

WIND SPEED FORECASTING USING MACHINE LEARNING MODELS

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering Degree in
Computer Science and Engineering

By

M.NIROSHINI (17SCS9004)

N.V.S.PAVANI (17SCS9003)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND
TECHNOLOGY (DEEMED TO BE
UNIVERSITY)**

**Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAGIV GANDHI
SALAI, CHENNAI – 600 119**

MARCH - 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC I 12B Status by UGC I

Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **M.NIROSHINI (17SCS9004)** and **N.V.S.Pavani (17SCS9003)** who carried out the project entitled "**WIND SPEED FORECASTING USING MACHINE LEARNING MODELS**" under my supervision from November 2020 to March 2021.

Internal Guide

Mrs. D. DEEPA, M.E., (Ph. D).,

Head of the Department

Submitted for Viva voce Examination held on

Internal Examiner

External Examiner

DECLARATION

We **M. NIROSHINI (17SCS9004)** and **N.V.S. PAVANI (R17SCS9003)** hereby declare that the Project Report entitled “**WIND SPEED FORECASTING USING MACHINE LEARNING MODELS**” is done by me under the guidance of **Mrs. D. DEEPA, M.E., (Ph.D.)**, Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of **Sathyabama Institute of Science and Technology** for their kind encouragement in doing this project and for completing its successfully. I am grateful to them.

I convey my thanks to **Dr. T. SASIKALA, Ph.D., Dean, School of Computing** and **Dr. S. VIGNESHWARI, Ph.D., and Dr. L. Lakshmanan, M.E., Ph.D., Head of the Department,** Department of **COMPUTER SCIENCE AND ENGINEERING** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide. **Mrs. D. DEEPA, M.E., (Ph.D) Assistant Professor,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of **COMPUTER SCIENCE AND ENGINEERING** who were helpful in many ways for the completion of the project.

ABSTRACT

Wind energy is the strongest renewable energy source which ensures clean and safe production of energy. The wind speed prediction has an important place in wind energy systems and to drive turbines that further helpful for generating electricity, but the issue with the system is that power generated from wind is uncertain. So, accurate wind speed forecasting is required to produce more electric power. To address this issue, many approaches are presented by various researchers. This project describes an empirical study of modeling and forecasting of wind speed of Chennai city. Thanks to the National Institute of Wind Energy (NIWE) under The Ministry of New and Renewable Energy (MNRE) for sharing their yearly wind speed data. This project aims to build two wind speed prediction models Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) and try to find out which model is best for wind speed forecast by evaluating each of the model performance with the mean square error (MSE) and root mean square error (RMSE), respectively.

Key words: wind power, NIWE, time series method, ARIMA, SARIMA, MSE, RMSE.

TABLE OF CONTENTS

CHAPTER No	TITLE	PAGE No
	ABSTRACT	v
	LIST OF FIGURES	viii
	LIST OF TABLES	ix
	LIST OF ABBREVIATIONS	x
1	INTRODUCTION	
	1.1. INTRODUCTION OF THE PROJECT	1
	1.2. WIND ENERGY	2
	1.3. GLOBAL WIND ENERGY FORECAST	3
	1.4. CURRENT INDIA WIND ENERGY SCENARIO	3
	1.5. OUTLINE OF THE PROJECT	4
	1.6. SYSTEM OVERVIEW	5
2	LITERATURE SURVEY	
	2.1. GENERAL	7
	2.2. REVIEW OF LITERAL SURVEY	7
	2.2.1. TITLE	7
	2.2.2. AUTHOR	7
	2.2.3. DESCRIPTION	8
3	ALGORITHMS AND METHODS	
	3.1. SYSTEM REQUIREMENTS	13
3.1.1.	HARDWARE REQUIREMENTS	13
3.1.2.	SOFTWARE REQUIREMENTS	13
	3.2. DATA COLLECTION AND VALIDATION	13
	3.3. DATA VISUALIZATION	14
	3.4. SPLITTING DATA	15
	3.5. TIME SERIES MODELS	16
	3.5.1. ARIMA MODEL	16
3.5.1.1	ARIMA BASED WSF METHODOLOGY	20

CHAPTER No	TITLE	PAGE No
	3.5.1.2.DATA REFINING	20
	3.5.1.3.MATHEMATICAL TRANSFORMATION	21
	3.5.1.4. MODEL ORDER IDENTIFICATION	21
3.5.2.	SEASONAL - ARIMA MODEL	22
	3.5.2.1. SARIMA METHODOLOGY	25
3.6.	AUTO CORRELATION	29
	3.7.PARTIAL AUTO CORRELATION	30
	3.8.MAKE PREDICTIONS	30
	3.9.ERROR ANALYSIS	30
4	RESULTS AND DISCUSSION	
	4.1.MODEL SUMMARY	31
	4.2.RESULT	33
	4.3.PERFORMANCE ANALYSIS	34
5	CONCLUSION AND FUTURE WORK	
	5.1.CONCLUSION	36
	5.2.FUTURE WORK	37
	REFERENCE	38
	APPENDICES	
	A. SOURCE CODE	41
	B. SCREEN SHOTS	47

LIST OF FIGURES

FIGURE No	FIGURE NAME	PAGE No
1.1	Flow diagram of the system	5
3.1	Dataset sample	13
3.2	Time plot for wind speed	15
3.3	Time plot for wind speed with rolling mean	15
3.4	Partial Autocorrelation plot of ARIMA model	19
3.5	Autocorrelation plot of ARIMA model	19
3.6	ARIMA based WSF Methodology	20
3.7	Autocorrelation plot of SARIMA model	23
3.8	Partial Autocorrelation plot of SARIMA model	23
3.9	AIC combinations	24
3.10	ADF test result	24
3.11	SARIMA methodology	25
3.12	Histogram plus estimated density plot	27
3.13	The Q-Q plot	27
3.14	Standardized residual plot	28
3.15	Correlogram plot	29
4.1	ARIMA MODEL SUMMARY	32
4.2	SARIMA MODEL SUMMARY	33
4.3	ARIMA prediction vs actual wind speed	34
4.4	SARIMA prediction vs actual wind speed	35

LIST OF TABLES

TABLE No	TABLE NAME	PAGE No
3.1	MEASURED PARAMETERS	14
4.1	SUMMARY OF TEST STATISTICAL ERRORS	34

LIST OF ABBREVIATION

ABBREVIATION		EXPANSION
ACF	-	Auto-Correlation Function
AI	-	Artificial Intelligence
AIC	-	Akaike Information Criteria
AR	-	Autoregressive
ARIMA	-	Autoregressive Integrated Moving Average
CSV	-	Comma Separated Values
LSTM	-	Long Short Term Memory
MA	-	Moving Average
MNRE	-	Ministry of New and Renewable Energy
MSE	-	Mean Square Error
NIWE	-	National Institute of Wind Energy
Numpy	-	Numerical Python
NWP	-	Numerical Weather Prediction
PACF	-	Partial Auto-Correlation Function
RE	-	Renewable Energy
RMSE	-	Root Mean Square Error
SARIMA	-	Seasonal Autoregressive Integrated Moving Average
WSF	-	Wind Speed Forecasting

CHAPTER 1

INTRODUCTION

INTRODUCTION OF THE PROJECT

The development of modern technology made 'electricity' as one of the essential element to lead a quality life. The electricity is mostly generated by using fossil fuels such as coal, petroleum and natural gas, these are the main contributors to Global warming. To reduce this we need to reduce the usage of fossil fuels and to find alternate energy sources. Due these reasons the importance of renewable energies has been increasing rapidly as it has less impact on environmental pollution.

One of the best among such renewable energies is wind energy which ensures clean and safe production of energy. Wind is nothing but "Air in motion". The kinetic energy in the wind is converted into mechanical energy and then converted into electrical energy by means of Wind Turbines. Wind stations are to be constructed for using wind energy and for constructing wind stations proper and accurate investigation on wind speed at different regions need to be done. So we can reduce the operational cost.

Wind energy plays an important role in generating electricity. The wind speed prediction has an important place in wind energy systems and to drive turbines that further helpful for generating electricity. Wind power forecasting is primarily depends of wind speed forecasting. Wind speeds will become uncertain due to climate change. So, accurate wind speed forecasting is required to produce more electric power. Yearly wind speed data of Chennai city were collected from National Institute of Wind Energy (NIWE) under The Ministry of New and Renewable Energy (MNRE).

Many approaches have been described by many researchers to address this issue. These approaches are basically divided into physical methods, statistical methods, time series methods and hybrid methods.

Physical method is basically numerical weather prediction (NWP) of wind speed data. These methods require description of wind turbines, wind forms and

other computational components. Traditional Statistical methods require historical

data to find relationship of the historical data and uses mathematical formulae to calculate the wind speed. Combination of different model to gain advantages of each model is known as hybrid method.

This paper proposes two time series models namely Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) for wind speed forecasting.

ARIMA is one of the most widely used time series method which aims to find autocorrelation between the data. The structure of this model is defined by p , d , q . Partial Auto-Correlation Function (PACF) and Auto-Correlation Function (ACF) were used to determine the parameters p (Autoregressive component, AR) and q (Moving Average component, MA).

Extension of ARIMA with seasonality function is known as Seasonal- ARIMA which has the following parameters auto regression (AR), differencing (I) and moving average (MA) as well as an additional parameter for the seasonality.

The predictive accuracy of these models was evaluated based on the mean square error (MSE) and root mean square error (RMSE). The model which tend to lower RMSE and MSE values were selected as better model for wind speed forecasting.

WIND ENERGY

Wind energy is provides as the most matured source of energy all over the world for clean and safe production of energy. Wind is nothing but “air in Motion”. A mere 2 to 3 percent of differential absorption of solar radiation on varied earth's surface causes the air to move.

During sunny hours, air in the atmosphere gets heated up and the air particles tend to move towards the low pressure regions, creating wind. Wind has kinetic energy and this is converted into mechanical energy and then to electrical energy by means of Wind energy conversion system or Wind turbines. With the modern technological developments, the Wind Turbines has been manufactured to convert kinetic energy available in the wind in to electrical energy.

GLOBAL WIND ENERGY FORECAST

As per IRENA's and other agencies forecast, wind power will remain a key renewable energy option in the coming decades and the global installed capacity of onshore wind power would increase three- fold by 2030 and ten- fold by 2050 compared to installations in 2020.

Asia – mainly China and India would continue to lead global onshore wind power installations, with the region accounting for more than half of the total global capacity by 2050 and is well aligned with energy transformation scenarios of various institutions, clearly highlighting the importance of scaling up the wind power generation share in order to decarbonize the energy system in the next three decades.

The energy transition will increase employment across the border the broader economy and specifically in the energy sector. Many governments have prioritized renewable energy development, primarily to reduce emission and meet international climate goals, but also in pursuit of broader socio- economic benefits. IRENA estimates that employment in the wind industry would continue to rise, exceeding 3.7 million jobs by 2030 and 6 million jobs by 2050. Of the more than 6 million jobs by 2020, 5 million would be in the onshore wind sector and the rest would be offshore jobs.

By project segment, of the more than 5 million onshore wind sector jobs by 2050, 1.7 million would be in construction and installation, 2.18 in manufacturing and 1.17 million in O&M. For offshore wind, out of the one million jobs by 2050, an estimated 0.45 million would be in construction and installation, 0.39 million in manufacturing and 0.17 million in O&M. The rising traction of the wind energy industry demands a growing array of skills, including technical, business, administrative, economic and legal, among others.

CURRENT INDIA WIND ENERGY SCENARIO

Recognizing the need to reduce import dependence and improve country's energy security combined with the need to move away from fossil fuels in the longer

run in view of climate change considerations, special efforts were made by Government of India to increase the supply of energy from renewable energy

sources. Decades of concerted efforts have started to yield gratifying results and today, Wind power contributes about 10% of the total Indian energy mix of 373435.58 MW as on September 2020 and stands fourth in terms of installed wind power capacity worldwide.

Among all renewable energy (RE) options, wind power, which accounts for over 44% of the installed RE capacity in the country, is the most commercially competitive source of renewable energy. This may be attributed to technological maturity, proven installed base and lower setup and running costs.

India is now a global manufacturing hub for wind turbines with more than 20 large wind turbine manufactures with wind turbines capacity ranging from 225kW to 3000 kW and several small wind turbine manufactures producing capacity ranging from 0.3 kW to 50 kW. Some of the turbine models produced in India are currently exported by several manufacturers and are expanding capacity to meet growing demand.

To assure timely evolution of wind industry and to meet India's demands and development of rural areas, human resource development is a matter of immense importance.

Wind energy can serve as a growth engine to provide new employments and to bring about economic development in many underdeveloped areas as most of wind farms are located in rural places.

