Predicting Daily Mean Solar Power Using Machine Learning Regression Techniques

Faizan Jawaid[1]
[1]Karachi Institute of Economics and Technology
Pakistan
faizanj@pafkiet.edu.pk

Khurum Nazir Junejo[2]
[2]Singapore University of Technology and Design
Pakistan
khurram@pafkiet.edu.pk

ABSTRACT: *Daily mean solar irradiance is the most critical parameter in sizing the installation of solar power generation units. The average solar irradiation on a specific location can help predict the amount of electricity that will be generated through solar panels and an accurate forecast can help in calculating the size of the system, return on investment (ROI) and system load measurements. To predict the mean solar irradiation Wh/m^2 various regression algorithms have been used in conjunction with various parameters related to solar irradiance. In this paper we present a comparative analysis of forecasting through artificial neural networks (ANN) against the standard regression algorithms. Furthermore, we show that incorporation of azimuth and zenith parameters in the model significantly improves the performance.*

Keywords: Solar Power Generation, Artificial neural networks

Received: 14 May 2016, Revised 29 June 2016, Accepted 3 July 2016

1. Introduction

With the dwindling fossil fuel resources, research in renewable energy has gained significant impetus. The leading source for renewable energy is solar power generation mainly through photovoltaic cells. Advantages of using solar energy include its immunity to imitative circumstances like the oil prices, a clean source of energy, and reduction of imports and dependability on external resources. Though photovoltaic cells are considered as a major source for future energy generation, their return on investment and upfront cost is hindering their deployments. One of the reasons for this is lack of predictable supply because of changing weather conditions. Since photovoltaic cells generate electricity by converting solar energy to electric current, the amount of solar energy being provided in a day is very important to size the photovoltaic system. Therefore, the amount of electricity produced depends upon solar irradiance in a particular day, which itself depends on various parameters such as location, time, and weather patterns. Solar irradiance is defined as the power per unit area received from the Sun in the form of electromagnetic radiation in the wavelength range of the solar cell being used.

Unfavourable weather reduces the output of the solar plant to a large extent. Therefore, in order to fulfil the energy requirements, a power supply company needs to supplement the remaining amount by purchasing power from different power generation

companies running on costlier fossil fuels. The power rates charged by these companies not only depend upon the amount of the power required but also on the timeliness of the order. A timely order placement with these companies not only helps to meet the promised power supply goals, but also helps reduce the cost. Hence prior knowledge of the power produced plays a crucial role in maintaining quality of service and reducing cost. With solar power, its possible to predict the production knowing current and the past information about the weather and the irradiance.

Various researchers have proposed forecasting mechanisms with good results however a room for improvement still exists. There are two orthogonal avenues of improvement in this domain, one is more efficient algorithm design for forecasting and second is identification and quantification of the effect of parameters on forecast. In this paper we attempt to improve the state of the art in both the dimensions. Our first contribution is evaluation of forecast of solar irradiance using verity of machine learning regression algorithms. Artificial neural network (ANN) in particular achieve a very high degree of accuracy. Our second contribution is the use of solar angles (Azimuth and Zenith) in conjunction with the weather data. We show that these two angles help improve the prediction performance. Particularly, we try to predict the mean daily solar energy Wh/m^2 that can be used by a solar plant at a given location based on the weather forecast. To accomplish this we consider various weather parameters which include, but not limited to, mean temperature, wind speed, visibility etc. We put together a 10 years of observed weather and irradiance data from Los Angeles, California Zip 722950 for this study. We evaluate the learned models cross-validation, and split validation.

In Section 2 we review the previous research work conducted on solar energy prediction. We discuss our methodology in Section 3. Section 4 pertains to the results based on two different validations techniques, while Section 5 concludes our work.

2. Related Work

The current eras' accelerating advances in all fields rely on the constant supply of the electric power. The fossil fuels are key source of the reliable and consistent energy provision, but they have high cost associated with them, emit dangerous gases, and are subject to uncertainty of the international oil prices. On the other hand, solar power is cheap, clean source of energy that can be produced by every country, but its dependence on weather conditions make it less reliable. To cope with this unreliability the prediction of the solar power contributes towards a consistent supply. The prediction of solar power is a multidisciplinary research that needs contribution from meteorology, solar cell engineering, electrical engineering, and machine learning computation.

