Machine Learning Methods for Solar Irradiation Forecasting: A Comparison in a Mediterranean Site

Alexis Fouilloy, Cyril Voyant, Gilles Notton, Marie Laure Nivet, Jean Laurent Duchaud

Machine Learning Methods for Solar Irradiation Forecasting: A Comparison in a Mediterranean Site

Alexis Fouilloy
University of Corsica
France
fouilloy-a@univ-corse.fr

Cyril Voyant
University of Corsica
France
voyant@univ-corse.fr

Gilles Notton
University of Corsica
France
notton@univ-corse.fr

Marie Laure Nivet
University of Corsica
France
nivet@univ-corse.fr

Jean Laurent Duchaud
University of Corsica
France
duchaud_jl@univ-corse.fr

ABSTRACT

In this survey, several statistical and machine learning tools are analyzed and compared in view to forecast the solar irradiation in Ajaccio (Corsica, France, 41°55 N, 8°44 E, 4m asl). The forecasting horizon is from 1 to 6 hours with an hourly time granularity. Eleven forecasting models are compared: persistence, scaled persistence, ARMA, MLP, regression trees, boosted regression trees, bagged regression trees, pruned regression trees, random forest, Gaussian processes and support vector regression. The models are compared in term of error metrics: nRMSE (normalized root mean squared error), MAE (mean absolute error) and skill score related to the smart persistence.

KEYWORDS

Time Series forecasting, machine learning, variability, ARMA, ANN, Regression trees, Gaussian process, SVR.

1 INTRODUCTION

Solar radiation is one of the principal energy sources for physical, biological and chemical processes, occupying the most important role in many engineering applications [1]. The process of converting sunlight to electricity without combustion allows to create power without pollution. The major problem of such energy source is its intermittence and its stochastic character which make difficult their management into an electrical network [2]. Thereby, the development of forecasting models is necessary to use ideally this technology [3]. By considering their effectiveness, it will be possible for example to identify the most optimal locations for developing a solar power project or to maintain the grid stability and security of a power management system [4]. Thus the solar energy forecasting is a process used to predict the amount of solar energy available for various time horizons [5]. Several methods have been developed by experts around the world and the mathematical formalism of Times Series (TS [6]) has been often used for the short term forecasting (among 6 hours ahead) [5].

In this study, the solar forecasting is realized with an hourly time granularity for 1 hour to 6 hours' time horizon. The comparison between the models is evaluated by statistical indexes and graphical methods, the models are: persistence (P) [3,7], scaled or smart persistence (SP) [3,10], auto regressive moving average (ARMA) [11], artificial neural network (ANN) [10,12], regression trees (RT) [13], pruned regression trees (RT-pruned) [13], boosted regression trees (RT-boosted), bagged regression trees (RT-bagged) [14], random forest (RF) [12], Gaussian processes (GP) [13] and support vector regression (SVR) [14].

2 METHOD

Several preliminary stage must be realized before the forecasting phase:

- A quality control of the solar data: an automatic quality check used in the frame of GEOSS project (Group on Earth Observation System of System) (http://www.earthobservations.org/geoss.php) was applied to the data;
- A data preprocessing: removing of the night hours; sunset and the sunrise data induce some problems (mask effects and bad response of pyranometers at high zenith angle): a filter is applied on the datasets which remove all the data with a solar elevation angle up to 10°. The solar irradiation time series

contains some properties like seasonality and periodicity which must be deleted to make the time series stationary; to do it, a clear sky model [15] is used and the clear sky index calculated (paragraph 2.1).

- At last, the choice of the input data must be realized and is described in paragraph 2.2.

2.1 Clear Sky Irradiation (CSI)

The clear sky model is a theoretical model which determines the maximum theoretical global horizontal irradiation (GHI) for any geographical point at the ground level. The model used is SOLIS model [16-17] based on radiative transfer models, Beer Lambert function, and their integration on the solar spectra uses the information's of news satellites ERS-2/ENVISAT. The calculation of the clear sly irradiation CS(t) is given by:

$$CS(t) = H_0 . e^{\frac{-\tau}{sin^b(h(t))}} . sin(h(t)) \quad (1)$$

h(t) is the solar height in degrees and H_0 is the extraterrestrial irradiation; τ is the global total atmospheric depth and the parameter b is a fitting parameter, both of them depending on the meteorological characteristics of the site [19]. τ and b have different values for the three studied sites (https://aeronet.gsfc.nasa.gov/). The clear sky index (CSI) is calculated by:

$$CSI(t) = \frac{GHI(t)}{CS(t)} \quad (2)$$

For all the machine learning tools used here, the stationary hypothesis is necessary and then the different time series, CSI are directly used as an input in the forecasting model. GHI is then obtained using Eq. 2 and all the errors metrics are given in term of GHI in absolute and relative values.

