2$/day) is converted into daily *clearness index*, which is the ratio between the global horizontal insolation and the extraterrestrial insolation (the sum of hourly extraterrestrial irradiance). The hourly extraterrestrial irradiance for each pixel is calculated using the Solar Positioning Algorithm (SPA), a c program developed by the National Renewable Energy Laboratory (2008). The daily clearness index is used as input for the network design algorithms. The total data processing time is around 8 h using a typical personal computer. Through the data processing, a 10,000 (100 × 100 pixels) by 731 (2 years with one of them being a leap

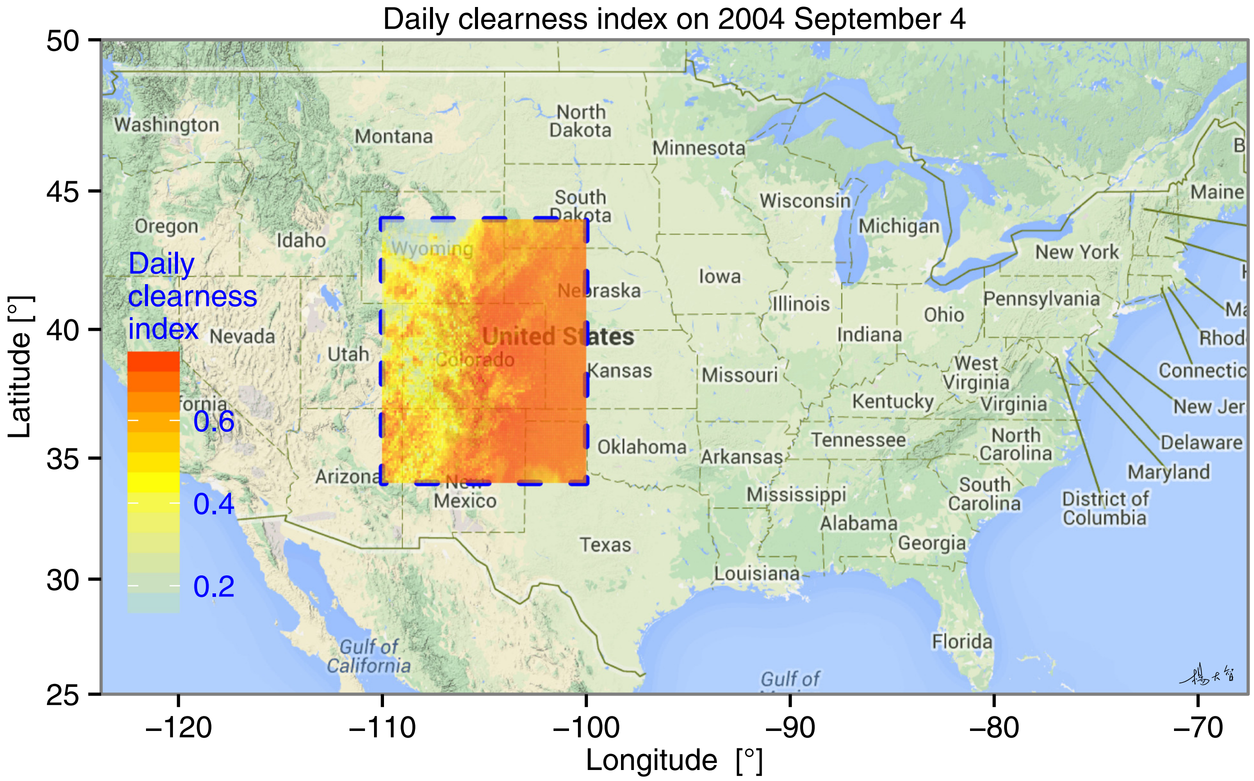

**Figure 1 SUNY gridded satellite-derived data.** Daily clearness index on 4 September 2004 is shown. The area of interest covers a 10° by 10° square over the state of Colorado and part of its surrounding states.

year) matrix of daily clearness index is produced. Figure 1 shows the daily clearness index on 4 September 2004.

