Spatial Data Dimension Reduction Using Quadtree: A Case Study on Satellite-Derived Solar Radiation

Dazhi Yang*, Gary S. W. Goh, Siwei Jiang and Allan N. Zhang
Singapore Institute of Manufacturing Technology (SIMTech)
*Agency for Science, Technology and Research (A*STAR)*
Singapore, Singapore
*Email: *yangdz@simtech.a-star.edu.sg, yangdazhi.nus@gmail.com*

***Abstract*—Satellite data is discrete in both space and time; it can be considered as temporal snapshots (time series) of lattice processes. As the raw datasets are often too large to host publicly, processed datasets with a coarse spatial resolution are often hosted as an alternative. Nevertheless, with a regular grid, the inhomogeneous variability in the lattice processes cannot be captured effectively. In this paper, a quadtree-based spatial data dimension reduction algorithm is demonstrated. Based on the stratum variance, this algorithm iteratively divides lattice data into strata of fours. In this way, the number of strata in an area can be correlated to the variability of that area. A satellite-derived surface solar radiation (SSR) dataset is used for the case study. Using parallel computing, the quadtree algorithm is applied on each temporal snapshot of SSR in the dataset. The processed data is then saved in a list structure. Finally, a solar resource assessment application, namely, optimizing the orientation of a photovoltaic array, is considered to demonstrate the effectiveness and efficiency of the dimension-reduced dataset.**

***Keywords*-variance quadtree; spatial data; dimension reduction; solar radiation**

I. Introduction

Satellite images, as well as their derivative products, are rich spatio-temporal data that are valuable to a variety of domains including, but not limited to, atmospheric sciences, oceanography and renewable energy engineering. The characteristics of satellite data align precisely with the well-known five Vs and the definition of *Big Data* given in Ref. [1]. Certainly, with today's data storage technology, maintaining a database of raw data is possible for government organizations such as NASA. However, the raw data are often too large to host publicly, especially when a community only requires certain parameters derived from the raw data. In such circumstances, appropriately processed data are preferred.

On this point, we consider surface solar radiation (SSR), a quantity that can be derived from weather satellite images. Solar resource assessment is a critical step in designing and investing a photovoltaic (PV) system. Like many other environmental quantities, SSR varies with geographical location and time. Therefore, during the design of an SSR database, one must consider two conflicting objectives: (1) minimizing the size of the database and (2) capturing the spatio-temporal variations in SSR.

In our present context, temporal variation refers to the variability in a solar radiation time series. In some circumstances, if a time series possesses seasonal cycles, its length could be reduced by only storing data points from a few cycles, or data points that can typify the variation (see Ref. [2] for the concept of typical meteorological year). However, in the field of solar energy engineering, although the annual and diurnal cycles can be somewhat modeled with a double-seasonal smoothing, it is generally recommended to consider a long enough (such as 20 years) dataset during resource assessment [3]. For such reasons, we do not consider temporal data compression here; this work focuses on the spatial variability.

A simple database would record spatial data on a regular grid. The only design parameter in this case would thus be the grid's spatial resolution. However, such design may not be efficient. Consider an area under an arid climate condition, the weather conditions in the neighboring satellite image pixels are similar; a single set of temporal data will most likely suffice for resource assessment of that area. On the other hand, for areas such as the state of Louisiana, which has a humid subtropical climate with unpredictable clouds, it is reasonable to include more spatial points (each associated with a time series) over those areas in the database. To that end, we demonstrate a dimension reduction technique for satellite-derived SSR data using the variance quadtree (VQT) algorithm. A database of SSR is constructed over an irregular grid stratified based on the spatial variability. An application of optimizing the PV installation orientation is used to demonstrate the effectiveness and efficiency of the designed database.

