Comparative Analysis of Markov Chain and Polynomial Regression for the Prognostic Evaluation of Wind Power

Mounika Yakasiri, M. Anuradha and B. K. Keshavan

January 29, 2020

Comparative Analysis of Markov Chain and Polynomial Regression for the Prognostic Evaluation of Wind Power

Mounika Yakasiri*, Anuradha.M†, Keshavan.B.K‡
Department of Electronics and Communication Engineering*†
Department of Electrical and Electronics Engineering‡
PES University
Bengaluru, India
Email: *rinku2197@gmail.com, †anuradha@pes.edu, ‡keshavanbk@pes.edu

***Abstract*—Humanity's quest to comprehend the nature has led to increasing demand in predictive sciences. Much of this is due to the exponential growth of energy production and the ever increasing power prices. Regions with power productions that are carried out with imported coal and oil will be more self-sustaining if alternative sources of power generation such as wind power is discharged. In this paper two predictive models, Markov Chain and Regression are implemented to predict wind power. In the first model, Markov Chains with 15 and 30 states has been constructed for the short- term forecast of the wind power. Following this a second order polynomial regression with independent variable as wind speed and dependent variable as wind power has been implemented for the medium-term forecast of wind power. A comparative study of both the models has been made to give a picture of the best model that suits the forecast. The geographical region under study is a wind farm in Chitradurga, Karnataka. The wind speed values have been sampled at an interval of 10 minutes for a period of 3 years starting from 1st January 2010 to 31st December 2012. Various forecasting errors have been enumerated to audit the credibility of the models.**

***Index Terms*—Stochastic models, Markov Chains, Transition matrix, Transition Diagram, Regression Analysis, Polynomial Regression, Correlation, Forward Selection Procedure.**

I. Introduction

Electricity grid demands a balance between the energy generation and consumption, incessant use of electricity and its applications seek to take complete advantage of its versatility and clean form. As long as sun exists major renewable energy forms continue to thrive perpetually. Wind power is the most prepossessing form of renewable energy, much of it is due to the wide range of benefits. Electricity generated from wind has no carbon dioxide emissions and hence does not account for greenhouse effects, it is labour intensive and thus creates more jobs. The total power that can be extracted using wind is considerably more than the current human power consumption, given that the wind power generation is a direct function of wind speed, focus on wind speed forecasts should be given the utmost importance. Considering the wind energy generation to be economical, India's major amount of energy that is being imported could be substituted with this renewable energy form. In this paper two models of wind power forecast has been implemented and each fall into different forecasting time-scales. The two models considered for forecast is Markov Chain and Regression. Markov models are omnipresent having wide range of applications due to its simplicity and the property of 'memorylessness'. This condition is used to make future forecasts in series of time-steps in each of which a random choice is made [1][10]. In this paper Markov Chain is exhibited using two states to compare the state wise accuracy. Increasing the number of states in a Markov model with a narrowed range set leads to a more precise state jump. This in turn leads to obtaining better accuracies with reduced MAPE. Wind power forecast has been made and compared with the two states implemented and the results are validated with the past data [1][2][10].Forecasts made using this model fall under short-term forecasts What follows this model is the Regression Analysis, a second order Polynomial Regression has been applied to the wind speed and formulated wind power which are considered as independent and dependent variables respectively [3]. An explanatory model that has a relationship between input and output facilitates a better understanding of the situation and permits experimentation with different arrangement of input to analyse their influence on the forecast i.e. the output. The forecasts made using this model on wind power falls under medium-term forecass [4]. Finally these two models are compared to know the better forecasting technique for wind power. The geographical area studied is a wind farm in Chirtadurga where wind speed values are sampled at an interval of 10 minutes over a period of three years. One year data that falls under the optimal range i.e. the cut-in and cut-off speed is used as the training data, the forecasts made are validated using various errors.