2.2 Dimension of the input matrix

We must choose the dimension of the input matrix of the forecasting tool. The approach used consists in predicting the future solar irradiation (at different time scales) based on the past observed data [18]. Mathematically, the researched formulation is (ϵ(t+h) is a random white noise):

$$GHI(t+h) = f(GHI(t), GHI(t-1), \dots, GHI(t-n)) + \epsilon(t+h) \quad (3)$$

The future time step (t+h) is forecasted based on the observed data at the times (t, t-1…, t-n). In other words, the objective is to calculate the value of n and to obtain ϵ as low as possible in absolute value. The choice of n, i.e. the dimension of the input matrix, is made by an auto mutual information method [19-20]. This auto mutual information is a property of the time series and depends on each dataset. It determinates the degree of statistical dependence of the variables specific to each site.

Another step of preprocess is the k-fold sampling [24]: the dataset is divided in ten samples and each sample is used at least one time for the training and one time for the test, the simulation is repeated as many time as necessary. This method allows to have results independent of the set of data used for the training (only one data set being able to have some particularities that disturbs the robustness of the conclusions).

3 METEOROLOGICAL STATION AND DATA

The datasets used in this study are time series of horizontal global solar irradiation measurements (GHI) in Ajaccio (Corsica, France, 41°55 N, 8°44 E, 4m asl) at about 100 m from the Mediterranean sea (Figure 1).

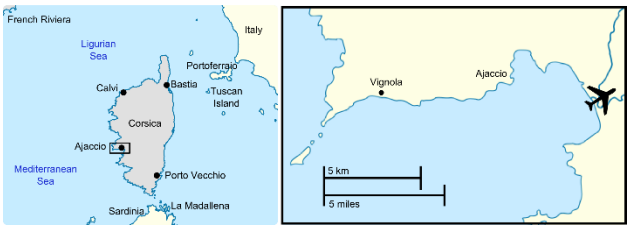

Figure 1: **Position of the meteorological station**

In this station, the period of available solar data are from 01/01/1998 to 12/31/2000 with 26280 validated data after the quality check.

After the calculation of the auto-mutual information between the solar radiation data [19-20], it has been decided that the number of input data n in the forecasting model will be n=8.Thus, for estimating the solar irradiation at time t, the solar irradiations at time (t-1), (t-2), (t-3)… (t-n) will be used.

4 FORECASTING MODELS

Eleven models are used to forecast the global horizontal irradiations and are briefly described here. These models are classified in 3 categories: naïve models, classical machine

learning models and regression trees based models. In this paper the symbol $\widehat{\ }$ indicates that the value is predicted, without this symbol the value is measured.

4.1 Naïve models

The two next models are generally used only as a reference in view to compare it with more sophisticated models in terms of performances.

The first model is the simplest model, the persistence, it is the repetition of the measure at the instant t to the instant t+h (h is the forecasting horizon) [18]:

$$\widehat{GHI}(t+h) = GHI(t) \quad (4)$$

$\widehat{GHI}(t)$ and $GHI(t)$ are respectively the predicted and measured hourly global horizontal solar irradiation at time t. This type of model is very simple to use but it gives results with very low accuracy.

In order to improve this model, the daily profile of the solar radiation (by clear sky) is taken into account and added in using the Solis clear sky model previously presented, then Eq. (4) becomes:

$$\widehat{GHI}(t+h) = GHI(t).\frac{CS(t+h)}{CS(t)} \quad (5)$$

This model is named scaled persistence (SP), a simple improvement of the persistence model [22]. The advantages of naïve models is the fact that no historical data are required what makes these models easily usable in all cases and for all sites even if no solar radiation measurements exits for a long time. The accuracy of such models decreases rapidly with the time horizon and is generally not adequate for a more than one hour ahead.

