**SOLAR RADIATION PREDICTION USING SARIMAX AND ANN(Hybrid Model)**

**Minor Project II**

Submitted by:

**Sanjeev Kumar (9917102033)**

**Aparajit Verma (9917102050)**

**Shuddhatm Jain (9917102064)**

Under the supervision of:

**Mrs. Deeksha Chandola**



**Forecasting Solar radiation using hybrid ARIMA-ANN model**

**Abstract:** Solar radiation prediction has a great importance in electricity generation from solar energy and helps to size photovoltaic power systems. Therefore, the Global Horizontal Irradiance (GHI) was predicted at 1-hour duration in this paper.

Direct Normal Irradiance (DNI), Direct Horizontal Irradiance (DHI), Temperature, Pressure, Relative Humidity, Dew Point, Wind Direction and Wind Speed parameters were used as atmospheric input variables for **time series model** by **Auto Regressive Integrated Moving Average(ARIMA) model**. ARIMA is one of the popular linear models in time series forecasting.

As time series also have non linearity hence, a hybrid model is made afterwards using residuals of ARIMA as input for Artificial Neural Networks. Statistical error measures such as the mean error (ME), the mean square error (MSE) and the mean absolute error (MAE) were calculated to compare the two methods.

The results showed that the Hybrid models predict the solar radiations with a higher accuracy than the ARIMA model in the four examined sites.

Keywords: SARIMAX; Box–Jenkins methodology; Artificial neural networks; Time series forecasting; Combined forecast, hybrid SARIMAX-ANN

Table of contents

Title Page No.

Acknowledgement 2

Abstract 3

List of tables/figures 5

Abbreviations and Nomenclature 6

Chapter 1: Introduction 7

Chapter 2: Software Analysis 10

2.1: Software

2.2: Hardware

2.3: Functional Requirements

2.4: Non-Functional Requirements

Chapter 3: Detailed Design 11

4.1: SARIMAX

4.2: Artificial Neural Network

4.3 Hybrid model

Chapter 5: Observation 12

Chapter 6: Implementation 13

Chapter 7: Testing Reports 14

Chapter 8: Conclusion 17

Chapter 9: The proposed work plan for the remaining period 18

Chapter 10: References 19

Chapter 11: PowerPoint Presentation 20

**ABBREVIATIONS AND NOMENCLATURE**

SARIMAX: Seasonal Auto Regressive Integrated Moving Average with Exogenous variable

ANN: Artificial Neural Networks

RMSE: Root Mean Squared Error

MAAPE: Mean Absolute Arc-tangential Error

**Chapter 1: Introduction**

Solar radiation prediction is an important problem with direct applications in renewable energy. Solar is one of the most important green sources of energy, that is currently under expansion in many countries of the world, especially in those with more solar potential, such as Rajasthan. An accurate estimation of the energy production in solar energy systems involves the accurate prediction of solar radiation, depending on different atmospheric variables. We have used humidity Direct Normal Irradiance (DNI), Direct Horizontal Irradiance (DHI), Temperature, Pressure, Relative Humidity, Dew Point, Wind Direction and Wind Speed as inputs ([Khatib](https://www.sciencedirect.com/science/article/abs/pii/S1364032112000767) [1]et al., 2012; [Inman](https://www.sciencedirect.com/science/article/pii/S0360128513000294) [2]et al., 2013; [Sozen](https://www.sciencedirect.com/science/article/pii/S0196890404000172) [3]et al., 2004; [Voyant](https://www.sciencedirect.com/science/article/abs/pii/S0360544210005955) [4]et al., 2011).