A. Background of the Application

Conventional flat-surface solar collectors, such as PV, are often fixed at a particular orientation (facing the Equator and with a tilt equal to the site's latitude) to maximize their

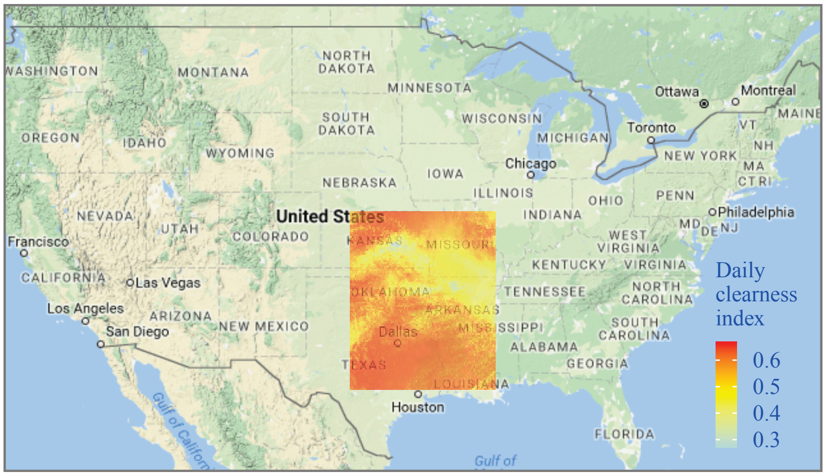

Figure 1. Daily clearness index on 2004 July 4 is shown. The region of interest covers a 10° by 10° square over some states in Southern US.

energy output [4]. However, it is often found that, due to the location-specific geographical and climatic conditions, such conventional orientation may not be optimal; simulations are used to further optimize the orientation and thus maximize the PV energy yield [4]–[6]. To obtain the optimal pair of tilt and azimuth angles, a set of solar radiation measurements at a horizontal surface is required. Using the so-called "transposition models" [7]–[9], the horizontal solar radiation can be projected to an arbitrary tilted plane. Let α and β be the azimuth angle and tilt angle respectively, the optimization problem can be written as:

$$\underset{\alpha,\beta}{\operatorname{argmax}} \sum_{t=1}^{T} \widehat{G}_c(t), \tag{1}$$

where $\widehat{G}_c(t)$ is the modeled tilted solar radiation over a time interval t (e.g., day, hour, minute), which is a function of α, β, $G_h(t)$ and $I(t)$; G_h is the global horizontal irradiance (GHI); I is direct normal irradiance (DNI); and $t = 1, \cdots, T$ denotes the time index of a long enough horizontal solar radiation time series. The tilted solar radiation is also a function of the so-called "diffuse horizontal irradiance" (DHI), denoted by D_h. As DHI can be deterministically calculated if GHI and DNI are known, we do not include it in (1).

There are two main ways to measure the above mentioned irradiance components, namely, using ground-based instruments and using satellite-derived data. Ground measurements are accurate and have high temporal resolution. However, setting up ground sensors (pyranometer for GHI and pyrheliometer for DNI) everywhere is costly and thus not practical for constructing a worldwide radiation database. Satellite-derived solar radiation thus becomes the major tool for solar resource assessment applications. As satellite-derived radiation data are often biased, site adaptation[1] is required. However, site adaptation is not within the scope of this paper; we refer interested readers to Ref. [10] for a detailed review.

B. The SUNY Data

The State University of New York (SUNY) gridded satellite-derived database, or SUNY database in short, is developed by Richard Perez [11]. The model used to derived the data is one of the most widely used operational satellite-based radiation models. The basic principle of this model is monitoring the dynamic range of the satellite image pixels and assigning radiation values based on the brightness of the pixels. The SUNY database contains hourly estimates of GHI, DHI and DNI data over a 10 km grid (about $0.1°$ in latitude and longitude) for all states in the United States except for Alaska, where satellite cannot resolve cloud information. The database covers a period of 12 years from 1998 to 2009. The original database is over 60 GB and can be obtained freely available at ftp://ftp.ncdc.noaa.gov/pub/data/nsrdb-solar.