To identify the present demand of human resource and to know the need of human resource in future, the Indian wind energy sector and market potential should be analyzed in detailed.

OUTLINE OF THE PROJECT

The objective of this project is to build two wind speed prediction models Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) and try to find out which model is best for wind speed forecast by evaluating each of the model performance with the mean square error (MSE) and root mean square error (RMSE), respectively.

SYSTEM OVERVIEW

This project helps wind farms to forecast wind speed and to gain more power to produce electric power. They have to find Accuracy of the training dataset, Accuracy of the testing dataset, Specification by comparing algorithm using python code.

The flow diagram of this project to clearly understand the implementation is depicted below.

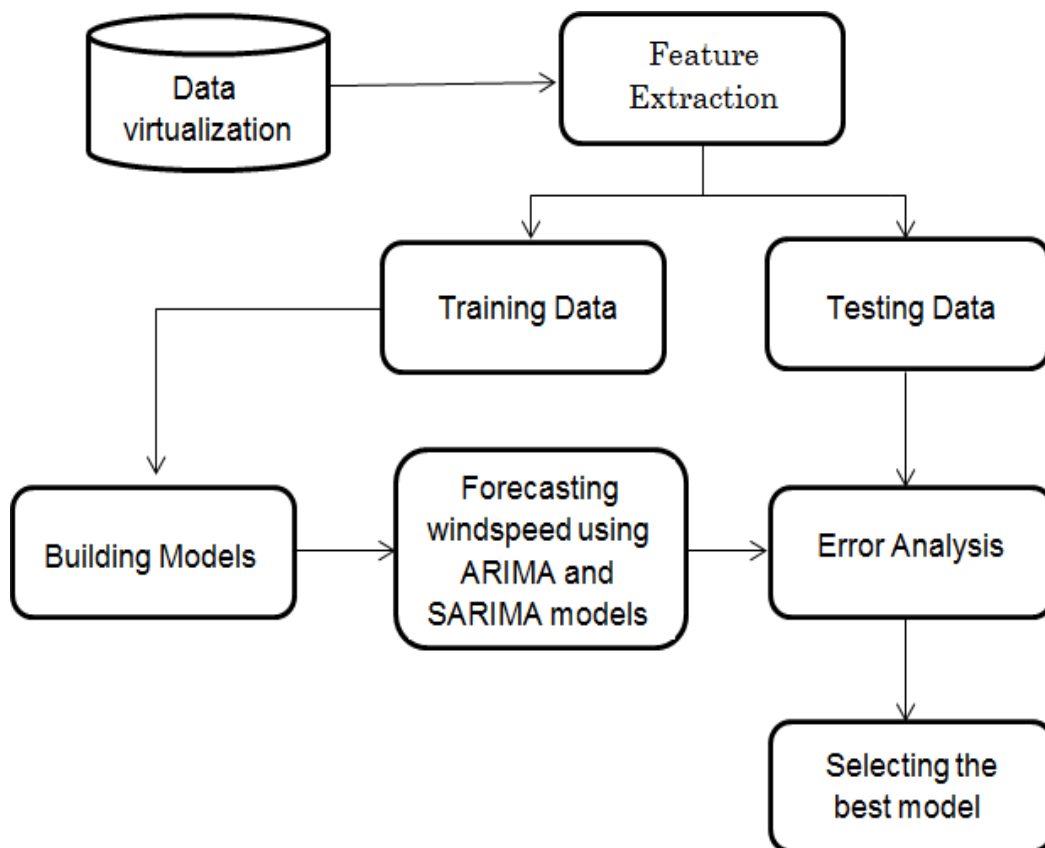


Fig 1.1: Flow diagram of the system

The data were collected and preprocessed to remove trends and seasonality using lag scatter plots and time plots. We split the data into two sets namely training and testing data sets. Two time series models

Autoregressive Integrated Moving Average (ARIMA)

Seasonal Autoregressive Integrated Moving Average (SARIMA)

Were built and trained with the help of training data sets. The necessary parameters to train the models were calculated by plotting the following graphs

Auto Correlation Graph

Partial Auto Correlation graphs.

The training summary of each models were noted. Then the future wind speed were predicted and created a new data set called predicted data set. The predicted accuracy of each model is calculated with the help of mean square error (MSE) and root mean square error (RMSE). The performance evaluation is done by comparing the errors and the better model was selected.

CHAPTER 2

LITERATURE SURVEY

GENERAL

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

REVIEW OF LITERATURE SURVEY

Title

Wind Energy Forecasting with Neural Network

Author

Jaume Manero¹, Javier B´ejar¹, Ulises Cort´es^{1,2}

2.2.2. Description

Time-series forecasting has been performed predominantly using statistical based methods, for example, the linear autoregressive (AR) models because of their flexibility to model many stationary processes. These include the well-known ARMA

(Autoregressive Moving Average) model. The ARMA model assumes a linear

relationship between the lagged variables and produces only a coarse approximation to real-world complex systems and generally fails to accurately predict the evolution of nonlinear and non-stationary processes. ARMA model performance frequently degrades considerably whenever time trends and seasonality features are present in the highly fluctuating time series data. Methods such as the linear Autoregressive Integrated Moving Average (ARIMA), which are based on the evolution of the increments are used at times to remove/reduce first-order (moment) non-stationary. ARIMA models are successfully applied for time series forecasting of wind speed, weather, birth rate etc

However, most of these methods tend to be limited for nonlinear and stationary time series forecasting by the local linearity assumption implicit with an AR-type structure.

TITLE

A novel compound wind speed forecasting model based on the back propagation neural network optimized by bat algorithm.

Author

Yanbin Cui & Chenchen Huang & Yanping Cui

Description

Wind power, a clean and renewable resource, is regarded as one of the most promising and economical resources during the transformation from fossil fuels to new energy resources. Thus, the accuracy of wind speed forecasting work is very important to integrate the wind resource into electrical power system on a large scale. To improve the short-term wind speed forecasting accuracy, a novel compound model is introduced in this paper. For the proposed model, the fast ensemble empirical mode decomposition method was employed to do the data preprocessing. After the data preprocessing, phase space reconstruction was used for choosing each sub-series' input and output vectors for the forecasting model dynamically. Then, the bat algorithm was applied to optimize the connection weights and thresholds of the traditional back propagation neural network. The forecasting

results can be obtained through the aggregation of sequential prediction. The

performance evaluation of this proposed model indicates that it can capture the nonlinear characteristics of the wind speed signal efficiently. The proposed model shows better performance when being compared with the parallel models.

TITLE

Prediction of evaporation in arid and semi-arid regions: a comparative study using different machine learning models.

Author

Zaher Mundher Yaseen, Anas Mahmood Al-Juboori, Ufuk Beyaztas, Nadhir Al-Ansari, Kwok-Wing Chau, Chongchong Qi, Mumtaz Ali, Sinan Q. Salih & Shamsuddin Shahid.

Description

The accurate prediction of evaporation is essential for multiple water resources engineering applications, particularly in developing countries like Iraq where the meteorological stations are not sustained and operated appropriately for in situ estimations. This is where advanced methodologies such as machine learning (ML) models can make valuable contributions. In this research, evaporation is predicted at two different meteorological stations located in arid and semi-arid regions of Iraq. Four different ML models for the prediction of evaporation – the classification and regression tree (CART), the cascade correlation neural network (CCNNs), gene expression programming (GEP), and the support vector machine (SVM) – were developed and constructed using various input combinations of meteorological variables. The results reveal that the best predictions are achieved by incorporating sunshine hours, wind speed, relative humidity, rainfall, and the minimum, mean, and maximum temperatures. The SVM was found to show the best performance with wind speed, rainfall, and relative humidity as inputs at Station I ($R^2 = .92$), and with all variables as inputs at Station II ($R^2 = .97$).

TITLE

A new artificial multi-neural approach to estimate the hourly global solar

radiation in a semi-arid climate site.

Author

Mohammed Ali Jallal & Samira Chabaa & Abdelouhab Zeroual.

Description

In this study, a new hybrid machine-learning algorithm called artificial multi-neural approach, for developing a multiinput single-output (MISO) model to estimate the hourly global solar radiation (HGSR) time series in Agdal site (latitude 31° 37' N, longitude 08° 01' W, elevation 466 m), Marrakesh, Morocco. To achieve this goal, three training algorithms (scaled conjugate gradient, Levenberg-Marquardt, and resilient back propagation) are selected to train the developed model, using some accessible hourly meteorological variables recorded during 7 years (from 2008 to 2014) as exogenous inputs. Furthermore, a pertinence determination test (PDT) is used to point out the most pertinent inputs for the accurate HGSR estimation. The best configuration of the multi-neural model includes five known input parameters as the best scenario, i.e., acquisition hour, air temperature, relative humidity, wind speed, and precipitation. Some statistical indicators are used to evaluate the performance of the developed model. The obtained results demonstrate the reliability and the precision of our proposed machine-learning algorithm, representing a good solution for estimating solar radiation time series needed to design and manage solar energy systems compared to the present standards.

Title

Wind Farm NWP Data Preprocessing Method Based on t-SNE.

Author

Jiu Gu , Yining Wang , Da Xie and Yu Zhang.

Description

The operation prediction of wind farms will be accompanied by the need for massive data processing, especially the preprocessing of wind farm meteorological data or numerical weather prediction (NWP). Because NWP data are strongly correlated with wind farm operation, proper processing of NWP data could not only

reduce data volume but also improve the correlations of wind farm operation

predictions. For this purpose, this paper proposes a data preprocessing algorithm based on t-distributed stochastic neighbor embedding (t-SNE). Firstly, the data collected were normalized to eliminate the influence caused by different dimensions. The t-SNE algorithm is then used to reduce the dimensionality of the NWP data related to wind farm operation. Finally, the wind farm data visualization platform is established. In this paper, 22 index variables in NWP data were taken as objects. The t-SNE method was used to preprocess the NWP historical data of a wind farm, and the results were compared with the results of the principal component analysis (PCA) algorithm. It outperformed PCA in error precision; in addition, t-SNE dimension reduction preprocessing also had a visual effect, which could be applied to big data visualization platforms. A long short-term memory network (LSTM) was used to predict the operation of the wind farm by combining the preprocessed NWP data and the operation data. The simulation results proved that the effect of the preprocessed NWP data based on t-SNE on the wind power prediction was significantly improved.

Title

An Online Hybrid Model for Temperature Prediction of Wind Turbine Gearbox Components.

Author

Qiang Zhao , Kunkun Bao , Jia Wang , Yinghua Han and Jinkuan Wang

Description

The temperature prediction model of wind turbine gearbox components is of great significance for monitoring the operation status of the gearbox. However, the complex operating conditions of wind turbines pose grand challenges to predict the temperature of gearbox components. In this study, an online hybrid model based on a long short term memory (LSTM) neural network and adaptive error correction (LSTM-AEC) using simple-variable data is proposed. In the proposed model, a more suitable deep learning approach for time series, LSTM algorithm, is applied to realize the preliminary prediction of temperature, which has a stronger ability to capture the

non-stationary and non-linear characteristics of gearbox components temperature series. In order to enhance the performance of the LSTM prediction model, the

adaptive error correction model based on the variational mode decomposition (VMD) algorithm is developed, where the VMD algorithm can effectively solve the prediction difficulty issue caused by the non-stationary, high-frequency and chaotic characteristics of error series. To apply the hybrid model to the online prediction process, a real-time rolling data decomposition process based on VMD algorithm is proposed. With aims to validate the effectiveness of the hybrid model proposed in this paper, several traditional models are introduced for comparative analysis. The experimental results show that the hybrid model has better prediction performance than other comparative models.

CHAPTER 3

ALGORITHMS AND METHODS

SYSTEM REQUIREMENTS

Hardware requirements

- ◆ Processor : Pentium 3.
- ◆ Hard Disk : 40 GB.
- ◆ Monitor : 15 VGA Color.
- ◆ Ram : 512 Mb.

Software requirements

- ◆ Operating system : Windows OS.
- ◆ Coding Language : Python
- ◆ Tool : Anaconda with Jupyter Notebook
- ◆ Libraries : pandas , matplotlib, sklearn , pmdarima,
auto_arima

DATA COLLECTION AND VALIDATION

Data for this project were collected with the help of National Institute of Wind Energy (NIWE) under the Ministry of New and Renewable Energy (MNRE) which is located at pallikaranai in Chennai. The dataset contain average wind speed of Chennai city. The head of the dataset is shown in the fig 3.1.

