Wind Energy forecasting using Time series Analysis

Student ID: X18113494 and X18104215

Abstract:

Wind energy is one of most green energy which can be used for various purposes and converted into electricity. The nature has given many ways to convert one form of energy into another without harming nature itself. Forecasting of electricity from wind energy will help human society to be prepared in many ways, such as to have equipment's to store the generated data, to alarm the society for any alerts, etc. The project aim here is to forecast electricity that can be generated over time based on previous data. San Francisco International Airport data is used here for time series analysis. Many factors are considered such as physical, statistical methods which affect the generation of electricity. For forecasting time series data from 2004 to 2018 is used, here regression models and advanced time series analysis models such as Linear regression, Polynomial regression with seasoning, Holt-Winters model and ARIMA model accuracy are compared and best model is determined. Results of models are evaluated using Level of significance, Residuals and MAPE values. Evaluation of the models is concluded with Holt-Winters model to be the best fit model with 68% accuracy.

Keywords- Wind Energy, Electricity, Forecasting, Linear Regression, Polynomial Regression, Holt-Winter, ARIMA.

1 Introduction

Wind energy is one of the major contributors in renewable energy for power generation. Wind energy is known as green energy because of its pollution free and nature friendly power generation. On the other hand, the non-renewable energy sources like oil, gas, petroleum, coal which are extruded from mines and are known as fossil fuel. [11] These causes pollution in the environment and once they used [12] cannot be reuse again hence these are considered in non-renewable resources. Wind is easily available natural resource. Wind energy is cost efficient as compare to Solar energy and geo thermal energy depending on the geographical conditions. With increasing demand for wind energy, wind industry is one of the growing industries globally. Short term wind speed prediction is important for wind energy production and transformation systems. As the wind speed is varying so the electricity generation from wind is not constant, the prediction of the energy generation can help the power station for distribution of the electricity load. [12] Although the wind is continuously changing factor in the environment, but certain speed of the wind follows specific periodic pattern. We have studied and exploit this wind energy prediction by using time series analysis. The energy generation from wind have lots of benefits like low energy generation cost, clean energy. This research highlights the time series model used for prediction of wind energy generation. The prediction has tested by various time series techniques like linear regression, polynomial regression, polynomial linear regression with seasoning, linear regression with Multiplicative and additive method, Holts winter, Auto regression Integrated Moving Average (ARIMA).

1.1 Research Question:

Does ARIMA and Holt's Winter Model increase the accuracy of the prediction of Wind energy generation?

2 Literature Review on Forecasting of Electricity Generation using Wind Energy:

2.1 Introduction

To find better alternatives for exorbitantly priced and pro-pollutant fossil fuels, investment in cheaper and cleaner energy resources is the need of the hour. As there are many clean energy alternatives are available, this project is keen on focusing Forecasting Methodologies for electricity generation using wind energy.

2.2 Review of Existing Models and Techniques used for Forecasting Wind Electricity

Forecasting wind speed and power is highly important due to the increase in the penetration of wind energy in electricity generation. The wind power forecasting methods are majorly classified based on 'Time-scales' and 'Overall Wind Power Forecasting'. They are further categorized as Ultra-short-term forecasting, Short-term forecasting, Medium-term forecasting, and Long-term forecasting on time-scales basis whereas Overall Wind Power Forecasting has categorized into six methods viz. persistence method, physical method, statistical method, spatial correlation method, artificial intelligence method, and hybrid approach [1]. Hybrid models give the highest level of forecasting accuracy which is a combination of AI method with other methods, but the information provided must be large enough.

Wind speed is the most important parameter in power generation and it gets varied easily due to factors like terrain, height, pressure, temperature and other obstacles. To increase wind power penetration, the power system regulators must make proper planning and set reserve capacity for better forecasting results [2]. Different forecasting models have varied results in different time-scales. In other words, few models perform well in short-term forecasting whereas other models perform well in long-term forecasting. According to Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., & Yan, Z. the conventional statistical model ARMA provides better sensitive output in predicting wind speed. It had been observed that for forecasting horizon of 1 hr, the persistence model had fewer errors than ARMA; whereas for 10 hr advance forecasting, ARMA outperformed the persistence model and had 12 % to 20% fewer errors than persistence model.

