Evaluation of Weather Forecasting Models and Handling Anomalies in Short-Term Wind Speed Data

Jayasri PA
d20z220@psgitech.ac.in

Manimegalai R
drrm@psgitech.ac.in

Reshmah CS
d20z213@psgitech.ac.in

Kaushik S
d21z607@psgitech.ac.in

Department of Computer Science and Engineering
PSG Institute of Technology and Applied Research, Coimbatore, India.

Abstract. The weather forecasting models are useful for understanding and prognosticating wind speed trends and their implications. Their usefulness varies across regions and there is a need to evaluate them before performing time-series analysis for wind energy production. This work evaluates the weather forecast models such as GFS and Meteoblue for the southern region of India at three places Aravoyal, Palakad and Sengotttai for two years. Live data from these wind farms is used in combination with the forecasts, to identify the suitable forecast model for the region. This work takes into consideration that the incoming live data is prone to environmental and mechanical disturbances. It analyzes linear and decision tree regression in handling anamolies in the dataset. The analysis reveals that Meteoblue forecasts are accurate and linear regression provides good imputation results for a period of three hours.

Keywords: Time-series Analysis, GFS, Meteoblue, Linear regression, Decision tree regression.

1 Introduction

The growing demand for wind energy worldwide which is driven by population growth, increasing energy demand and environmental concerns, require accurate wind power forecast. Recent advances in wind energy forecasting systems [14] have demonstrated their ability to optimize power generation by increasing system reliability and reducing operating costs. Wind forecast models play a crucial role in time-series analysis of wind. Effecient power planning for a region can be done by harnessing the insights. These models are available to customers in open source and propreitary versions. The current study focuses on comparison of forecast values generated by GFS and Meteoblue models betweeen 2019 and 2021 in the region of Aravoyal, Palakad and Sengottai. Meteoblue is a

Switzerland based company which provides forecast using ECMWF and its own private models. It is propreitary in nature, whereas, GFS is open-source. This analysis helps wind energy producers to choose the right model for further forecast analysis. This analysis helps them to effeciently manage their operational costs. European Center for Medium-Range Weather Forecasts (ECMWF) and Global Forecast System (GFS) are weather models of two regions of the world, Europe and the US. The first one is a consolidated service and the second is a single service. The resolution of ECMWF is twice as high as that of GFS - 14 km vs 27 km, but GFS is updated twice as often as every six hours. ECMWF forecast updates every hour, which is three times more often than GFS.

2 Literature Review

The evaluation metrics for wind power forecasts presented in [1] acts as the foundation for the current study. It reviews several factors affecting prediction such as location, size of the system and forecasting horizon. It highlights the relevance of input data to forecast models supplied in the format of Numerical Weather Prediction (NWP) and time-series. It describes the performance of forecasting models using metrics such as Root Mean Square Error (RMSE) which is integral to the current study.

The existing literature usually considers wind speed observations taken at 15 minute intervals for a certain duration in years. This makes the dataset very much cumbersome to handle with lots of time intervals taken as attributes. This puts forward the issue of handling the size of the dataset. The importance to use Principal Component Analysis (PCA) as a tool to reduce the dimentionality of the meteorological data and extract useful features for futher statistical analysis is highlighted in [2-4]. The principal components obtained reduces the noise and unwanted correlation in the dataset. It acts as a useful method in cleaning the dataset. It is highly effective when combined with machine learning models for forecast.The current study has a necessity to compare the linear combinations of wind speeds obtained through PCA, to decide suitable forecast model that is effective in the regions considered. The inspiration to take in spearmann correlation to compare parameters and draw results is taken from [5]. The work in [5] utilises this statistical techniques to perform a correlation study between various meteorological factors and COVID-19 pandemic. This highlights how parameters composed of dynamic data is handled effectively through correlation. This motivates the current study to consider observed data from wind farms to correlate with the forecast models.

The importance of forecast models comparison is discussed in [6]. It highlights the statistical and machine learning models that can prove effective with the forecast models. It put forwards techniques like linear regression (LR), Autoregressive Integrated Moving Average (ARIMA), k-Nearest Neighbours (k-NN), etc. The factors that affect the real-time data obtained from the SCADA systems of wind farms and the anamolies encountered are presented in [7, 8]. The study in [8] elaborates the parameters affecting wind turbine performance which in turn affects the wind speed data generated. This puts forward the concern of treating the incoming data to improve the accuracy for futher machine learning analysis. This rises the need for imputation of anomalies. The motivation to impute anomalies through linear regression is derived from [9]. Linear regression acts as a good solution for very small percentage of missing data in the dataset. The usage of decision tree as an alternative for wind forecast is discussed in [10]. It considers several decision trees through the technique of Bagged Regression Trees, whereas current study employs a single decision tree.