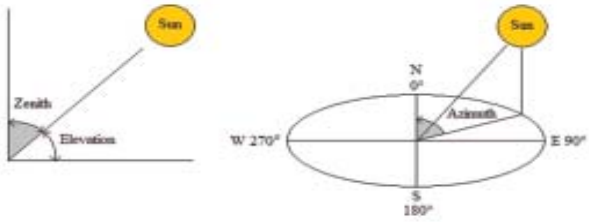

Figure 1. Solar azimuth and zenith angles

[1] used the radial basis function (RBF) to model the solar radiation based on sunshine duration and air temperature data. They use Multi-Layer Perceptron (MLP) to model the hourly forecast of the solar radiation using present values of air temperature and mean daily solar irradiance. [2] compare different models including Fuzzy Logic, Neural Networks and Autoregressive to predict the half daily values of solar radiation. In [3], 3-day forecast of the solar irradiance has been modelled using the forecast data provided by European Centre for Medium-Range Weather Forecasts (ECMWF). In [4], different weather parameters such as solar hours, latitude, longitude, elevation, maximum and minimum air temperature,humidity, and rainfall has been used to predict the solar power in Indonesia. It also used ANN to generate a model that could predict the solar power of a particular region. A recent addition is the use of computer algorithm based on fuzzy logic control (FLC) to estimate the wind and solar energies in a hybrid renewable energy system from natural factors [5]. The solar power was estimated using the temperature and the lighting as input parameters.

A detail introduction to the current research on forecasting solar irradiance is presented in [6]. It is a in-depth review to facilitate selection of the appropriate forecast method according to needs. They also comment on the statistical approaches and techniquesbased on images from satellite imagery. They also discuss numerical weather prediction (NWP) and hybrid models.

Online forecasting of power production from PV systems is discussed in [7], to predict hourly values of solar power for horizons of up to 36 h. A two-stage method of using adaptive linear time series models for clear sky model and autoregressive (AR) models NWPs is elaborated. [8] compare multiple regression techniques for generating prediction models for solar power, including linear least squares, and support vector machines using multiple kernel functions.

3. Methodology

3.1 Dataset

Reliable data availability and the choice of right attributes from this data are crucial to the accurate prediction in general and for solar power in specific. In this work we relied on the dataset provided by National Solar Radiation Database (NSRD)[1] and National Climate Data Center (NCDC)[2]. The former is provided by the National Renewable Energy Laboratory (NREL) which collects data for various resources including solar related energy for all cities of USA. The later is now known to be managed by National Oceanic and Atmospheric Administration(NOAA).

We chose 10 year weather information dataset over Los Angeles, California ZIP 722950 from above stated two different sources. The dataset available is based on hourly values of weather parameters. Average of the 24-hour data has been taken to convert it to the mean values per day. Different weather parameters have been collected from 1991 to 2000 to explore the relationship between the mean solar irradiance and weather data for the accurate prediction of mean solar irradiance. The collected data comprises of the average daily values of air temperature, min. air temperature, max. air temperature, wind speed, dew-point, visibility, solar azimuth angle, and solar zenith angle. The solar azimuth angle gives the direction of the sun measured in degrees. Whereas, solar zenith tells how high the sun is. It is also measured in degrees. The figure 1 shows the solar azimuth and zenith angles. To increase the prediction accuracy, month and week of the year is also made part of the dataset. Figure 2 shows the pattern of different weather parameters measured on daily basis for Year 2000. These are the mean values of hourly data measured on daily basis.

3.2 Regression Modelling for Prediction

We have tested four commonly used machine learning techniques in this work with the selected dataset to evaluate individual performances with the choice of weather attributes. The first one is K-Nearest Neighbor(K-NN) approach [9]. KNN is a lazy learning approach i.e. it does not explicitly build a model over the training data, instead, given a unseen record from the test sample, it finds its closest K matches in the training data. Since our prediction variable is continuous valued, mean of these K closest matches is predicted as the output of the unseen test sample. We experimented with different values of K but report results for only when K=3 and K=5. RMS error increases for values of K>3.

Linear Regression (LR) is the most basic and widely used technique for regression [10]. It models the relationship between the input and output variables using linear predictor functions whose unknown model parameters are estimated from the data using a least squares approach. The parameter values can be estimated either by solving a set of linear equations or using an iterative method such as gradient descent.

