Bayesian Mixed Effect Model and Classical Time Series Forecasting on Wind Power

Kaiyan Xu

Abstract:

In this paper, I tried to model the natural factors that influence wind power and conducted a forecast of wind power generation. Firstly, I built a Bayesian nonlinear model to capture the general contribution of wind speed to wind power. Then I tried to add random effects on hour and month to capture more variability and account for a lack of independence in the original model. The second part of the paper is a classical time series model that forecasts future wind power generation by using ARIMA method.

1. Introduction:

The efficiency of using traditional energy has approached its threshold so keep relying on traditional energy produces more cost than benefit. As a result, it is an imperative to gradually turn to renewable energy like wind power. From 2015 to 2021, World renewable energy investment increases from 14.8% in 2015, the year of Paris Agreement, to 19.8% in 2021 (IED). Data from IED also shows that net electricity produced by wind in the US increased 97.4% since 2015, in total 31378.5 GWh of electricity in 2021. Therefore, correctly modeling and estimating wind power is becoming more important. Statistical models play a significant role for windmill companies to schedule, dispatch and balance the demand and supply in the electricity system. Accurate prediction minimizes the need for additional waste to reserve energy and improve profit (Aoife et al. 2011).

However, capturing the variability of potential wind power generated is difficult. Wind is weather dependent and highly stochastic. Wind is influenced by many other natural factors including season, temperature, pressure, humidity, cloud cover, rainfall, etc. (Wang et al. 2011). Moreover, topographic characteristics like turbine position, turbine size, tower height, elevation under different environments also influence the efficiency of generating power. In general case, hybrid models with training data updated regularly have been found to be more accurate model (Saroha et al. 2020). However, since in my dataset from a windmill in Texas only has significant relationship with wind speed, it is sufficient to build a univariate model.

The second effect that leads to huge variability of wind power is the daily and seasonal pattern. The hypothesis of this paper is that such daily and seasonally pattern have significant underlying effects to the variability of wind power despite its high stochasticity. In my hourly wind power data, I incorporate the hourly and monthly random effects to the original fixed effect model and evaluated their effects. In the second part of the paper, the dataset is 10-year monthly generated wind power data in the US. By including past seasonal pattern, the accuracy of predicting future wind power is also assumed to have a huge improvement.

2. Materials and Methods

2.1 Wind Power Regression Analysis

In the first part of the paper, hourly generated wind power along with other weather data is measured in a windmill in Texas. By drawing correlation plot of wind power with each of the four measurements: wind speed, wind direction, pressure, and temperature, only wind speed has an obvious relationship with wind power, and the variability along with its curve is small. Therefore, it is safe to only include wind speed as the regressor and determine there is little omitted variable biase to do so. The shape of the curve between wind power and wind speed is exponential. Therefore, I proposed a logistic shaped curve to simulate their relationship. By calculating the correlation between wind speed and wind direction, pressure, temperature, other weather measurements are uncorrelated with wind speed either. As a result, there is almost no endogeneity in the model, i.e., E(residual|wind speed) = 0. There is no need to add these variables as instruments to isolate the movements in wind speed that are uncorrelated with model residuals, which in turn permits consistent estimation of coefficient of wind speed (Stock and Francesco 2003).

The process model of the first fixed effect Bayesian model is:

$$\mu = a \times \frac{e^{b+c*x}}{1+e^{b+c*x}} ,$$

where a, b, c are parameters derived from the mean of posterior probability in the Bayesian model. While logistic regression only constrains its range between [0,1], it is required to have a multiplicator that magnify the range of logistic regression to roughly fit the range of wind power, and parameter a has such an effect. By implementing MCMC using 3 chains, the chains converge well so the results in posterior are valid. A plot of fitted curve along with 95% confidence and prediction interval is shown in Figure 1. The fitted line fits the observation well, and 95% confidence interval is quite close to the fitted line, which means the fitted line is very accurate.

