Solar Radiation Prediction Based on Machine Learning for Istanbul in Turkey

Veysel Çoban[1(✉)] and Sezi Çevik Onar[2]

[1] Bozok University, Erdoğan Akdag Campus, Yozgat, Turkey
veysel.coban@bozok.edu.tr
[2] Management Faculty, Istanbul Technical University, Istanbul, Turkey

Abstract. The correct installation of solar energy systems is important for the energy efficiency of the system. The total solar radiation values reaching the system have an important role in determining the energy production potential of the solar energy system. In this study, statistical and machine learning methods used in solar radiation estimation are discussed. Forecasting methods are evaluated with the application on Istanbul region. The variability of the data collected for the Istanbul region is examined and the inappropriate data in the data are extracted. The data that are checked and approved are applied to the forecasting models and the models are compared and evaluated according to their error values. Models are evaluated according to variability values and error values over temporal horizons. Variability has an important role in determining the most appropriate forecasting model.

Keywords: Forecasting methods · Machine learning · Solar irradiation

1 Introduction

Electrical network management is important for grid-independent structures and systems. Although the supply-demand balance is important in electricity management, the inclusion of intermittent energy sources in the system is an important challenge. Energy production needs to be planned in a way to meet consumption for efficient use of energy. Forecasting methods have been developed to incorporate the intermittent generation of renewable energy sources into the traditional energy generation system [1]. Forecasting methods are used in determining the amount of energy to be produced from solar energy systems which have an important place in renewable energy systems. The forecasting methods used in solar energy systems are used to estimate solar radiation values according to different environmental conditions and input values. These prediction models are defined in three main classes as sky imaging, numerical weather prediction (NWP) and machine learning models [2]. Short, medium and long-term forecasts are performed with sky imaging models, machine learning methods, and NWP models respectively. In this study, forecasts based on machine learning models are performed by evaluating the estimated solar energy applications for Turkey. The models to be discussed in this study are tested main models such as persistence, auto regressive moving average, artificial neural network, regression tree, random forest method, Gaussian processes, and support vector regression [3]. The continuation of the study is planned as

C. Kahraman et al. (Eds.): INFUS 2019, AISC 1029, pp. 197–204, 2020.
https://doi.org/10.1007/978-3-030-23756-1_25

follows: Sect. 2 explains the clear sky model and the forecasting tools and sampling. Section 3 describes the forecasting methods that are divided into three parts as naive models, classical machine learning models and regression trees-based models. Section 4 presents the measurement of the performance and variability of the models. Section 5 explains the application and evaluation about Istanbul province. In the conclusion section, the results obtained from the study are generally evaluated.

2 Methods

Collecting the data, defining the accuracy and debugging errors to be used in solar radiation estimations is the most important and the initial stage. From these data, values which do not meet the usage period of solar radiation systems are extracted. Data should be reviewed according to machine learning methods that require stationary time series [4, 5]. Thus, the use of data based on open sky models is suitable for machine learning models.

2.1 Clear Sky Models

Clear sky models calculate global horizontal irradiation (GHI) values at ground level, taking into account atmospheric scattering and absorption. Machine learning approaches based on the stationary hypothesis of time series generate data based on the assumption that they do not change over time. The periodicity of the time series is removed by dividing the measured irradiation values by the clear sky value. Time series defined between 0 and 1 are used to predict cloud formations. The CS model based on the SOLIS (SOLar Irradiance Scheme) model that provides direct, global, and diffuse radiation calculations using advanced atmospheric information is preferred because of the appropriate results [1]. The applied clear sky model is defined as follows:

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

where $h(t)$ denotes the solar height, H_0 denotes extraterrestrial irradiance. τ denotes global total atmospheric deep and b denotes the fitting parameter that they represent meteorological characteristics of the area [2]. Clear sky index based on the clear sky value is defined as:

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

2.2 Forecasting Tools and Sampling

The data is formalized under the time series (TS) that are indexed by time. The mathematical representation of the approach based on the estimation of future open sky data using the data observed in the past is as follows [7]:

$$CSI(t+h) = f(CSI(t), CSI(t-1), CSI(t-2), \ldots, CSI(t-n)) + \varepsilon(t+h) \quad (3)$$

where $\varepsilon(t+h)$ represents the random white noise and $(t+h)$ refers to the future time step to be forecasted according to the observed data at a given time $(t, t-1, t-n)$. The size of the input matrix (n) is determined by the auto mutual information method, which is a property of a time series. Time series are divided into two groups as training and test group in order to train the model and test the model in the machine learning method. The k-fold method is used to prevent training and test data from being affected by meteorological events. Separation of training and test data takes place using the k-fold method [8].