Support Vector Machines (SVM) are non-probabilistic binary linear classifiers. They project the data in a higher dimensional feature space by means of a kernel function that separates the points. A hyperplane is learned in this feature space to discriminate between the points of the two classes. The hyperplane is such that it maximizes the margin between the closest points of the two classes, thus achieving a better generalization over the unseen data. This has led SVM to exhibit high performance on most of the real world classification problems [11]. [12], [13] propose a method to use SVM for regression. After projecting the data into a higher dimensional feature space a linear regression is performed.

Artificial Neural Network (ANN) are inspired by the functionality of neurons in animals. A neuron is a processing unit having output and inputs and an activation function. There are many variations of the ANN, for this work we have used the simple feed forward neural network with back-propagation [14]. The weather parameters are given as inputs to the ANN and it gives the predicted output as the solar power. The neuron functions are learned iteratively using training dataset. For the ANN model, learning cycles are 1000 while learning rate is 0.2 and momentum is 0.1. It is observed that as the number of learning cycles were increased, the RMS error decreased. Increasing the learning rate and momentum effected the RMS error as well.

[1] http://rredc.nrel.gov/solar/old data/nsrdb/1991-2010/hourly/list by state.html
[2] ftp://ftp.ncdc.noaa.gov/pub/data/gsod/