II. Models For Prediction

The Wind speed values have been sampled at an interval of 10 minutes for a period of 3 years starting from 1st January 2010 to 31st December 2012. The three years wind speed data obtained has been used to compute synthetic wind power using the formula,

$$P = 0.5\rho AV^3 \tag{1}$$

Several factors related to the site of interest shown in Table 1 have been considered to compute wind power.

TABLE I
ENERCON E40 MACHINE SPECIFICATIONS

Rated Power	600KW
Rotor Diameter	44m
Swept area	$1521m^2$
Cut-in Wind Speed	2.5 m/s
Cut-off Wind Speed	29 m/s
Air Density	$1.188Kg/m^3$
Average Temperature	24^oC

A. Markov Chain

Markov Chain is the simplest form of stochastic model showing its simplicity in its property which states that the historic events do not affect the future given the present event. It is a stationary model and hence independent of the initial value. Using a discrete valued Markov Chain, forecasts are made such that the system only changes during those discrete time values [1]. A discrete valued Markov Chain has states ‘s’ all belonging to a countable set ‘S’ and is represented as,

$$S = (1, 2, 3.......n) \tag{2}$$

Two state categories – 15 states and 30 states are implemented using a second order Markov Chain. If a state ‘s’ moves from state ‘i’ to ‘j’ in one step, two steps so on till ‘r’ steps then the transition probability is of the order 1, 2, ..., ‘r’ [2] [3]. The transition probability is given by the equation,

$$P_{ij} = P[X_s + 1 = j|X_s = i] \tag{3}$$

Where Pij is independent of-the time period, any Markov Chain with this property is said to be a time homogeneous Markov Chain or a Stationary Markov Chain. This phenomenon can also be called as the Steady-State Markov Chain as it is a steady state vector-of-the Transition matrix. A steady-state is an eigenvalue for a stochastic matrix, where a probability vector multiplied by a probability transition step matrix results in the same probability vector [1]. For all the stochastic processes the eigenvalues are fixed to $\lambda = 1$. The transition probabilities are encapsulated in the form of a matrix thereby forming the Markov matrix. A more comprehensive explanation using Chapman-Kolmogorov equation has been made in the paper written by us on Markov Chains [10].

The Markov matrix forms the basis of the forecasts, each row in the matrix is a conditional distribution and each column gives the probability of landing in state ‘j’ from some initial state ‘i’. The matrix formed is a right transition matrix whose single rows add up to ‘1’. The second order transition probabilities are used to generate a Markov matrix of 15x15 and 30x30 and Fig. 1 represents a state transition diagram schematically. Most importantly the transition matrix gives an adequate idea about the behaviour of the Markov Chain [3]. The following points have been observed from our transition matrix in Table 2.

- The highest transition probabilities fall in the diagonal of the matrix thereby indicating that a wind value is highly likely to remain in the same state once in that state.
- The probabilities increase till the highest probability is attained and then decreases drastically.
- The farthest a wind values can increase is just two – three states beyond the present state and the least it can come down is just two – three preceding states.

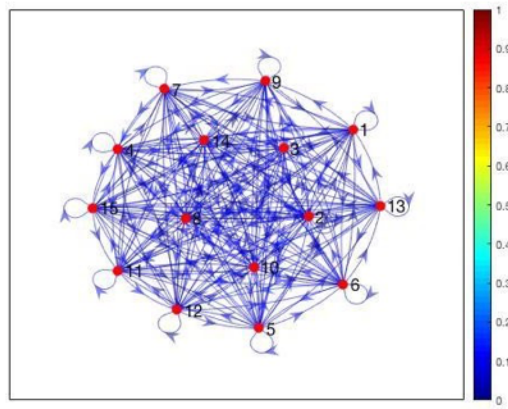

Fig. 1. State Transition Diagram Of 15x15 Markov Chain

One year data excluding the values falling outside optimal range that is the cut-in speed 2.5 m/s and cut-out speed 29 m/s is used as the training data. Varying the range of training data the best accuracy is chosen for a short term forecast for 15 states and 30 states and validated using Various errors like MSE, RMSE, MAE, MAPE.