	Temperature	Relative Humidity	Pressure	Wind direction	Wind speed
Date					
2017-01-01	295.60	82.41	1009.75	26.35	3.89
2017-01-02	295.37	88.94	1010.66	18.64	4.03
2017-01-03	294.93	86.88	1010.35	17.55	3.88
2017-01-04	294.78	81.62	1009.31	21.78	3.37
2017-01-05	295.37	78.82	1009.58	19.24	3.14

Figure 3.1.Dataset sample

The data were collected with the help of Met Mast of height 45 meters which is located at NIWE. All the parameters were collected from the same Met Mast. The

different devices fixed in different heights of the Met mast measures all the parameters. We can get the clear picture from the table 3.1.

Table 3.1.Measured parameters

Parameters	Devices	Units
Wind Speed	Anemometer	m/s
Wind Direction	Wind Vane	Degree
Temperature	Temperature Sensor	Degree Celsius
Humidity	Hygrometer	g/m ³
Pressure	Pressure Sensor	Mbar

Before starting with the validation, the analyst should confirm the monitoring configuration and settings, including anemometer heights, wind vane dead band orientation, anemometer transfer function and time stamps.

Data validation has two phases

1. Automated screening
2. Manual review

The automated screening flags potentially suspect data records, while the manual review verifies the results. Validation routines are designed to screen each parameter for suspect values before they are incorporated into the database and used for site analysis.

DATA VISUALIZATION

An effective first and foremost step is visualizing the time series data to make sure the data which we imported is same as what we plan to use for forecasting and also we can detect initial parameters, identify its components and detect the problems by visualizing the time series.

The best method for visualizing the data is Time plot which is presented in Fig. 3.2. Time plot is nothing but a graph that shows observations against time. Time plot is used to find whether the time series have trends and seasonality, if so then that must be removed. If the time series has too much variation, then we need to resample that by using rolling mean in time plot to determine trend. The time plot which uses rolling mean is presented in Fig. 3.3.

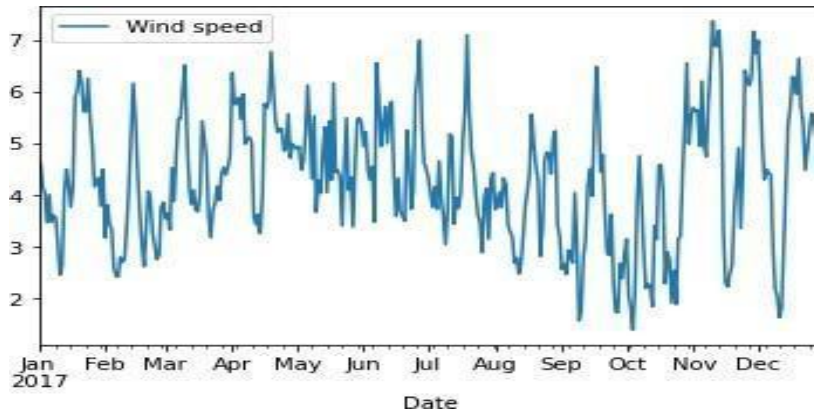


Figure 3.2. Time plot for wind speed

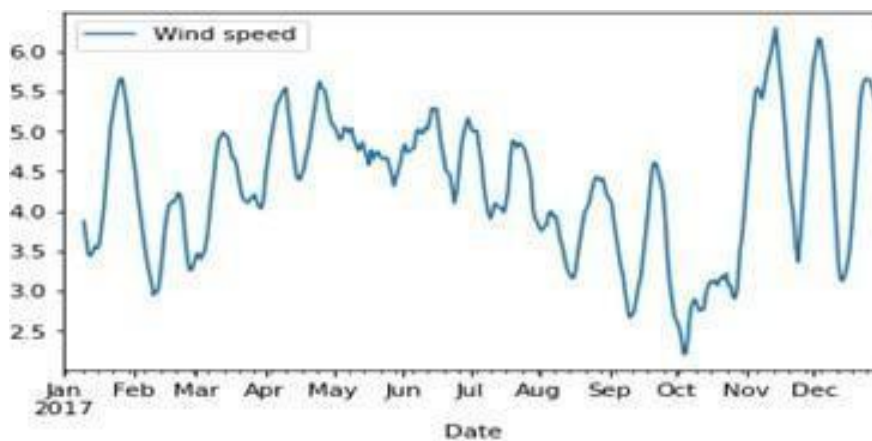


Figure 3.3. Time plot for wind speed with rolling mean

Another method for data visualization is lag scatter plot. In time series the previous observation is referred as lag. The lag scatter plot is used to explore the relationship between an observation and the lag observation of that observation.

SPLITTING DATA

The data were segregated into two sets called training set and testing data set. The training data set is a subset of a data which is used to train the model, the models look and learn from this training set to predict the future value. The subset of the data set which is used for evaluation of the final model which learnt from training set is known as testing data set. In addition to the training and testing data, another set of observation called predicted data set is required to validate the behavior of the time series models.

Wind speed data set is basically a table which contains information about average wind speed of a day and the date. Dataset should be in CSV format. Upload the data into Jupyter Notebook. Then we encapsulate `pd.read_csv("windspeed.csv")` method using pandas. Further we divide the dataset into training data and test data using the following method.

```
speed_train = w_speed [y] method and speed_test=w_speed [x:y]
```

The X prefix in variable denotes the feature values and y prefix denotes target values. This method divides dataset into training and test data randomly in ratio of 80:20. Then we encapsulate any algorithm. In the next line, we fit our training data into this algorithm so that computer can get trained using this data.

TIME SERIES MODELS

A continues observations that varies based on time is called as Time series and the models which uses time series for multiple purposes are called time series models. A time series model reproduces the patterns of the prior movements of a variable over time and uses this information to predict its future movements. It is possible in this way to construct a simplified model of the time series that represents its randomness, so that it is useful for prediction. There are four modules based on the timescales, and for each of them, various forecasting techniques used. The present study uses an Autoregressive Integrated Moving Average (ARIMA) model and Seasonal Autoregressive Integrated Moving Average (SARIMA) for the forecasting of Wind speed.

ARIMA Model

ARIMA is one of the frequently used time series model, including Jian Jiao approach [9], because ARIMA models are robust and easy to learn and implement. The ARIMA model sprang up with a point of performing various differencing methods to get stationary observations from the non-linear and non-stationary observations.

ARMA model understands the significance of using the Auto Regression (AR) and Moving Average (MA) terms for accurate forecasting. But the time series for the wind speed is not stationary. For this purpose, the ARIMA model comes into

forecasting scenario.

ARIMA model is illustrated as ARIMA (p, d, q), where:

p - order of the auto regression

d - degree of differencing

q - order of the moving window

We can divide the ARIMA model into three terms AR (p), I (d), MA (q).

Auto Regression Model (AR):

Auto Regression model is only applied on the data which do not have trend and seasonality. It tells as how many lagged values are used for forecasting that means p. It is basically a linear regression model. With the help of training data our software will train the model automatically and find the value p. Once we found the value of p, then we can start prediction.

This Auto-Regressive model estimates a dependent variable Y_t at any timing instant based on some number of lagged observations.

The variables can be summarized as follows and formulated by the below given equation:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \varepsilon_t \quad (1)$$

Where,

- $Y_{(t-1)}$: Lag 1 of the time series
- β_1 : Coefficient of lag1
- α : Intercept term

Differencing part (I):

This is used to find how many differencing order used to make the time series stationary. In order to find a right order of differencing ACF plots are used. The ARIMA model sprang up with a point of performing various differencing methods to get stationary observations from the non-linear and non-stationary observations.

Moving Average Model (MA):

This model inspects the relationship by applying moving average to lag values between the true value and a forecast error by applying moving average to lag values. It used to find number of lagged error (q).

The MA model also estimates a dependent variable Y_t based on the lagged forecast errors.

$$Y_t = \alpha + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \dots + \phi_p \varepsilon_{t-p} \quad (2)$$

The optimal parameter model can be obtained through the AIC criterion. After the process of identification and estimation, it is necessary to test whether the model is applicable or not. Only when the test is passed, can the model be reflected to meet the requirements and be qualified for the prediction task. Conversely, if it does not meet the requirements, the model must be modified or re-identified, and the process must be repeated until the requirements are met. In the estimation process and the recognition process of the above models, the premise is to assume that the disturbance term is white noise.

The time series equation goes through some mathematical transformations, like differencing, to analyze any non-stationary time series by making it stationary. Thus, the degree of differencing d is included along with the AR and MA lags to include the order of ARIMA model as (p, d, q) . The final equation of ARIMA (p, d, q) is given by:

$$Y_t^d = c + e_t + \sum_{k=1}^p \phi_k Y_{t-k}^d + \sum_{j=1}^q \theta_j e_{t-j} \quad (3)$$

Where, c is the constant drift or the average of the changes between the consecutive observations, and e_t is the white noise or error of time series forecasting, j and k are the indices of MA and AR time lags, $\phi_k, k=1 \dots p$ are the coefficient of historical values of time series, $\theta_j, j=1 \dots q$ are the coefficient $\phi_1, \phi_2, \dots, \phi_p$ and $\theta_1, \theta_2, \dots, \theta_q$ multiple time series are obtained. Also, the error term e_t effects only the scale of the time series. The pattern of the series remains unaltered with any variation in e_t .

The order of the model can be preliminarily determined by ACF and PAC. Parameter identification of time series can be obtained by least square estimation, that is, the estimated parameters $\phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q$.

ARIMA – Differencing + Combination of Parameters from AR and MA model

The equation of ARIMA model with the parameters p, d, q also written as

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_p y_{t-p} + \varepsilon_t \quad (4)$$

Where, a is called the Backshift operator such that

$$a_k(Y_t) = Y_{t-k}$$

The points (3, 0, 2) were taken as the best ARIMA (p, d, q) structure to forecast the wind speed, the points were calculated by plotting partial autocorrelation and autocorrelation graphs which are presented in Fig. 3.4 and 3.5.

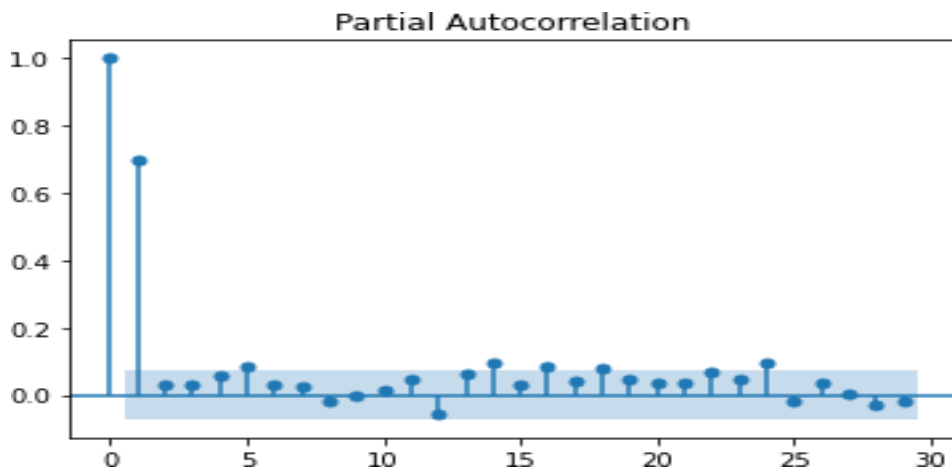


Figure 3.4. Partial Autocorrelation plot of ARIMA model

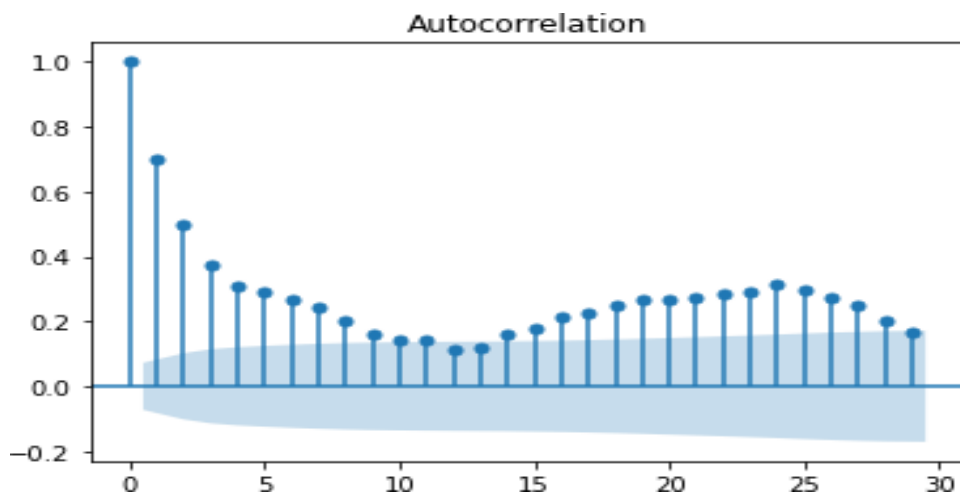


Figure 3.5. Autocorrelation plot of ARIMA model

ARIMA based WSF methodology

The ARIMA model based wind speed forecasting methodology is shown in the below fig.3.6.