In recent years, several works have been developed to try to predict solar radiation using machine learning techniques and environmental parameters. They used different input geographical and atmospheric parameters like latitude, longitude, temperature, wind speed and direction, daily global irradiation, sunshine duration or precipitation ([Mellit and Kalogirou](https://www.sciencedirect.com/science/article/pii/S0360128508000026)[5] ,2008; [Mubiru](https://www.sciencedirect.com/science/article/abs/pii/S0960148108000074)[6], 2008). According to [Bilgili and Ozoren](https://link.springer.com/article/10.1007%2Fs00703-011-0137-9)[7] (2011), sunshine duration, air temperature and relative humidity are the most widely used meteorological parameters to predict daily solar radiation and its components

An initial category to distinguish is time-series models, which can be further divided into Box and Jenkins techniques, smoothing techniques, Kalman-ﬁltering theory, and spectral analysis. Early applications of Box and Jenkins techniques in the ﬁeld of traffic forecasting were implemented by [Ahmed and Cook](https://trid.trb.org/view/148123)[8] and [Nihan and Holmesland](https://sci-hub.se/https:/link.springer.com/article/10.1007/BF00167127)[9]. More advanced techniques have been applied recently, including autoregressive integrated moving average (ARIMA) models with intervention x-variables (ARIMAX) by [Williams, B. M., and L. A. Hoel](https://sci-hub.se/https:/ascelibrary.org/doi/abs/10.1061/(ASCE)0733-947X(2003)129:6(664))[10], seasonal ARIMA models (SARIMA) by [Williams, B. M](https://sci-hub.se/https:/ascelibrary.org/doi/abs/10.1061/(ASCE)0733-947X(2003)129:6(664))[11], seasonal ARIMA with intervention x-variables(SARIMAX) by [SI Vagropoulos, GI Chouliaras](https://sci-hub.se/https:/ieeexplore.ieee.org/abstract/document/7514029/)[12].

Neural network models are the second category of techniques that can be identiﬁed. [Y Jiang](https://sci-hub.se/https:/www.sciencedirect.com/science/article/pii/S0301421508003133)[13] was among the first who applied ANN on solar radiation prediction. [Amit Kumar Yadav , Hasmat Malik](https://sci-hub.se/https:/www.sciencedirect.com/science/article/pii/S1364032113008228)[14] wrote about the most relevant input parameters for artificial neural network based solar radiation prediction models.

A suitable combination of linear and nonlinear models provides a more accurate prediction modelthan an individual linear or nonlinear model for forecasting time series data originating from various applications. The linear autoregressive integrated moving average (ARIMA) and nonlinear artificial neural network (ANN) models are explored in this paper to devise a new hybrid ARIMA–ANN model for the prediction of time series data. The results obtained from all of these data sets show that for both one-step-ahead and multistep-ahead forecasts, the proposed hybrid model has higher prediction accuracy.

A hybrid ARIMA–ANN model was proposed by [Zhang](https://www.sciencedirect.com/science/article/abs/pii/S0925231201007020?via%3Dihub) [15], which was shown to give more accurate predictions than the individual models. On Wolf’s sunspot data, Canadian lynx data, and exchange rate time series data, this hybrid model was shown to outperform individual ARIMA and ANN models in the case of one-step-ahead prediction. Another hybrid ARIMA–ANN method was proposed by [Khashei and Bijari](https://www.sciencedirect.com/science/article/abs/pii/S1568494610002759?via%3Dihub) [16], which was shown to give better performance for one-step-ahead forecasting than the method proposed by Zhang. The hybrid method proposed by Zhang was also used for [electricity price forecasting](http://dx.doi.org/10.1049/iet-gtd.2012.0263) in [17] and for [water quality time series prediction](https://www.sciencedirect.com/science/article/abs/pii/S0952197609001390?via%3Dihub) in [18].

For error detection,we have used MAAPE and RMSE values. The mean absolute percentage error (MAPE) is one of the most widely used measures of forecast accuracy, due to its advantages of scale-independency and interpretability. However, MAPE has the significant disadvantage that it produces infinite or undefined values for zero or close-to-zero actual values. In order to address this issue in MAPE, we have used a new measure of forecast accuracy called the mean arctangent absolute percentage error (MAAPE) by [Sungil Kima and Heeyoung Kim](https://sci-hub.se/https:/www.sciencedirect.com/science/article/pii/S0169207016000121)[19]. MAAPE has been developed through looking at MAPE from a different angle. In essence, MAAPE is a slope as an angle, while MAPE is a slope as a ratio, considering a triangle with adjacent and opposite sides that are equal to an actual value and the difference between the actual and forecast values, respectively. MAAPE inherently preserves the philosophy of MAPE, overcoming the problem of division by zero by using bounded influences for outliers in a fundamental manner through considering the ratio as an angle instead of a slope.