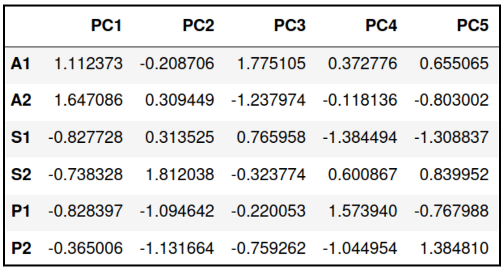
Decision tree is popular machine learning approach for classification and regression. In this paradigm, a set of decisions and their outcomes are represented by a tree-like structure [18]. The algorithm poses a query about a particular aspect of the data at each node of the tree. Depending on the response it moves on to the next node until it reaches a leaf node that signifies the outcome [11]. By choosing the optimal characteristic to divide the data into subsets that maximize the homogeneity of the subgroups, the decision tree is constructed. Recursively, this process is carried out up until the tree is finished or a halting requirement is satisfied. In machine learning, decision trees have various advantages. They can handle categorical and numerical data and are simple to visualize and analyze. They can be applied for feature selection and can deal with missing values [11]. Decision trees may not always be the best accurate model for complicated datasets since they are prone to overfitting if not appropriately pruned.

Linear regression is a supervised machine learning algorithm. It is one of the most popular statistical methods for modelling and data analysis [12]. The technique searches for the best-fitting linear function by assuming that the input and output variables have a linear connection [19]. In two dimensions, the function is represented as a straight line; in higher dimensions, it is represented as a hyperplane. The discrepancy between the output variable's actual value and its anticipated value are reduced by using linear regression. A cost function, such as the Mean Squared Error (MSE), is typically used to quantify this discrepancy. The two primary varieties of linear regression are: *i*). Simple linear regression: When there is only one input variable to predict the output variable, simple linear regression is performed. In this instance, a straight line serves as the linear function. *ii*). Multiple linear regression: This method is used to predict the output variable when there are numerous input variables. In this instance, a hyperplane serves as the linear function. In machine learning, linear regression offers several benefits. It is straightforward, simple to comprehend, and applicable to a variety of issues. It is computationally effective and may be applied to both classification and regression problems. However, it presumes that the input and output have a linear connection [12].

3 Analysis of Weather Forecast Models and Imputation

This study makes use of historical wind speed data gathered in the southern areas of India between 2020 and 2021, from Aravoyal, Palakad, and Sengottai. The records with historical wind speed data are obtained from Leap Green Energy, a reputed green energy distribution firm with headquarters in Coimbatore. These data sources are used to analyze the models and do a thorough wind speed research. The data is collected at 15 minute intervals, throughout the year through the SCADA systems.For dimensionality reduction in data analysis, Principal Component Analysis (PCA) is a frequently used method. By breaking the data set down into a more manageable group of variables known as primary components, it is possible to find patterns in the data and reduce the complexity of the data set [16]. Through PCA, the original data's most crucial details are preserved while a new set of variables is sought for to account for the greatest amount of variance. The first principal component is the direction in which there is a greater variation in the data, and the second principal component is the direction that is orthogonal to the first principal component and accounts a greater variation in the data, and so on [12].

Data compression, data visualization, and data exploration are few uses for PCA. It is frequently used to find hidden patterns in huge data sets and to cut down on the number of variables required to represent the data in areas including economics, biology, image processing, and marketing. PCA is a potent tool that, in the end, can help to simplify complicated data sets, enhance data visualization, and offer insights into underlying patterns and correlations between variables.The wind datasets of forecast models are extremely large as they are recorded at 15 minute interval, contributing 96 slots per day over two years. Therefore, the dimensions of the dataset was reduced through Principal Component Analysis (PCA) and five useful features are extracted. The dataset is then standardized to bring uniformity.



	PC1	PC2	PC3	PC4	PC5
A1	1.112373	-0.208706	1.775105	0.372776	0.655065
A2	1.647086	0.309449	-1.237974	-0.118136	-0.803002
S1	-0.827728	0.313525	0.765958	-1.384494	-1.308837
S2	-0.738328	1.812038	-0.323774	0.600867	0.839952
P1	-0.828397	-1.094642	-0.220053	1.573940	-0.767988
P2	-0.365006	-1.131664	-0.759262	-1.044954	1.384810

This denotes that it is relatively accurate than the other forecast model such as model2. Model Chosen column denotes the forecast model with the higher correlation value for a given slot. Mode of the Model Chosen column is taken to conclude the results.

The incoming data from the SCADA systems of the wind farms have the risk of being errorneous, as they are prone to environmental conditions such as varying humidity, temperature and pressure levels. The conditions in turn affect the SCADA systems leading to wrong or missing data. To perform imputation of erroneous data, two machine learning algorithms such as linear and decision tree regressions are considered and the optimal one is chosen. The given observed dataset contains only wind speeds recorded at each turbine and not any other entities. This rises the need to choose features. To choose features, heatmap is constructed for each place considering the rows and columns as 96 time slots for a period of two years. This creates a 96 x 96 correlation matrix with a gradient. On considering the heat maps for the six turbines, it is identied that, for any given time slot's wind speed on average is highly correlated with its previous four wind slots and its next four wind slots. The eight time slots are considered as features for a regression model. For example, time slot 11:30's value recorded for two years, which creates 730 entries will act as input for a regression model. The regression model is built for each time slot, and therefore creating 96 regression models for both linear and decision tree regressions, in each place. The linear and decision tree regression for each time slot are compared using MSE and r2_scores. This is essential to test the feasability of the models built. Threshold of 0.05 for MSE and 0.7 for r2_score are considered for this work.