4.2 Machine Learning models

A. Auto Regressive Mobile Average (ARMA)

The ARMA model [23] is a predictive model with two separate parts, the auto regressive one and the mobile average one. This model is able to predict the future values of time serie from a combination of past values of the time series and a white noise:

$$\widehat{CSI}(t+h) = \varepsilon_t + \sum_{i=0}^{p} \varphi_i \,.\, CSI(t-i) + \sum_{i=0}^{q} \theta_i \,.\, \varepsilon(t-i) \quad (6)$$

CSI (t+h) the clear sky index time series at time t+h, φ and θ the parameters of the autoregressive and moving average part deduced by a least square method, p and q are the orders of the model and ε(t) is an error term distributed like a Gaussian white noise for the time t.

B. Multi-Layer Perceptron (MLP)

The MLP is a type of artificial neural networks, here, a feed forward MLP with two layers (one hidden layer and one output layer) is used with, in input, the solar irradiation time series [9]. The input matrix size was defined according to the results of the auto mutual method applied to the solar data. The mathematical formula for a MLP with one hidden layer of m neurons, one output neuron and p input variables is a function described by:

$$\widehat{CSI}(t+h) = \sum_{j=1}^{m} \omega_{j,s} \,.\, (g.\left(\sum_{i=0}^{n-1} \omega_{i,j}.CSI(t-j) + b_j\right) + b_s \quad (7)$$

with CSI the input vector of clear sky index constituted by the values of n variables, $\widehat{CSI}(t+h)$ the output vector with the predicted values of the model, bj the bias of the hidden neuron j and ω(i,j) the weight affected at the measured input CSI(t). g is the transfer function of the hidden neurons, bs the bias of the output neuron and ω(j,s) his weight affected to the output of the hidden neuron j. The optimization of the MLP is made by a classical technique: several configurations with a different number of hidden nodes are tested (with a number of hidden nodes varying between 3 and n+2) and the most efficient is selected.

C. Gaussian Process (GP)

The Gaussian process (GP) is a nonlinear modeling. It's a generalized Gaussian distribution with an infinity of variables [13,24]. As the solar data are often noisy, every observation is represented by a function ($CSI(\boldsymbol{\tau})$), with an independent Gaussian noise $\mathcal{N}(O, \sigma_n^2)$ characterized by a variance σ_n^2 with CSI(τ)=(CSI(t),CSI(t-1),⋯,CSI(t0); thus,

$$\widehat{CSI}(t+h) = f\big(CSI(\boldsymbol{\tau})\big) + \mathcal{N}(O, \sigma_n^2) \quad (8)$$

A GP is defined by a mean function m(CSI(τ)) and a covariance function. The covariance function k relies $\widehat{CSI}(t_p+h)$ and $\widehat{CSI}(t_q+h)$ and is exponential

squared function, tp and tq being two successive instants. The noise in the prediction allows to note the covariance function as follow:

$$k\left(\widehat{CSI}(t_p+h), \widehat{CSI}(t_q+h)\right) = \sigma_f^2 exp\left[\frac{-\left(CSI(t_p)-CSI(t_q)\right)^2}{2l^2}\right] + \delta_{pq}\sigma_n^2 \quad (9)$$

δ_{pq} is Kronecker delta, σ_f^2 are the hyper parameters of the covariance function and are responsible of the complexity of the model, l is a length parameter. They are deducted and optimized during the training phase [13].

D. Support Vector Regression (SVR)

The Support vector regression (SVR) is a Kernel based model (as Gaussian process), it was firstly developed for regression problems and the application to time series forecasting is another successful evolution [25]. Equation 10 gives the formulation of the SVR applied on forecasting problems, with a training dataset $D = \{CSI(\tau), CSI(t+h)\}$ with the test vector CSI(t+h) and h the prediction horizon:

$$\widehat{CSI}(t+h) = \sum_{\tau=1}^{t-1} \alpha_i \, . k_{rbf}(CSI(t+h), CSI(t-\tau)\ \ ) \quad (10)$$

and the Kernel radial basis function defined by:

$$k_{rbf}\left(CSI(t_p), CSI(t_q)\ \ \right) = exp\left[\frac{-\left(CSI(t_p)-CSI(t_q)\right)^2}{2\sigma_f^2}\right] \quad (11)$$

α_i is the difference between Lagrange multipliers which are the solutions of a quadratic problem, b is the bias determined by specific conditions and Eq. (10).