The data model of the Bayesian model is

$$y = dnorm(\mu, s),$$

where s is standard deviation, and it denotes the error in the model. Here, it is assumed that observations are normally distributed around the predicted value because in most cases, the observed value follows normal distribution with the predicted curve. Although when wind speed is smaller than 3m/s, no power is generated, and confidence interval and predictive interval in low wind power might produces wind power that is lower than 0 in this case under normal distribution, it is worthwhile to notice that this part of data only accounts for a very rare situation and the distance lower the horizontal axis is very small. Therefore, a slight negative predicted interval can be treated as 0 that there is no wind power at all.

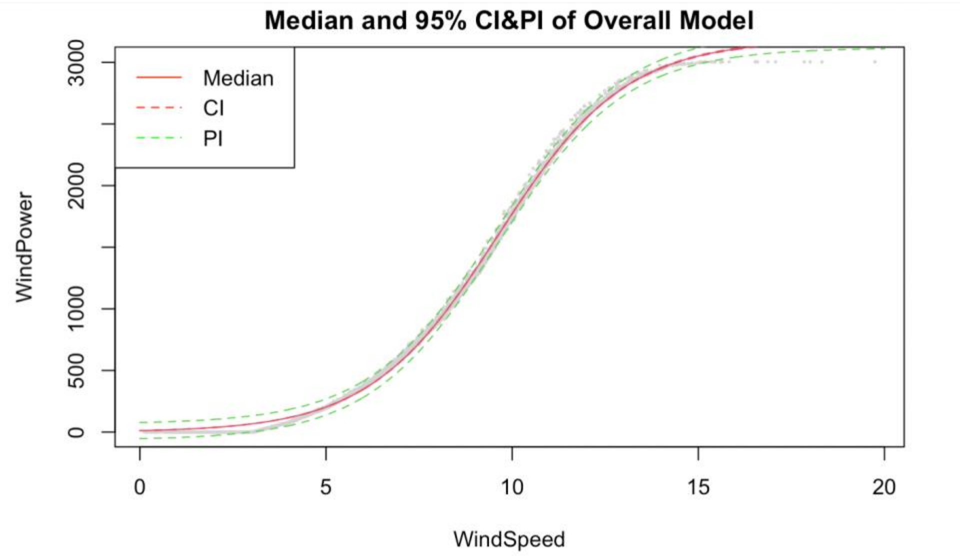


Figure 1. Overall Bayesian Nonlinear Model

Then, I divided the dataset to make sure if high wind power and low wind power have significant different or opposite parameters. In such a case, I need to separate them into two subgroups for further analysis. It is observed that wind power from 13:00 PM to 21:00 PM and from January to June are much higher on average, so I set two subsets to run the previous Bayesian model again. One subset consists of data from January to June at 13PM to 21PM; another subset consists of data from July to December at 23PM to 11AM. After comparing the parameters, I found that they are similar in two models, so there is no need to model high wind power and low wind power separately.

The next step is to add hourly and seasonal random effect to the fixed effect model. Random effects account for a lack of independence and unexplained variance within the same group after repetition (Liu et al. 2016). Firstly, I added hour of the measurement as the first random effect. Since it is a nonlinear model, every parameter can be set as random effect (Barrowman et al. 2003). I added random effect on each of the three parameters a, b, c and on the residual of the nonlinear model. The four process models become:

$$\mu_2 = a_i \times \frac{e^{b+c*x}}{1+e^{b+c*x}},$$

$$\mu_3 = a \times \frac{e^{b_i + c*x}}{1+e^{b_i+c*x}},$$

$$\mu_4 = a \times \frac{e^{b+c_i*x}}{1+e^{b+c_i*x}},$$

$$\mu_5 = a \times \frac{e^{b+c*x}}{1+e^{b+c*x}} + \alpha.h_i,$$

where a_i, b_i, c_i, $\alpha.h_i$ denotes that the hourly random effects added. In such an effect, wind power measure in the same hour a day have the same parameter in the model. By comparing DIC of the fixed effect model with the four random effect model above, it turns out that setting residuals as random effect has lowest DIC (Table 1). Therefore, I considered to keep adding random effects of month on the residual to see if it still improves DIC. The process model becomes:

$$\mu_6 = a \times \frac{e^{b+c*x}}{1+e^{b+c*x}} + \alpha.h_i + \alpha.m_i,$$

where $\alpha.m_i$ denotes the monthly random effect added. Now the process model can capture both the uncertainty from the same hour in a day and days within the same month. This model with two random effects on residual turns out to have the lower DIC. In an example, Figure 2 is a plot of best fitted line for January data shown in red. Still, confidence interval is narrow so the uncertainty in estimate is small and it indicates that the best fitted line is accurate. Most January data falls within the prediction interval, so the model captures most variability of wind power in January.

Process Model	DIC
$\mu = a \times \frac{e^{b+c*x}}{1+e^{b+c*x}}$	43165
$\mu_2 = a_i \times \frac{e^{b+c*x}}{1+e^{b+c*x}}$	43133
$\mu_3 = a \times \frac{e^{b_i+c*x}}{1+e^{b_i+c*x}}$	43142

$\mu_4 = a \times \frac{e^{b+c_i*x}}{1+e^{b+c_i*x}}$	43134
$\mu_5 = a \times \frac{e^{b+c*x}}{1+e^{b+c*x}} + \alpha.h_i$	42373
$\mu_6 = a \times \frac{e^{b+c*x}}{1+e^{b+c*x}} + \alpha.h_i + \alpha.m_i$	41760

Table 1. DIC of Models

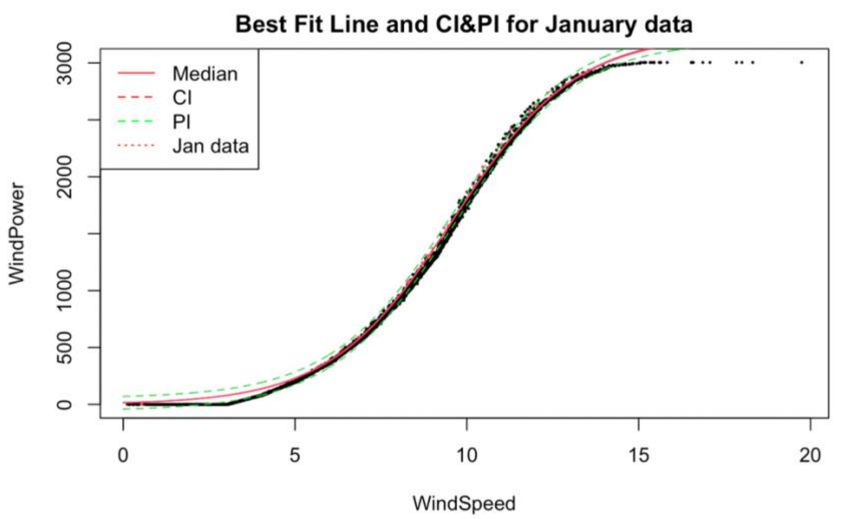


Figure 2. Hourly and Monthly Random Effect Model January Data

There is still high autocorrelation over 100 lags between relevant measurements in the residuals (Appendix a). Therefore, it is estimated that DIC will be much lower if accounting for the remaining autocorrelation between closed measurements.

2.2 Wind Power Forecast

The second part of this paper is to forecast US wind power generation for the first 6 month of 2022 using classical ARIMA method in a frequentist context. The dataset for this project consists of US monthly wind power generation from 2010-2021. It is extracted from the larger dataset, "*Monthly Electricity Statistics - Monthly electricity production and trade data for 47 countries*," from IEA.

The time series plot shows a clear upward trend and seasonal pattern in the data as expected (Figure 3). Also, variance of time series is increasing so there are more variability of wind power generation as average generation increases. Therefore, it is obvious that the process is nonstationary, and hence it requires transformation before catching the correct model.