B. Regression Analysis

A quantitative analysis which determines the relationship between the estimator and the predictor that are categorized into dependent and independent variables respectively is known as Regression Analysis. Here, the known data is regressed to its primitive form which is mathematical. The goal of Regression analysis is to fit a line in such a way that the difference between the data points and the regressed points is minimized [4] [5]. Correlation between the data under consideration determines the extent of dependency, linking the dependent and independent variable. We have ‘n’ pairs of values $(Y_i^{'}, Y_i)$ which are the forecasted values and real values respectively. It is important to know how these two terms are related to each other and their degree of dependency. The correlation between these two values is designated as ‘R’ which is usually represented in the squared form as ‘R2’ and is known as the ‘Coefficient of Determination’. Correlation is in squared form because it can explain the meaning of proportional variation in ‘$Y^{'}$’ which is explained by ‘Y’.

$$R^2 = \frac{\sum(Y^{''} - Y^{'})}{\sum(Y - Y^{'})} \tag{4}$$

TABLE II
A 15X15 TRANSITION MATRIX SHOWING THE INDIVIDUAL PROBABILITIES OF THE STATES

0.528	0.400	0.054	0.011	0.004	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.156	0.604	0.205	0.028	0.005	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.014	0.187	0.578	0.193	0.022	0.004	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.002	0.022	0.215	0.564	0.173	0.021	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.001	0.004	0.028	0.245	0.533	0.169	0.017	0.003	0.001	0.001	0.000	0.000	0.000	0.000	0.000
0.001	0.003	0.003	0.037	0.255	0.514	0.160	0.023	0.003	0.001	0.000	0.000	0.000	0.000	0.000
0.000	0.001	0.001	0.008	0.039	0.295	0.457	0.162	0.031	0.005	0.002	0.000	0.000	0.000	0.000
0.000	0.000	0.001	0.000	0.008	0.074	0.281	0.423	0.169	0.036	0.005	0.003	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.003	0.011	0.083	0.306	0.397	0.154	0.044	0.001	0.001	0.000	0.000
0.000	0.000	0.002	0.002	0.000	0.004	0.012	0.095	0.307	0.390	0.177	0.009	0.000	0.002	0.000
0.000	0.000	0.000	0.000	0.000	0.009	0.011	0.020	0.087	0.242	0.513	0.116	0.002	0.000	0.000
0.000	0.000	0.009	0.000	0.009	0.000	0.000	0.026	0.043	0.470	0.342	0.094	0.009	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.143	0.429	0.250	0.143	0.036
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.455	0.364	0.182
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.167	0.500	0.000	0.333	0.000

Because the 'Y' values are explained in association to the estimated regression equation, It can also be expressed as,

$$R^2 = \frac{\text{Explained variance of Y}}{\text{Total variance of Y}} \tag{5}$$

The closer the value of R to '1' more the dependency and better the results. The data under consideration is wind speed and power and are categorized as follows:-
-Wind speed - Independent variable
-wind power - Dependent variable.
Our initial understanding was that the two variables under study had a relationship that is directly proportional, hence a linear model was applied to the data set. However it was identified that Linear Regression was not a best fit for the data. Very high errors in MAPE and scattered values in the run plot against wind speed were among the first few indications of the inappropriate model. Nevertheless merely looking at the original scatterplot does not confirm non- linearity, it is the Residual graph with no pattern, i.e. no increasing or decreasing trend curve that confirms if the model is linear or non-linear. Run plots generated for linear and non-linear regression shown in Fig. 2 and Fig. 3 explain the shift from linear to non-linear comprehensively. The blue circle in Fig. 2 below represents the defined section for which the Linear regression line fits the data almost perfectly resulting in predicted values very close to actual values. Hence, the error in this region is less, but the vertical deviation of the predicted values at the extreme ends of the curve are very far from the regression line resulting in large error. From the graph it is evident that the region best explained by the model is very less in comparison to the totality of the model. In some cases it is seen that though the prediction is made for a small duration the error is much higher, this is because the predicted regression line passing through the graph only defines a part of the curve, in some cases it is seen that the linear graph does not even account for even one point of the original data set and hence has an accuracy of 0%. Consequently a Non-linear Regression has been employed to forecast the wind power because of the definite curvature observed in the residual plot representing a non-constant variance spread across the graph [6]. Fig. 3 shows a second order Polynomial Regression applied to the same set of data as seen in Fig. 2.