E. Regression Trees

Decision trees based on "If-Then" rules are one of the most popular methods used in machine learning for classification, since they offer results that can be easily interpreted. Thus, this approach obtains ordered sets of If-Then rules for prediction that produces understandable models. In the last years, learning algorithms generating decision trees for classification problems have been extended for predicting values of attributes when they are numeric. These extensions have led to regression trees. A Regression Tree (RT) is a decision tree in which the leaf nodes have been set as regression models, and therefore, continuous numeric values can be predicted [35].

Standard and pruned regression trees (RT and RT-pruned)

A model of regression trees is directly derived from the classification trees, Hastie and Tibshirani [27] propose a formalization of the models:

$$\widehat{CSI}(t+h) = \sum_{i=1}^{t-1} k_i \times I(CSI(t-i)) \quad (12)$$

with ki constant factors, I is a function which return 1 if input is used and 0 if it's not used; The trees are built by splitting the data based on the values of predictive attributes. Once the tree has been constructed, a regression model is computed for each node. The error for each node is the mean of the absolute difference between the predicted and the actual value of each instance of the training set that reaches the node. Pruning aspect of regression tree is operated with an elevation of the quadratic error tolerance per node. Splitting nodes stops when quadratic error per node drops below tolerance. For normal RT, the tolerance is close to zero, while for the pruned RT, a higher value is chosen using a heuristic method based on the minimizing of the global error of prediction. In pruned RT, the function I has more 0 than in the normal mode.

Boosted and bagged regression trees (RT-boosted and RT-bagged)

For RT, two types are boosted [28] and bagged of classification trees [29]. In boosting, successive trees give extra weight to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. In bagging, successive trees do not depend on earlier trees each is independently constructed using a bootstrap sample of the data set. In the end, a simple majority vote is taken for prediction.

The boosting method consists in assembly weak RT classifiers and take the average of predictions in order to improve the efficiency. In this case a weak predictor is a simple single split regression tree. The main technique is that the next trees give more weights to the data which have a bad prediction at the precedent point, at the end of the simulation, a weighted vote is taken for the prediction.

The Bagging method is another version of the prediction with regression trees, it was described by Breiman [11]. Bagging mean bootstrap aggregating, the model is an aggregation of regression trees which grown from samples of dataset. The subtrees are employed for the prediction and a vote is taken for the prediction.

Random Forest

Breiman [12] proposed the random forests, which add an additional layer of randomness to bagging. In addition to a construction for each tree using a different bootstrap sample of data, random forests change how the regression trees are constructed [12]. In standard trees, each node is split using the best split among all variables. In a random forest, the dataset is equally divided in samples but each regression tree grows differently, each node is split using the best among a subset of predictors randomly chosen at that node [12]. This improvement by randomness give robustness to the model and decreases the over-training risks.

5 ACCURACY PARAMETERS

In order to evaluate the accuracy of the different models, three different statistics are used.

The mean absolute error (MAE), defines the absolute value of the gap between the observed and the predicted value:

$$MAE = \frac{\sum_{i=1}^{N} |\widehat{GHI}(i) - GHI(i)|}{N} \quad (13)$$

N is the number of variables.

The normalized root mean squared error (nRMSE) is more sensitive to big forecast errors, and hence is suitable for applications where small errors are more tolerable and larger errors cause disproportionately high costs, as for example in the case of utility applications. It is probably the reliability factor that is most appreciated and used: it's a good statistical index to evaluate the accuracy of a model, the aim of an operator is to minimize it in order to improve model performances:

$$nRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{GHI}(i) - GHI(i))^2}}{\overline{\widehat{GHI}}} \quad (14)$$

where $\overline{\widehat{GHI}}$ is the algebraic average of the observed values.

The skill score is an index calculated in order to compare the performance of a given model with a reference model, here the reference model is the smart persistence model (SP) described in 4.1.).

$$Skill\ Score = \frac{Metric_{forecast} - Metric_{reference}}{Metric_{perfect_{forecast}} - Metric_{reference}}$$

$$= 1 - \frac{nRMSE_{forecast}}{nRMSE_{reference}} \quad (15)$$

The skill score is always inferior at 1, negative if the forecaster is less performant than the reference, 0 if the performances are similar, and positive if it is better.

6 FORECASTING MODELS PERFORMANCE

Tables 1 and 2 show the results of the nRMSE and MAE calculations for the eleven models and a forecast horizon from 1 to 6 hours with an hourly resolution.