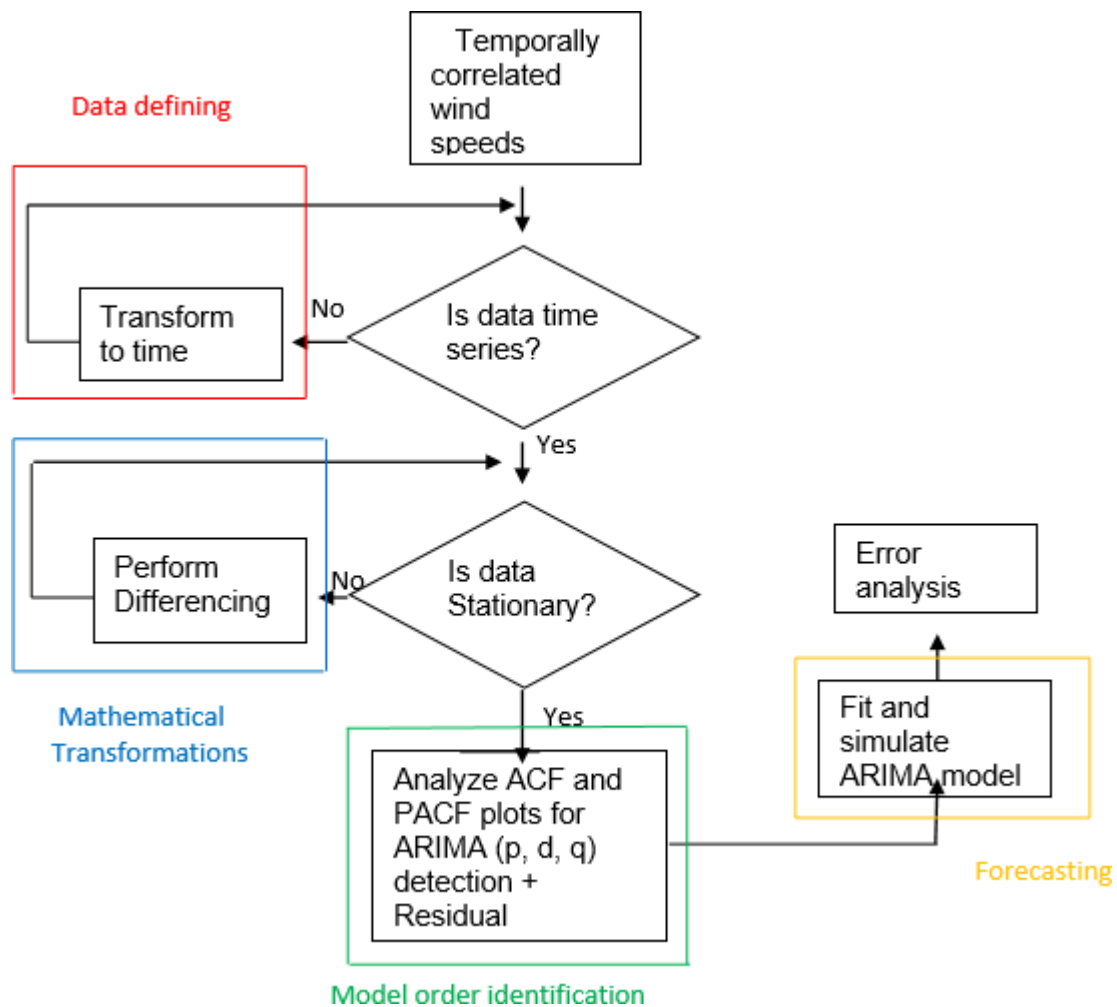


Figure 3.6.ARIMA based WSF Methodology

Data Refining

The temporally correlated wind speed data for the given location transformed into time series format for performing ARIMA model simulations .The obtained wind speeds are then plotted and analyzed for any non– Stationary , i.e., seasonality and trends, etc. If there is any seasonality or trends then they will be removed by finding mean value

Mathematical Transformation

To make the data stationary, logarithmic or differencing transformations are done to stabilize the variance. If the time series has too much variation, then we need to resample that by using rolling mean.

Model Order Identification

Next, the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) of the stationary data are observed to identify the model order (p,d,q).

The mathematical formulations for ACF and PACF are provided in (5) and (6) respectively. If y_t is the original series and $\bar{y}$ is the mean of the data, then the auto correlations of order $k=0, 1, 2, \dots, n$, where n is the total number of lags possible is computed as:

$$ACF = \frac{\frac{1}{n-k} \sum_{t=k+1}^n (F_t - \bar{F})(F_{t-k} - \bar{F})}{\sqrt{\frac{1}{n} \sum_{t=1}^n (F_t - \bar{F})^2} \sqrt{\frac{1}{n-k} \sum_{t=k+1}^n (F_{t-k} - \bar{F})^2}} \quad (5)$$

$$PACF = \frac{\frac{Cov[F_t, F_{t-k} | F_{t-1}, \dots, F_{t-k+1}]}{\sqrt{var[F_t | F_{t-1}, \dots, F_{t-k+1}]}}}{\sqrt{var[F_{t-k} | F_{t-1}, \dots, F_{t-k+1}]}} \quad (6)$$

The model order is selected for which the Akaike Information Criteria (AIC) value is least or minimum. The formula for AIC determination is stated in (7).

$$AIC = 2P - \ln(L) \quad (7)$$

Where P is the number of parameters estimated for the model and L is the maximum value of the likelihood function for the model. The model order chosen is correct if the ACF and PACF plots of the residual are similar to that of the white noise

pattern.

Next, the ARIMA model is fitted and simulated for the model order (p,d,q) selected and then the forecasts are obtained for the ahead values.

Seasonal - ARIMA Model

SARIMA is one of the frequently used time series model that uses past observations to forecast future values.

This model is an extension of the ARIMA with seasonality function is known as Seasonal - ARIMA which has all the parameters which is used in the ARIMA model and an additional parameter for the seasonality.

The disadvantage of ARIMA model is that doesn't support seasonal data. But Seasonal-ARIMA identifies the trend and seasonality in the time series with the help of a new seasonal component.

SARIMA model is illustrated as SARIMA (p, d, q,) x (P, D, Q, s) which consists of trend and seasonal elements.

p - order of the auto regression

d - degree of differencing

q - order of the moving average

P - order of the auto regression with seasonality

D - seasonal degree of differencing

Q - order of the moving average with seasonality

s - number of periods per season

The auto correlation and partial autocorrelation graphs were plotted which were presented in Fig. 3.7 and 3.8 to identify the seasonal and non- seasonal parameters which are required to build and train the Seasonal Auto Regressive Integrated Moving Average (SARIMA) model.

Seasonal - ARIMA = Differencing (I) + combination of AR and MA parameters +
seasonal component

A seasonal ARIMA model uses differencing at a lag equal to the number of seasons (s) to remove additive seasonal effects. As with lag 1 differencing to remove a trend, the lag s differencing introduces a moving average term. The seasonal ARIMA model includes autoregressive and moving average terms at lag s.

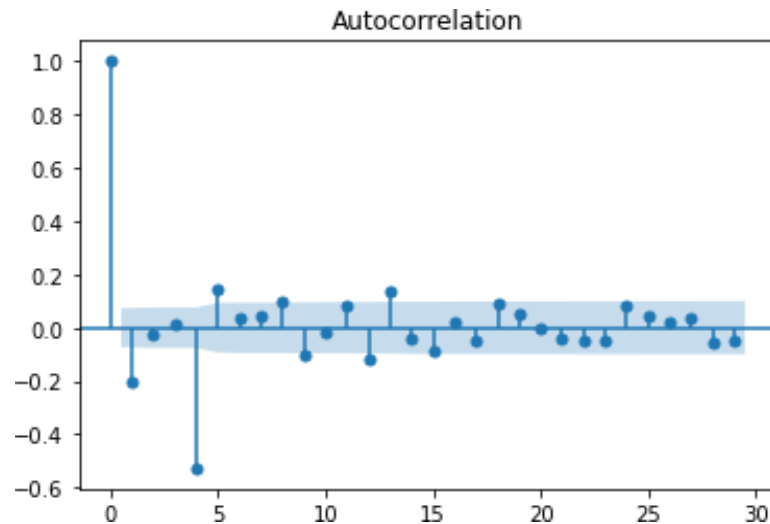


Figure 3.7. Autocorrelation plot of SARIMA model

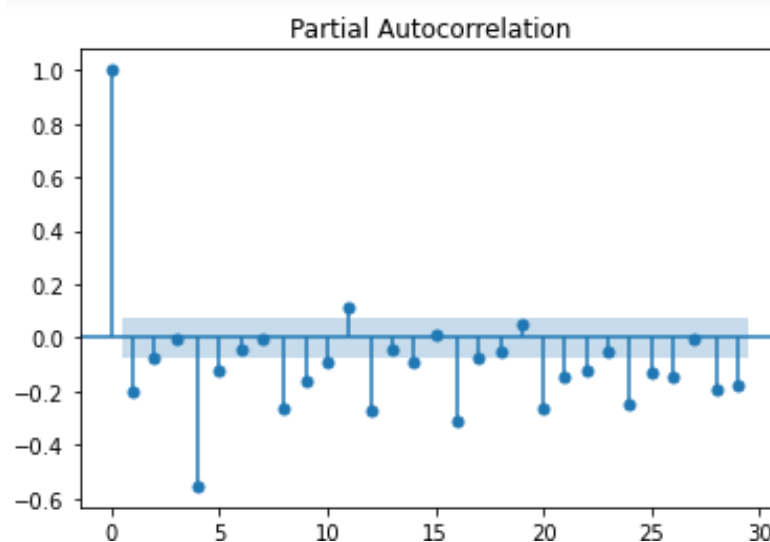


Figure 3.8. Partial Autocorrelation plot of SARIMA model

Looking at the ACF plot, these plots can give us a rough idea of the processes in play, it is better to test multiple scenarios and choose the model that yield the lowest AIC.

Therefore, let's write a function that will test a series of parameters for the SARIMA model and output a table with the best performing model at the top.

Note that we will only test different values for the parameters p , P , q and Q . We

know that both seasonal and non-seasonal integration parameters should be 1,

and that the length of the season is 4. Therefore, we generate all possible parameters combination.

We get 256 unique combinations. Now, our function will fit 256 different SARIMA models on our data to find the one with the lowest AIC. The AIC combinations are presented in fig 3.9.

	(p,q)x(P,Q)	AIC
0	(1, 2, 1, 3)	616.498651
1	(2, 3, 1, 2)	619.490145
2	(1, 2, 2, 3)	621.673155
3	(1, 2, 0, 3)	627.853442
4	(1, 2, 0, 2)	628.437777
...	...	...
251	(0, 0, 1, 0)	1747.503888
252	(3, 0, 0, 0)	1747.543752
253	(2, 0, 0, 0)	1775.399208
254	(1, 0, 0, 0)	1869.674791
255	(0, 0, 0, 0)	2076.083825

Figure 3.9.AIC combinations

Augmented Dickey-Fuller test again to see if we have a stationary time series is represented in fig 3.10. The Dickey–Fuller test tests the null hypothesis that a unit root is present in an autoregressive model. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity.

```

1. ADF : -11.39558587779346
2. P-Value : 7.873303025006261e-21
3. Num Of Lags : 0
4. Num Of Observations Used For ADF Regression and Critical Values Calculation : 730
5. Critical Values :
   1% : -3.4393396487377155
   5% : -2.865507363200066
  10% : -2.5688826684180897

```

Figure 3.10.ADF test result

The equation of SARIMA model is

$$(1 - \phi_1 \beta) (1 - \phi_1 \beta^4) (1 - \beta) (1 - \beta^4) Y_t = (1 + \Theta_1 \beta) (1 + \Theta_1 \beta^4) \varepsilon_t \quad (8)$$

The points (0, 1, 2) x (0, 1, 2, 4) were taken as the best Seasonal - ARIMA (p, d, q) x (P, D, Q, s) structure to forecast the wind speed, the points were calculated by plotting partial autocorrelation and autocorrelation graphs.

3.5.2.1. SARIMA methodology

The SARIMA model based wind speed forecasting methodology is shown in the below fig.3.11.