4 Results

On taking the mode of the Model Chosen column, Meteoblue has higher correlation than GFS for 73 times out of 96 time slots, accounting for 76.042% and GFS has higher correlation than Meteoblue for 23.958% [6]. Thus, for the taken region Meteoblue is the most reliable forecast model when compared to GFS. But Meteoblue is propreitary in nature, whereas, GFS is open-source. For imputation, we consider the MSE and r2_scores of the regression models for each place. On considering the threshold, linear regression performs better for 98.96% of the slots across all places. Whereas, decision tree regression performs better for 2.083% and 3.125% of slots in Aravoyal turbines 1 and 2 respectively. It performs better for 1.042% and 11.458% of slots in Sengottai turbines 1 and 2 respectively. It performs better for 4.167% and 0% of slots in Palakad turbines 1 and 2 respectively. From this it is observed that the linear regression suits better than decision tree regression for imputation.

References

1. Piotrowski, P., Rutyna, I., Baczyński, D., Kopyt, M.: Evaluation Metrics for Wind Power Forecasts: A Comprehensive Review and Statistical Analysis of Errors. Energies. 15, 9657, 2022.
2. Geng, D., Zhang, H., Wu, H.: Short-Term Wind Speed Prediction Based on Principal Component Analysis and LSTM. Applied Sciences. 10, 4416 ,2020.
3. Ling, Z., Gao, Y., Chen, Q.: Application of Principal Component Analysis in Meteorological Forecast - IOPscience, 2019
4. Deepika, K.K., Varma, P.S., Reddy, Ch.R., Sekhar, O.C., Alsharef, M., Alharbi, Y., Alamri, B.: Comparison of Principal-Component-Analysis-Based Extreme Learning Machine Models for Boiler Output Forecasting. Applied Sciences, 2022.
5. Kumar, G., Kumar, R.R.: A correlation study between meteorological parameters and COVID-19 pandemic in Mumbai, India. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 14, pp. 1735–1742, 2020.
6. Vijayvargia, Archita, Kailash Chand Sharma, and Rohit Bhakar. "A Comparative study of short-term wind speed forecasting models", 2019.
7. Wen, W., Liu, Y., Sun, R., Liu, Y.: Research on Anomaly Detection of Wind Farm SCADA Wind Speed Data. Energies. 15, 5869, 2022.
8. Bashir MBA. Principle Parameters and Environmental Impacts that Affect the Performance of Wind Turbine: An Overview. Arab J Sci Eng. 47(7):7891-7909, 2022
9. Kamisan, Nur Arina Bazilah, Siti Mariam Norrulashikin, and Siti Fatimah Hassan. "Missing Values Imputation For Wind Speed." Applied Mathematics and Computational Intelligence (AMCI), pp. 319-327, 2021.
10. Radmila Mandzhieva, Rimma Subhankulova - Data-driven applications for wind energy analysis and prediction: The case of "La Haute Borne" wind farm, Digital Chemical Engineering, Volume 4, 2022.
11. Sushmitha Kothapalli - A real-time weather forecasting and analysis, IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017.
12. Madan, Shubham & Kumar, Praveen & Rawat, Seema & Choudhury, Tanupriya, Analysis of Weather Prediction using Machine Learning & Big Data, pp.259-264, 2018.
13. Srinivasan, Kathiravan & Nema, Anant & Huang, Chao-Hsi & Ho, Tung. Weather Forecasting Application Using Web-Based Model-View-Whatever Framework. 1-2, 2018.
14. Lawrence, A., Data Analytics and machine learning: Let's talk basics. AnswerRocket. https://www.answerrocket.com/data-analytics-machine-learning/ , 2021.
15. Jijo, Bahzad & Mohsin Abdulazeez, Adnan, Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends. 2. 20-28, 2021.
16. S. Naveen, A. Omkar, J. Goyal and R. Gaikwad, "Analysis of Principal Component Analysis Algorithm for Various Datasets," International Conference on Futuristic Technologies (INCOFT), Belgaum, India, 2022, pp. 1-7, 2022.
17. S. Angra and S. Ahuja, "Machine learning and its applications: A review," International Conference on Big Data Analytics and Computational Intelligence (ICBDAC),Chirala,Andhra Pradesh, India, pp. 57-60, 2017.
18. M. Bhaskar, A. Jain and N. Venkata Srinath, "Wind speed forecasting: Present status," International Conference on Power System Technology, Zhejiang, China, pp. 1-6, 2010.
19. A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, "Overview of use of decision

tree algorithms in machine learning," IEEE Control and System Graduate Research Colloquium, Shah Alam, Malaysia, pp. 37-42, 2011.

20. M. Huang, "Theory and Implementation of linear regression," International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China, pp. 210-217, 2020.