## Methods

The variance quadtree algorithm (VQA) is applied on the SUNY dataset described above. VQA was originally designed for purely spatial sampling of the normalized difference vegetation index, an index for the observable live green vegetation over an area (Minasny et al. 2007). We transfer this application into a spatio-temporal framework. In VQA, the spatial or spatio-temporal data is split into many strata with each one having a similar degree of variation within the stratum. This is suitable for spatial or spatio-temporal sampling of solar irradiance as some areas may experience larger variations than the others due to geographical and meteorological reasons.

### Introduction to the VQA

In a purely spatial framework, the VQA is designed as follows:

1. Frame the spatial data in a rectangle.
2. Split the encapsulated rectangle into four equally partitioned strata. For each stratum $h$, a dispersion measure called stratum variance, $Q_h$ is calculated:

$$Q_h = \sqrt{n_h^2 \times \bar{\gamma}(A_h, A_h)} \tag{2}$$

where $A_h$ is the area of the stratum $h$ and $\bar{\gamma}(\cdot)$ is the average semivariance of the stratum. For discrete points $s_i$, with $i = 1, 2, \cdots, n_h$, the average semivariance is calculated by:

$$\begin{aligned}\bar{\gamma}(A_h, A_h) &= \frac{1}{n_h^2}\sum_{i=1}^{n_h}\sum_{j=1}^{n_h}\gamma(s_i - s_j)\\ &= \frac{1}{n_h^2}\sum_{i=1}^{n_h}\sum_{j=1}^{n_h}[z(s_i) - z(s_j)]^2 \end{aligned}\tag{3}$$

where $z(\cdot)$ is the variable of interest, in our case the clearness index.

3. Select the stratum with the largest $Q_h$ value. The selected stratum is further split into four strata. We can therefore say that at iteration $i$, the number of strata is $3i + 1$.
4. Repeat step 3 until the algorithm stops. The stopping criterion can be either when a fixed number of iteration is reached or when the maximum $Q_h$ for all strata is smaller than a threshold, say $\varepsilon$, i.e., $\max(Q_h) < \varepsilon, \forall h$. Other stopping criterion can be used, tuning to the application.

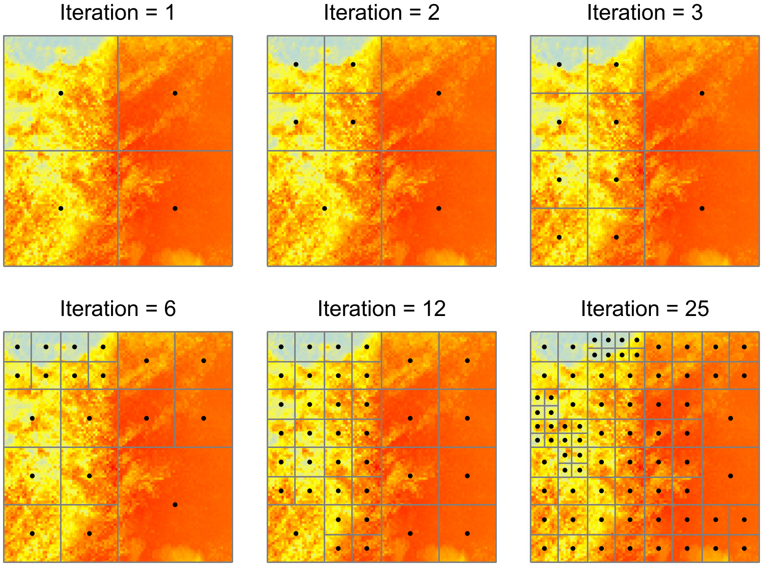

**Figure 2 The variance quadtree algorithm.** Evolution of the variance quadtree algorithm for SUNY Colorado data on 4 September 2004. The background color scheme displays the daily clearness index across Colorado. At iteration *i*, the number of strata is given by $3i + 1$. The black dots are the centers of the strata.