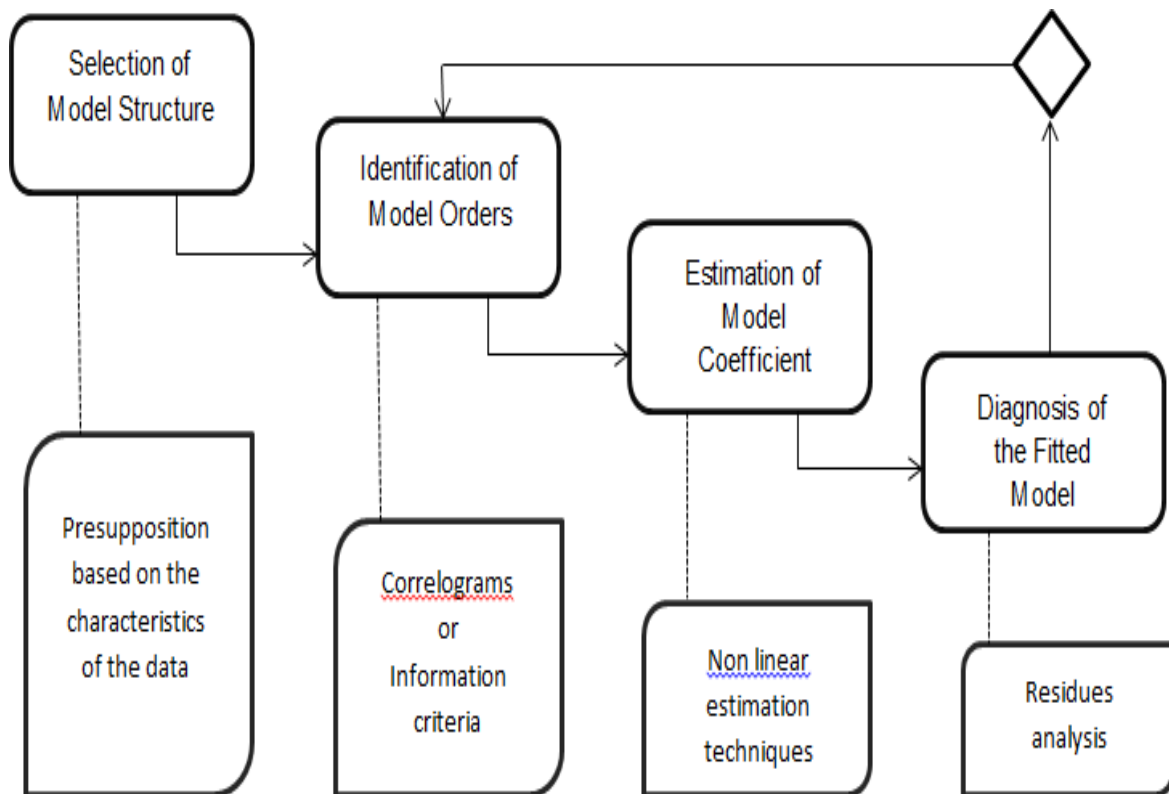


Figure 3.11. SARIMA methodology

The basic steps of the methodology are as follows

1. Sequence stationary process

Eliminate possible cyclical and seasonal behavior. In order to improve the forecast with low order models, it is convenient to eliminate this behavior. We usually

remove this seasonal behavior by means of differentiation. To identify possible seasonal and cyclical components present in the time series, the auto correlation function (ACF) is used. If the sequence is non- stationary, generally one or two time differences can make the sequence smoother by difference processing.

2. Identify model

After the elimination of seasonal and cyclical components, the time series can be transformed into a stationary one. According to the process of smooth stationary sequence number, the difference from the auto correlation (AC) and partial auto correlation (PAC) function can primarily determine the order number of the preliminary model.

In general, this kind of method often cannot complete model identification. The method of trial- and- error is estimated for several different models based on the information criterion (AIC). The final model would be established by comparison.

3. Parameter estimation

By using the non- linear least squares method for the selected model preliminary estimates are obtained for optimizing parameters. A parameter estimated method of non- linear least squares is put forward.

4. Prediction model testing

Adopting the χ^2 test method handles the residual series to verify whether the series is a white-noise series. If the residual sequence is not a white-noise sequence, this means that there is no useful information extraction, and the model needs to be further improved.

Histogram plus estimated density plot

A histogram is a plot that lets you discover, and show, the underlying frequency distribution of a set of continuous data. This allows the inspection of the data for its underlying distribution. The red KDE line follows closely with the N (0, 1) line. This is a good indication that the residuals are normally distributed. Histogram plus estimated density plot is shown in fig 3.12.

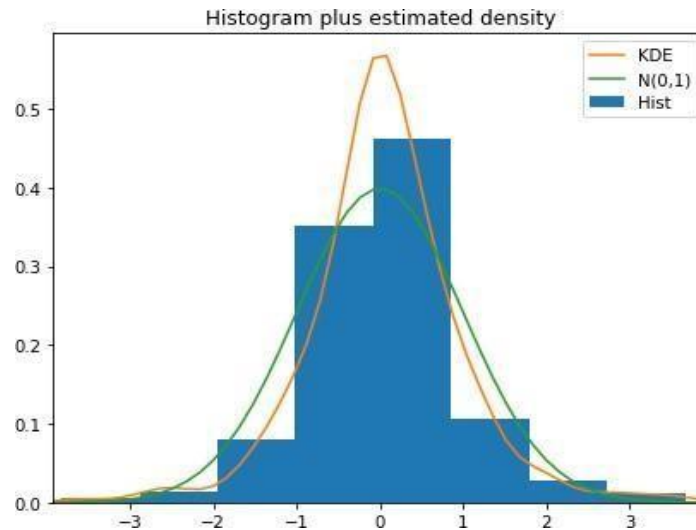


Figure 3.12. Histogram plus estimated density plot

The Q-Q plot

A Q-Q plot is a scatterplot created by plotting two sets of quantities against one another. If both sets of quantities came from the same distribution, we should see the points forming a line that's roughly straight. This plot shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with $N(0, 1)$. This is an indication that the residuals are normally distributed. The Q-Q plot is represented in fig 3.13.

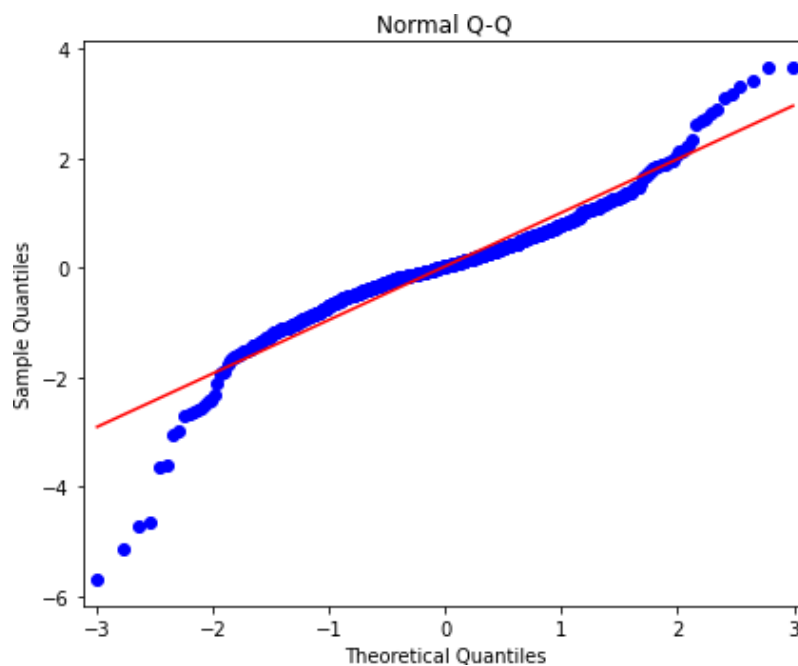


Figure 3.13. The Q-Q plot

The standardized residual plot

The standardized residual is a measure of the strength of the difference between observed and expected values. It's a measure of how significant your cells are to the chi-square value. The residuals over time don't display any obvious seasonality and appear to be white noise. The standardized residual plot is shown in fig 3.14.

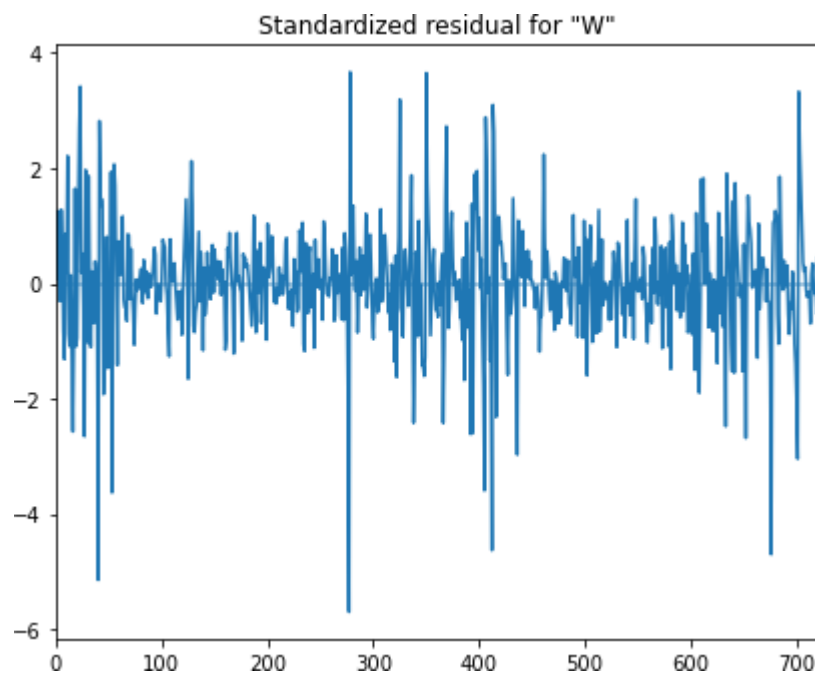


Figure 3.14. Standardized residual plot

The Correlogram plot

In the analysis of data, a correlogram is a chart of correlation statistics. The correlogram is a commonly used tool for checking randomness in a data set. If random, autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero. In addition, correlograms are used in the model identification stage for autoregressive moving average time series models. Autocorrelations should be near-zero for randomness; if the analyst does not check for randomness, then the validity of many of the statistical conclusions becomes suspect. The correlogram is an excellent way of checking for such randomness.

Sometimes, corrgrams, color- mapped matrices of correlation strengths in multivariate analysis are also called correlograms. This plot shows that the time series residuals have low correlation with lagged versions of itself. The correlogram plot is represented in fig 3.15.

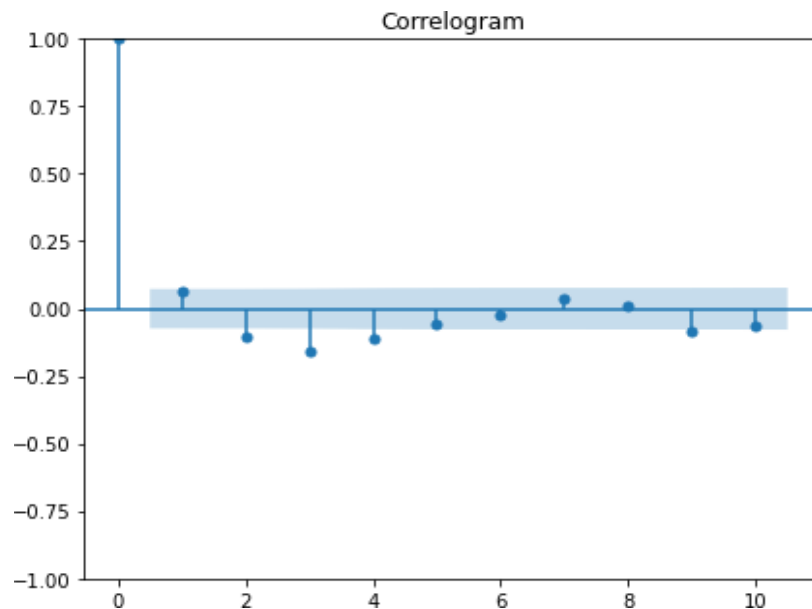


Figure 3.15. Correlogram plot

AUTO CORRELATION

The word autocorrelation is taken from Greek language. From the name itself we can find the meaning that term autocorrelation means the data corresponds with itself instead with some other data. The other name of Autocorrelation is serial correlation and it is a function that represents the time lag between various observations. The autocorrelation analysis is mainly done for finding the repetitive series or patterns and also the missed frequencies of a signal.

The term autocorrelation is generally termed as a variable using time series data. Generally it can be found by plotting ACF plots were tested with Durbin-Watson test.

Finally we can say that Autocorrelation is the degree of correspondence among same variable's values in different observations of the given data.