There is another method called Grey Model (GM) is used for 1-hour forecasting in which the actual data series is transformed into a new data series of less noise and randomness. Then the differential equation is formed, and the coefficients of this equation are determined by the least square method. Forecasted values of new series are calculated using these coefficients, and actual forecasted values of the original series are retrieved by inverse operation [3]. But it is observed that it presents overshoots and to eliminate those overshoots an alpha-GM is presented in which the coefficients have weighing factors of a differential equation. But it has poor time-series tracking feature. So, two new models 'Improved GM' and 'Averaged GM' are developed. According to El-Fouly and others the improved grey model, which is based on generating two shifted prediction models from the traditional GM (1,1) model. This model increases the prediction accuracy of wind speed and wind power [4]. Another model is averaged GM, which combines the features of both the traditional GM (1,1) and that of Improved GM which helps in reducing the overshoots and reducing the prediction errors at intervals. "The averaged Grey rolling model discloses an improvement in the prediction accuracy, compared with the persistent model, of wind speed up to 26.89% for the MAE, 20.25% for the RMSE and for wind power prediction up to 36.31% for the MAE, 25.83% for the RMSE and 36.34% for the average percentage error” [4, p.937].

Long-term wind speed prediction and power forecasting using the local recurrent neural network (RNN) is carried out by Barbounis and others based on the available meteorological data. In this experiment, an hourly forecast up to 72 hours ahead is generated from the wind park sites which are 30km apart from the wind turbine cluster. Three types of local recurrent neural networks are used as forecasting models, viz., the infinite impulse response multilayer perceptron (IIR-MLP), the local activation feedback multilayer network (LAF-MLN), and the diagonal recurrent neural network (RNN) [5]. The outputs generated from these experiments were compared with static models, a finite impulse response NN (FIR-NN) and a conventional static MLP network. To establish a fair comparison basis, the structure of the competing networks is selected so that they contain approximately the same number of parameters as the recurrent networks. The simulation results showed that it outperforms the FIR-NN and the static MLP by 11.82% and 12.7% in terms of MAE. In this case of NN based experiment, the input selection is quite crucial as the algorithmic complexity raises the computational cost and memory demands of the learning algorithm. The data has to be classified and sort in training and testing batches for performing experimentation. Neural networks have proven better accuracy results than those of traditional and conventional models. A study performed using ANN [6], shows that the sampling time of the input values should be shorter than the time ahead of forecasting in order to get better results while considering differences of wind speeds from their moving average should be preferred as inputs.

Another study is experimented to forecast wind speed using the empirical mode decomposition (EMD) and support vector regression (SVR) methods. The EMD is used to decompose the wind speed time series into several intrinsic mode functions (IMFs) and a residue. Then, a vector combining one historical data from each IMF and the residue is obtained to train the SVR. This EMD-SVR model is assessed with a wind speed data set. It is observed that the EDM-SVR model has performed better than the other models in terms of accuracy and computational complexity [7]. Yet another similar study is done, in which a hybrid model is developed using EMD-ANN for wind speed prediction. The purpose of this work is to find a simple and reliable hybrid intelligent forecasting model for the small wind farms. Liu.H and others studied that to estimate the performance of the EMD-ANN model, two forecasting cases are completed and compared with the ANN model and the Autoregressive Integrated Moving Average (ARIMA) model, respectively. The simulations in this experiment were repeated around 30 times to avoid the randomness caused by the ANN model or the ANN part of hybrid EMD-ANN model. The results show that the EMD-ANN model is robust in dealing with jumping samplings in non-stationary wind series [8].

