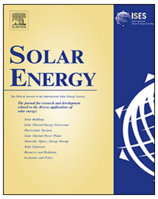
Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener



Spatial prediction using kriging ensemble

Dazhi Yang

*Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research (A*STAR), Singapore*

ARTICLE INFO

Keywords:
Kriging
Ensemble
Spatial interpolation
Solar irradiance

ABSTRACT

It is surprising how a commonly used concept in temporal prediction—combining forecasts, or rather combining predictions—has not really been brought forward in spatial prediction. Analogous to forecasting, where forecasts made using models such as exponential smoothing or neural networks are combined through regressions, the various prediction combination methods are herein transferred to spatial prediction problems. Through a series of empirical studies, the advantage and potential of kriging ensemble, or more generally, spatial-interpolator ensemble, are demonstrated. Both geostatistical and lattice data (solar irradiance) are considered. Although in theory, the improvement in predictive performance is not guaranteed, just like how we cannot guarantee that ensemble improves forecasts, in practice, a validated ensemble performs at least as good as the best component model, just like how the ensembles in forecasting would behave.

1. Introduction

Classical spatial statistics is concerned with three types of data: (1) geostatistical data, e.g., heavy metal concentrations in the top soil in a flood plain, (2) lattice data, e.g., crime rate in the United States by county, and (3) spatial point patterns, e.g., locations of oaks in a forest.[1] Given the "single snapshot" nature, traditional predictive studies on spatial data often aim at identifying a single most appropriate model for a particular dataset. Although there are a handful of studies available (e.g., Toal and Keane, 2013; Dutta et al., 2006), ensemble spatial prediction has not really been brought forward. These studies received only minimal attention. With no exception, such concept seems to be missing from the large body of literature on spatial interpolation of solar irradiance, notwithstanding our persistent interest in the topic since at least the early 80s (e.g., Maclaine-cross, 1980; Hay, 1986; Bland and Clayton, 1994; Perez et al., 1997; Merino et al., 2001; Righini et al., 2005; Alsamamra et al., 2009; Wu et al., 2013; Rodríguez-Amigo et al., 2017; Lorenzo et al., 2017).

Regardless of temporal or spatial data, the motivation for modeling is the same: finding a model that appropriately describes the data, and thus making inference and predictions. The famous statement by Box and Draper (1987)—all models are wrong, but some are useful—suggests that modeling error is ubiquitous and unavoidable. To that end, forecasters have come to a consensus that using multi-modeling, combinations, reanalysis, reconciliation, or collectively known as ensemble, is likely to improve forecast accuracy. Ensembles have been well studied in solar forecasting, especially in the recent years (e.g., Yang et al., 2017a,b; Jiang et al., 2017; Sperati et al., 2016; Sanfilippo et al., 2016; Chu et al., 2015). For a comprehensive review on ensemble forecasting in renewable energy applications, the reader is referred to the review by Ren et al. (2015). This paper aims at extending the concept of ensemble to spatial prediction problems.

In the remaining part of the paper, Section 2 gives an overview of the ensemble approach adopted in this paper, namely, regression-based prediction combination. A series of empirical studies are then used to demonstrate the advantage and potential of ensemble spatial prediction in Section 3. Various forms of kriging—the optimal prediction—alongside with several distance-weighted interpolators are used as *component* models. Section 4 concludes this work.

2. Methodology

Without loss of generality, the ensemble approach used in this paper takes a regression setting, as seen in Yang and Dong (2018):

$$Z(\boldsymbol{s};t) = \boldsymbol{\beta}(\boldsymbol{s})\boldsymbol{X}(\boldsymbol{s};t) + \varepsilon(\boldsymbol{s};t), \tag{1}$$

where $Z(\boldsymbol{s};t)$ denotes the "true" value[2] of the random process Z indexed by time t and location $\boldsymbol{s}$; $\varepsilon(\boldsymbol{s};t)$ is the regression error, indexed in the same way; $\boldsymbol{\beta}(\boldsymbol{s})$ is a row vector of regression coefficients, indexed only by spatial locations but not by time; and $\boldsymbol{X}(\boldsymbol{s};t)$ is the column vector of estimates made by N prediction models, i.e.,

$$\boldsymbol{X}(\boldsymbol{s};t) \equiv (\widehat{Z}_1(\boldsymbol{s};t), \cdots, \widehat{Z}_N(\boldsymbol{s};t))^{\top}. \tag{2}$$

E-mail address: yangdazhi.nus@gmail.com.

[1] The distinction among these three types of data is not always clear. For example, aggregating the exact locations where crimes happened (point patterns) by county leads to the lattice data mentioned earlier, and satellite-derived solar irradiance data (lattice) can be considered as the result of integrating an irradiance process (geostatistical) over small areas, i.e., pixels.

[2] Due to measurement error, the true value can never be known. In the present case, the words "true value" refer to the measurement.

https://doi.org/10.1016/j.solener.2018.06.105
Received 15 March 2018; Received in revised form 24 May 2018; Accepted 26 June 2018

Under this regression setting, once the value of $\boldsymbol{\beta}(\boldsymbol{s})$ is estimated, the combined prediction can be obtained via:

$$\widehat{Z}(\boldsymbol{s};t') = \widehat{\boldsymbol{\beta}}(\boldsymbol{s})\boldsymbol{X}(\boldsymbol{s};t'). \tag{3}$$

In this regard, all methods for regression parameter estimation can be used to construct ensembles of this kind, and each construction differs only in terms of its $\boldsymbol{\beta}(\boldsymbol{s})$ estimate. In this paper, the construction is exemplified by three methods: (1) ordinary least squares (OLS), (2) least absolute deviations (LAD), and (3) least absolute shrinkage and selection operator (lasso).[3] For other constructions, the reader is referred to Timmermann (2006) for a review. In addition to the regression-based combinations, two basic combination methods, namely, simple averaging of the component predictions, and variance-based combination, are also considered, see Yang and Dong (2018) for details. To denote these methods, SmallCaps[4] is used—Ols, Lad, Lasso, Avg, and Var.

The ensemble approach shown in Eq. (1) is location specific. Since spatial predictions are often made at a collection of locations, this approach can be extended to model the regression weights globally, i.e., the $\boldsymbol{\beta}$ is estimated for all locations, so that its index $\boldsymbol{s}$ can be dropped. In this way, the correlation structure in $\varepsilon(\boldsymbol{s};t)$ must be exploited, e.g., investigating whether the correlation structure follows the spatial correlation of the data. Subsequently, linear models with heterogeneous variance, linear models with correlated errors, or linear mixed-effect models (Gałecki and Burzykowski, 2013) can be used. However, it is thought that the model building and parameter estimation for those models require a substantial amount of work, this paper thus does not go beyond the framework shown in Eq. (1).

3. Performance of kriging ensemble

3.1. NSRDB PSM data

Physical solar model (PSM) version 3 data from the national solar resource database (NSRDB) are half-hourly, regularly-gridded (4 km × 4 km), satellite-derived irradiance—and other meteorological variables, such as air temperature, surface albedo, or relative humidity—with a spatial coverage for most of America, over the years 1998–2016 (Habte et al., 2017; Sengupta et al., 2015, 2014). In a parallel work by Yang (2018), a spatio-temporal subset of the PSM data, namely, the 0.1° × 0.1° gridded data over California for 2016, was used. Furthermore, 8 regular and irregular lattices with different spatial resolutions were designed to study the predictive performance of kriging at 50 noncoincident (with the lattices) validation locations (see Fig. 1). A total of 6 kriging methods, namely, simple kriging (Simple), ordinary kriging (Ordinary), universal kriging (Universal), simple local kriging (LocalS), ordinary local kriging (LocalO), and universal local kriging (LocalU), were tested on each lattice, along with 3 benchmarking methods, namely, thiessen polygon (1nn), multiple nearest neighbor (5nn), and inverse distance weighting interpolation (Idw). Since the PSM data has been previously discussed and validated[5] in detail (Yang, 2018; Habte et al., 2017), I do not discuss the data and component kriging models further, but refer the reader to the original publications. Nevertheless, the results of ensembles are tabulated and presented in the form of Yang (2018), so that a comparison can be easily made.

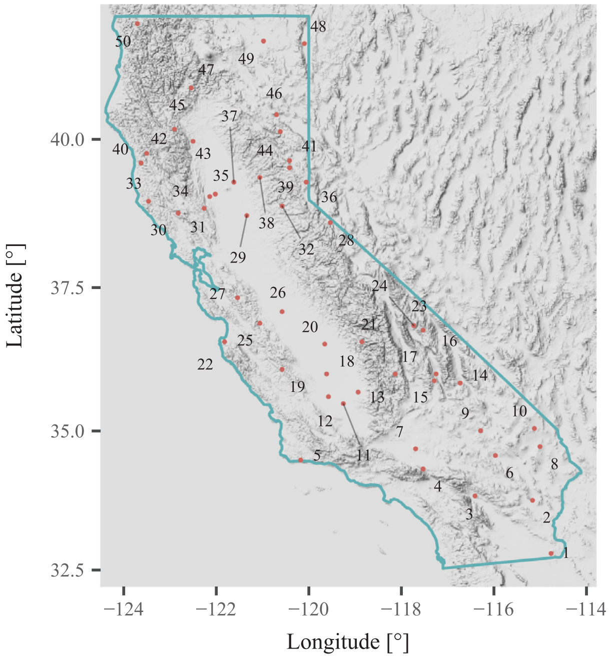


Fig. 1. Google Map (topographic) of the California region with the 50 validation locations. Data from 8 different lattices (not shown) are used to predict half-hourly GHI at each of these locations.
Source: Google Maps.

It should be noted that the regression-based ensemble models require fitting, i.e., the regression coefficients needs to be estimated. On this point, a 10-fold cross validation (CV) is used. The results of the component predictions are assigned to one of the folds via random sampling. When a fold is being evaluated, the remaining 9 folds are used to fit the regression coefficients. In this way, all the component predictions can be used to generate true out-of-sample ensemble predictions. Since the prediction results are often hard to reproduce, and thus limiting the uptake of the methodology (Yang et al., 2018), to enhance the understanding and facilitate future development, I provide the data and code in the supplementary material section.

The normalized root mean squares errors (nRMSEs), in percentage, of the 5 kriging ensembles and 9 components models are shown in Table 1 and Fig. 2. It is evident that the performance of kriging ensemble dominates that of the component models. It is also observed that the ensemble methods can effectively integrate the advantages of various component models. For example, 1nn has high accuracies at some locations that is not achieved by kriging (see Lattice1, Lattice2 and Lattice5 in Fig. 2). Ensembles can reach similar accuracies at these locations while maintaining good performance across other locations. Furthermore, the *risk* of using an ensemble method is smaller than choosing the best component model (see Yang and Dong, 2018, Hibon and Evgeniou, 2005, for a discussion on such risk). Through this case study, the performance of kriging ensemble is empirically validated on lattice data of different spatial resolutions. The following case studies investigate the performance on geostatistical data. More specifically, the validation of geostatistical data is carried out in both sub-continental and metropolitan scales.

3.2. SONDA data

The data from a sparse ground-based irradiance monitoring network, namely, Sistema de Organização Nacional de Dados Ambientais (SONDA),[6] is used to test the performance of kriging ensemble in a sub-continental scale. The SONDA network, with a spatial coverage of

[3] OLS is the most basic approach for regression parameter estimation. Whereas OLS penalizes heavily on large errors due to the squared loss, LAD minimizes the sum of absolute errors, which is robust to outliers. When the number of component predictions gets large, prediction accuracy and interpretability become important. To that end, lasso is used for variable selection and shrinkage; it is chosen here to represent the class of penalized regression methods.

[4] I learned about such notation from Gueymard and Ruiz-Arias (2016), found it quite useful when a number of models are being referenced and discussed frequently, and used it in several earlier works (e.g., Yang, 2016).

[5] Since satellite-derived irradiance often poses some bias, site adaptation, or at least checking the data accuracy, is essential (Polo et al., 2016).

[6] Data courtesy of Dr. André Nobre, who was a colleague of mine in the Solar Energy Research Institute of Singapore during 2012–2015.