PARTIAL AUTO CORRELATION

From the name itself we can find that it is partially correlated not complete unlike autocorrelation function.

In data analysis part partial autocorrelation has an important role and is used for identifying how far the lag is extended in auto regressive model. At first time this function is used in time series modeling for determining the lags p in AR (p) model by plotting PACF function.

MAKE PREDICTIONS

Using Time series model, you can forecast a time series using the series past values. In this project, we build an optimal ARIMA model from scratch and extend it to Seasonal ARIMA (SARIMA) model. Now that we have a model built, we want to use it to make forecasts. First I am using the model to forecast for time periods that we already have data for, so we can understand how accurate the forecasts.

The smaller values of the AIC criteria indicate the better statistical estimate model to the fitting. The adjusted root squared overcome the influence of the argument number, used to discriminate goodness-of-fit more effectively than the root squared test. The greater value of the adjusted root squared represent better model fitting. A distinctive feature of the criterion values suggests the appropriate form of the original time series is ARIMA (3, 0, 2) and SARIMA (0, 1, 2) $\times$ (0, 1, 2, 4) as the final prediction model.

ERROR ANALYSIS

After forecasting the time series using ARIMA and Seasonal - ARIMA models, the comparison is made between testing data set and the predicted data set to obtain the accuracy of these models.

The mean square error (MSE) and root mean square error (RMSE) were calculated to perform the evaluation.

Mean Square Error (MSE) is the measure of squared difference between the

true values and forecast values.

$$MSE = \frac{1}{n} \sum_{j=1}^n (\tilde{Y}_j - Y_j)^2 \quad (9)$$

Root Mean Square Error (RMSE) is the measure of square root of difference between the true values and forecast values.

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (\tilde{Y}_j - Y_j)^2} \quad (10)$$

CHAPTER 4

RESULTS AND DISCUSSION

MODEL SUMMARY

The ACF and PACF graphs of both the ARIMA and SARIMA models were inspected inward to decide model fitting parameters. Then each model is fitted with parameters which were found by plotting ACF and PACF plots and predicted upcoming wind speed. The ARIMA and Seasonal- ARIMA models summaries were presented in fig.4.1 and fig.4.2.

Dep. Variable:	Wind speed	No. Observations:	701
Model:	ARMA(3, 2)	Log Likelihood	-1047.932
Method:	csm-mle	S.D. of innovations	1.078
Date:	Tue, 02 Feb 2021	AIC	2109.865
Time:	22:18:48	BIC	2141.732
Sample:	01-01-2016	HQIC	2122.182
	- 12-01-2017		

	coef	std err	z	P> z	[0.025	0.975]
const	3.9444	0.331	11.930	0.000	3.296	4.592
ar.L1.Wind speed	0.7439	0.226	3.298	0.001	0.302	1.186
ar.L2.Wind speed	0.7088	0.356	1.992	0.046	0.011	1.406
ar.L3.Wind speed	-0.4647	0.139	-3.347	0.001	-0.737	-0.193
ma.L1.Wind speed	-0.1113	0.219	-0.509	0.611	-0.540	0.317
ma.L2.Wind speed	-0.7845	0.207	-3.790	0.000	-1.190	-0.379

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.2228	+0.0000j	1.2228	0.5000
AR.2	1.0160	+0.0000j	1.0160	0.0000
AR.3	1.7319	+0.0000j	1.7319	0.0000
MA.1	1.0603	+0.0000j	1.0603	0.0000
MA.2	-1.2022	+0.0000j	1.2022	0.5000

Figure 4.1.ARIMA MODEL SUMMARY

SARIMAX Results						
=====						
Dep. Variable:	Wind speed		No. Observations:	726		
Model:	SARIMAX(0, 1, 2)x(0, 1, 2, 4)		Log Likelihood	-429.491		
Date:	Wed, 24 Feb 2021		AIC	868.982		
Time:	11:49:08		BIC	891.885		
Sample:	0		HQIC	877.823		
	- 726					
Covariance Type:	opg					
=====						
	coef	std err	z	P> z	[0.025	0.975]

ma.L1	-1.3865	0.024	-58.723	0.000	-1.433	-1.340
ma.L2	0.3956	0.020	19.428	0.000	0.356	0.436
ma.S.L4	-1.9963	0.090	-22.223	0.000	-2.172	-1.820
ma.S.L8	0.9973	0.090	11.117	0.000	0.821	1.173
sigma2	0.1801	0.017	10.803	0.000	0.147	0.213
=====						
Ljung-Box (L1) (Q):	2.83	Jarque-Bera (JB):	859.03			
Prob(Q):	0.09	Prob(JB):	0.00			
Heteroskedasticity (H):	0.92	Skew:	-0.63			
Prob(H) (two-sided):	0.55	Kurtosis:	8.20			
=====						

Figure 4.2.SARIMA MODEL SUMMARY

To find the best model, the parameters p, q, d, P, Q, D and s had different values assigned. For each of the parameters, a new model was constructed. The highest value of parameters is 10 because the model should predict with the lowest error values and should be the simplest one. The higher values of p and q are inefficient, and the model becomes too complex to calculate and analyze.

RESULT

The data that are considered for the study were yearly data of wind data from the wind-monitoring systems installed by National Institute of Wind Energy (NIWE), Chennai, from the period 2015–2017. Data set which is collected from NIWE contains the parameters such as temperature, relative humidity, pressure, wind direction and the wind speed.

Two tests were performed with two different datasets to verify performances each model. ARIMA and Seasonal - ARIMA methods are applied in each test and the prediction results of the two methods are compared.

Table 4.1. Summary of test statistical errors

Test	Model	MSE	RMSE
Test 1	ARIMA	1.9469	1.3953
	SARIMA	2.2361	1.4953
Test 2	ARIMA	0.8840	0.9402
	SARIMA	0.9375	0.9682

The predictive accuracy of these models was evaluated based on the mean square error (MSE) and root mean square error (RMSE) which were listed in table 1. From the results we concluded that the ARIMA model is more accurate than the SARIMA model.

PERFORMANCE ANALYSIS

To evaluate the performance of both the ARIMA and SARIMA models graphs were plotted, and represent the relation between actual wind speed and the model predicted wind speed.

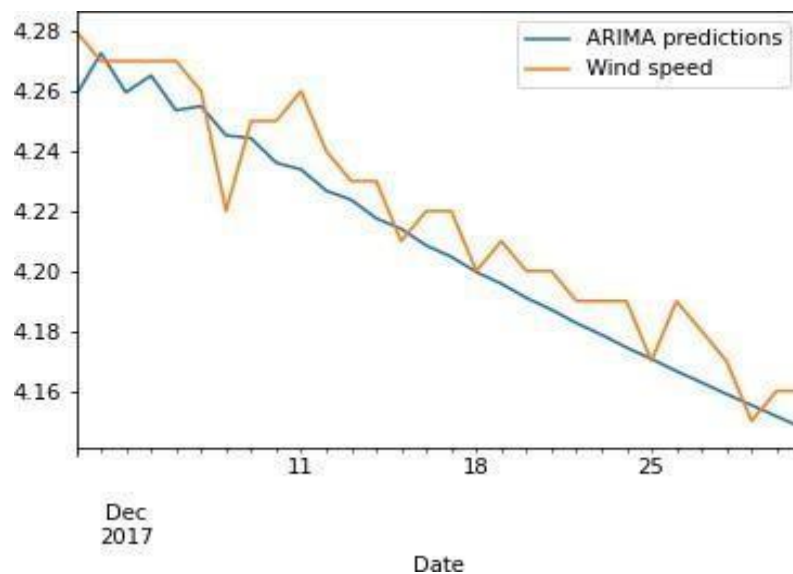


Figure 4.3. ARIMA prediction vs actual wind speed

From fig 4.3 we can understand that the prediction data of ARIMA model is consistent with the trend of real data. The forecast data of the ARIMA model is very close to the real data. The difference between the forecast data of ARIMA model and the real data is small.

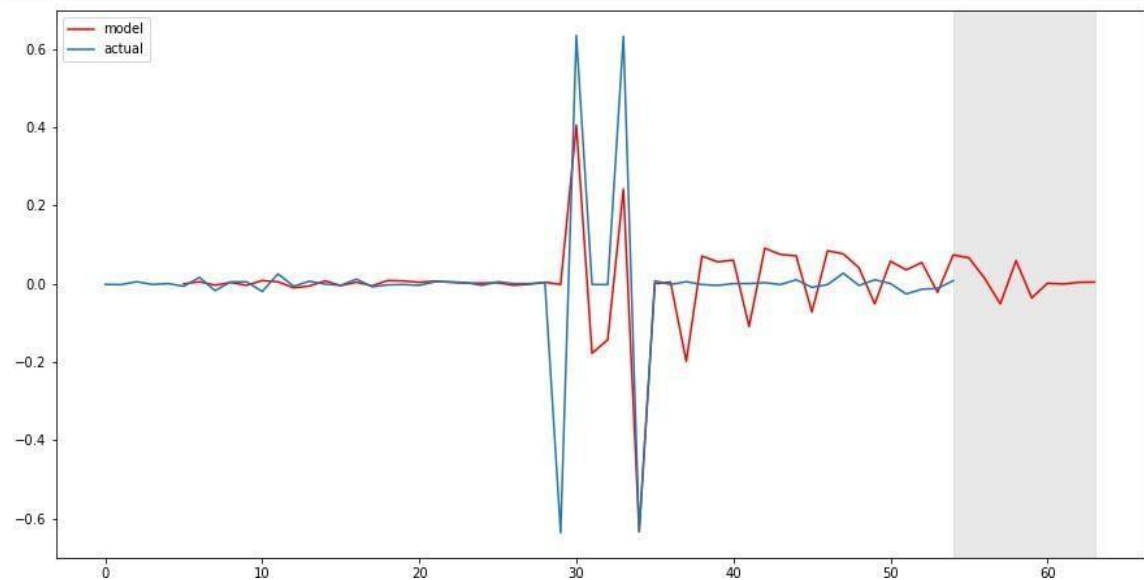


Figure 4.4. SARIMA prediction vs actual wind speed

It can be seen from fig 4.4 that the predicted data of the SARIMA model is highly consistent with the real data. There is a slight delay in the data obtained from the SARIMA model, while there is a significant delay between the data predicted by this model and the real data.

CHAPTER 4

CONCLUSION AND FUTURE WORK

CONCLUSION

Electric power generation is mainly depends on the wind energy. The wind speed prediction has an important place in wind energy systems and to drive turbines that further helpful for generating electricity. Wind power forecasting is primarily depends on the forecasting of wind speed.

.Accurate forecasting results have significant influence on the economy. Recently, academia and industry have paid more attention to wind speed forecasting. More accurate forecasting could reduce costs and risks, improve the security of power systems.

The models to analyze and predict wind speed were developed and tested. One is with the help of Autoregressive Integrated Moving Average (ARIMA) model and, another one is by using Seasonal Autoregressive Integrated Moving Average (SARIMA) model.

The time interval of data is shorter, the accuracy is higher. Results showed the accuracy of daily wind speed forecasting is the best.

Autocorrelation and partial autocorrelation plots can be made to check the authenticity of these models.

The points (3, 0, 2) were taken as the best ARIMA (p, d, q) structure to forecast the wind speed and (0, 1, 2) x (0, 1, 2, 4) for SARIMA (p, d, q) x (P, D, Q, s).

The predictive accuracy of these models was evaluated based on the mean square error (MSE) and root mean square error (RMSE). The ARIMA model poses with low RMSE and MSE values, so we determined that the ARIMA is the selected as better model for the forecasting of wind speed.

To reduce errors obtained, the use of a hybrid model (ARIMA or SARIMA

combined with other techniques) should be addressed.

FUTURE WORK

The main aim for improving the prediction performance for the time series wind speed prediction model is designed and developed in this work. The implemented technique is effective and precise for wind speed prediction but some limitations of the model is also observed thus in near future need to be review before use of the proposed technique.

In the future, we plan to develop our own fuzzy logic techniques for wind speed forecasting. Finally, we need to compare those techniques to find the best one.

To automate this process by show the prediction result in web application or desktop application.

To optimize the work to implement in Artificial Intelligence (AI) environment.

To get the more accurate forecasted values Hybrid model should be used.