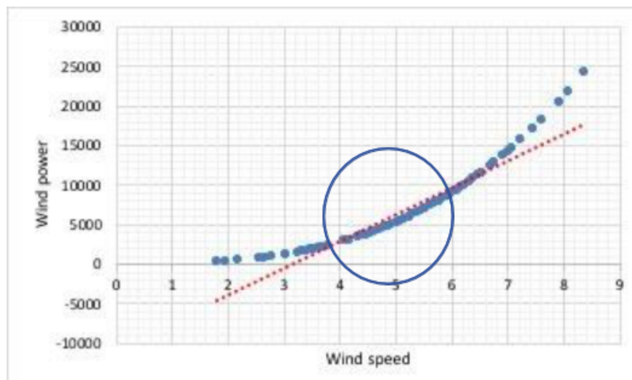

Fig. 2. Linear Regression Applied On Windspeed And Power

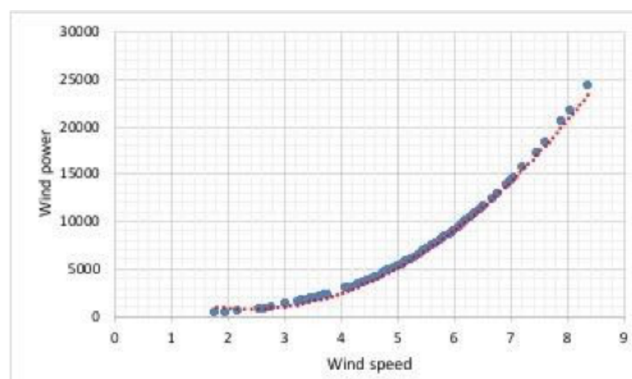

Fig. 3. Second Order Polynomial Regression Applied On Windspeed And Power

It can be observed that this model has adapted itself to fit the data curve hence explaining all points. However it is important to keep the order of the polynomial low, setting arbitrary orders to the data can lead to misuse of the Regression model. The model building strategy employed is of 'Forward Selection Procedure'. In this method the order of the model is successively increased to attain an appreciable value of the regression coefficient and order that fits the data well. Polynomial Regression is such that as the order of the model is increased the curve is altered in such a way that it fits the data covering even the outliers leading to an overfitting model. The overfitting model may be suitable to the present computing data but gives erroneous values when new data points are added, this results in unreliable future predictions. Hence a second order polynomial regression has been applied to the data set to obtain more reliable balanced curves. The second order polynomial is represented as,

$$\alpha = a_0 + a_1 T + a_2 T^2 \tag{6}$$

Where 'α' is the independent variable and 'T' is the dependent variable. The coefficients in equation 7 is calculated using the formula,

$$\begin{bmatrix} n & \sum T_i & \sum T_i^2 \\ \sum T_i & \sum T_i^2 & \sum T_i^3 \\ \sum T_i^2 & \sum T_i^3 & \sum T_i^4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum \alpha_i \\ \sum \alpha_i T_i \\ \sum \alpha_i T^{(}2) \end{bmatrix} \tag{7}$$

Application of the second order quadratic equation for the forecast of wind power with very high accuracy for 1 day and a considerably good accuracy for three days has been presented in Table 5 with the graph in Fig. 5. The presence of a dependent variable increases the model accuracy in Regression Analysis giving a detailed interpretation of the correlation between the estimator and predictor. Due to the high accuracies obtained Regression Analysis can be used for long term forecast.