Modern machine learning techniques like neural networks have been used for forecasting hourly wind speed time series, but it has verified that it is hard to make significant improvements into the performance of the simple persistence model. An alternative approach is experimented using group method data handling (GMDH) based abductive networks. This method gives an advantage of simplified and more automated model synthesis and transparent analytical input-output models. “Models described include a single generic model to forecast next-hour speed from the previous 24-hourly measurements and an hour index, which give an overall mean absolute error (MAE) of 0.85 m/s and a correlation coefficient of 0.83 between actual and predicted values. The model achieves an improvement of 8.2% reduction in MAE compared to hourly persistence" [9, p.1686,1698].

Probing into the core fundamentals of wind power generation, it will be clearly perceived that there are also factors other than wind speed, which make a significant impact in forecasting wind power generation. Those factors are hub height turbulence intensity and wind speed shear across the rotor disk. The generator power, turbine diameter, and blade shape are configured based on-site elements such as annual average wind speed and the wind speed distribution. Turbine manufacturers measure their turbine's 'power curve'—the relationship between power output and wind speed—at turbine test sites. This study seeks a robust method for incorporating turbulence intensity and wind shear into power prediction tools [10]. [10] shows that we need to consider other parameters while forecasting wind speed and wind power. Most studies focused primarily on wind speed mechanism and showed that is the important parameter which affects wind power generation significantly.

In our project we have used various statistical algorithms namely linear regression (LR), polynomial LR, Polynomial LR with seasoning, LR with a multiplicative and additive method, Holds Winter and ARIMA. The detailed explanation of these models is in the following sections of methodology, evaluation, and results.

3. Methodology

There are various scientific data mining methodologies available like CRISP-DM (Cross Industry Process of Data Mining), KDD (Knowledge discovery and data mining technique). This research follows KDD data mining technique for wind energy prediction. KDD consist of 5 different stages.

Fig 1 shows the hierarchical model of KDD.

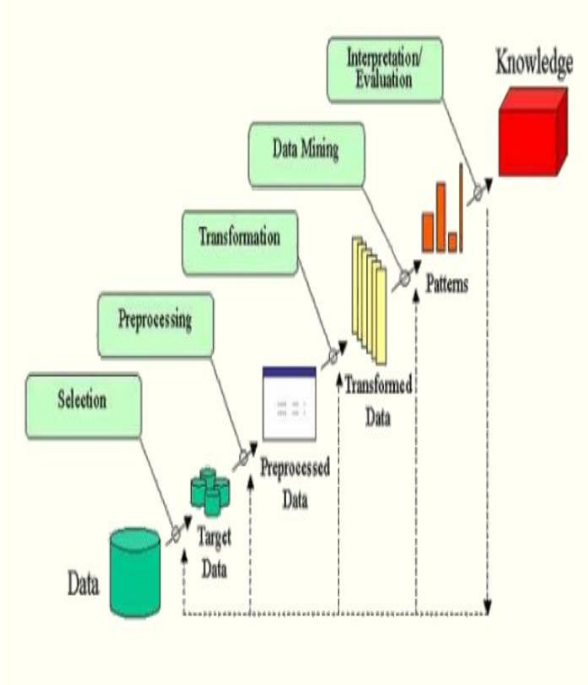


Fig1. KDD Process flow

The steps in the projects are as follows

1. Understanding the domain:

Depending on the business requirement and the proper solution for the business problem the main research in this domain is seasonal wind energy generation prediction. By considering various factors like wind velocity, air density, temperature and calculated power by the equation.

2. Selection of sample data: Depending upon the research question the data for this research model is selected. Data is gathered from California airport weather station data. The main reason behind the selection of California weather station is because California is the state in USA which consist of the greatest number of wind energy generation farm. The data is gathered from https://www.ncdc.noaa.gov/cdoweb/datatools/lcd website. The temperature data is downloaded separately from the same source. Wind velocity data consist of 46 columns and 119821 rows as it is hourly calculated data. The various attribute consists of temperature, wind velocity, air density relative humidity and other weather factors which are varying in the minimum, maximum and average manner. As for this project we need only selected range of data which we have obtained during transformation of the data. The main motto behind taking temperature data is we can calculate air pressure by using the formula.