To demonstrate the purely spatial sampling selection using VQA, SUNY satellite-derived irradiance over Colorado on 4 September 2004 is used. Figure 2 shows the evolution of VQA through iterations. Iteration steps 1, 2, 3, 6, 12, and 25 are selected for display. We can visually identify from iteration 1, the upper left stratum has the largest variance; thus, it is selected for iteration 2. At iteration 25, a total of $3 \times 25 + 1 = 76$ strata are defined. It is clear that at iteration 25, the quadtree has already identified some areas with larger spatial variations than the rest.

### VQA for spatio-temporal data

As introduced earlier, we aim to extract the spatio-temporal similarities among the pixels (locations) of the SUNY satellite-derived data. A natural progression of a purely spatial VQA is to add subsequent data into the algorithm. Daily clearness index data from year 2004 are used. The algorithm thus considers 366 temporal images, as year 2004 is a leap year.

Solar insolation received each day may be significantly different from the previous day. To capture the temporal variations at each pixel, the arithmetic mean is used,

$$Q_h = \frac{1}{T}\sum_{j=1}^{T} Q_{hj} \tag{4}$$

i.e., the stratum variance $Q_h$ for each stratum $h$ is the average of the stratum variance $Q_{hj}$ for each time stamp $j$ ranging from 1 to $T$.

An alternative to the summation operation is to use dissimilarity measures. A dissimilarity measure $D_{ij}$ describes the difference between time series collected at $s_i$ and $s_j$. For instance, the spatio-temporal dispersion defined in (Sampson and Guttorp 1992) as:

$$D_{ij} = d_{ij}^2 = \mathrm{var}\left[z(s_i,t) - z(s_j,t)\right], \quad t \in 1 \cdots T \tag{5}$$

can be used. However, given the large dataset, such statistics (using softwares such as MATLAB and R) will contribute significantly to computation time and random access memory usage, due to the subtraction operator prior to the variance calculation. To reduce the computation complexity, correlation is used. Correlation/covariance is defined as similarity measures. Following

$$D_{ij} = \zeta - S_{ij} \tag{6}$$

where $\zeta$ is a constant of choice, the similarity $S_{ij}$ is transformed to dissimilarities. We define the dissimilarity as follows:

$$D_{ij} = 1 - \mathrm{cor}\left[z(s_i,t), z(s_j,t)\right], \quad t \in 1 \cdots T \tag{7}$$

where cor[ ·] denotes correlation and $\zeta = 1$. The stratum variance $Q_h$ for spatio-temporal data is thus expressed as follows:

$$Q_h^{\mathrm{ST}} = \sqrt{\sum_{i=1}^{n_h}\sum_{j=1}^{n_h} D_{ij}} \tag{8}$$

where superscript ST denotes spatio-temporal, and the definition of $D_{ij}$ follows Equation 7. We note that the inclusion of the square root is not necessary for the algorithm.

### Conventional VQA stopping criteria

Two types of stopping criteria are commonly used, namely, the fixed iterations criterion and the maximum variance criterion. In solar monitoring network design, the economical considerations are important. If the financial budget is the primary concern, the fixed iterations criterion should be adopted. For example, if the maximum number of sensors allowed is $n_s$, the VQA should stop after the $\lfloor (n_s - 1)/3 \rfloor$ iterations, where $\lfloor \cdot \rfloor$ is the floor operator.

More often, the designers are concerned with the sampling efficiency of the network. In this case, the maximum variance criterion should be used. Maximum variance criterion means that the algorithm stops when the maximum for all the stratum variance is less than a threshold. To demonstrate this, we use the method described previously. SUNY daily clearness index data from year 2004 are used, i.e., the number of temporal observations at each pixel, $T$, in Equation 7, is 366. We perform the VQA iteratively; at each step, the mean, maximum, and minimum $Q_h^{ST}$ is noted. The results are plotted against the iteration. From Figure 3, it is clear that after 21 iterations, the decrease in maximum $Q_h$ saturates. Thus for Colorado data, 64 stations shall be a reasonable design.