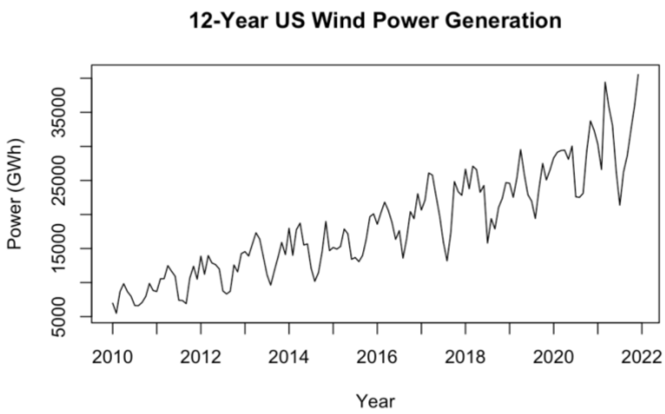


Figure 3. Time Series Plot of US Monthly Wind Power Generation

To transform the process to be stationary, I firstly applied Box-Cox transformation to stabilize the variance. I applied order 1 differencing to eliminate the trend and order 12 differencing to eliminate the seasonality as it is monthly data. Starting from here, I can fit models based on the ACF and PACF plot of residuals (Appendix b).

The method is to fit a model based on autocorrelation and partial autocorrelation. From the ACF and PACF plot, for nonseasonal part, ACF cuts off after lag 1 and PACF can be treated as either decaying to 0 or cut off after lag 4. For seasonal part, ACF either cuts off after seasonal lag 1 (lag12) or decays to zero, the same as PACF. To find the model with lowest AICc, I fit 6 relevant models derived from the correlation plots. The table (Appendix c) shows the 6 models and their AICc.

Based on the Table 1, $SARIMA(4,1,1) * (1,1,0)_{12}$ and $SARIMA(0,1,1) * (1,1,0)_{12}$ have lowest AICc. However, a diagnostic show that the normality assumption of fitting an ARIMA model is violated in these two models, so they are inadequate models. The model with third lowest AICc, on the other hand, shows normality on its residuals and that its standardized residuals follow white noise process (Appendix d). Therefore, it is an adequate and relative accurate model for prediction. The result of prediction is shown in Figure 5. By taking out the last 6-month data as testing set and refit the data using the first 138 data as training set, all 6 testing data is contained in the 80% confidence interval. Therefore, the ARIMA model here is accurate in predicting the future wind power generation.

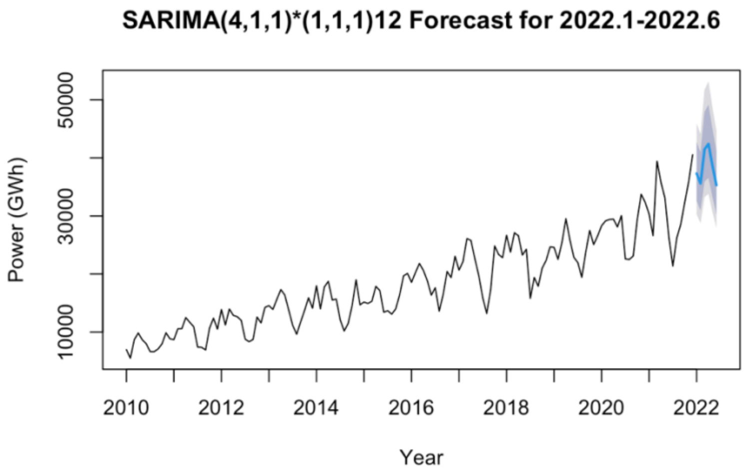


Figure 5. $SARIMA(4,1,1) * (1,1,1)_{12}$ Forecast

3. Results

In Texas Windmill Dataset, the model with the lowest DIC is the mixed model with hour and month as random effects. It verifies that there is hourly and seasonal pattern in wind power. Within the same hour a day and same month, the variability can be captured by the same random effect. When modeling seemingly highly stochastic data, considering random effects improves the accuracy of the model by borrowing strength across the dataset.