3. Cleaning and pre-processing: The dataset obtained which consist of some missing values initially. As the data is hourly measures so there were not much noticeable variations in the numbers. The missing values are replaced with total average values. As mentioned earlier for this project we only required Month, Day, year, relative humidity, temperature, air pressure and wind speed columns were important and there were total 46 columns. The columns which are not useful in this project were removed by using R code. For this project we have considered the windmill with same constant area and rotor size.

4. Transformation of the data: The various transformation has been taken are as follows. The first and main transformation is the available temperature was in degree Fahrenheit which is converted in to degree kelvin using formula ((Fahrenheit - 32) * (5/9)) + 273.15. Using this scientific formula, the temperature has been converted in to kelvin. Temperature here we required to calculate air density. Again, in the second transformation takes placed for pressure column, pressure was given in hg which is mercury mm but for calculation of air density we required the pressure in pascal which is converted by using the ratio that 1 hg is equals to 3386 pascals. As mentioned initially that the dataset present was hourly dataset but for time series model we have converted the data in to monthly average data which then reduced to two hundred rows and eight columns.

5. Selecting the technique for Model: Out of all other methods for prediction like neural networks, Regression tree, K-nearest neighbours, multiple linear regression we have selected time series analysis for this project. As the data is suitable for time series forecasting the different time series techniques has been applied for this data.

6.Data Mining Algorithm Selection: As concluded in literature review, data mining techniques such as Linear Regression, Polynomial Linear Regression, Holt-Winters and Auto regressive moving average (ARIMA) were used.

a) Linear Regression

A dependable method which can be used to identify which variable have an impact on a topic is called regression analysis. The relationship between 2 variables by fitting a linear equation to recorded values this process is called linear regression. One variable is a response or dependent variable and the other to be explanatory variable. The idea is to get the best line fit to data that has been used for analysis. There are different values that need to be analyzed in order to get the best fit like residual plot, influential observation, scatter plot etc. Once the model is fit to group of data residual plot will help observer to determine the unused data [13].

Equation for Linear Regression is $Y = a+bX$ where X is explanatory variable, Y being dependent variable and b is the slop line and a is the intercept [14].

b) Polynomial Regression

The regression analysis where the relationship between independent variable and dependent variable is modelled as an nth degree polynomial. Where x being the independent variable and dependent variable is y and the nth degree polynomial is E(y | x). Polynomial regression is a nonlinear model to any data, as a statistical estimator for linear data. For this reason, polynomial regression is a special case of multiple linear regression. Under the condition of Gauss-Markov theorem the least-square method reduces the variance of unbiased estimates of the co-efficient [15].

c) Holt-Winters

Holt-Winters model forecasting is a model used to predict the behaviour of a sequential values or data over time – time series. It is one of the most popular and decade's old model, but still used vastly due to its ubiquitous in many applications including monitoring. Holt-Winters have three aspects of time series value, trend and seasonality, these are expressed as 3 types of exponential smoothing. The model predicts combining all 3 influences provided the data points are in series i.e. seasonal [16].

d) ARIMA

Auto Regressive Integrated Moving Average is termed as ARIMA in short. ARIMA model is used to forecast data or future values based on historical records which is called time series data. It has its own lags, lagged errors and so that the question can be utilized to forecast future values. In order to predict the future values, the data should not be seasonal and there should not be any white noise in the data [17]. ARIMA is characterized by three terms (p,d,q):

p – order of the AR term

d – differences required to make the data stationary

q – order of the MA term

To find the data correlation in the time series ACF (Autocorrelation function) plot is used to find lags and irregularity of data points. Once the shorter lags are removed PACF (Partial Autocorrelation Function) is used to determine the lag k at a point after removing the correlation. After all the above tests a model is built to forecast data points till N and compare the actual to predicted values [18].

4. Results and Evaluation:

Result evaluation is performed based on various iterations and factors data trend, linear curve, residual data analysis, level of significance, RMSE, MAPE etc. p and r-square values are calculated from interval mean values and time. In order to perform time series analysis using any regression model's historical data set with readings are taken and based on different parameter obtained results are interpreted and validated to get the right model for forecasting of values.