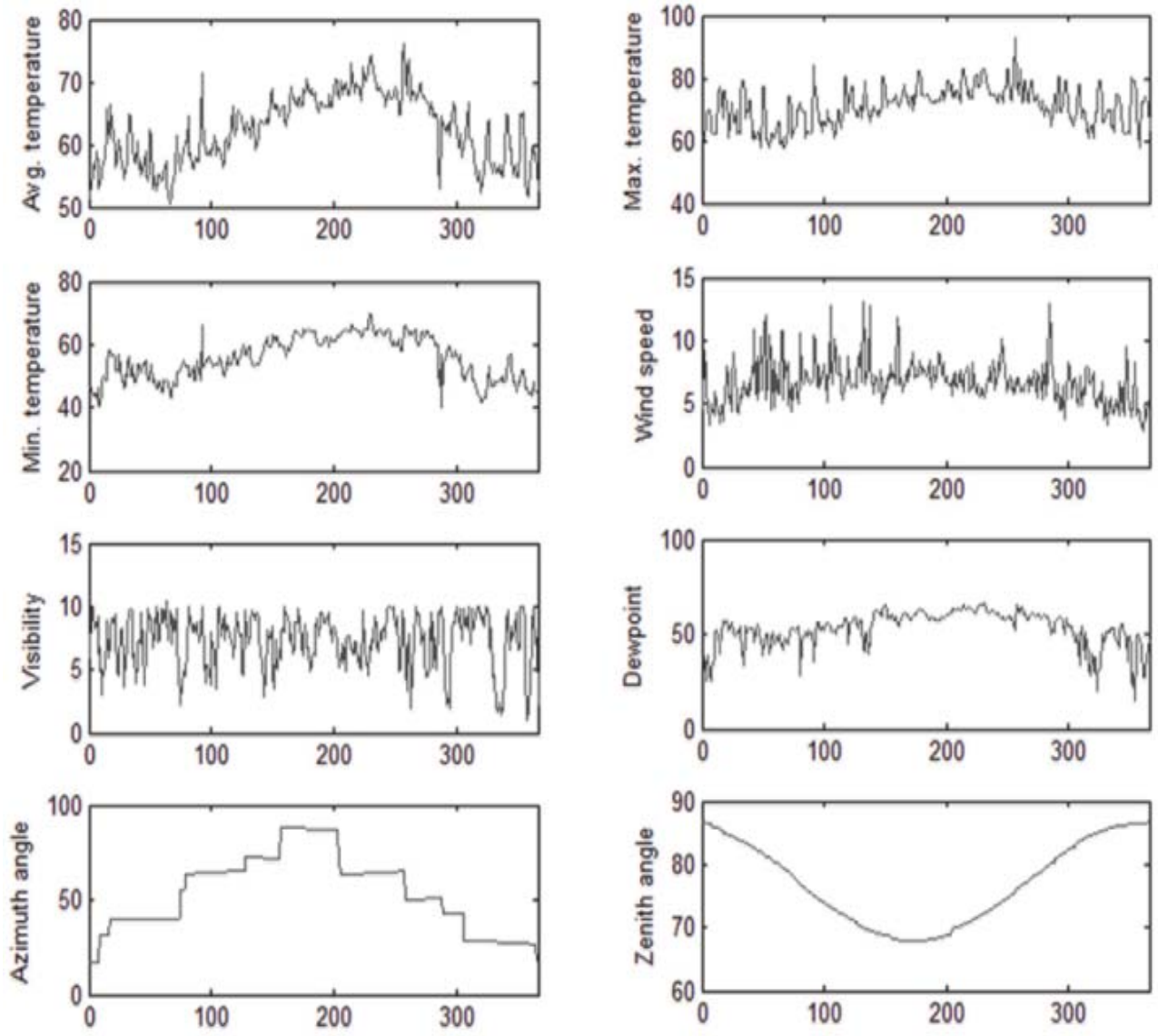

Figure 2. Mean daily values of weather parameters of the Year 2000, X-Label=Day of Year 2000

The discussed techniques are designed and tested with RapidMiner [15]. The dataset is split in testing and training dataset, the techniques used to split along with the results are discussed in 3.3.

3.3 Validation

To better understand and evaluate the trained models, two different validation techniques are used namely, K-fold Cross-Validation (CV) and Split Validations (SV). At first, the CV technique is used which orderly divides the data into K segments. The training sequence is performed on K-1 segments and testing is done on the remaining segment. This whole cycle is repeated K times. Each validation cycle yields a root means squared (RMS) error value. The final result is an average of these K RMS values. RMS error [3] is defined as follows:

$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(model_i - observed_i)^2} \quad (1)$$

where n=10 years * 365 = 3650 instances. Due to the chronological nature of the data, we also gather results using split validation (SV) approach. Therefore we divide the first nine years of data (years from 1991 to 1999) for training while testing only on the last one year of data (year 2000) (Figure 3). This scenario is more realistic as well as more challenging as the distribution of the data in the training data might be different than that of the test data because of the concept shift that occurs over time.

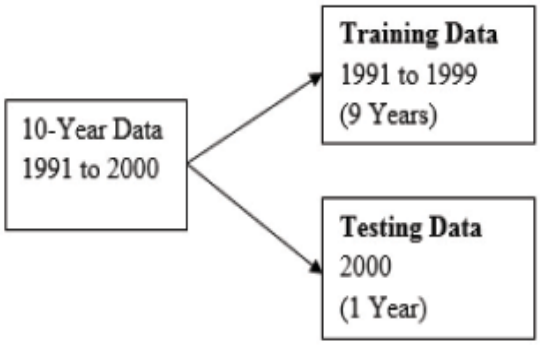

Figure 3. Split validation of the dataset

4. Results And Discussion

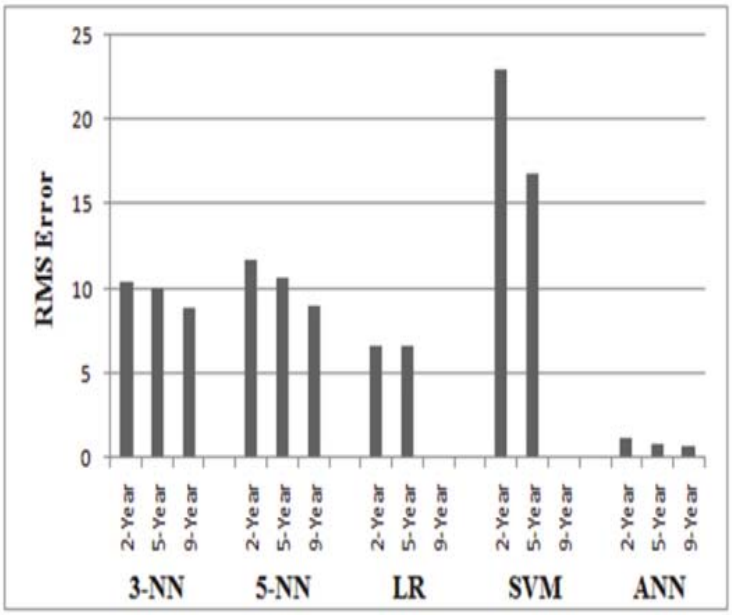

Figure 4. RMS Error of different models

As discussed previously, two validation methods, K-fold cross-validation and split-validation were used, and we present their results separately. In order to ascertain how much training data is sufficient for achieving a good prediction, we vary the size of the data used. We used three sizes of the data, i.e. two, five, and nine years. Figure 4 compares the performance of the four models using the K-fold cross validation approach on these three dataset sizes. It can be observed that ANN outperforms the rest of the three models by a sufficient margin. Secondly, increasing the size of the data decreases the RMS error. It might be noted that RMS error reported here for daily mean solar power Wh/m^2 , is an overestimate of the average difference of the predicted value from the actual value. This is because RMS error metric is sensitive to outliers because it squares the difference of the predicted and the observed value. Therefore, the actual average difference between the observed and predicted value is

lesser. RMS error is used because of its nice convergence properties. The result for LR, and SVM 9-year are missing because of the scalability issues.