Chapter 2: Requirement Analysis

3.1 Software Requirements

3.1.1 Library used:

• Numpy

• Pandas

• OpenCV

• Keras

• Sklearn

3.1.2 Other Requirements:

• Anaconda Platform (spyder, Jupyter Notebook)

• Python 3 or higher

3.2 Hardware Requirements

• Microsoft Windows 10

• Processor: Intel ® Core (TM) i5 -6200U CPU @2.30GHz 2.40GHz

• Ram : 4 GB and above

• Disk Space : 1 TB

3.3 Functional Requirements

• Appropriate data set to work on.

• An eligible software to implement our ideas.

• Suitable libraries for using algorithm in the source code.

3.4 Non-functional Requirements

• Validation: Applying benchmark functions for checking the accuracy and performance of algorithms.

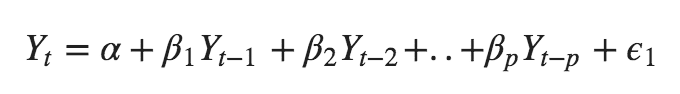
**SEASONAL Autoregressive Integrated Moving Average**

**process (SARIMAX) :-**

SARIMAX is one of the most traditional methods of non-stationary time series analysis. In contrast to the regression models, the ARIMA model allows rt to be explained by its past, or lagged values and stochastic error terms. A SARIMAX model is usually

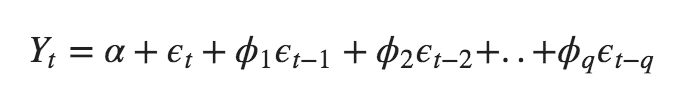
stated as SARIMAX (p, d, and q).

A pure **Auto Regressive (AR only) model** is one where Yt depends only on its own lags. That is, Yt is a function of the ‘lags of Yt’.

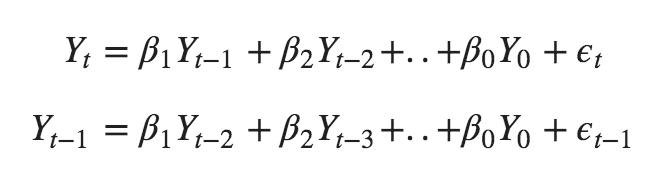
[](https://www.machinelearningplus.com/wp-content/uploads/2019/02/Equation-1-min.png)

where, Y*{t-1} is the lag1 of the series, beta*1 is the coefficient of lag1 that the model estimates and alpha is the intercept term, also estimated by the model.

Likewise a pure **Moving Average (MA only) model** is one where Yt depends only on the lagged forecast errors.

[](https://www.machinelearningplus.com/wp-content/uploads/2019/02/Equation-2-min.png)

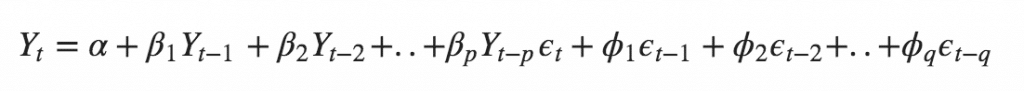
where the error terms are the errors of the autoregressive models of the respective lags. The errors Et and E(t-1) are the errors from the following equations :

[](https://www.machinelearningplus.com/wp-content/uploads/2019/02/Equation-3-min.png)

That was AR and MA models respectively.

So what does the equation of an ARIMA model look like?

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. So the equation becomes:

[](https://www.machinelearningplus.com/wp-content/uploads/2019/02/Equation-4-min.png)

**SARIMAX model in words:**

Predicted Yt = Constant + Linear combination Lags of Y (upto p lags) + Linear Combination of Lagged forecast errors (upto q lags) .

**Construction Model :-**

ACF= Autocorrelation Function

PACF= Partial Autocorrelation Function

Exogenous Variable

**ACF**

The autocorrelation function (**ACF**). Intuitively, a stationary **time series** is defined by its mean, variance and **ACF**. A useful result is that any function of a stationary **time series** is also a stationary **time series**.