Table 1: **nRMSE vs forecast horizon, the two best models are highlighted (results in %)**

Horizon	h+1	h+2	h+3	h+4	h+5	h+6
Persistence	26.60	42.62	54.10	61.39	64.51	63.86
SP	19.26	26.46	31.18	34.15	36.92	38.93
ARMA	18.35	29.27	31.38	32.25	33.18	33.69
MLP	18.26	29.26	31.31	32.47	32.98	33.84
RT	24.64	36.88	38.47	39.74	39.95	41.24
Boosted RT	18.75	29.55	31.89	32.51	33.55	33.98
Bagged RT	18.76	29.80	31.10	32.17	33.35	34.02
Pruned RT	18.72	30.88	32.27	33.76	34.01	35.00
RF	18.97	29.63	31.62	32.38	33.37	33.91
GP	18.97	30.08	31.96	33.29	33.55	34.44
SVR	18.55	38.78	41.03	41.56	41.66	41.60

Table 2: **MAE values (in Wh/m²) vs forecast horizon, the two best models are highlighted**

Horizon	h+1	h+2	h+3	h+4	h+5	h+6
Persistence	104.6	176.0	220.2	252.2	259.2	256.0
SP	55.5	79.5	96.0	107.1	116.4	124.0
ARMA	60.7	88.8	95.0	97.4	100.6	102.3
MLP	60.6	89.4	95.2	98.9	100.7	103.3
RT	73.3	117.7	125.0	129.8	132.0	135.6
Boosted RT	61.4	91.8	98.1	100.1	103.3	104.6
Bagged RT	61.4	93.3	97.3	101.3	104.6	106.0
Pruned RT	60.8	95.4	100.0	105.0	106.1	108.9
RF	61.1	92.9	99.1	102.0	105.0	106.5
GP	61.9	96.1	101.1	104.8	106.3	108.3
SVR	54.6	126.6	133.3	135.5	135.1	134.6

The naïve models (persistence and smart persistence) give always bad performances in all the sites and is the worst model.

For Ajaccio with a low variability dataset the better models are ARMA and MLP, followed by scaled persistence and bagged regression trees; however the calculated error is similar for all the models excepted for the persistence and the simple regression tree. With such a variability, the machine learning models allow a good forecasting.

Concerning the MAE, excepted for the naïve model, all the models are similar in term of absolute error, the results are in the same magnitude.

The skill score was computed in the same conditions for all the models, Figure 2 is the graphical representation of the skill scores (related to the smart persistence reference model).

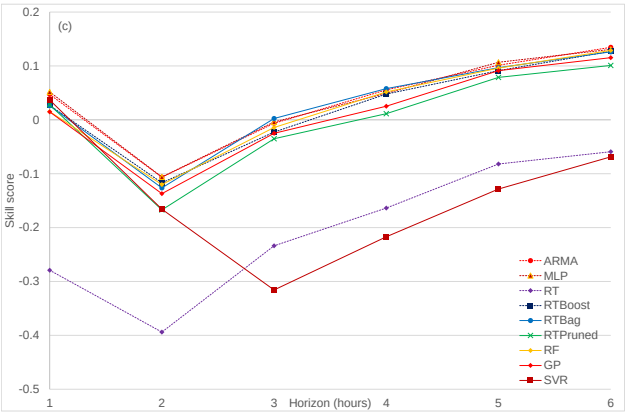

Figure 2: **Skill score vs forecast horizon for Ajaccio**

6 CONCLUSION

Several statistical and machine learning tools (benchmark related to 11 forecasting models) were analyzed and compared in view to forecast the horizontal solar irradiation in Ajaccio. For a weak variability as Ajaccio, the two classical methods: ARMA and MLP seem to be the most efficient.

In a future work, we will integer solar data from other sites in order to improve the conclusions and maybe to modify the number of variability classes using clustering methods.

REFERENCES

[1] Hoff TE, Perez R, Modeling PV fleet output variability. Sol Energy, 2012, 86, pp. 2177–89.

[2] Voyant C, Paoli C, Muselli M, Nivet M, Multi-horizon solar radiation forecasting for Mediterranean locations using time series models. Renew Sustain Energy Rev, 2013, 28, pp. 44–52.

[3] Lauret P, Voyant C, Soubdhan T, David M, Poggi P., A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. Sol Energy, 2015, 112, pp 446–57.