REFERENCES

- 1) Yanbin Cui, Chenchen Huang, Yanping Cui. A novel compound wind speed forecasting model based on the back propagation neural network optimized by bat algorithm. Springer-Verlag GmbH Germany, part of Springer Nature 2019.
- 2) Zaher Mundher Yaseen, Anas Mahmood Al-Juboori, Ufuk Beyaztas, Nadhir Al-Ansari, Kwok-Wing Chau, Chongchong Qi, Mumtaz Ali, Sinan Q. Salih & Shamsuddin Shahid. Prediction of evaporation in arid and semi-arid regions: a comparative study using different machine learning models. *Engineering Applications of Computational Fluid Mechanics*, 14:1, 70-89.
- 3) Mohammed Ali Jallal, Samira Chabaa, Abdelouhab Zeroual. A new artificial multi- neural approach to estimate the hourly global solar radiation in a semi-arid climate site. Springer- Verlag GmbH Austria, part of Springer Nature 2019.
- 4) Qiang Zhao, Kunkun Bao, Jia Wang, Yinghua Han and Jinkuan Wang. An Online Hybrid Model for Temperature Prediction of Wind Turbine Gearbox Components. *Energies* 2019, 12, 3920.
- 5) Jiu Gu, Yining Wang, Da Xi, and Yu Zhang. Wind Farm NWP Data Preprocessing Method Based on t-SNE. *Energies* 2019, 12, 3622.
- 6) Mengning Wu, Christos Stefanakos, Zhen Gao, Sverre Haver. The prediction of short- term wind and the wave conditions for marine operations using a multi-step-ahead decomposition- ANFIS model and quantification of its uncertainty. *Ocean Engineering* 188(2019) 106300.
- 7) P. A. Costa Rocha, J. L. Fernandes, B. Modolo, R. J. Pontes Lima, M.E.Vieira da Silva, C. A. Dias Bezerra. Estimation of daily, weekly and monthly global solar radiation using ANNs and a long data set: a case study of Fortaleza, in Brazilian Northeast region. *International Journal of Energy and Environmental Engineering* (2019) 10:319–334.
- 8) Majid Jamil, Mohammad Zeeshan. A comparative analysis of ANN and chaotic approach-based wind speed prediction in India. *Neural Computing and Applications* (2019) 31:6807–6819.
- 9) Jian Jiao. A Hybrid Forecasting Method for Wind Speed. *MATEC Web of*

Conferences 232, 03013 (2018).

- 10) DAVID B. ALENCAR, CAROLINA M. AFFONSO, (Member, IEEE), ROBERTO C. L. OLIVEIRA, AND JOSE C. R. FILHO. Hybrid Approach Combining SARIMA and Neural Networks for Multi-Step Ahead Wind Speed Forecasting in Brazil. Digital Object Identifier 10.1109/ACCESS.2018.2872720.
- 11) Ernesta Grigonytė, Eglė Butkevičiūtė. Short-term wind speed forecasting using ARIMA model. ENERGETIKA. 2016. T. 62. Nr. 1–2. P. 45–55.
- 12) Erasmo Cadenas, Wilfrido Rivera, Rafael Campos- Amezcua and Christopher Heard. Wind Speed Prediction Using a Univariate ARIMA Model and a Multivariate NARX Model. Energies 2016.
- 13) Liu H., Tian H., Liang X., Li Y. New wind speed forecasting approaches using fast ensemble empirical model decomposition, genetic algorithm, Mind Evolutionary Algorithm and Artificial Neural Networks. Renewable Energy. 2015. Vol. 83. P. 1066–1075.
- 14) M. Bouzerdoun, A. Mellit, and A.M. Pavan, "A hybrid model (SARIMA-SVM) for short-term power forecasting of a small scale grid connected photovoltaic plant," Solar Energy, vol.98, pp.226–235, 2013.
- 15) Cassola, F., Burlando, M. Wind speed and wind energy forecast through Kalman filtering of numerical weather prediction model output. Appl. Energy 2012, 99, 154–166.
- 16) Zhang. W, Wu. J, Wang. J, Zhao. W, Shen. L. Performance analysis of four modified approaches for wind speed forecasting. Appl. Energy 2012, 99, 324–333.
- 17) Qu X., Shi J. Bivariate modelling of wind speed and air density distribution for long-term wind energy estimation. International Journal of Green Energy. 2010. Vol. 7(1). P. 21–37.
- 18) Liu, H.; Tian, H.; Chen, C.; Li, Y. A hybrid statistical method to predict wind speed and wind power. Renew. Energy 2010, 35, 1857–1861.
- 19) Cadenas, E.; Rivera, W. Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA-ANN model. Renew. Energy 2010, 35, 2732–2738.
- 20) Ucar, A., & Balo, F. (2009). Evaluation of wind energy potential and electricity

generation at six locations in Turkey. *Applied Energy*, 86(10), 1864-1872.

- 21)C. Monteiro, R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzel mann, "Wind Power Forecasting: State-of-the-Art 2009", Decision and Information Sciences Division, Argonne National Laboratory, Argonne, Illinois, Nov. 2009.
- 22)Zhang, G., Patuwo, B.E., Hu, M.Y., "Forecasting with Artificial Neural Networks: The State of the Art", Inter. Journal of Forecasting,(14), 35- 62, 1998.

APPENDICES

A. SAMPLE

CODE pip install

pmdarima

```
import pandas as pd
```

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
%matplotlib inline
```

```
w_speed=pd.read_csv("windspeeddata.csv",index_col=[0],parse_dates=True)
```

```
w_speed=w_speed.dropna()
```

```
print('Shape of data',w_speed.shape)
```

```
w_speed.head()
```

```
w_speed['Wind speed'].plot(figsize=(12,5))
```

```
color_table= ['blue', 'orange', 'green', 'red']
```

```
fig, axes = plt.subplots(4,1, sharex=True, figsize = (6,18))
```

```
for i, mat in enumerate(w_speed.columns):
```

```
    pd.plotting.lag_plot(w_speed[mat],label=mat, ax = axes[i],c=color_table[i])
```

```
    axes[i].legend(loc='upper left')
```

```
plt.title('Lag plots for each of the parameters')
```

```
_=pd.plotting.scatter_matrix(w_speed, alpha=0.2, figsize=(8, 8), diagonal='kde')
```

```
w_speed_m = w_speed.resample('MS').mean()
```

```
w_speed_m = w_speed_m.fillna(w_speed_m.bfill())
```

```
w_speed_w = w_speed.resample('W').mean()
```

```
w_speed_w = w_speed_w.fillna(w_speed_w.bfill())
```

```
color_table= ['blue', 'orange', 'green', 'red']
```

```
fig, axes = plt.subplots(4,1,figsize = (6,28))
```

```
for i, mat in enumerate(w_speed_w.columns):
```

```

pd.plotting.autocorrelation_plot(w_speed_m[mat].dropna(),label=mat, ax =
axes[i],c=color_table[i])
from statsmodels.tsa.stattools import adfuller

def adf_test(dataset):
    dfctest = adfuller(dataset, autolag = 'AIC')
    print("1. ADF : ",dfctest[0])
    print("2. P-Value : ", dfctest[1])
    print("3. Num Of Lags : ", dfctest[2])
    print("4. Num Of Observations Used For ADF Regression and Critical Values
Calculation :", dfctest[3])
    print("5. Critical Values :")
    for key, val in dfctest[4].items():
        print("\t",key, ": ", val)

adf_test(w_speed['Wind speed'])
from pmdarima import auto_arima
import warnings
warnings.filterwarnings("ignore")

stepwise_fit = auto_arima(w_speed['Wind speed'],suppress_warnings=True)
stepwise_fit.summary()

from statsmodels.tsa.arima_model import ARIMA
train=w_speed.iloc[:-30]
test=w_speed.iloc[-30:]
print(train.shape,test.shape)
print(test.iloc[0],test.iloc[-1])

model=ARIMA(train['Wind speed'],order=(3,0,2))

model=model.fit()

```

```
model.summary()
```

```
start=len(train)
end=len(train)+len(test)-1
pred=model.predict(start=start,end=end,typ='levels').rename('ARIMA predictions')
pred.plot(legend=True)
test['Wind speed'].plot(legend=True)
```

```
pred.plot(legend='ARIMA Predictions')
test['Wind speed'].plot(legend=True)
```

```
from sklearn.metrics import mean_squared_error
from math import sqrt
arima_mse=mean_squared_error(pred,test['Wind speed'])
arima_rmse=sqrt(arima_mse)
print("Mean Square error for ARIMA prediction = ",arima_mse)
print("Root Mean Square error for ARIMA prediction = ",arima_rmse)
```

```
model2=ARIMA(w_speed['Wind speed'],order=(3,0,2))
model2=model2.fit()
w_speed.tail()
```

```
index_future_dates=pd.date_range(start='2017-12-01',end='2017-12-31')
print("Date\t Wind speed")
pred=model2.predict(start=len(w_speed),end=len(w_speed)+30,typ='levels').rename
('ARIMA Predictions')
pred.index=index_future_dates
print(pred)
```

```
pred.plot(figsize=(12,5),legend=True)
```

```
from statsmodels.graphics.tsaplots import plot_pacf
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.tsa.statespace.sarimax import SARIMAX
```

```
from statsmodels.tsa.holtwinters import ExponentialSmoothing
from statsmodels.tsa.stattools import adfuller
```

```
import matplotlib.pyplot as plt
from tqdm import tqdm_notebook
import numpy as np
import pandas as pd
from itertools import product
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

plot_pacf(data['Wind speed']);
plot_acf(data['Wind speed']);

ad_fuller_result = adfuller(data['Wind speed'])
print(f'ADF Statistic: {ad_fuller_result[0]}')
print(f'p-value: {ad_fuller_result[1]}')

data['Wind speed'] = np.log(data['Wind speed'])
data['Wind speed'] = data['Wind speed'].diff()
data = data.drop(data.index[0])

plt.figure(figsize=[15, 7.5]); # Set dimensions for figure
plt.plot(data['Wind speed'])
plt.title("Log Difference")
plt.show()

data['Wind speed'] = data['Wind speed'].diff(4)
data = data.drop([1, 2, 3, 4], axis=0).reset_index(drop=True)

ad_fuller_result = adfuller(data['Wind speed'])
print(f'ADF Statistic: {ad_fuller_result[0]}')
print(f'p-value: {ad_fuller_result[1]}')
```

```
plot_pacf(data['Wind speed']);  
plot_acf(data['Wind speed']);
```

```

def optimize_SARIMA(parameters_list, d, D, s, exog):
    results = []
    for param in tqdm_notebook(parameters_list):
        try:
            model=SARIMAX(exog,order=(param[0],d,param[1]),seasonal_order=(param[
2], D, param[3], s)).fit(dis=-1)
        except:
            continue
        aic = model.aic
        results.append([param, aic])

    result_df = pd.DataFrame(results)
    result_df.columns = ['(p,q)x(P,Q)', 'AIC']
    result_df=result_df.sort_values(by='AIC',ascending=True).reset_index(drop=True)
    return result_df

```

```

p = range(0, 4, 1)
d = 1
q = range(0, 4, 1)
P = range(0, 4, 1)
D = 1
Q = range(0, 4, 1)
s = 4
parameters = product(p, q, P, Q)
parameters_list = list(parameters)
print(len(parameters_list))

```

```

result_df = optimize_SARIMA(parameters_list, 1, 1, 4, data['Wind speed'])
result_df

```

```

best_model = SARIMAX(data['Wind speed'], order=(0, 1, 2), seasonal_order=(0, 1,

```

```
2, 4)).fit(dis=-1)
print(best_model.summary())
```

```
best_model.plot_diagnostics(figsize=(15,12));
```

```
data['sarima_model'] = best_model.fittedvalues
```

```
data['sarima_model'][:4+1] = np.NaN
```

```
forecast = best_model.predict(start=data.shape[0], end=data.shape[0] + 8)
```

```
forecast = data['sarima_model'].append(forecast)
```

```
plt.figure(figsize=(15, 7.5))
```

```
plt.plot(forecast, color='r', label='model')
```

```
plt.axvspan(data.index[-1], forecast.index[-1], alpha=0.5, color='lightgrey')
```

```
plt.plot(data['Wind speed'], label='actual')
```

```
plt.legend()
```

```
plt.show()
```

```
print("Date\t\t Wind speed")
```

```
print(pred)
```

```
print("Mean Square error for SARIMA prediction = ",sarima_mse)
```

```
print("Root Mean Square error for SARIMA prediction = ",sarima_rmse)
```

```
if arima_rmse>sarima_rmse:
```

```
    if arima_mse>sarima_mse:
```

```
        print("Seasonal ARIMA gives less error values, so Seasonal-ARIMA is the best  
model")
```

```
    else:
```

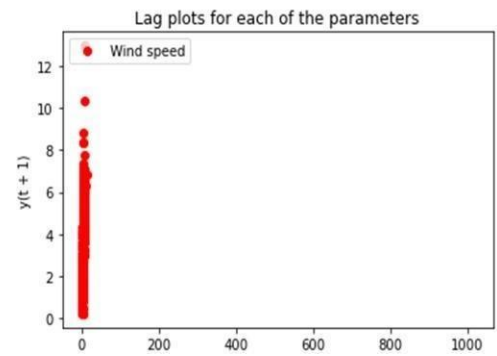
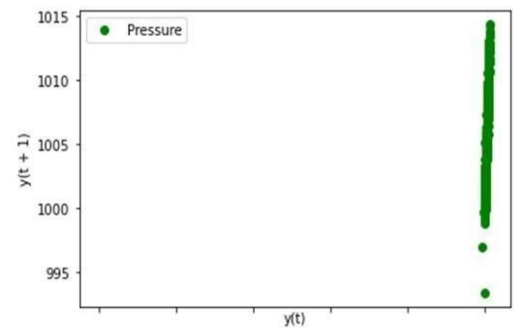
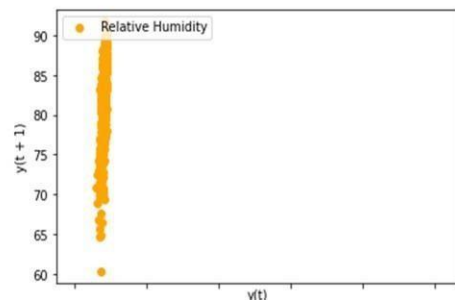
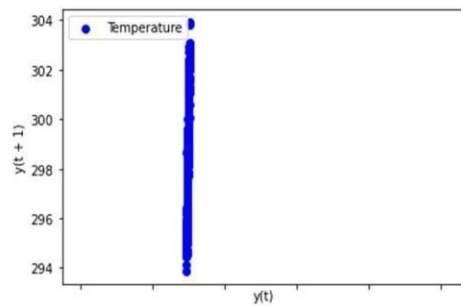
```
        print("Both ARIMA and SARIMA has same error values, So we can't decide  
which is the best model")
```

```
else:
```

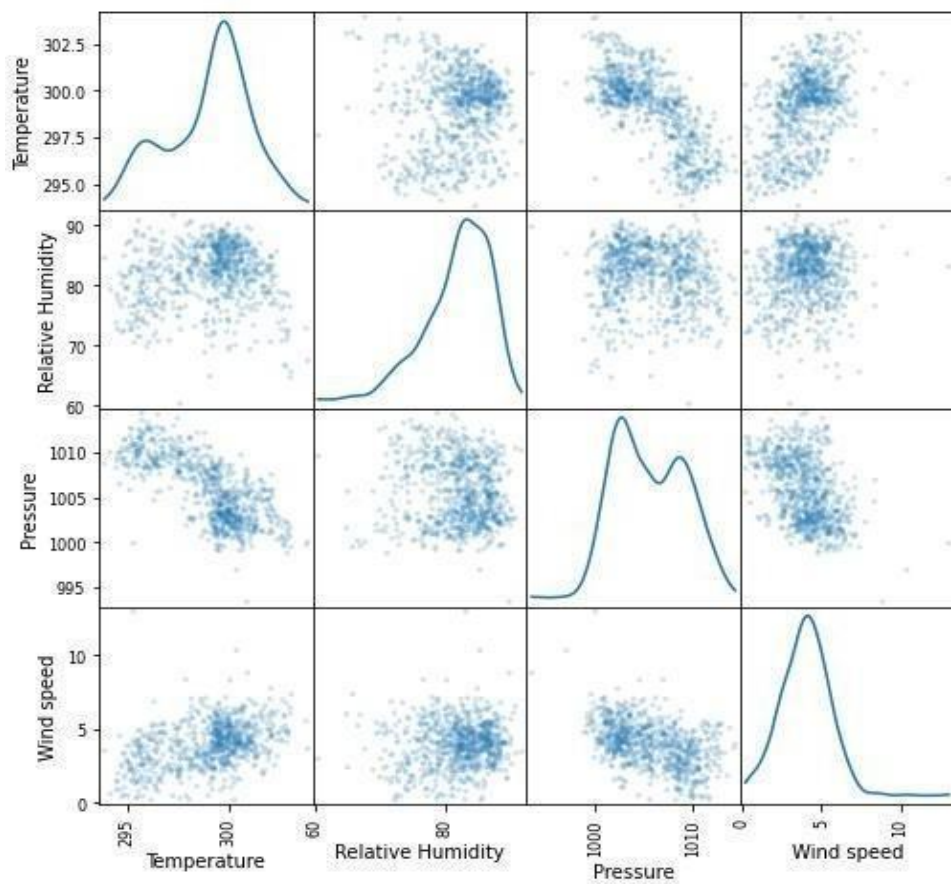
```
    print("ARIMA gives less error values, so ARIMA is the best model")
```

B. SCREEN SHOTS

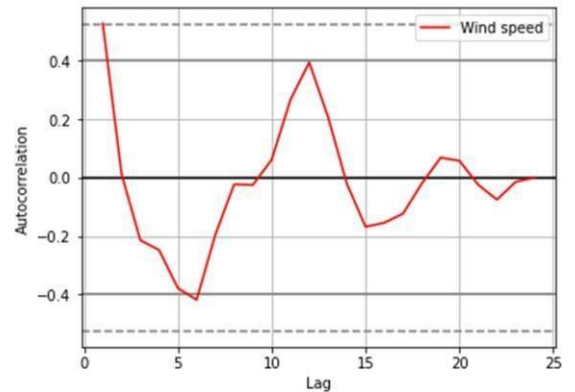
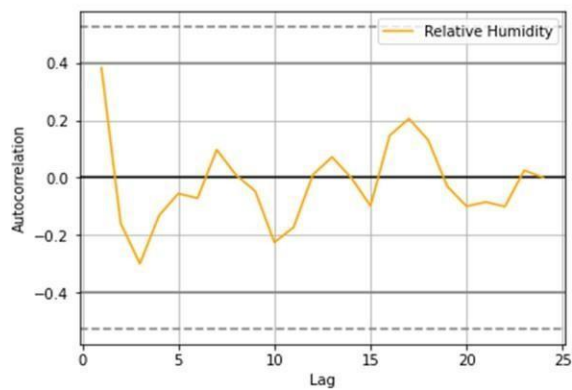
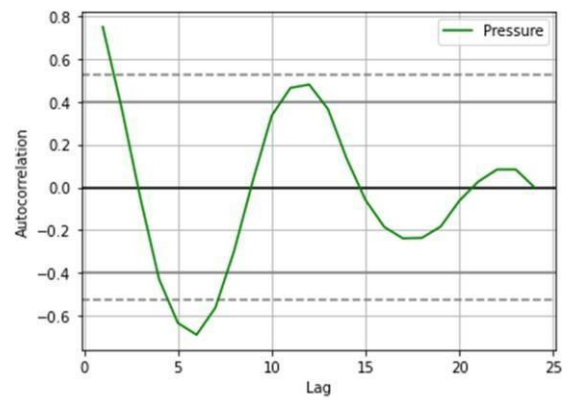
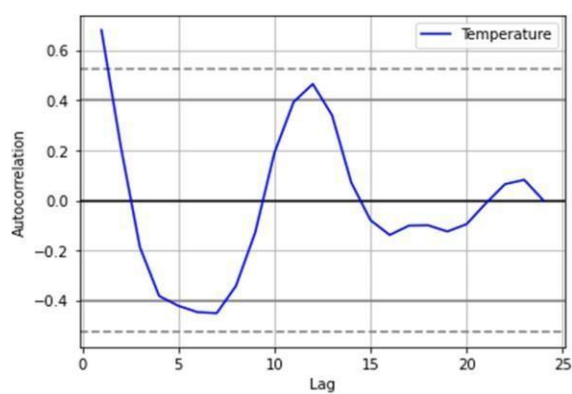
Text(0.5, 1.0, 'Lag plots for each of the parameters')



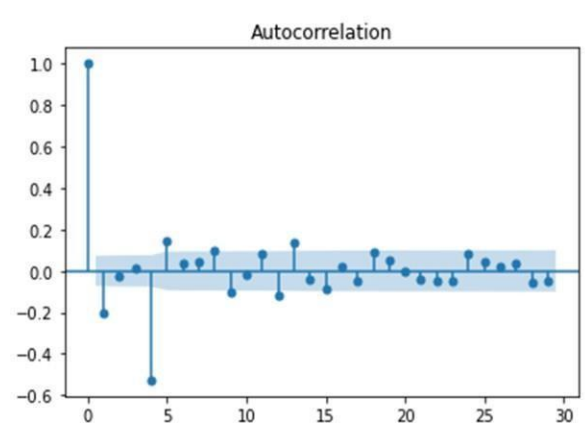
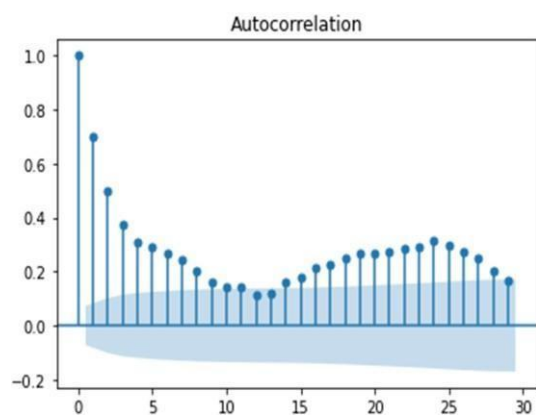
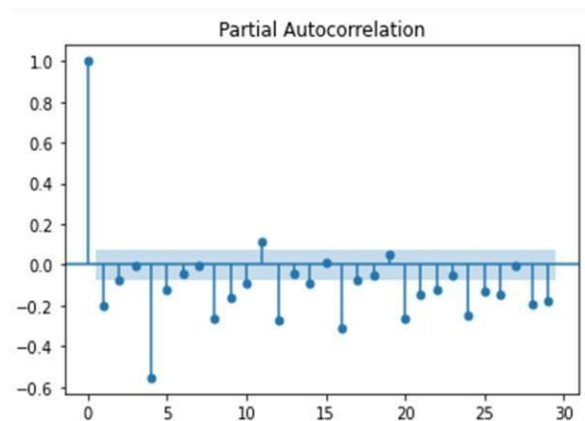
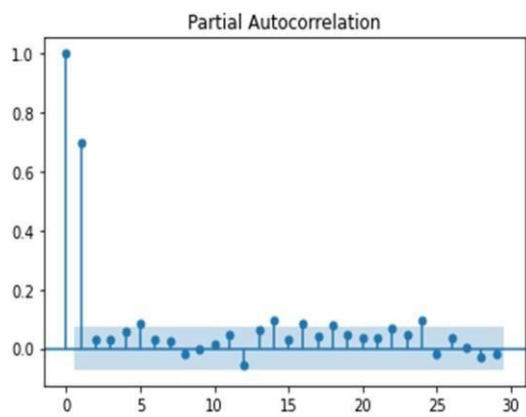
Lag plots for each of the parameters



SCATTER PLOT



Autocorrelation plots for each of the parameters for ARIMA



Autocorrelation and Partial Autocorrelation plots for SARIMA

```

(701, 4) (30, 4)
Temperature          298.15
Relative Humidity     78.84
Pressure             1007.20
Wind speed           7.06
Name: 2017-12-02 00:00:00, dtype: float64 Temperature          296.14
Relative Humidity     86.40
Pressure             1008.16
Wind speed           4.28
Name: 2017-12-31 00:00:00, dtype: float64

```

```

Date
2017-12-01    4.244187
2017-12-02    4.184545
2017-12-03    4.177819
2017-12-04    4.147923
2017-12-05    4.148794
2017-12-06    4.131698
2017-12-07    4.133500
2017-12-08    4.122472
2017-12-09    4.123459
2017-12-10    4.115621
2017-12-11    4.115583
2017-12-12    4.109587
2017-12-13    4.108716
2017-12-14    4.103864
2017-12-15    4.102406
2017-12-16    4.098304
2017-12-17    4.096462
2017-12-18    4.092873
2017-12-19    4.090795
2017-12-20    4.087570
2017-12-21    4.085359
2017-12-22    4.082400
2017-12-23    4.080126
2017-12-24    4.077368
2017-12-25    4.075078
2017-12-26    4.072478
2017-12-27    4.070201
2017-12-28    4.067730
2017-12-29    4.065485
2017-12-30    4.063123
2017-12-31    4.060922
Freq: D, Name: ARIMA Predictions, dtype: float64

```

Mean Square error for ARIMA prediction = 1.9469236492548214
Root Mean Square error for ARIMA prediction = 1.3953220593306843

Mean Square error for SARIMA prediction = 2.2361236492548215
Root Mean Square error for SARIMA prediction = 1.3953220593306843

ARIMA gives less error values, so ARIMA is the best model