C. Comparative Analysis

The two Stochastic models that have been implemented and described for the estimation and forecast of wind power are second order Markov Chain and Second order Polynomial Regression. It has been noticed that increasing the number of states in a Markov model with a narrowed range set leads to a more precise state jump. This in turn leads to obtaining better accuracies with reduced MAPE. In regression Analysis the forecast is given by a function with certain number of factors that affect its output. An explanatory model that has a relationship between input and output facilitates a better understanding of the situation and permits experimentation with different arrangements of input to analyze their influence on the forecast. By doing this the explanatory models can be geared toward intervention, influencing the future through decisions made today. It has been observed that the forecasts made using Markov Chains are not applicable for long term prediction given the acceptable MAPE and accuracy being 25% and 75% respectively. However it can be used for short-term forecast up to 14 hours with acceptable MAPE being 25%. The Polynomial Regression of order two better explains the data than the linear model, therefore, the nonlinear model makes better predictions and also predicts for a good amount of time, hence it can be used for a medium-term forecast. Table 6 shows the comparative results of the two models.

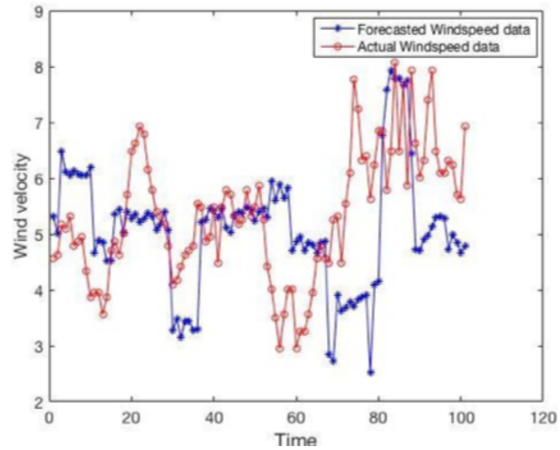

Fig. 4. 14 Hour Forecast Of Wind Power Using 30x30 Markov Chain

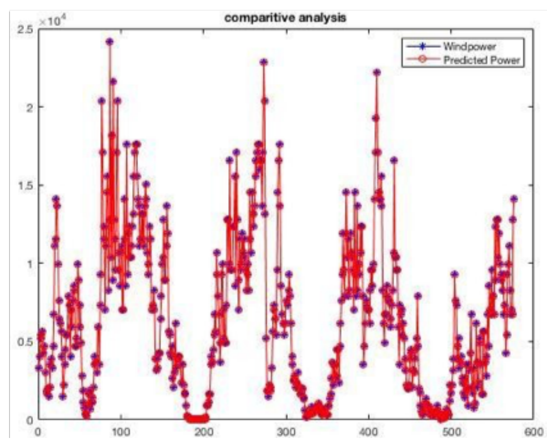

Fig. 5. 72 Hours Forecast Of Wind Power Using second order Polynomial Regression

CONCLUSION

Most of the times there is a time lag between awareness of an impending event or need and occurrence of that event. This lag in time is the main reason for planning and forecasting. If the lead time is zero or very small there is no need for planning else the outcome of the final event is conditional on identifiable factors, planning can perform an important role . In such situations , forecasting is needed to determine when an event will occur and a need arises, so that appropriate action can be taken.Employing Stochastic and Statistical models for the prediction of wind energy can be a game changer in the history of energy production minimizing all the negative effects the current form of energy production has. Wind energy itself has the potential to suffice the current electricity requirements without any other forms of energy and has to be given noteworthiness. In this paper two models are implemented for the forecast of wind power and the results are compared to give the better forecasting model. The aim of this paper is to