1) Linear Regression Model

Linear regression is the most basic model which needs to be tested against any time series data for model fit. After the linear regression analysis LOS (Level Of Significance) obtained was 0.588.

```
Residuals:
    Min      1Q  Median      3Q     Max
-1827.1 -1267.4  -219.8   970.8  4584.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2026.678    208.074   9.740   <2e-16
power$time    -1.082      1.994  -0.543    0.588
```

Figure 2: Linear Regression Residuals

LOS should always be less than equal to 0.05 for any analysis for the model to be the right fit. The RMSE and MAPE values for the same data was obtained as 1.3834 and 1.6106 which are too high for any model.

```
         mae          mse         rmse
1.183308e+03 1.910614e+06 1.382250e+03 1.61
```

Figure 3: Linear Regression RMSE and MAPE

2) Polynomial Regression

Polynomial Regression is the advanced version of Linear regression where the data is further analyzed based on seasoning and trend to get the best fit for the data. In polynomial regression residuals are almost same as linear regression and LOS is 0.648 which is way more than the previous results.

```
Residuals:
    Min      1Q  Median      3Q     Max
-1803.2 -1253.8  -219.4   965.8  4483.6

Coefficients:
                              Estimate Std. Error t val
(Intercept)                 2134.48474  314.99882   6.7
poly(power$time, 2, raw = TRUE)1   -4.63602    8.03546  -0.5
poly(power$time, 2, raw = TRUE)2    0.01964    0.04300   0.4
```

Figure 4: Polynomial regression residuals

Level of significance which should be less than 0.05 is high in the results which we have obtained. The RMSE and MAPE values for Polynomial regression are 1.3814 and 1.6068, which are almost equal. The model fit plot is also obtained to see the line fit of both models.

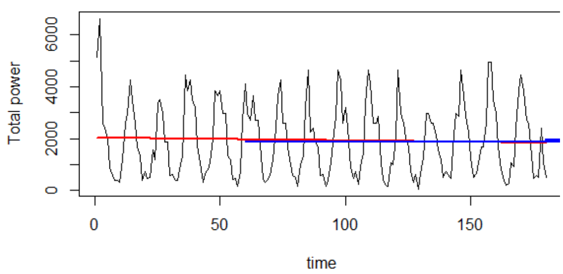


Figure 5: Red – Linear Regression & Blue – Polynomial regression

After the obtained results from iteration 1 & 2 few more iteration with other regression models were applied with seasoning, additive and multiplicative methods but the values were almost same so data was introduced to advanced models.

3) Holt-Winters time series forecasting

This is one of the popular and suitable method for time series data in which specific seasonal trend is present.

```
         xhat    level     trend    season
[1,] 3021.544 1838.327 -1.802557 1185.0191
[2,] 4308.048 1833.992 -1.802557 2475.8582
[3,] 3518.894 1830.323 -1.802557 1690.3741
[4,] 2567.931 1827.961 -1.802557  741.7723
[5,] 1639.128 1826.386 -1.802557 -185.4554
[6,] 1363.318 1826.339 -1.802557 -461.2188
```

Fig4: Holts Winter Outputs

The execution shows xhat is the value which come up with the evaluation of Holts winter model. xhat is not but a calculated prediction of the wind energy by considering seasonality.

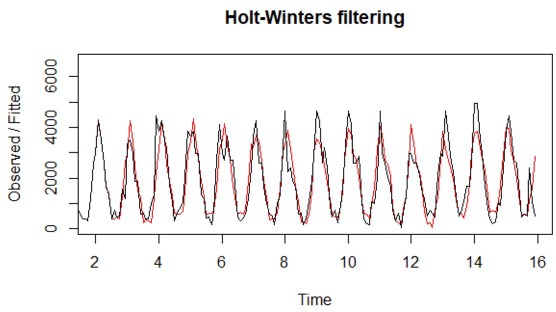


Fig 6: Time series graph Holt Winter

This graph represents the values of energy generation predicted by holt winter model. The values predicted are in red colour whereas the original values are present in black colour. Holt winter shows the accuracy of 67 percentage with error rate of 33. For this time series data we assume that holts winter is one of the suitable model for time series prediction.

4) ARIMA

ARIMA is best fitted to the model when there is seasonal trend and seasonally variable values in the dataset. As in our dataset the values were seasonally varying and from graph its has little increasing trend so we have used ARIMA in our project.

```
ARIMA(1,0,0)(1,1,0)[12] with drift

Coefficients:
         ar1     sar1   drift
      0.2254  -0.4648  0.5611
s.e.  0.0764   0.0722  3.8436

sigma^2 estimated as 448601:  log likelihood=-1331.52
AIC=2671.04   AICc=2671.29   BIC=2683.54
```

Fig7: ARIMA Output

This is the output of the ARIMA execution. As the values are within expected limits further we proceed for prediction graph.

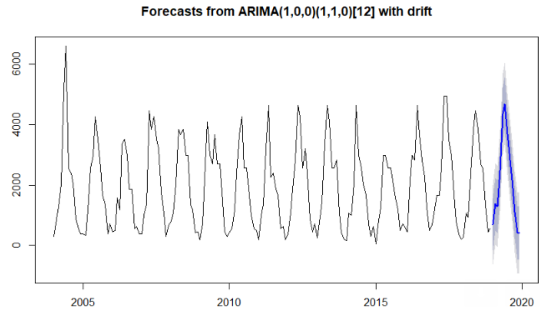


Fig8: Time series graph

The blue colour indicates the predicted values of wind energy generation.

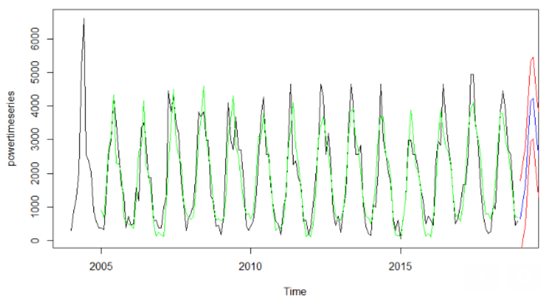
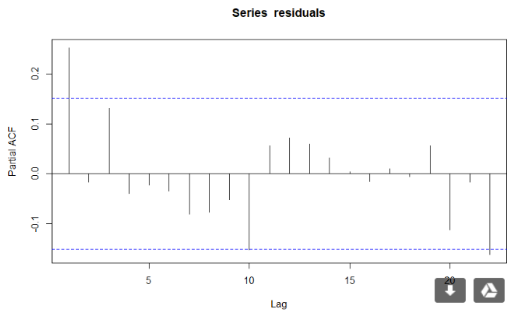


Fig 9: Time series Graph

Here green lines are predicted outputs of ARIMA models are which are very closely like actual output. The blue graph indicates the future prediction with upper and lower limits indicated by red colour.



ARIMA MAPE Value:

```
> MAPE
[1] 34.28878
```

The residual chart indicates that the values are within limits which is indicated by blue line. First are last values are ignored in this chart as a part of time series data.

MAPE value obtained here is 34.28 which is an error rate. We can say that ARIMA showed 65.72 percentage of accuracy. There are many external factors like change in air pressure, temperature which affects on the wind speed hence it is difficult to predict actual wind speed accurately But by using ARIMA model we have achieved around 65 percentage accuracy.

Conclusion

By analysis performed on wind energy data obtained from different sources, Holt-Winters and ARIMA models have more accuracy than any other time series models. The accuracy for both models is almost equal which is 68% and 67% respectively. But data chosen to have more appealing factors for the Holt-Winters model as a model can be even more improved going forward. Therefore, it can be concluded that wind energy forecasting can be done most accurately using Holt-Winters model and it is the best fit considering the physical methods which are not in control of predictive factors.

References

[1] Chang, W.-Y., "A Literature Review of Wind Forecasting Methods". *Journal of Power and Energy Engineering*, 2, 161-168., 2014. Available at: http://dx.doi.org/10.4236/jpee.2014.24023.

[2] Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., & Yan, Z." A review on the forecasting of wind speed and generated power". *Renewable and Sustainable Energy Reviews*, 13(4), 915–920, 2009. doi: 10.1016/j.rser.2008.02.002.

[3] Soman, S. S., Zareipour, H., Malik, O., & Mandal, P. "A review of wind power and wind speed forecasting methods with different time horizons". *North American Power Symposium 2010*, 2010 doi:10.1109/naps.2010.5619586.

[4] El-Fouly, T. H. M., El-Saadany, E. F., & Salama, M. M. A. "Improved Grey predictor rolling models for wind power prediction". *IET Generation, Transmission & Distribution, 1(6), 928.*, 2007. doi:10.1049/iet-gtd:20060564.

[5] Barbounis, T. G., Theocharis, J. B., Alexiadis, M. C., & Dokopoulos, P. S., "Long-Term Wind Speed and Power Forecasting Using Local Recurrent Neural Network Models". *IEEE Transactions on Energy Conversion*, 21(1), 273–284, 2006. doi:10.1109/tec.2005.847954.

[6] Alexiadis, M., Dokopoulos.P, Sahsamanoglou, H., & Manousaridis, I. "Short-term forecasting of wind speed and related electrical power". *Solar Energy,* 63(1), 61–68., 1998. doi:10.1016/s0038-092x(98)00032-2.

[7] Ren, Y., Suganthan, P. N., & Srikanth, N., "A Novel Empirical Mode Decomposition with Support Vector Regression for Wind Speed Forecasting". *IEEE Transactions on Neural Networks and Learning Systems,* 27(8), 1793–1798, 2016. doi:10.1109/tnnls.2014.2351391.

[8] Liu, H., Chen, C., Tian, H., & Li, Y., "A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks". *Renewable Energy*, 48, 545–556, 2012. doi: 10.1016/j.renene.2012.06.012.

[9] Abdel-Aal, R. E., Elhadidy, M. A., & Shaahid, S. M., "Modeling and forecasting the mean hourly wind speed time series using GMDH-based abductive networks". *Renewable Energy,* 34(7), 1686–1699., 2009. doi: 10.1016/j.renene.2009.01.001.

[10] A Clifton, L Kilcher, J K Lundquist and P Fleming, "Using machine learning to predict wind turbine power output". 2013. Available: https://iopscience.iop.org/article/10.1088/1748-9326/8/2/024009/meta. DOI: https://doi.org/10.1088/1748-9326/8/2/024009. [Accessed on: Apr. 26, 2019]

[11] F. M. Guangul, "SWOT Analysis of Wind Energy as a Promising Conventional Fuels Substitute," in *IEEE*, Muscat, Oman, Oman, 2019.

[12] L. Cifarelli, "Introduction to wind energy systems.," *Wind Electric Power Generation,* p. 17, Aug 2018.

[13] Nagy, G. (2018). Sector Based Linear Regression, a New Robust Method for the Multiple Linear Regression. *Acta Cybernetica*, 23(4), pp.1017-1038.

[14] Raghuvanshi, M. (2016). Knowledge and Awareness: Linear Regression. *Educational Process: International Journal*, 5(4), pp.279-292.

[15] Schhnbrodt, F. (2015). Testing Fit Patterns with Polynomial Regression Models. *SSRN Electronic Journal*.

[16] Ortiz, R. (2016). The Accuracy Rate of Holt-Winters Model with Particle Swarm Optimization in Forecasting Exchange Rates. *Journal of Computers*, 11(3), pp.216-224.

[17] Kannan, K. and E. Sakthivel, E. (2011). Fuzzy Time Series Model and ARIMA Model - A Comparative Study. *Indian Journal of Applied Research*, 4(8), pp.624-636.

[18] Bandyopadhyay, G. (2016). Gold Price Forecasting Using ARIMA Model. *Journal of Advanced Management Science*, pp.117-121.

Work Distribution

- Vaibhav (50%) [**x18104215**] – Data Gathering, Literature Review, Holt-Winter and ARIMA Execution.
- Dharshan (50%) [**X18113494**]– Data Pre-processing, Literature Review, Linear Regression and Polynomial Regression Model.