Istanbul's time series of horizontal global solar irradiation measurements (GHI) is defined as a dataset. The dataset that reflects the meteorological conditions is defined temporally and the inappropriate inputs in the dataset are extracted. Data for the Istanbul region were collected during the 01/01/2017–12/31/2017 time period and the number of valid data was determined as 8510.

3 Forecasting Methods

In this study, basic prediction models described in the literature are discussed under naive models, classical machine learning models and regression trees-based models.

3.1 Naive Models

The naive models used to compare advanced models also decide whether to use these complex models. The predicted and measured hourly global horizontal solar irradiation for persistence model are represented as $\widehat{GHI}$ and GHI respectively [9].

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

Although the model has a simple application, the accuracy of the results is low. The daily profile of solar radiation can be added using the SOLIS clear sky model.

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

Models that decrease their accuracy parallel to time horizon are not suitable for horizons higher than 1 h.

3.2 Classical Machine Learning Models

- Auto regressive mobile average (ARMA): The ARMA model forecasts the future energy consumption with the linear combination generated by using historical data in two main stages [10, 11].

$$\widehat{CSI}(t+h) = \varepsilon(t) + \sum_{i=0}^{k} \varphi_i CSI(t-i) + \sum_{i=0}^{m} \theta_i \varepsilon(t-i) \tag{6}$$

where $CSI(t-i)$ represents the clear sky index at time $(t+h)$, φ and θ obtained from least square method are ARMA parameters, k and m denote the model orders and $\varepsilon(t)$ denotes the error for a normal distribution.

- Artificial neural network (ANN): MultiLayer perceptron (MLP): The feed forward MLP method that uses the hidden layer and output layer is a convenient tool for solar energy systems [10, 11].

$$\widehat{CSI}(t+h) = \sum_{i=1}^{p} \omega_i g\left(\sum_{j=0}^{q-1} \omega_{j,i} CSI(t-i) + b_i\right) \tag{7}$$

where CSI represents the input vector of n clear sky indexes, $\widehat{CSI}(t+h)$ represents the predicted value, b_i represents the biases of the hidden neuron i, $\omega_{j,i}$ represents the weights between the input value j and the hidden node i, g represents the transfer function, ω_i represents the weight between the output and the hidden neuron i.

- Gaussian process (GP): The Gaussian process is a nonlinear model with Gaussian distribution based on the infinity of variables [10, 11].

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

where $f(CSI(\tau))$ represents the summation function, $N(O, \sigma_n^2)$ represents the independent Gaussian noise, σ_n^2 denotes the variance. Mean function, $m(CSI(\tau))$ and covariance function, c are used to define the Gaussian process.

$$c\left(\widehat{CSI}\left(t_q+h\right), \widehat{CSI}\left(t_p+h\right)\right) = \sigma_f^2 e^{\left[\frac{-\left(CSI(t_q)-CSI(t_p)\right)^2}{2l^2}\right]} + \delta_{qp}\sigma_n^2 \tag{9}$$

where δ_{qp} denotes the Kronecker delta, σ_f^2 and σ_n^2 represent the hyper parameters of the covariance function and l denotes the length parameter.

- Support vector regression (SVR): The SVR model developed for the solution of the regression problems is used as a prediction method based on Kernel [10, 11].

$$\widehat{CSI}(t+h) = \sum_{\tau=1}^{t-1} \alpha_\tau k_{rbf}(CSI(t+h), CSI(t-\tau)) + b \tag{10}$$

where k_{rbf} represents the Kernel radial basis function:

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

where α_τ denotes the Lagrange multipliers and b denotes the bias.