[4] Mellit A, Kalogirou SA, Hontoria L, Shaari S., Artificial intelligence techniques for sizing photovoltaic systems: A review, Renew Sustain Energy Rev, 2009, 13, pp. 406–19.

[5] Lorenz E, Remund J, Müller SC, Traunmüller W, Steinmaurer G, Pozo D, et al., Benchmarking of different approaches to forecast solar irradiance, 24th Eur. Photovolt. Sol. Energy Conf. Hambg. Ger., vol. 21, 2009, p. 25.

[6] Join C, Fliess M, Voyant C, Chaxel F., Solar energy production: Short-term forecasting and risk management. ArXiv160206295 Cs Q-Fin 2016.

[7] Voyant C, Motte F, Fouilloy A, Notton G, Paoli C, Nivet M-L., Forecasting method for global radiation time series without training phase: Comparison with other well-known prediction methodologies, Energy, 2017, 120, pp. 199–208.

[8] Gooijer De JG, Hyndman RJ., 25 years of time series forecasting, Int J Forecast, 2006, 22, pp. 443–73.

[9] Kalogirou SA., Applications of artificial neural-networks for energy systems, Appl Energy, 2000, 67, pp. 17–35.

[10] Breiman L., Classification and regression trees, Wadsworth International Group, 1984.

[11] Breiman L,. Bagging predictors, Mach Learn, 1996, 24, pp. 123–40.

[12] Breiman L., Random forests, Mach Learn, 2001, 45, 5–32.

[13] Rasmussen CE, Gaussian Processes in Machine Learning. In: Bousquet O, Luxburg U von, Rätsch G, editors. Adv. Lect. Mach. Learn., Springer Berlin Heidelberg; 2004, pp. 63–71.

[14] Smola AJ, Schölkopf B., A tutorial on support vector regression. Stat Comput, 2004, 14, 199–222.

[15] Hornik K, Stinchcombe M, White H., Multilayer feedforward networks are universal approximators. Neural Netw, 1989, 2, 359–66.

[16] Mueller RW, Dagestad KF, Ineichen P, Schroedter-Homscheidt M, Cros S, Dumortier D, et al., Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module. Remote Sens Environ, 2004, 91, pp.160–74.

[17] Ineichen P. Comparison of eight clear sky broadband models against 16 independent data banks, Sol Energy, 2006, 80, pp.468–78.

[18] Diagne M, David M, Lauret P, Boland J, Schmutz N. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renew Sustain Energy Rev, 2013, 27, pp. 65–76.

[19] Jiang A-H, Huang X-C, Zhang Z-H, Li J, Zhang Z-Y, Hua H-X, Mutual information algorithms, Mech Syst Signal Process, 2010, 24, 2947–60.

[20] Parviz RK, Nasser M, Motlagh MRJ. Mutual Information Based Input Variable Selection Algorithm and Wavelet Neural Network for Time Series Prediction. In: Kůrková V, Neruda R, Koutník J, editors. Artif. Neural Netw. - ICANN 2008, Springer Berlin Heidelberg; 2008, p. 798–807.
[21] Wiens TS, Dale BC, Boyce MS, Kershaw GP. Three way k-fold cross-validation of resource selection functions. Ecol Model 2008, 212, pp. 244–55.
[22] Voyant C, Soubdhan T, Lauret P, David M, Muselli M. Statistical parameters as a means to a priori assess the accuracy of solar forecasting models. Energy, 2015, 90, Part 1:671–9.
[23] De Gooijer JG, Hyndman RJ. 25 years of time series forecasting. Int J Forecast, 2006, 22, pp. 443–73.
[24] Salcedo-Sanz S, Casanova-Mateo C, Munoz-Mari J, Camps-Valls G. Prediction of Daily Global Solar Irradiation Using Temporal Gaussian Processes. IEEE Geosci Remote Sens Lett, 2014, 11, pp.1936–40.
[25] Vapnik V. The Nature of Statistical Learning Theory. Springer Science & Business Media, 2013.
[26] Burrows WR. CART Regression Models for Predicting UV Radiation at the Ground in the Presence of Cloud and Other Environmental Factors. J Appl Meteorol, 1997, 36, 531–44.
[27] Hastie T, Tibshirani R. Generalized additive models. Stat Sci, 1986, 1, pp.297–318.
[28] Breiman L. Bagging Predictors. Mach Learn 1996;24, pp. 23–40.
[29] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat, 2001, 29, pp. 1189–232.