## Results

### L-method and benchmarking

The maximum variance criterion discussed above is used by Minasny et al. (2007). The idea is to identify a particular iteration where the decrease in stratum variance $Q_h$ becomes small in all subsequent iterations. In other words, the decrease in stratum variance is expected to saturate through the iterations. Therefore, we seek to identify the knee of the curve shown in Figure 3. A so-called 'L-method' can be used here. The L-method was originally designed for the detection of anomalies in time series (Salvador and Chan 2005). In this section, we apply the L-method following the network design procedure using the SUNY data.

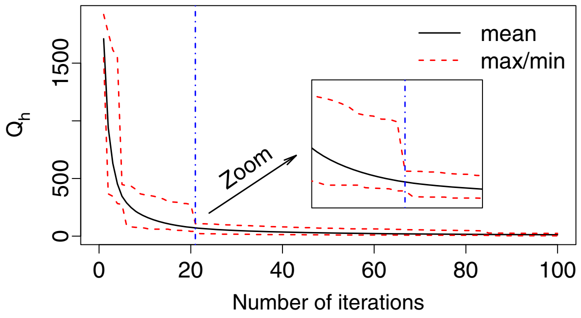

**Figure 3 VQA convergence.** Stratum variance $Q_h$ versus number of iterations of the VQA. Maximum $Q_h$ is stable after 21 iterations (see the blue dash line and the zoomed view).

The L-method can be described using the following equation:

$$\mathrm{RMSE}_c = \frac{c-1}{b-1} \times \mathrm{RMSE}(L_c) + \frac{b-c}{b-1} \times \mathrm{RMSE}(R_c) \quad (9)$$

Suppose the total number of points in a scatter plot is $b$, see Figure 4a, $b = 80$. For every choice of $c$, we can separate the data into two parts, namely, the sequence of points on the left side of $c$, $L_c$, with indices $i = 1, \cdots, c$ and the sequence of points on the right, $R_c$, with indices $i = c+1, \cdots, b$. Two linear regression lines are fitted using the two sets of points, respectively; their fitting root-mean-square errors (RMSE) are denoted by $\mathrm{RMSE}(L_c)$ and $\mathrm{RMSE}(R_c)$. The total fitting error can thus be expressed by Equation 9. Figure 4b shows the $\mathrm{RMSE}_{\mathrm{VQA}}$ for all possible $c$ values. When $c = 7$ iterations, i.e., $3 \times 7 + 1 = 22$ stations give the minimum $\mathrm{RMSE}_{\mathrm{VQA}}$, thus can be considered as the best design following the L-method. The two regression lines at $c = 7$ is shown in Figure 4c. In what follows, we verify the estimation using a former method used by Zagouras et al. (2013).

### Network design using the k-means clustering

In (Zagouras et al. 2013), k-means clustering was used together with principle component analysis (PCA) and the L-method to design the solar irradiance network for Greece. In that work, an instantaneous cloud modification factor (CMF) map over Greece is derived from the daily images collected by the Spinning Enhanced Visible and Infrared Imager (SEVIRI) on the Meteosat Second Generation (MSG) at 10:30 UTC each day. Following the outline of that paper, we apply the techniques using the SUNY dataset.