In US wind power generation dataset, the forecasting wind power from this model does not deviate from the historical pattern a lot. It is on a high position and follows past average seasonal pattern: from December to February, wind power decreases; from March to April, wind power is high; from May to June, wind power goes down again. Therefore, seasonal pattern is significant. This forecast assumes that no intervention would occur. Based on the forecast, windmill companies can plan their production and storage of wind power accordingly, and government can manage production of other renewable energy when during the months when wind power generation is low.

4. Discussion

There are many constraints in this paper. Firstly, the Texas Windmill Dataset only has wind speed that has obvious relationship with wind power, which is not representative of most wind power generation pattern in real life. It is necessary to find another dataset with more useful covariates and build a more complex model to comprehensively understand effects of natural factors on wind power.

Secondly, I did not try enough combination of random effects on nonlinear model parameters. I only assumed one of the parameters are random effects. When adding the second random effect, month, I only added to the best model after adding the first random effect. Therefore, it is possible that I missed the model with lowest DIC. To solve this I need to use greedy method to compare every combination of parameters.

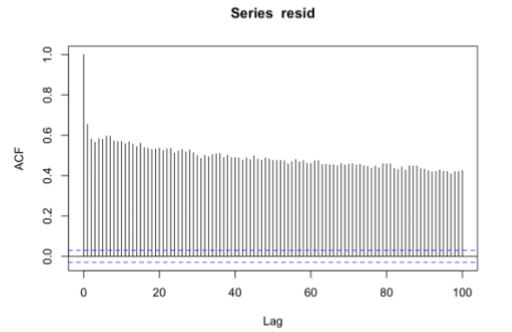
Thirdly, I did not consider the remaining autocorrelation after adding the random effects. There is still high autocorrelation between relevant data points over 100 lags. I need to add a covariance matrix as the variance of data model to account for different covariance at each point.

5. Acknowledgement

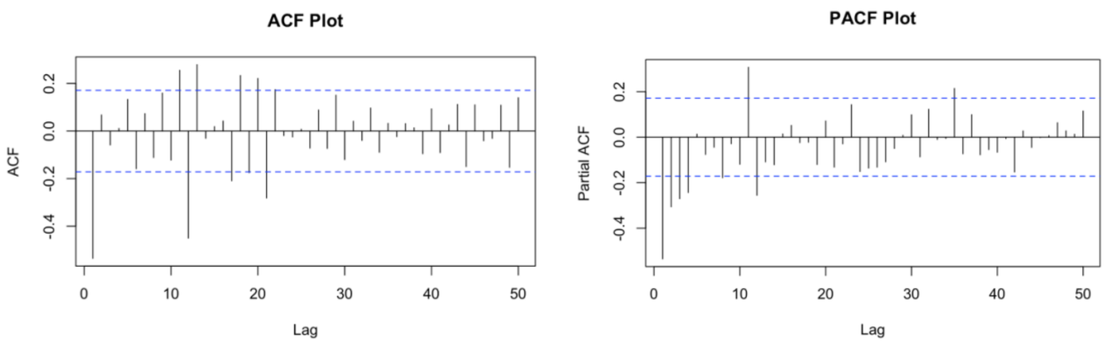
I thank Professor Michael Dietze for his instruction and help for the whole semester.