TABLE III
RESULTS OF 15 X 15 AND 30X30 MARKOV CHAIN

No. of Wind speed input	**No. of Predictions**	**Time Period (hrs)**	**Error Calculation**									
			MAE		**MSE**		**RMSE**		**MAPE%**		**Accuracy%**	
States			**15**	**30**	**15**	**30**	**15**	**30**	**15**	**30**	**15**	**30**
45995	288	48	0.662	0.547	0.181	4.761	0.621	0.169	62.1	53.7	37.88	46.28
45995	144	24	0.326	0.304	0.129	3.002	0.349	0.118	34.9	22.5	65.07	77.48
45995	100	14	2.445	0.171	0.115	1.926	0.265	0.103	26.5	19.9	73.41	80.05

TABLE IV
RESULTS OF SECOND ORDER POLYNOMIAL REGRESSION

No.of Wind Speed inputs	**No.of Predictions**	**Time Period (hrs)**	R^2	**MAPE**	**Accuracy**
10000	100	14	0.9987	2.06	97.93
10000	144	24	0.9935	7.99	92.01
10000	288	48	0.9916	11.12	88.87
10000	432	72	0.9201	15.61	82.16

TABLE V
COMPARATIVE RESULTS OF THE TWO PREDICTIVE MODELS

No. of Predictions	**Stochastic Model**	**MAPE**	**Accuracy**
14 hours	Second Order Markov Chain	19.9%	80.05%
14 hours	Second order Polynomial Regression	2.06%	97.93%

present a fundamental and simple- minded survey. The models under consideration fall under two forecasting time-scales, Short-term and Medium-term which are second order Markov Chain using 15 and 30 states and second order Polynomial Regression respectively. It has been observed that the second order Polynomial Regression can predict wind power values up to 3 days with excellent accuracy. Though third and higher order polynomials result in almost zero errors it is not considered as it may lead to over-fitting.

REFERENCES

[1] P. Bremaud, "Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues (Texts in Applied Mathematics)", Springer 2001 edition.

[2] A. Carpinone, R. Langella, A. Testa and M. Giorgio, "Very short-term probabilistic wind power forecasting based on Markov chain models," 2010 IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems, Singapore, 2010, pp. 107-112.

[3] IonuAlexandruPetre, MihaiRebenciuc, tefan Cristian Ciucu, "The use of Markov chains in forecasting wind speed: Matlab source code and applied case study".

[4] H. Li and S. Yamamoto, "Polynomial regression based model-free predictive control for nonlinear systems," 2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Tsukuba, 2016, pp. 578-582.

[5] C. M. van der Walt and N. Botha, "A comparison of regression algorithms for wind speed forecasting at Alexander Bay," 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), Stellenbosch, 2016, pp. 1-5.

[6] D. Wang, Y. Gao and Z. Tian, "One-Variable Linear Regression Mathematical Model of Color Reading and Material Concentration Identification," 2017 International Conference on Smart City and Systems Engineering (ICSCSE), Changsha, 2017, pp. 119- 122.

[7] Chengcheng Zhang and Fei Wang, "Research on correlation analysis between test score and classroom attendance based on linear regression model," 2010 The 2nd International Conference on Industrial Mechatronics and Automation, Wuhan, 2010, pp. 545-548.

[8] H. M. Al-Hamadi, "Long-term electric power load forecasting using fuzzy linear regression technique," 2011 IEEE Power Engineering and Automation Conference, Wuhan, 2011, pp. 96-99.

[9] HafzullahAksoy, Z. FuatToprak, Ali Aytek, N. Erdeml: "Stochastic generation of hourly mean wind speed data". Renewable Energy 29(14), 2004, pp. 2111–2131.

[10] M. Yakasiri, J. Avrel, S. Sharma, M. Anuradha and B. K. Keshavan, "A Stochastic Approach for the State-Wise Forecast of Wind Speed Using Discrete-Time Markov Chain," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 575-580.