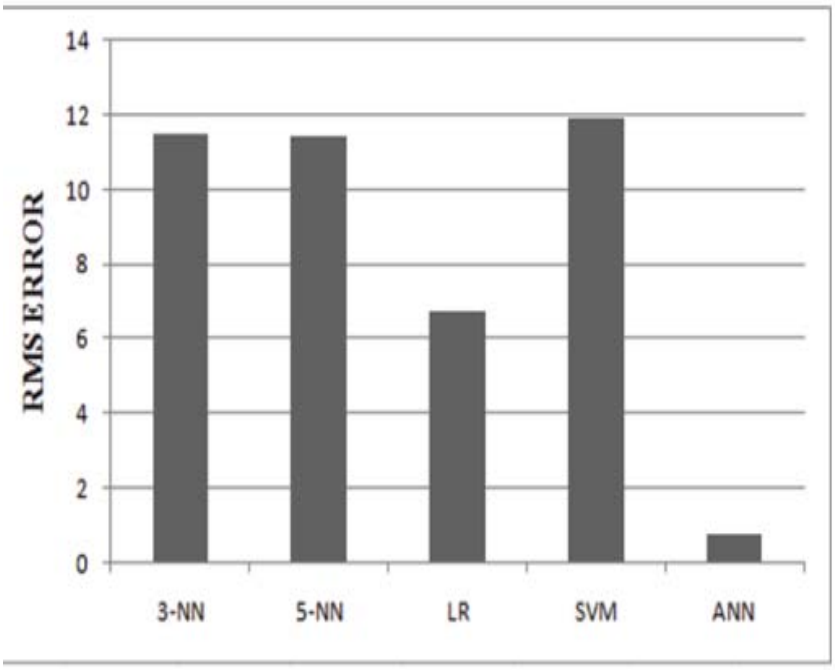

Figure 5. RMS Error of different models using SV

Figure 5 shows the results for the split validation approach. The figure shows the result on the one year data used in the test set. The results for SV tell a similar story. There is no significant difference in the errors from CV approach. This indicates that there were no significant change in weather pattern for the Los Angeles area over the ten years.

4.1 RMS Error Comparison

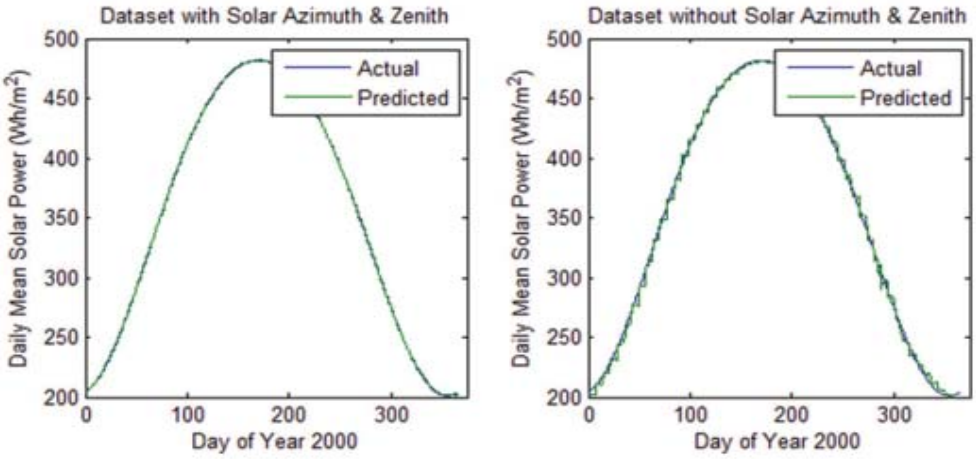

Figure 6. Predicted Vs Observed (Actual) Daily Mean Solar Power

We plot out the predicted values against the actually observed values using the SV approach in Figure 7. K-NN overestimates the daily mean solar power deviation for the first half of the year 2000, and underestimates it for the second half of the year.