6. Appendix

a. Autocorrelation between Measurements



b. ACF and PACF of Differenced Time Series

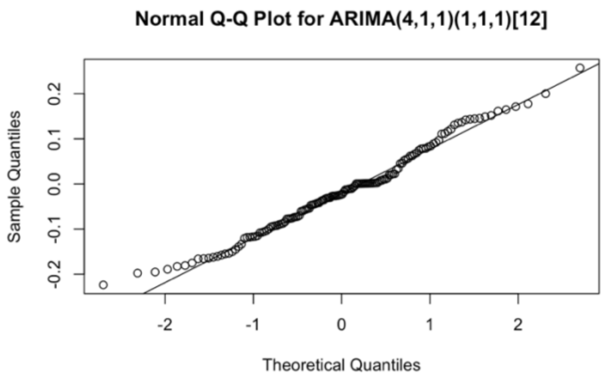


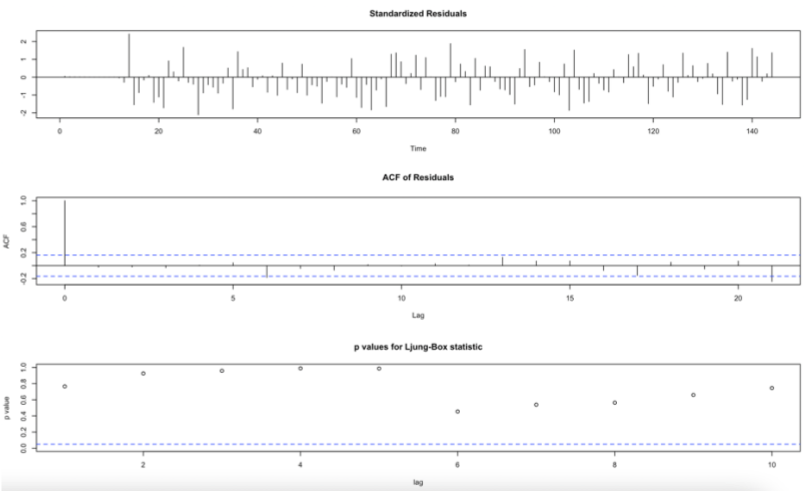
c. AIC_C of ARIMA Models

Model	AIC_C
SARIMA(0,1,1)(1,1,1)[12]	-198.31
SARIMA(0,1,1)(0,1,1)[12]	-200.43

SARIMA(0,1,1)(1,1,0)[12]	-174.25
SARIMA(4,1,1)(0,1,1)[12]	-192.09
SARIMA(4,1,1)(1,1,0)[12]	-166.68
SARIMA(4,1,1)(1,1,1)[12]	-189.88

d. Diagnostics of ARIMA Model





7. References

World Energy Investment 2021 Datafile, IED, https://www.iea.org/data-and-statistics/data-product/world-energy-investment-2021-datafile

Monthly Electricity Statistics, IED,

https://www.iea.org/data-and-statistics/data-product/monthly-electricity-statistics

Aoife M. Foleyabd, Paul G. Leahyab, Antonino Marvugliac, Eamon J.McKeoghab. Current methods and advances in forecasting of wind power generation. 2011. https://doi.org/10.1016/j.renene.2011.05.033

Xiaochen Wang, Peng Guo, Xiaobin Huang, A Review of Wind Power Forecasting Models, Energy Procedia, Volume 12, 2011, Pages 770-778, ISSN 1876-6102, https://doi.org/10.1016/j.egypro.2011.10.103.

Sumit Saroha, Sanjeev Kumar Aggarwal and Preeti Rana. Wind Power Forecasting, 2020, DOI: 10.5772/intechopen.94550. https://www.intechopen.com/chapters/74076

Stock, James H., and Francesco Trebbi. 2003. “Who Invented Instrumental Variable Regression?” Journal of Economic Perspectives 17: 177–194.

Liu, Xiaolei et al. “Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies.” PLoS genetics vol. 12,2 e1005767. 1 Feb. 2016, doi:10.1371/journal.pgen.1005767

Barrowman, Nicholas J., et al. “The Variability among Populations of Coho Salmon in the Maximum Reproductive Rate and Depensation.” Ecological Applications, vol. 13, no. 3, 2003, pp. 784–93, http://www.jstor.org/stable/4134695. Accessed 7 May 2022.