3.3 Regression Trees-Based Models

The decision trees that work according to the "If-Then" rules provide ease of classification work with offering the graphical representation. The regression trees developed for numerical value estimations are used in solar radiation forecasts. The formulations of the Regression trees based methods that differ according to the application methods are as follows [3, 4]:

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

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

  where k_i denotes the constant factors, I denotes the binary return function. The RT-pruned method was operated by increasing the second-order error tolerance per node.
- Boosted and bagged regression trees (RT-boosted and RT-bagged): Boosting and bagging operations are used to improve the classical RT method.
  RT-boosted:

$$\widehat{CSI}(t+h) = \sum\nolimits_m \beta_m b\left(\widehat{CSI}(t+h), \gamma_m\right) \tag{13}$$

  where b denotes the individual trees with the split variable, γ_m and weight, β_m for each node.
  RT-bagged:

$$\widehat{CSI}(t+h) = a\vartheta_k \varphi_k\left(\widehat{CSI}(t+h)\right) \tag{14}$$

  where φ_k represents the predictors, $a\vartheta_k$ represents the mean of the predictors.
- Random forests (RF): Random forests operation is a method developed by adding randomness layer to the bagging method. Thus, the new model with robustness reduces the risk of excessive exercise.

4 Measuring the Performance and Variability of the Models

Mean absolute error (MAE), the normalized root mean squared error (nRMSE) and the skill score are used to show the accuracy of the models [5].

$$MAE = \frac{\sum_{k=1}^{N} \left| \widehat{GHI}(k) - GHI(k) \right|}{N} \tag{15}$$

where $\widehat{GHI}(k)$ and $GHI(k)$ represent the predicted and observed values with N number of data respectively.

$$nRMSE = \frac{\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\widehat{GHI}(k) - GHI(k)\right)^2}}{\overline{GHI}} \tag{16}$$

where $\overline{GHI}$ represents the average of the observed values.

$$SS = \frac{m_{frc} - m_{ref}}{m_{ref_{frc}} - m_{ref}} = 1 - \frac{nRMSE_{frc}}{nRMSE_{ref}} \tag{17}$$

where m_{frc} and m_{ref} denote the forecast and reference metrics. The skill score (SS) evaluates the performance of the models in comparison with the reference model. The mean absolute log return ($meanabs(logr)$) that is one of the statistical methods used to evaluate the variability of data sets is used for solar irradiation time series [6].

$$meanabs(logr) = \frac{\sum_{k=1}^{N} |\log(CSI(k)) - \log(CSI(k-1))|}{N} \tag{18}$$

The variability of the data set of Istanbul region is calculated as 0.1742 with the Mean absolute log return value and the variability is defined as weak.

5 Application and Evaluation

The pre-processing step takes place with the calculation of the clear sky solar radiation values using the SOLIS model, the calculation of the clear sky indices and the extraction of inappropriate data. The number of inputs used in the models is determined as 6 (Istanbul = 6) for each time series with the auto-mutual information method. This means that the number of past measurements to be used in solar radiation forecasting is 6. 80% of the dataset is defined as the training set and 20% as the testing set. The nRMSE and MAE values of the models and the forecast horizon from 1 to 6 h values are calculated (see Table 1). Differentiation of factors affecting model calculations prevents the generalization of evaluation results. However, some results are seen to be prominent: naive models generally show poor performance, ensemble learning models (bagged-RT and RF) perform well in situations involving high variability, machine learning models perform relatively well.

These inferences reveal the ability of the ensemble learning models to make solar radiation estimates with complex situations. High dataset variability is seen to affect the performance of classical models (ARMA and MLP). While the MLP, ARMA and scaled persistence models perform well, GP, SVR and RT models perform poorly for the low-variability of Istanbul's data set. Absolute error and estimation horizon are variable according to the variability of the data set.

Table 1. Comparison of forecast horizon with nRMSE and MAE values.

Methods	Horizon											
	h+1(%)		h+2(%)		h+3(%)		h+4(%)		h+5(%)		h+6(%)	
	nRMSE	MAE	nRMSE	MAE	nRMSE	MAE	nRMSE	MAE	nRMSE	MAE	nRMSE	MAE
Persistence	23.4	106.8	37.21	161.8	51.95	211.7	61.02	271.7	62.31	262	62.74	265.3
Scaled Persistence	18.63	53.4	25.45	87.8	31.42	103.1	32.84	113	42.05	109.9	41.52	133.5
ARMA	17.48	56.1	30.62	82.7	30.76	95.2	32.56	102.1	31.93	102.6	33.36	103.1
MLP	17.71	62.2	27.48	85.1	31.01	94.9	34.13	102.7	32.05	110.8	33.27	112.5
GP	18.31	61.3	30.74	103.6	34.48	112.7	35.72	106.3	32.55	111.7	33.54	117.2
SVR	17.94	63.8	38.19	112.2	46.27	124.6	43.81	125.5	43.13	133.3	45.49	145.9
RT	25.13	74.7	37.82	121.6	45.69	126.8	44.52	144.7	45.27	145.6	46.67	114.5
Pruned -RT	19.15	63.4	30.96	94.3	33.8	96.4	32.91	105.6	34.45	111.8	34.75	114.2
Boosted -RT	19.26	62.5	29.07	93.6	32.92	95.7	33.76	103.3	33.3	110.3	33.83	115.4
Bagged -RT	19.86	65.6	31.34	95.4	31.05	95.4	34.25	103.5	33.5	111.2	37.91	113.6
RF	19.78	66.8	30.67	98.9	32.36	97.5	33.28	103.9	33.79	110.9	35.22	113.8

6 Conclusion

Statistical methods and machine learning tools used for global solar radiation forecasts are discussed in this study. The described method and tools are evaluated by solar radiation forecasting of Turkey's Istanbul province. Calculations based on meteorological characteristics are used in performance measurement of models. The solar data time series measured in Istanbul region are characterized by evaluating the variability. The measurement and calculation values for Istanbul region show that the variability is low and ARMA and MLP tools are the most appropriate estimation methods. In addition, the method used for forecasting becomes complex as the variability increases. In future studies, it is aimed to make a comparison between models by applying estimation methods in regions with different meteorological and environmental characteristics, and to make generalizations about the results obtained.

References

1. 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. **27**, 65–76 (2013)
2. 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 **112**, 446–457 (2015)
3. Fouilloy, A., Voyant, C., Notton, G., Motte, F., Paoli, C., Nivet, M., Guillot, E., Duchaud, J.: Solar irradiation prediction with machine learning: forecasting models selection method depending on weather variability. Energy **165**, 620–629 (2018)
4. Iqbal, M.: An Introduction to Solar Radiation. Elsevier, Amsterdam (2012)
5. Badescu, V.: Modeling Solar Radiation at the Earth's Surface: Recent Advances (pp. 486–492). Springer, Heidelberg (2008)
6. Mueller, R.W., Dagestad, K., Ineichen, P., Schroedter-Homscheidt, M., Cros, S., Dumortier, D., Kuhlemann, R., Olseth, J.A., Piernavieja, G., Reise, C.: Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module. Remote Sens. Environ. **91**(2), 160–174 (2004)
7. Ineichen, P.: Comparison of eight clear sky broadband models against 16 independent data banks. Sol. Energy **80**(4), 468–478 (2006)
8. Parviz, R.K., Nasser, M., Motlagh, M.J.: Mutual information based input variable selection algorithm and wavelet neural network for time series prediction. In: International Conference on Artificial Neural Networks. Springer (2008)
9. 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 **90**, 671–679 (2015)
10. de Oliveira, E.M., Oliveira, F., Luiz, C.: Forecasting mid-long term electric energy consumption through bagging ARIMA and exponential smoothing methods. Energy **144**, 776–788 (2018)
11. Kalogirou, S.A.: Applications of artificial neural-networks for energy systems. Appl. Energy **67**(1–2), 17–35 (2000)
12. Burrows, W.R.: CART regression models for predicting UV radiation at the ground in the presence of cloud and other environmental factors. J. Appl. Meteorol. **36**(5), 531–544 (1997)
13. Pedro, H.T.C., Coimbra, C., David, M., Lauret, P.: Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. Renew. Energy **123**, 191–203 (2018)
14. Perez, R., Kivalov, S., Schlemmer, J., Hemker, J.K., Hoff, T.: Short-term irradiance variability: Preliminary estimation of station pair correlation as a function of distance. Sol. Energy **86**(8), 2170–2176 (2012)