As the goal of this paper is to demonstrate the VQT algorithm, we only consider a spatio-temporal subset of the data, from the year 2004, covering an area (100×100 pixels) in Southern United States (see Figure 1). In principle, the VQT can be applied to each hourly snapshot of SSR. However, the difference in sunrise time and the bell-shaped diurnal

[1]The term "site adaptation" is used to refer to the improvement that can be achieved in satellite-derived radiation data when ground measurements are used to correct the bias in the original dataset [10].

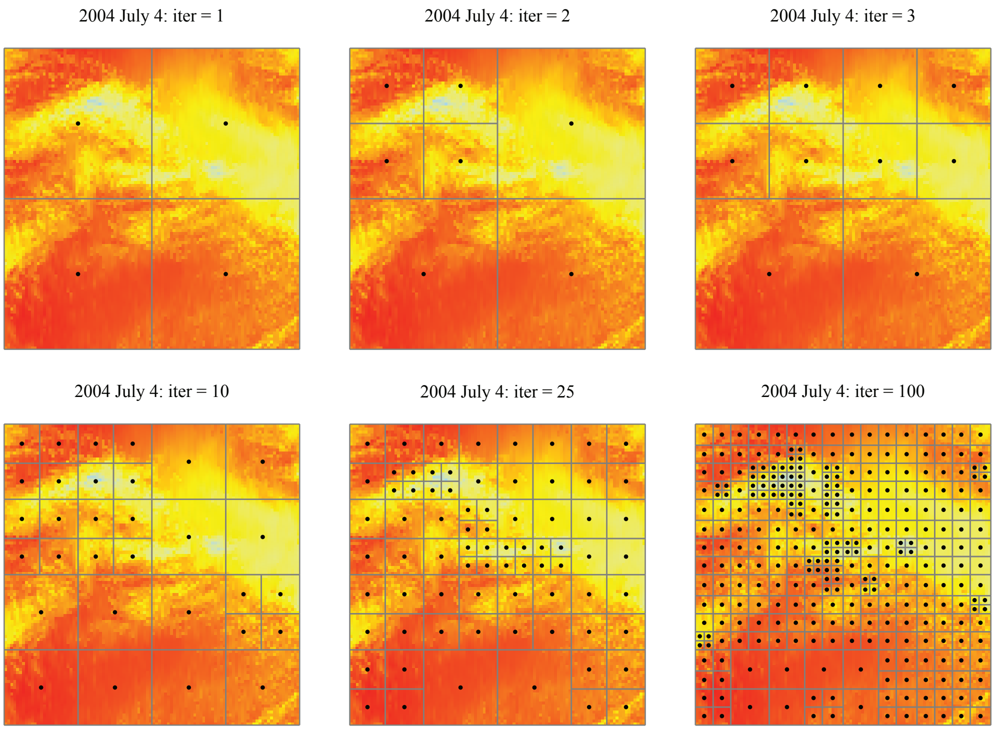

Figure 2. An illustration of the variance quadtree algorithm. The background heat map shows the daily clearness index over the area of interest on 2004 July 4. The strata generated by VQT at iterations 1, 2, 3, 10, 25 and 100 are shown.

trend in SSR affect the stratum variance calculation and thus the variability. On that note, VQT is applied on daily *clearness index*, a normalized measure of SSR (the ratio between the actual radiation and the radiation just outside of the Earth's atmosphere). The value of the clearness index ranges from 0 to 1. In this way, the strata are divided based only on the "intrinsic" spatial variability of SSR. Nevertheless, hourly SSR data will be used to optimize the PV orientation in our application (see Section III). Figure 1 shows the daily clearness index on 2004 July 4 over the area of interest. It can be seen that some areas, such as the bottom-left corner, are less variable than other areas.

II. Variance Quadtree

The variance quadtree algorithm was designed for spatial sampling problem; its earliest application was to sample the normalized difference vegetation index [12]. The algorithm is described as follows:

1) Encapsulate the spatial data in a rectangular box.
2) Split the framed area into four equally partitioned strata. For each stratum h, a variability 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; n_h is the total number of pixels in the statum; and $\bar{\gamma}(\cdot)$ is the *average semivariance* of that stratum. For discrete points s_i, where $i = 1, 2, \cdots, n_h$, $\bar{\gamma}(\cdot)$ is calculated by:

$$\bar{\gamma}(A_h, A_h) = \frac{1}{n_h^2} \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} \gamma(s_i - s_j) \tag{3}$$

where $\gamma(s_i - s_j) = [z(s_i) - z(s_j)]^2$ and $z(\cdot)$ is the parameter of interest.
3) Based on all the Q_h values available at the current iteration, the stratum with the largest Q_h is further split into four smaller strata with equal size.
4) Repeat step 3 until the algorithm satisfies some stopping criteria.

The iteration procedure of VQT is illustrated in Figure 2. The clearness index data on 2004 July 4 is plotted as a heat map. The strata produced by VQT at iterations 1, 2, 3, 10, 25 and 100 are shown. After the first iteration, the top-left stratum is found to have the highest variability; it is thus split into four new strata. After the algorithm is stopped at the 100th iteration, we can see that the bottom-left strata are the biggest over the area of interest. This agrees with our earlier observation, namely, the bottom-left corner has small variability.

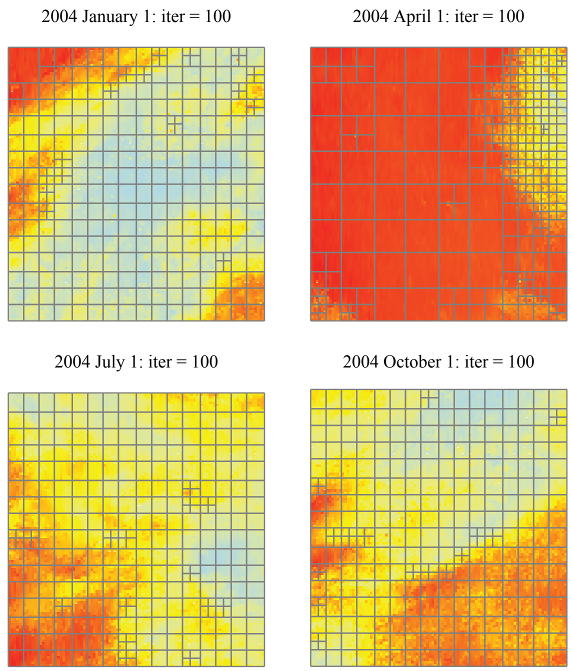

Figure 3. More examples of strata generated by the VQT algorithm.

Using the VQT algorithm, we can generate the stratified data for each temporal snapshot of the clearness index process, i.e., we repeat the procedure shown in Figure 2 366 times (2004 is a leap year). For each snapshot, the algorithm stops after 100 iterations. It is noted that other stopping criteria can be employed, however, we leave that to future discussions. Figure 3 provides addition examples of the VQT. The data from the first day of each quarter of the year are shown. It can be seen that the VQT is dynamic. This distinguishes our present application of VQT from the previous ones, where VQT was used for spatial sampling design tasks [12]–[14].

The strata generated by the VQT algorithm can be saved as a list object. Each item in the list comprises information of a particular temporal snapshot. It should be pointed out that at iteration i, the number of strata is $3i + 1$. Each stratum is bounded by 4 corners which can be represented by 4 floating-point numbers (2 longitudes and 2 latitudes). Together with the 24 hourly GHI values and 24 DNI values averaged over each stratum, the total number of floating-point numbers to save an ith iteration VQT is thus $52\times(3i+1)$. When i is smaller than a few hundreds, which would be sufficient for most applications, the VQT representation of the spatial data requires much less storage space than the raw data. In our case study here, the raw data has a size of 140 MB, whereas the dimension reduced dataset only takes 5.5 MB.

III. Application

In this section, the application of optimizing the PV orientation is tested on both the dimension-reduced and raw datasets. To determine the optimal PV orientation (tilt and azimuth angle) at a location, we need to solve the maximization problem described in (1). Recall that $G_c(t)$ can be modeled as a function of α, β, $G_h(t)$ and $I(t)$, and there are many choices of this function. For instance, if the classic isotropic transposition model [15] is employed, $G_c(t)$ can be expressed as:

$$G_c(t) = I(t)\cos\theta(t) + \frac{1+\cos\beta}{2}D_h(t) + \frac{0.2(1-\cos\beta)}{2}G_h(t), \tag{4}$$

where $\theta(t)$ denotes the average solar incidence angle over a time interval t, which can be calculated if α and β are known; and $D_h(t) = G_h(t) - I(t)\cos Z(t)$, where $Z(t)$ is the average zenith angle over t. For a general description on the modeling of $G_c(t)$, we refer the readers to Ref. [7].

The maximization problem is solved using the general-purpose optimization routine (function `optim`) in R [16]. For every pixel of the selected area, the maximization is performed twice using G_h and D_h values before and after VQT, respectively. Parallel computing is used to speed up the process. The optimal α and β values are recorded as the maps shown in Figure 4. From a visual comparison, we can conclude that the optimal orientation maps produced from the dimension-reduced dataset agree well with the ones produced from the raw data. In terms of percentage root-mean-square error (RMSE), the RMSEs for α and β are 0.54% and 0.95%, respectively. These errors are negligible considering the data uncertainties.

IV. Conclusions and Future Works

A quadtree-based spatial data dimension reduction algorithm is applied on a satellite-derived surface solar radiation dataset. The VQT algorithm iteratively stratifies the space into strata of fours. By using the VQT algorithm, the original 140 MB of data is reduced to 5.5 MB; the dimension-reduced dataset can perform solar resource assessment tasks, or more specifically, optimize PV orientation, with negligible errors.

In the current version of the algorithm, a fixed-step stopping criterion (after 100 iterations) is used. Other stopping criteria should be considered in the future. Furthermore, more elaborate transposition modeling can be employed to improve the accuracies of solar resource assessment.

To promote the uptake of our results, the R code and datasets used in this paper are provided and can be obtained by contacting the corresponding author.

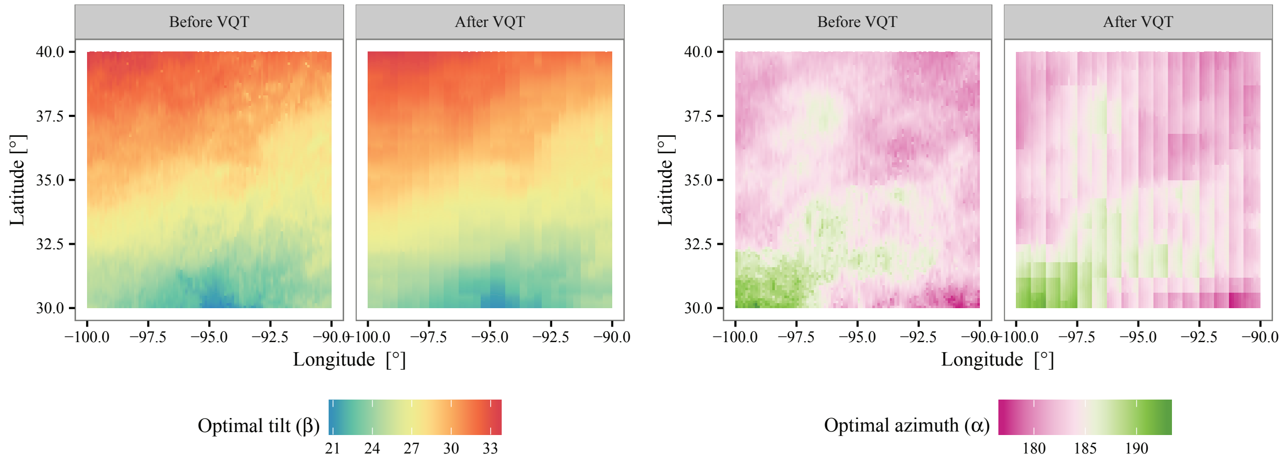

Figure 4. Comparison of optimal PV orientation maps before and after applying the VQT algorithm for data compression.

Acknowledgment

This work is partially supported under the A*STAR TSRP fund 1424200021 and Antuit–SIMTech Supply Chain Analytics Lab.

References

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[2] S. Wilcox and W. Marion, "Users manual for TMY3 data sets," National Renewable Energy Laboratory, Golden, Colorado, Tech. Rep. NREL/TP-581-43156, May 2008. [Online]. Available: http://www.nrel.gov/docs/fy08osti/43156.pdf

[3] C. A. Gueymard and S. M. Wilcox, "Assessment of spatial and temporal variability in the US solar resource from radiometric measurements and predictions from models using ground-based or satellite data," *Solar Energy*, vol. 85, no. 5, pp. 1068–1084, 2011.

[4] Y. S. Khoo, A. Nobre, R. Malhotra, D. Yang, R. Ruther, T. Reindl, and A. G. Aberle, "Optimal orientation and tilt angle for maximizing in-plane solar irradiation for PV applications in Singapore," *IEEE Journal of Photovoltaics*, vol. 4, no. 2, pp. 647–653, March 2014.

[5] C. J. Smith, P. M. Forster, and R. Crook, "An all-sky radiative transfer method to predict optimal tilt and azimuth angle of a solar collector," *Solar Energy*, vol. 123, pp. 88–101, 2016.

[6] M. Lave and J. Kleissl, "Optimum fixed orientations and benefits of tracking for capturing solar radiation in the continental United States," *Renewable Energy*, vol. 36, no. 3, pp. 1145–1152, 2011.

[7] D. Yang, "Solar radiation on inclined surfaces: Corrections and benchmarks," *Solar Energy*, vol. 136, pp. 288–302, 2016.

[8] D. Yang, Z. Ye, A. M. Nobre, H. Du, W. M. Walsh, L. I. Lim, and T. Reindl, "Bidirectional irradiance transposition based on the perez model," *Solar Energy*, vol. 110, pp. 768–780, 2014.

[9] D. Yang, Z. Dong, A. Nobre, Y. S. Khoo, P. Jirutitijaroen, and W. M. Walsh, "Evaluation of transposition and decomposition models for converting global solar irradiance from tilted surface to horizontal in tropical regions," *Solar Energy*, vol. 97, pp. 369–387, 2013.

[10] J. Polo, S. Wilbert, J. Ruiz-Arias, R. Meyer, C. Gueymard, M. Súri, L. Martín, T. Mieslinger, P. Blanc, I. Grant, J. Boland, P. Ineichen, J. Remund, R. Escobar, A. Troccoli, M. Sengupta, K. Nielsen, D. Renne, N. Geuder, and T. Cebecauer, "Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets," *Solar Energy*, vol. 132, pp. 25 – 37, 2016.

[11] R. Perez, P. Ineichen, K. Moore, M. Kmiecik, C. Chain, R. George, and F. Vignola, "A new operational model for satellite-derived irradiances: description and validation," *Solar Energy*, vol. 73, no. 5, pp. 307–317, 2002.

[12] B. Minasny, A. B. McBratney, and D. J. Walvoort, "The variance quadtree algorithm: Use for spatial sampling design," *Computers & Geosciences*, vol. 33, no. 3, pp. 383–392, 2007.

[13] D. Yang and T. Reindl, "Solar irradiance monitoring network design using the variance quadtree algorithm," *Renewables: Wind, Water, and Solar*, vol. 2, no. 1, pp. 1–8, 2015.

[14] A. B. McBratney and B. Minasny, "Spacebender," *Spatial Statistics*, vol. 4, pp. 57–67, 2013.

[15] B. Y. H. Liu and R. C. Jordan, "Daily insolation on surfaces tilted towards the equator," *ASHRAE Transactions*, vol. 67, no. 3, pp. 526–541, 1961.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/