**PACF**

In **time series** analysis, the partial autocorrelation function (**PACF**) gives the partial correlation of a **time series** with its own lagged values, controlling for the values of the **time series** at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags.

**Terminologies in SARIMAX**

ARIMA model can be (almost) completely summarized by three numbers:

**p = the number of autoregressive terms**

p is the number of autoregressive terms (AR part). It allows to incorporate the effect of past values into our model. Intuitively, this would be similar to stating that it is likely to be warm tomorrow if it has been warm the past 3 days.

**d = the number of nonseasonal differences**

d is the number of nonseasonal differences needed for stationarity. Intuitively, this would be similar to stating that it is likely to be same temperature tomorrow if the difference in temperature in the last three days has been very small.

**q = the number of moving-average terms**

q is the number of lagged forecast errors in the prediction equation (MA part). This allows us to set the error of our model as a linear combination of the error values observed at previous time points in the past.

These are the three integers (p, d, q) that are used to parametrize ARIMA models. Hence, this is called an “ARIMA (p, d, q)” model.

**Exogenous Variables :-**

An **exogenous variable** is one whose value is determined outside the model and is imposed on the model. In other words, **variables** that affect a model without being affected by it. Read more about **exogenous variables** here. Many models can be used to solve a task like this, but **SARIMAX** is the one we'll be working with.

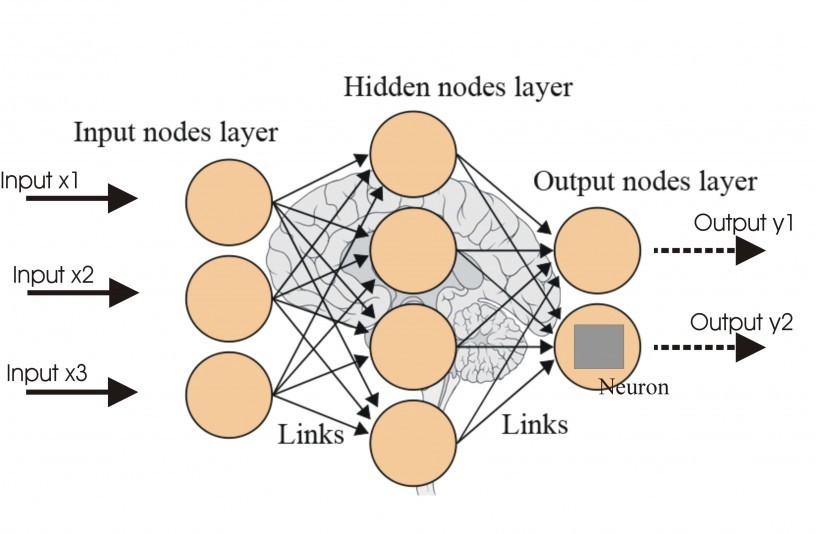
**Seasonality :-**

Seasonality in a time series is a regular pattern of changes that repeats over S time periods, where S defines the number of time periods until the pattern repeats again.

In a seasonal ARIMA model, seasonal AR and MA terms predict xtusing data values and errors at times with lags that are multiples of *S* (the span of the seasonality).

* With monthly data (and *S* = 12), a seasonal first order autoregressive model would use **xt−12** to predict  xt . For instance, if we were selling cooling fans we might predict this August’s sales using last August’s sales. (This relationship of predicting using last year’s data would hold for any month of the year.)
* A seasonal second order autoregressive model would use  **xt−12** and  **xt−24** to predict xt. Here we would predict this August’s values from the past two Augusts.
* A seasonal first order MA(1) model (with *S* = 12) would use  **wt−12** as a predictor. A seasonal second order MA(2) model would use  **wt−12** and**wt−24 .**

**ANN (Artificial Neural Network) :-**



ANNs Approach to Time-Series Modeling ANNs are flexible computing frameworks for modeling a broad range of nonlinear problems. One significant advantage of the ANN models over other classes of nonlinear model is that ANNs are universal approximators that can approximate a large class of functions with a high degree of accuracy. Their power comes from the parallel processing of the information from the data. No prior assumption of the model form is required in the model building process. Instead, the network model is largely determined by the characteristics of the data. The network architecture of ANNs consists of the three-layer fully connected feed forward neural network. The model is characterized by a network of three layers of simple processing units connected by acyclic links. The relationship between the output 𝑦𝑡 and the inputs **(𝑦𝑡−1 , 𝑦𝑡−2 , ⋯ , 𝑦𝑡−𝑝)** has the following mathematical representation as in Eq. (5).