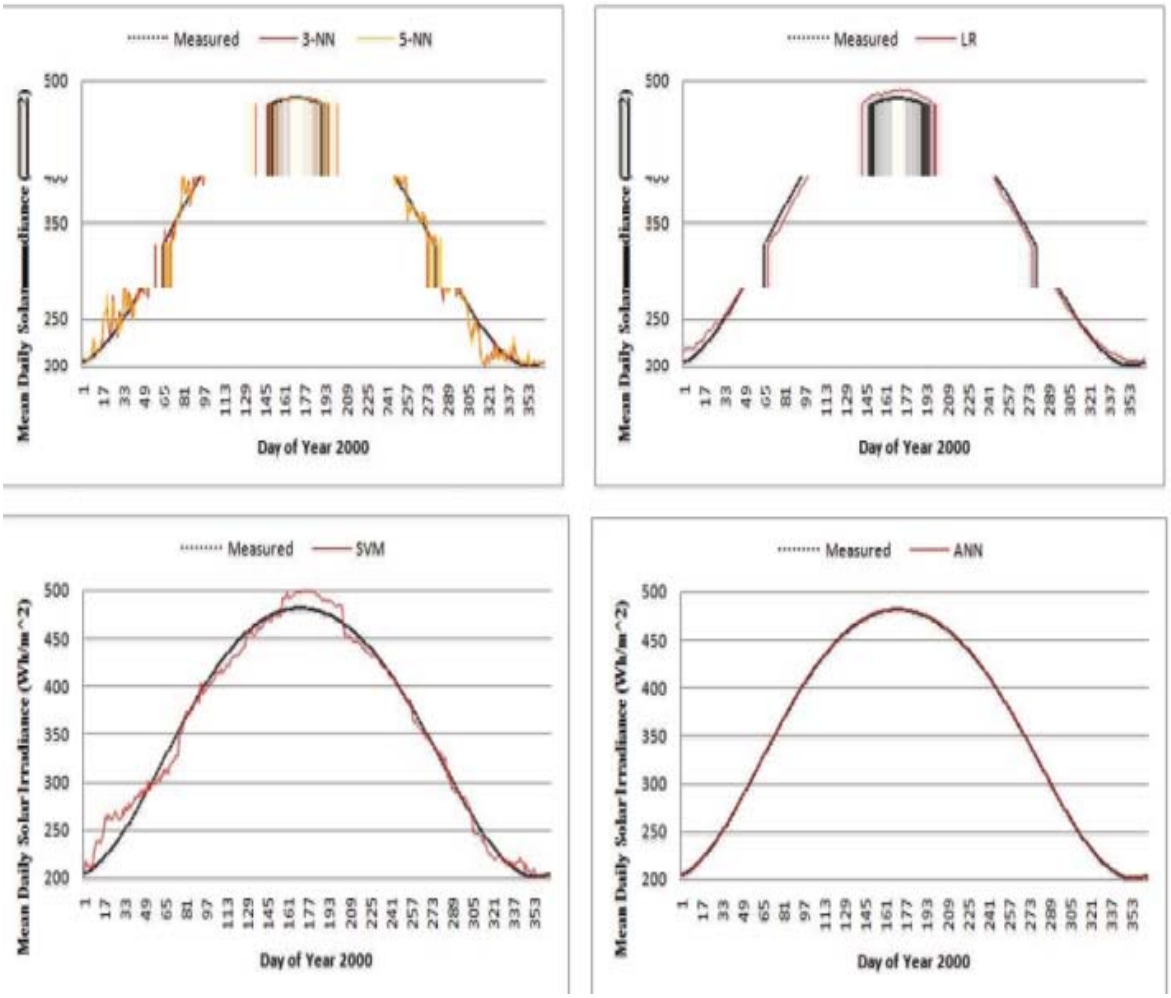

Figure 7. Predicted vs Observed Daily Mean Solar Power using SV

Whereas the prediction of SVM remains fluctuating in the first half but predicts better in the second half. LR, overall, performs better than K-NN, and SVM. It also exhibits a lesser deviation than the two. ANN on the other hand fits the observed curve almost perfectly.

Finally, we try to determine the most important attribute in prediction of the daily mean solar power. For this purpose we excluded the attributes one by one from the dataset and relearned the whole model. The maximum drop in the performance of the model were because of the exclusion of solar azimuth and zenith. Figure 6 shows the difference in the performance of ANN using the SV approach with and without these two attributes.