The data matrix used here has a dimension of 10000 × 731, containing 2 years of daily clearness indices at all pixels. PCA is used to identify the principle components (PCs). We reduce the 731 initial dimensions down to 144 eigenvectors of PCA that preserve a portion of up to 90% of the initial variance. A k-means clustering algorithm is then applied repeatedly to perform the clustering using the reduced PCs. The reason for multiple k-means is to avoid the problem of the initial centroids. Unlike the VQA, the number of clusters using the k-means algorithm needs to be predefined. We evaluate the algorithm using a number of clusters ranging from 5 to 70. Twenty k-means runs are performed at each number of clusters. Two evaluation indices, the Davies-Bouldin (DB) (Davies and Bouldin 1979) index and the Caliński-Harabasz (CH) (Caliński and Harabasz 1974) index are used for clustering validation. Figure 4d,e,f shows the evaluation graphs using the DB index and Figure 4g,h,i

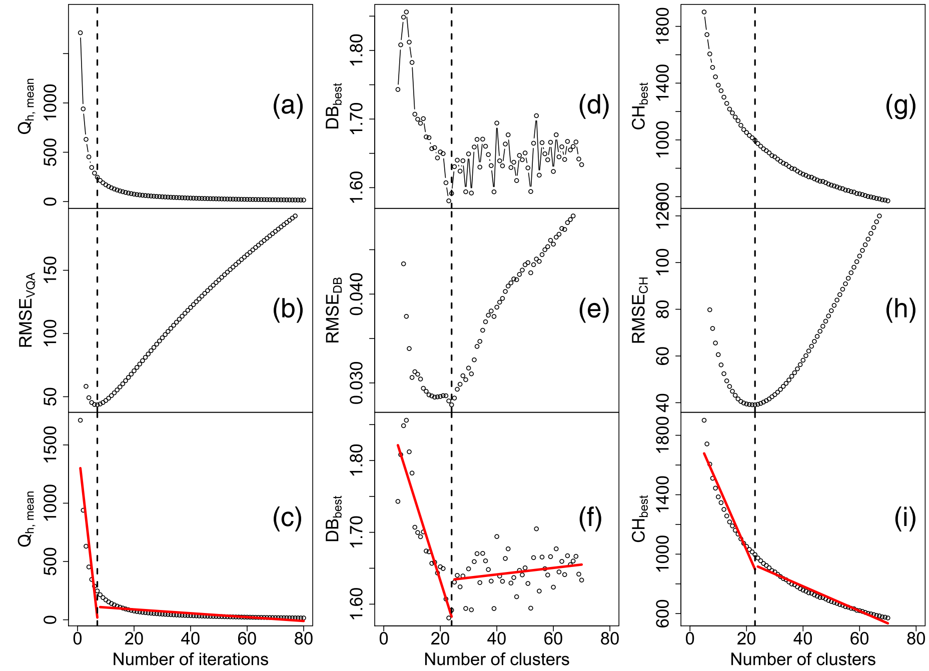

**Figure 4 L–method benchmarking.** Evaluation graphs for determining the optimal number of clusters using the stratum variance, Davies-Bouldin (DB) index, the Caliński-Harabasz (CH) index, and the L-method. **(a)** shows the mean stratum variance $Q_h$ at each iteration. **(b)** shows the overall root-mean-square error using the L-method for each possible separation. **(c)** shows the best fit partition of 7 iterations (i.e., 22 stations). **(d)** shows the best DB index for different number of clusters over a total of 20 repeats of the k-means clustering. **(e)** and **(f)** show $RMSE_{DB}$ and the best fit L-method partition given by the DB index, respectively. Twenty-four clusters give the smallest $RMSE_{DB}$. Subplots **(g-i)** show the respective plots for the CH index. Twenty-three clusters give the smallest $RMSE_{CH}$.

shows the graphs for using the CH index. The results give 24 and 23 clusters as the optimal choice, respectively. The estimations on the number of monitoring stations using the earlier methods agree with our estimation using the VQA.

The solar irradiance monitoring network designs shown above identified final networks of approximately 23 stations. Considering the similarities between the SUNY dataset and the SEVIRI dataset, our result agrees with the earlier estimates of using 22 stations for irradiance monitoring in Greece (Zagouras et al. 2013). However, the internal validation indices such as the DB index and CH index only measure the goodness of a clustering structure without respect to external information. It is almost obvious in these applications that 20+ stations are far too sparse to sample the highly-variable irradiance spatio-temporal random fields in the US and/or Greece.

Perez et al. (2012) simulated that for 15-min along-wind irradiance measurements, the de-correlation distance is around 10 km at a mid-latitude site. A de-correlation distance is the distance which the irradiance measurements at two locations are first becoming uncorrelated. In a later work, Lonij et al. (2013) verified the de-correlation distance using actual power output data from 80 rooftop PV systems over a 50 by 50 km area in Tucson, Arizona. In general, if correlations in all directions (instead of considering only the along-wind direction) are considered, de-correlation distance is usually not observed (Murata et al. 2009); the distance can then be referred to as the threshold distance (Yang et al. 2014). The estimated threshold distance in Singapore is about 10 km. In every case, the inter-station distances of the designed monitoring networks are much larger than both the de-correlation distance and the threshold distance. In other words, network design using the L-method alone does not warrant good spatio-temporal predictability.

### Predictive performance validation

A monitoring network should have good predictability at the unobserved locations. Kriging and other spatial interpolation techniques are suitable tools in assessing the spatial predictability of a network. In this section, SUNY data from the year 2005 is used to assess the predictive performance of the designed networks. Therefore, all predictions are true out-of-sample predictions. Three interpolation methods are used, namely, Thiessen polygon interpolation, inverse distance weighted interpolation, and simple kriging.

Let $z(s_j)$ denote the spatial process observed at point $s_j$, where $j = 1, 2, \cdots, n$ are $n$ observation points or monitoring stations; the general setup of spatial interpolation is as follows:

$$z(s_0) = \sum_{j=1}^{n} w_j z(s_j) \tag{10}$$

where $w_j$ is the weight of sampling point $s_j$. For simplicity, we write $z(s_0)$ as $z_0$ and $z(s_j)$ as $z_j$ hereafter.

### Thiessen polygon interpolation

Thiessen polygon (TP) interpolation is also called the nearest neighbor method; it assumes that the attribute of interest at an unobserved location is equal to the measurements from its nearest observation points. Suppose location $s_0$ has a nearest neighbor $s_i$ where the observations are made, the interpolation weights are as follows:

$$w_j = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases} \tag{11}$$

### Inverse distance weighted interpolation

Another commonly used interpolation method is the inverse distance weighted (IDW) interpolation. IDW assumes the interpolation weights follow

$$w_j = \frac{f(d_{0j})}{\sum_{j=1}^{n} f(d_{0j})} \tag{12}$$

where $f(d_{0j})$ is a general function of $d_{0j}$, the distance between points $s_0$ and $s_j$. A commonly used $f(\cdot)$ is

$$f(d_{0j}) = d_{0j}^{-\beta}, \quad \beta > 0 \tag{13}$$

where $\beta$ is a constant of choice. Here we choose $\beta = 2$ for example.

### Simple kriging

Simple kriging (SK), ordinary kriging, universal kriging, and their variants are perhaps the most commonly used geostatistical interpolation methods. We use only simple kriging in this work as an example. Simple kriging aims to minimize the variance of interpolation error $z_0 - \hat{z}_0$,

$$\sigma_e^2 = \text{var}[\, z_0 - \hat{z}_0] = \text{var}\left[ z_0 - \sum_{j=1}^{n} w_j z_j \right] \tag{14}$$

where $\hat{z}_0$ denotes the estimates of $z_0$. By expanding the above, we have

$$\sigma_e^2 = \sigma^2 - 2\sum_{j=1}^{n} w_j \text{cov}\left(z_0, z_j\right) + \sum_{j=1}^{n}\sum_{i=1}^{n} w_i w_j \text{cov}\left(z_i, z_j\right) \tag{15}$$

where $\sigma^2$ is the variance of $z_0$ and $\text{cov}(\cdot)$ represents the covariance. By setting the first-order derivative (w.r.t. $w_j$) of the above expression to zero, we have

$$\sum_{i=1}^{n} w_i \text{cov}(z_i, z_j) = \text{cov}(z_0, z_j), \quad j = 1, 2, \cdots, n \tag{16}$$

If the homogeneity assumption can be satisfied, Equation 16 can be written as

$$\sum_{i=1}^{n} w_i \text{cor}\left(z_i, z_j\right) = \text{cor}(z_0, z_j), \quad j = 1, 2, \cdots, n \tag{17}$$

where $\text{cor}(\cdot)$ denotes correlation. Homogeneity means that the standard deviation $\sigma_i$ and $\sigma_j$ are equal for all $i$ and $j$. One step further could be taken by assuming isotropy in the spatial process, so that the correlation can be written as a function of distance only, i.e.,

$$\sum_{i=1}^{n} w_i \rho\left(d_{ij}\right) = \rho\left(d_{0j}\right), \quad j = 1, 2, \cdots, n \tag{18}$$

$\rho(\cdot)$ is a correlation function. The interpolation weights can be obtained by solving this linear system of equations. For our implementation, an exponential correlation function with a nugget effect

$$\rho(d) = (1 - \nu)\exp(-c \cdot d) + \nu \mathbf{I}_{d=0} \tag{19}$$

is used.

### Validation results

All three selected interpolation methods are used to assess the predictive performance iteratively. At each iteration $i$, VQA will output a particular design with $3 \times i + 1$ stratum centers. Daily clearness indices from the year 2005 at these centers are used to interpolate the clearness indices at all other locations. For example, for the seventh iteration, 22 centers are produced by VQA, each interpolation method will thus generate $N = (10000 - 22) \times 365$ predictions.

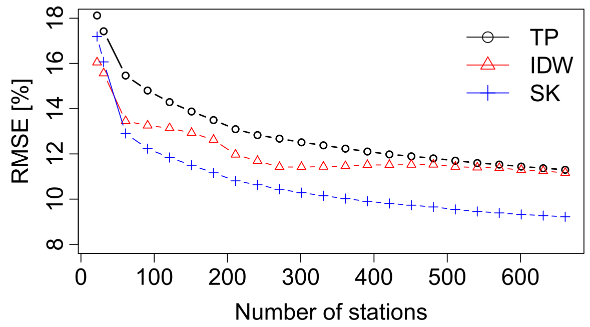

**Figure 5 VQA validation.** Interpolation validation root-mean-square errors as functions of number of stations. All three interpolation methods, namely, Thiessen polygon (TP), inverse distance weighted (IDW) interpolation, and simple kriging (SK), are used for number of stations up to 661 (220 iterations).

After the predictions of clearness index are made, these predictions are adjusted back to daily insolation for error calculation. The percentage RMSE

$$\text{RMSE} = \frac{\sqrt{\frac{1}{N}\sum\left(\hat{G}-G\right)^2}}{\frac{1}{N}\sum G} \times 100\% \tag{20}$$

is used on daily insolation $G$. Figure 5 shows the RMSE as a function of the number of stations. It is evident from the plot that earlier estimate of approximately 23 stations will result in large errors when we use the designed network for prediction.

Following the above discussions, it should be clear now that the design of an irradiance monitoring network can be subjective. Various termination criteria for VQA will lead to different designs. The trade-off between the number of stations and the network's predictive performance needs to be considered. We do not recommend any 'optimal' setting, instead, the object-oriented design should be promoted.

## Conclusions

Spatio-temporal variance quadtree algorithm is proposed for solar irradiance monitoring network design. As the algorithm itself is elegant and flexible, the termination criterion becomes the focus. If we monitor the change in stratum variance, 64-station setup provides a stable stratum variance. When the L-method is used, the optimal setup only consists of 23 stations. The small network obtained through the L-method introduces large spatial prediction errors. Therefore, both internal and external validations are required for network design; a properly designed network should consider its predictive performance.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

TR revised the manuscript critically and proposed the inclusion of the L-method as a benchmarking model. The remaining work is done by DY. Both authors read and approved the final manuscript.

Received: 9 July 2014 Accepted: 11 September 2014
Published online: 20 January 2015

### References

Caliński, T, & Harabasz, J (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27.

Cressie, N, & Wikle, CK (2011). *Statistics for spatio-temporal data*. Hoboken, New Jersey: John Wiley & Sons.

Davies, DL, & Bouldin, DW (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*(2), 224–227.

Lonij, VPA, Brooks, AE, Cronin, AD, Leuthold, M, Koch, K (2013). Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors. *Solar Energy*, *97*(0), 58–66.

MacQueen, JB (1967). Some methods for classification and analysis of multivariate observations. In Cam, LML, Neyman, J (Ed.), *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1* (pp. 281–297).

McBratney, AB, Webster, R, Burgess, TM (1981). The design of optimal sampling schemes for local estimation and mapping of of regionalized variables–i: theory and method. *Computers & Geosciences*, *7*(4), 331–334.

McKenney, DW, Pelland, S, Poissant, Y, Morris, R, Hutchinson, M, Papadopol, P, Lawrence, K, Campbell, K (2008). Spatial insolation models for photovoltaic energy in Canada. *Solar Energy*, *82*(11), 1049–1061.

Minasny, B, McBratney, AB, Walvoort, DJJ (2007). The variance quadtree algorithm: use for spatial sampling design. *Computers & Geosciences*, *33*(3), 383–392.

Moreno, A, Gilabert, MA, Martinez, B (2011). Mapping daily global solar irradiation over Spain: a comparative study of selected approaches. *Solar Energy*, *85*(9), 2072–2084.

Murata, A, Yamaguchi, H, Otani, K (2009). A method of estimating the output fluctuation of many photovoltaic power generation systems dispersed in a wide area. *Electrical Engineering in Japan*, *166*(4), 9–19.

National Renewable Energy Laboratory (2008). Solar Position Algorithm. <http://rredc.nrel.gov/solar/codesandalgorithms/spa/>. Accessed 02.01.2012.

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*(5), 307–317.

Perez, R, Kivalov, S, Schlemmer, J, Jr, KH, Hoff, TE (2012). Short-term irradiance variability: preliminary estimation of station pair correlation as a function of distance. *Solar Energy*, *86*(8), 2170–2176.

Righini, R, Gallegos, HG, Raichijk, C (2005). Approach to drawing new global solar irradiation contour maps for argentina. *Renewable Energy*, *30*(8), 1241–1255.

Sampson, PD, & Guttorp, P (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, *87*(417), 108–119.

Salvador, S, & Chan, P (2005). Learning states and rules for detecting anomalies in time series. *Applied Intelligence*, *23*(3), 241–255.

van Groenigen, JW, Siderius, W, Stein, A (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, *87*(3–4), 239–259.

Vignola, F, Harlan, P, Perez, R, Kmiecik, M (2007). Analysis of satellite derived beam and global solar radiation data. *Solar Energy*, *81*(6), 768–772.

Yang, F, Cao, S, Liu, X, Yang, K (2008). Design of groundwater level monitoring network with ordinary kriging. *Journal of Hydrodynamics, Ser. B*, *20*(3), 339–346.

Yang, D, Gu, C, Dong, Z, Jirutitijaroen, P, Chen, N, Walsh, WM (2013). Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging. *Renewable Energy*, *60*(0), 235–245.

Yang, D, Dong, Z, Reindl, T, Jirutitijaroen, P, Walsh, WM (2014). Solar irradiance forecasting using spatio-temporal empirical kriging and vector autoregressive models with parameter shrinkage. *Solar Energy*, *103*(0), 550–562.

Zagouras, A, Kazantzidis, A, Nikitidou, E, Argiriou, AA (2013). Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations. *Solar Energy*, *97*(0), 1–11.

Zagouras, A, Inman, RH, Coimbra, CFM (2014). On the determination of coherent solar microclimates for utility planning and operations. *Solar Energy*, *102*(0), 173–188.