**𝑦𝑡 = ∝0+ ∝𝑗 𝑔 (𝛽0𝑗 + 𝛽𝑖𝑗 𝑦𝑡−1 𝑝 𝑖=1 𝑞 𝑗=1) +∈𝑡**

**∝𝑗 (𝑗 = 1,2,3, … , 𝑞) and 𝛽𝑖𝑗 (𝑖 = 1,2,3, . . ; = 1,2,3, … , 𝑞)** are the model parameters, often called the connection weights, p is the number of input nodes, and q is the number of hidden nodes. The logistic function is often used as the hidden-layer transfer function, and is given by Eq. (6).

**𝑔 𝑥 = 1 /1 + exp −𝑥 6 .**

Hence, the ANN model of Eq.(2), in fact, performs a nonlinear functional mapping from the past observations 𝑦𝑡−1 , 𝑦𝑡−2 , ⋯ , 𝑦𝑡−𝑝 to the future value 𝑦𝑡 , given by Eq. (7).

**𝑦𝑡 = 𝑓 𝑦𝑡−1 , 𝑦𝑡−2 , ⋯ , 𝑦𝑡−𝑝 , 𝜔 +∈𝑡**

Where 𝜔 is a vector of all parameters, and 𝑓 𝑦𝑡−1 , 𝑦𝑡−2 , ⋯ , 𝑦𝑡−𝑝 , 𝜔 is a function determined by the network structure and connection weights and ∈ is the random error.

The choice 𝑞 of is datadependent and there is no systematic rule in deciding this parameter. In addition to choosing an appropriate number of hidden nodes, another important task of ANN modeling of a time series is the selection of the number of lagged observations p and the dimension of the input vector. This is perhaps the most important parameter to be estimated in an ANN model because it plays a major role in determining the (nonlinear) autocorrelation structure of the time series. However, there is no theory that can be used to guide the selection of p.

**Hybrid Model** :-

ARIMA and ANN are good in linear and nonlinear domains but none of them is a universal model that is suitable for all circumstances. The approximation of ARIMA model for complex nonlinear problems may not yield good results similarly, ANN will give mixed results in the linear domain. Since it is difficult to completely know the characteristics of the data in a real problem, a hybrid methodology that has both linear and nonlinear modeling capabilities can be a good strategy for practical use. By combining different models, different aspects of the underlying patterns may be captured.

Practically a time series may be considered to have a linear component and a non-linear component as shown in the Eq. (8).

**𝑦𝑡 = 𝑙𝑡 + 𝑛𝑡**

Where 𝑙𝑡 is the linear part and 𝑛𝑡 is the nonlinear part. The two parts can be estimated separately from the data. First the linear part is determined that is seperated from the time series. This residual is fitted with the ANN model . This is fundamental strategy behind this model. Let 𝑒𝑡 be the residuals which can be obtained from the time series by substracting forecasted value 𝑙 𝑡 from arima model as shown in Eq.(9).

**𝑒𝑡 = 𝑦𝑡 − 𝑙 𝑡**

If there are linear correlations left in residuals, then the linear models are not sufficient in forecasting the data. The residual analysis is sufficient enough to capture the nonlinear patterns in the data. Therefore, even if a model has passed the diagnostic checking, the model may still not be adequate in that nonlinear relationships have not been appropriately modeled. Any significant nonlinear pattern in the residuals will indicate the limitation of the ARIMA. By modeling residuals using ANNs, nonlinear relationships can be discovered. **The mean absolute percentage error (MAPE), mean absolute error (MAE) and root mean square error (RMSE) are used evaluate the accuracy of forecasting.**