5. Conclusion

Over the past years solar power has gained a significant importance as a renewable, clean, and alternative source of energy. However, the output of a solar power plant heavily depends on the weather patterns and time of the year. An accurate prediction of daily solar power by modelling these varying patterns can help solar power companies complement their production shortfalls in a timely manner, thus saving high cost of purchasing power from the market at the last moment. We therefore, demonstrate the ability of machine learning regression techniques, especially artificial neural networks to accurately predict the daily mean solar power up to a very high degree of accuracy. In addition to using weather and time of year specific attributes, we also propose and establish the effectiveness of using solar angles (azimuth and zenith) to increase the accuracy of such predictions. We validated our models on 10 years of solar

irradiance and weather data collected for the Los Angeles city, USA. This study is a first step in suggesting sites that are most suitable for solar power generation.

References

[1] Mellit, A., Menghanem, M., Bendekhis, M. (2005). Artificial neural network model for prediction solar radiation data: application for sizing stand-alone photovoltaic power system, *In*: Power Engineering Society General Meeting, 2005. IEEE, June 2005, p. 40–44 Vol. 1.

[2] Martn, L., Zarzalejo, L. F., Polo, J., Navarro, A., Marchante, R., Cony, M. (2010). Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning, *Solar Energy*, 84 (10) 1772 – 1781. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X10002379

[3] Lorenz, E., Hurka, J., Heinemann, D., Beyer, H (2009). Irradiance forecasting for the power prediction of grid-connected photovoltaic systems, *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 2 (1) 2–10, March 2009.

[4] Prastawa, A., Dalimi, R (2013). New approach on renewable energy solar power prediction in indonesia based on artificial neural network technique: Southern region of sulawesi island study case, *In*: QiR (Quality in Research), 2013 International Conference on, June 2013, p. 166–169.

[5] Faquir, S., Yahyaouy, A., Tairi, H., Sabor, J (2015). Implementing a fuzzy logic based algorithm to predict solar and wind energies in a hybrid renewable energy system, *Int. J. Fuzzy Syst. Appl.*, 4 (3) 10–24, Jul. 2015. [Online]. Available: http: // dx.doi.org/10.4018/IJFSA.2015070102

[6] Diagne, M., David, M., Lauret, P., Boland, J., Schmutz, N (2013). Review of solar irradiance forecasting methods and a proposition for small-scale insular grids, *Renewable and Sustainable Energy Reviews*, V. 27, p. 65 – 76, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032113004334

[7] Bacher, P., Madsen, H., Nielsen, H. A (2009). Online short-term solar power forecasting, *Solar Energy*, 83 (10) 1772 – 1783. [Online]. Available: http://www.sciencedirect.com/science/article/ pii/S0038092X09001364

[8] Sharma, N., Sharma, P., Irwin, P.E., Shenoy, P.J (2011). Predicting solar generation from weather forecasts using machine learning. *In*: SmartGridComm. IEEE, 2011, pp. 528–533. [Online]. Available: http://dblp.uni-trier.de/db/conf/smartgridcomm/ smartgridcomm2011.html#SharmaSIS11

[9] Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, The American Statistician, 46 (3) 175–185, 1992. [Online]. Available: http://dx.doi.org/10.2307/2685209

[10] Montgomery, D. C., Peck, E. A., Vining, G. G. (2015). Introduction to linear regression analysis. John Wiley & Sons.

[11] Ahmad, I., Hussain, M., Alghamdi, A., Alelaiwi, A (2014). Enhancing svm performance in intrusion detection using optimal feature subset selection based on genetic principal components, *Neural Computing and Applications*, 24 (7-8) p. 1671–1682.

[12] Smola, A., Vapnik, V (1997). Support vector regression machines, Advances in neural information processing systems, vol. 9, p. 155–161, 1997.

[13] Basak, D., Pal, S., Patranabis, D. C.(2007). Support vector regression, *Neural Information Processing-Letters and Reviews*, 11 (10) 203–224.

[14] Rumelhart, D. E., Hinton, G. E. Williams, R. J. (1998). Neurocomputing: Foundations of research, J. A. Anderson and E. Rosenfeld, Eds. Cambridge, MA, USA: MIT Press, 1988, ch. Learning Representations by Back-propagating Errors, p. 696–699. [Online]. Available: http://dl.acm.org/citation.cfm?id=65669.104451

[15] Hofmann, M., Klinkenberg, R. (2013). RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC.