Data Analytics: Assignment 2

by Anthony Rizkallah

Table of contents

1 Python Packages

Pandas

Matplotlib

Numpy

Sklearn

Statsmodels

Scipy

Arch

Itertools

2 Question 1

2.1 Steps

To verify whether seasonality in the Wind Generation data exists, plotting the Generation in MWh over time should provide a good starting point to visually identify any seasonalities. However, not one time scale is satisfactory in the sense that if the plot of “time in hours” does not show any seasonality, maybe “time in days” might and/or “time in weeks” etc. To test for this caviat, the wind generation can be averaged over the time period of interest.

For example, if on Day 1 of the dataset, there were 24 hours of reported generation, then averaging the generation through the 24 hours yields the value for Day 1 that can be plotted for a daily time series. Similar to daily, the different timescales can be computed in a similar method.

Then, visually identifying seasonality will be highly dependent on the visibility of the figures and the captured data. In other words, the hourly plots capture all the available raw data, but the daily data, which might provide a better visual when plotted, capture the averages for each hour of generation in each day. In general, the general trend is not affected, but hiding the granularity in the data affects trends on shorter timescales.

2.2 Results

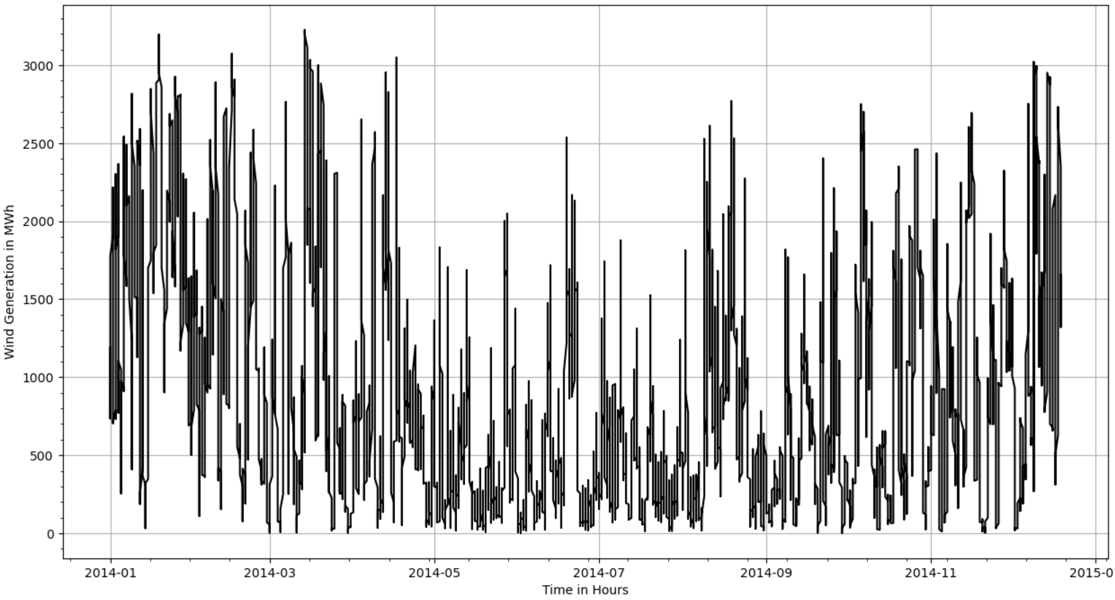

Figure 1. Hourly Wind Generation in MWh

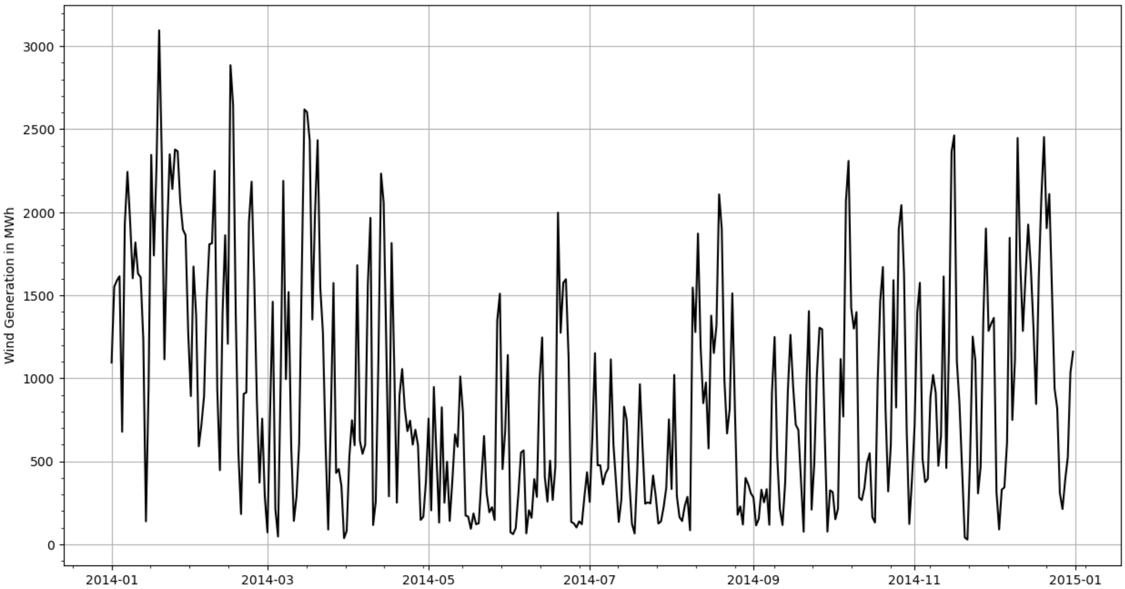

Figure 2. Daily Wind Generation in MWh

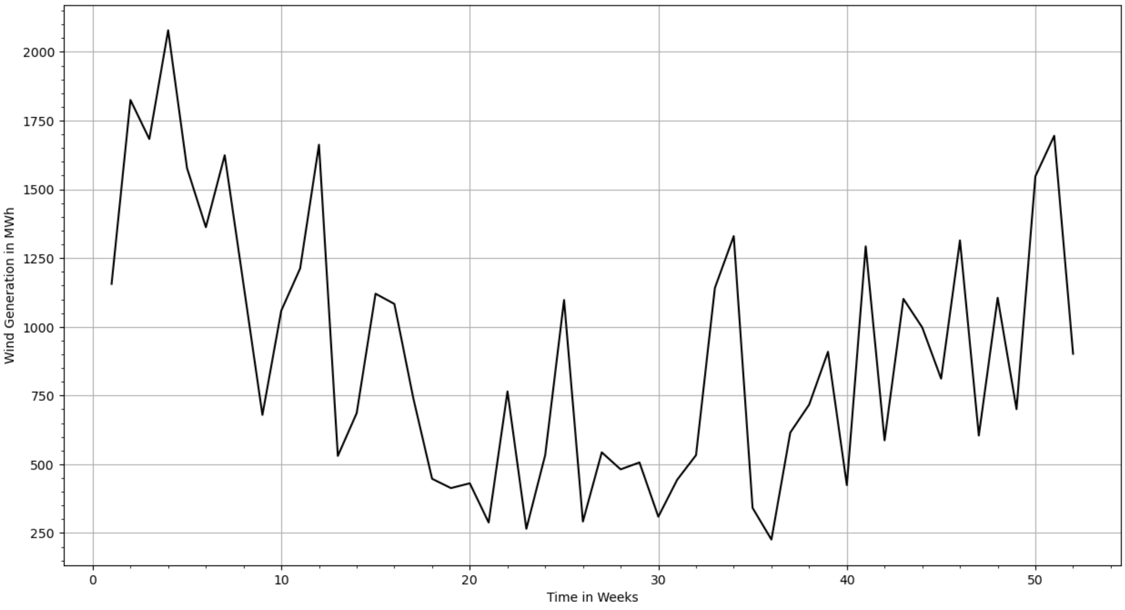

Figure 3. Weekly Wind Generation in MWh

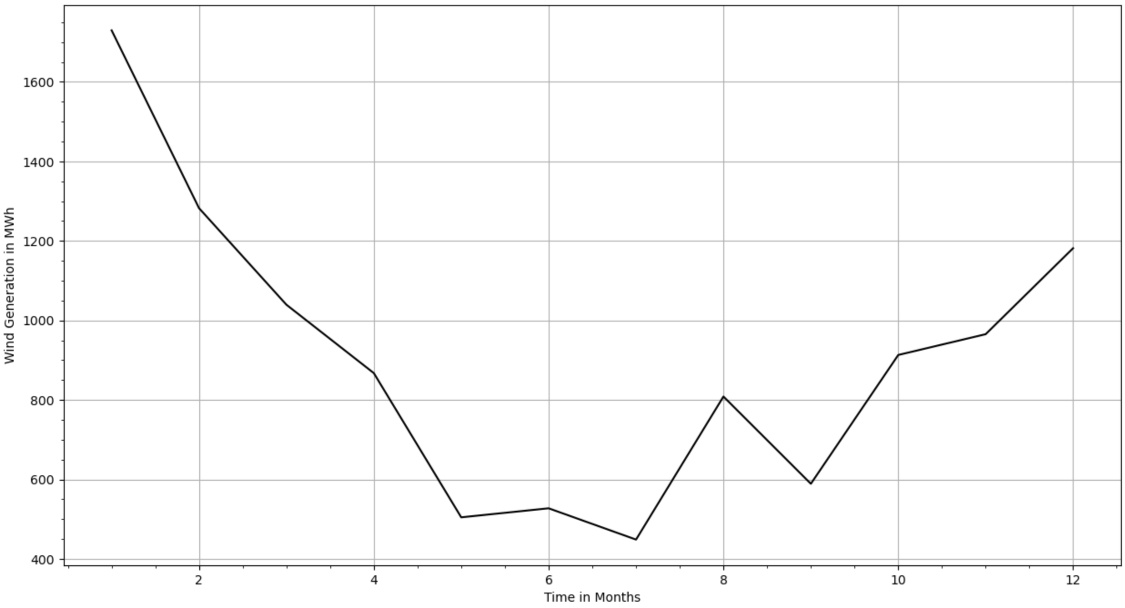

Figure 4. Monthly Wind Generation in MWh

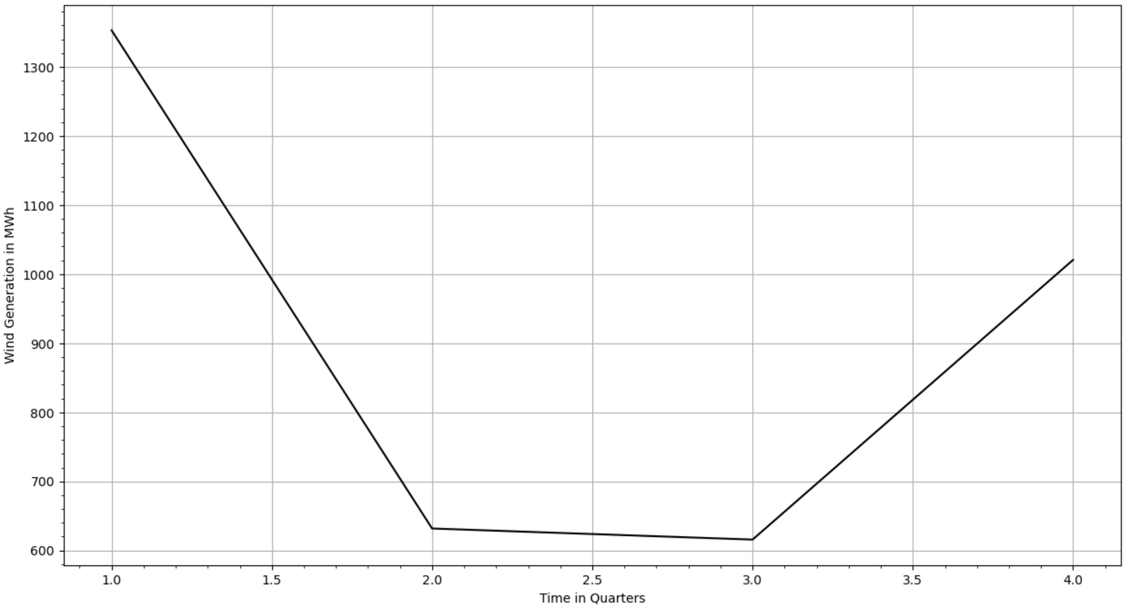

Figure 5. Quarterly Wind Generation in MWh

2.3 Inference

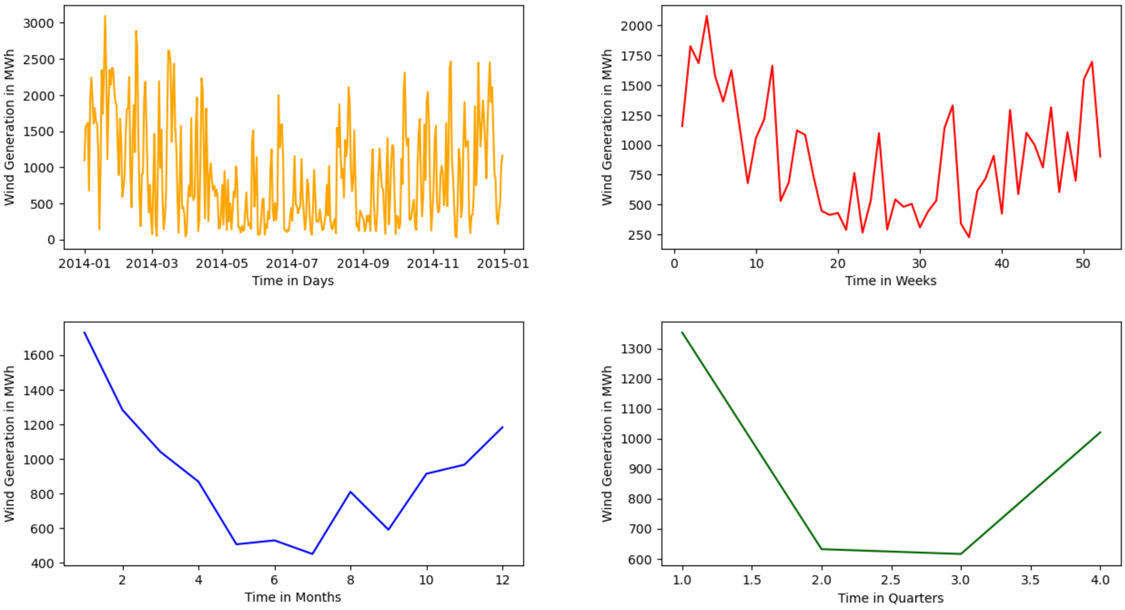

Figure 6. Wind Generation in MWh over different timescales

As mentioned in Section (2.1), trends will be captured regardless, and this is mainly evident in the Daily and Weekly graphs. Although the daily graph is denser to read, when averaged and plotted over weeks, it becomes evident that there is evidence of weekly seasonality that can be leveraged to predict or model wind generation. In fact, there **might** exist seasonality, but there is no sufficient insight yet to make this decision.

3 Question 2

3.1 Steps

Another test to identify seasonality is to calculate the *relative difference* or *range ratio* of wind generation over the different timescales. The relative difference is computed as

$$\text{Relative Difference} = \frac{x(t+d) - x(t)}{x_{\text{MAX}}} \tag{1}$$

where $x(t)$ is the value of wind generation at time t, $x(t+d)$ is the value at the next timestamp d, and x_{MAX} is the maximum wind generation value. This equation is defined as the *ramp formula* in Section (4.1). For now, it is worth plotting the relative difference as a function of the different timescales where $d = 1$ for each of the hourly, daily, weekly, monthly, and quarterly datapoints.

3.2 Results

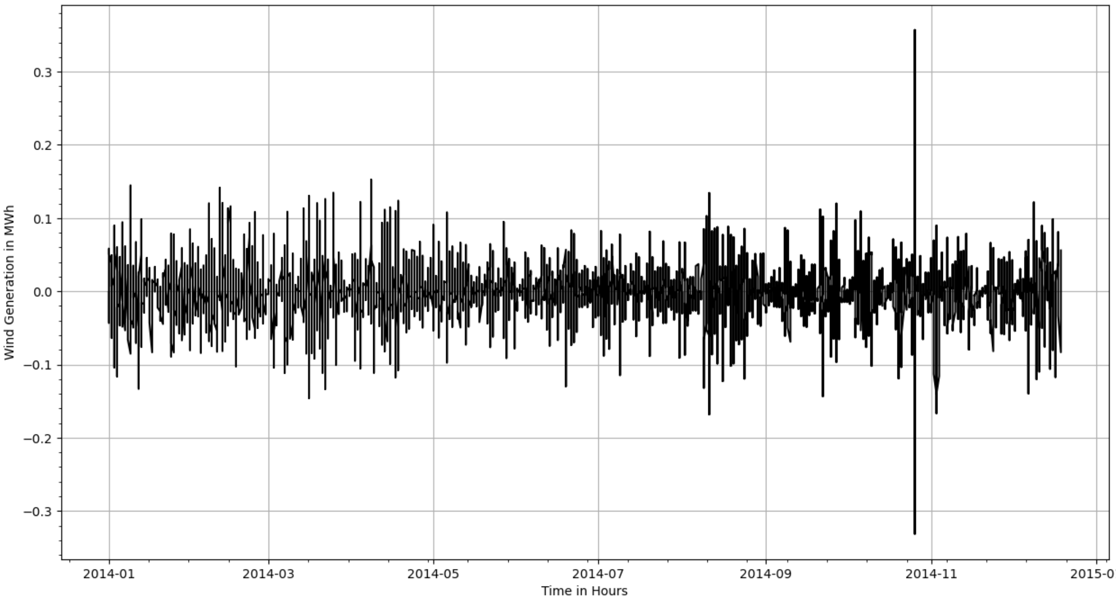

Figure 7. Hourly Relative Difference in Wind Generation in MWh

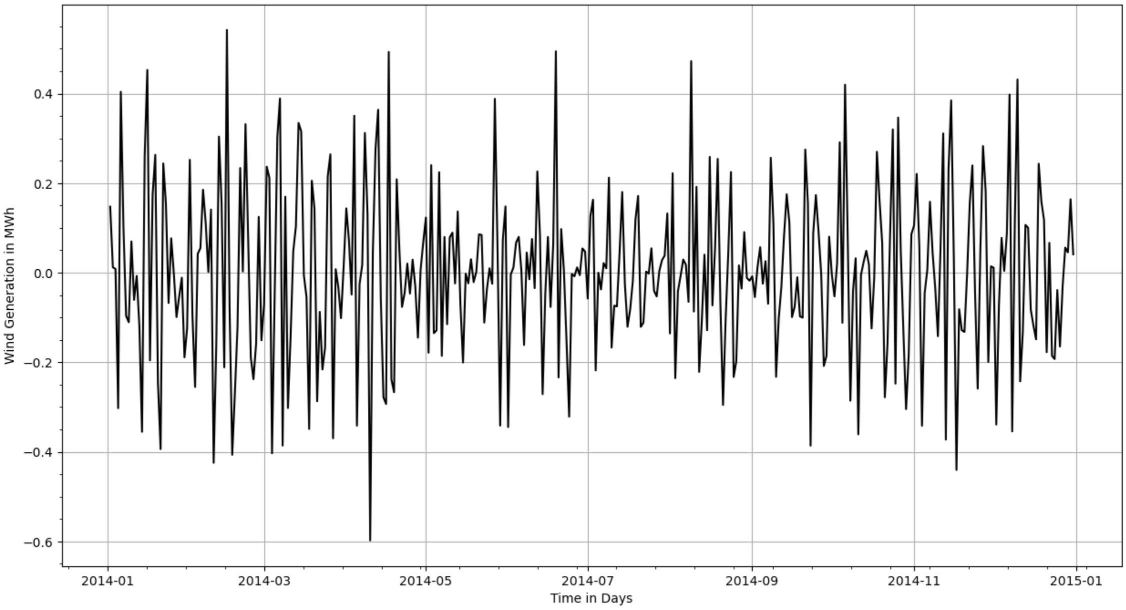

Figure 8. Daily Relative Difference in Wind Generation in MWh

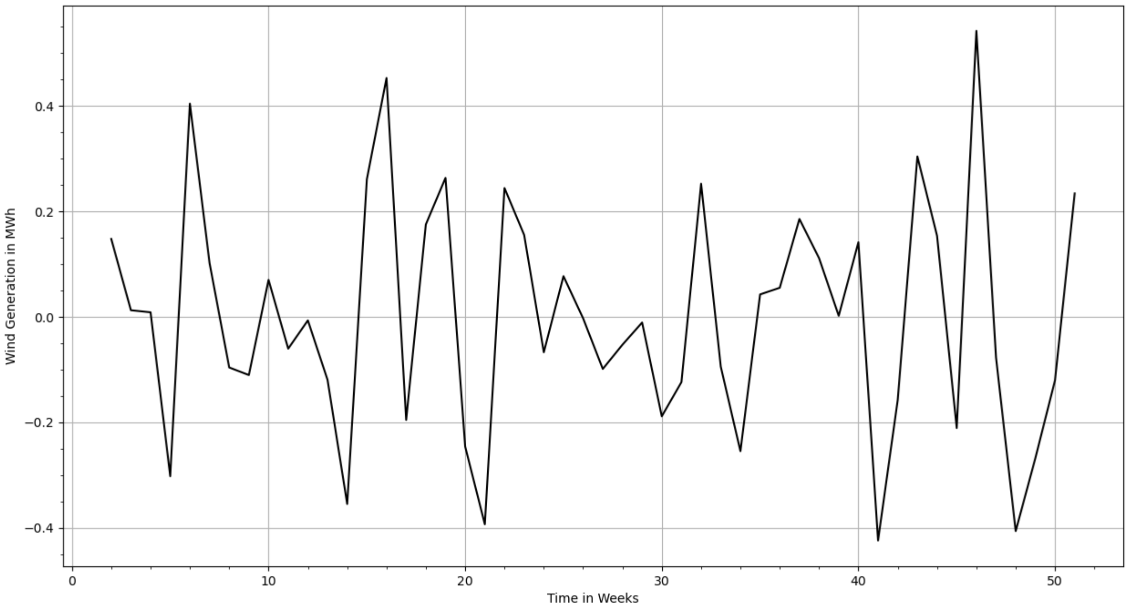

Figure 9. Weekly Relative Difference in Wind Generation in MWh

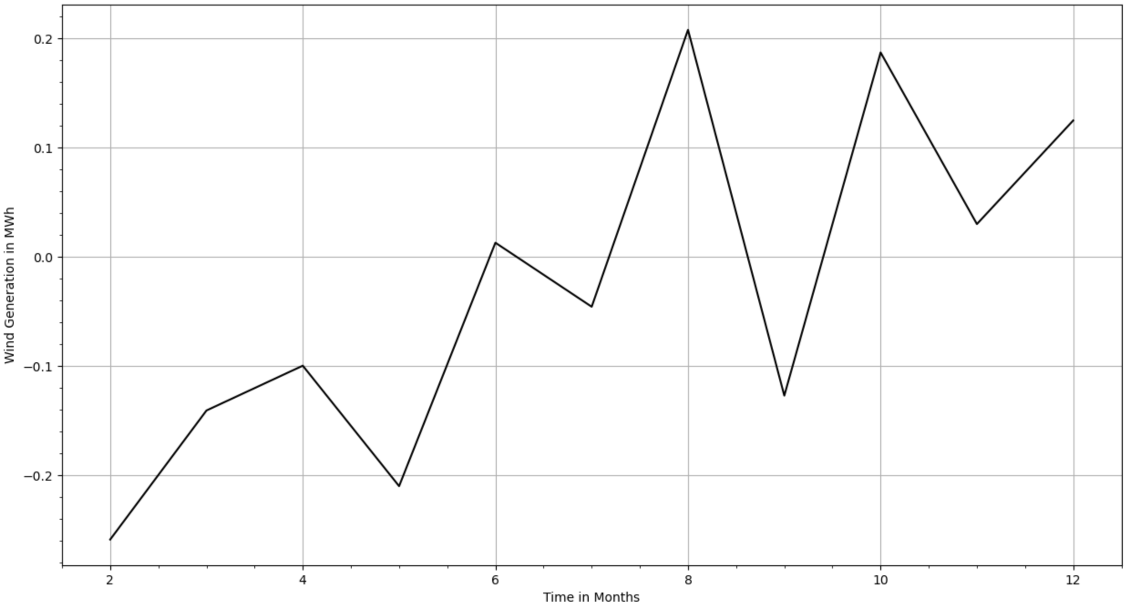

Figure 10. Monthly Relative Difference in Wind Generation in MWh

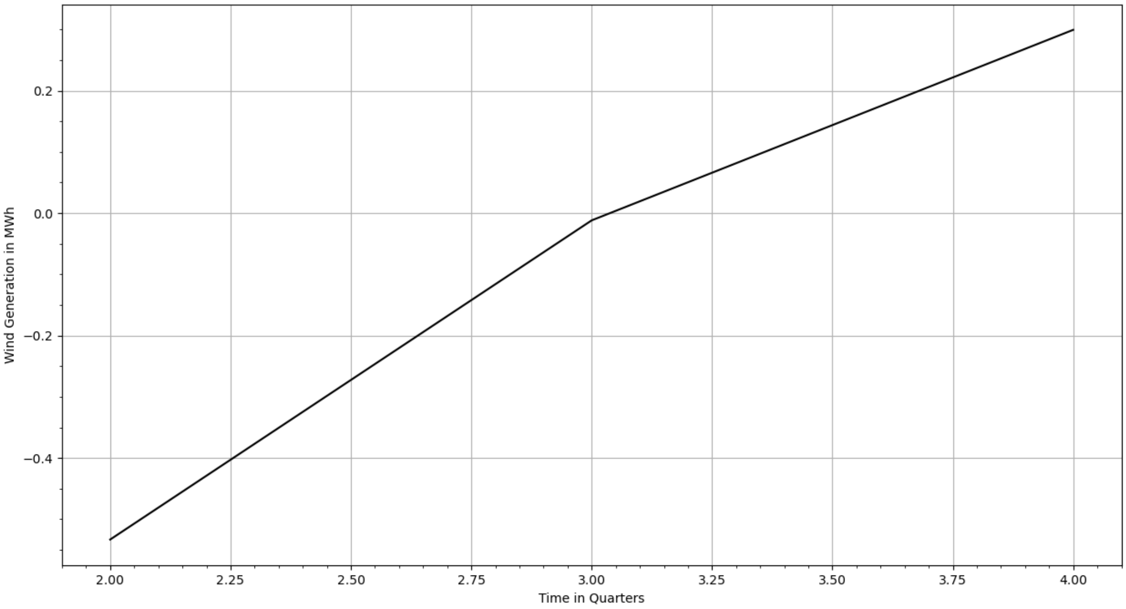

Figure 11. Quarterly Relative Difference in Wind Generation in MWh

3.3 Inference

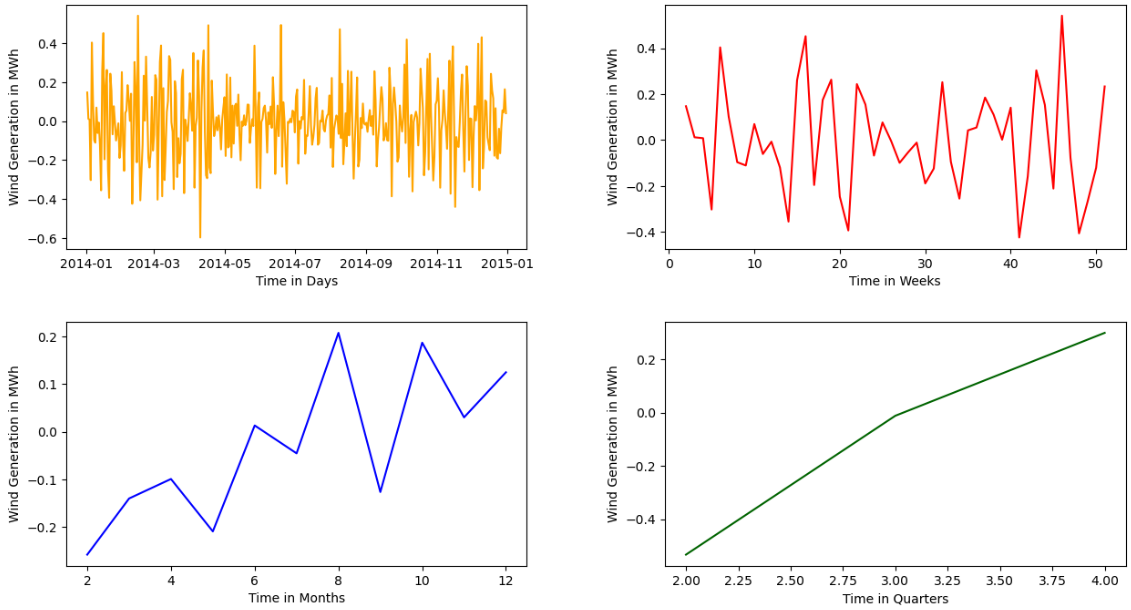

Figure 12. Relative Difference in Wind Generation in MWh over different timescales

One can argue that there is some evidence of *daily* and *weekly* seasonality when identified visually, but it is hard to distinguish seasonality from noise in the graph. If one looks at the peaks and how often they recur, it is hard to distinguish significant recurrences from the ones that perhaps happen by chance.

4 Question 3

4.1 Steps

As stated previously in Section (3.1), the ramp function or formula is defined as

$$r(t,d) = 100 \times \frac{x(t+d) - x(t)}{x_{\mathrm{MAX}}} \tag{2}$$

As also evident from the seasonality graphs plotted in Section (3.2), one can attempt to find how often, and the range, at which wind power fluctuations happen. The ramp function can then be used to calculate the values of those fluctuations, and plotting $r(t,d)$ yields graphs similar to Section (3.2). So, instead of re-plotting those graphs, an interesting practice that helps transform the data into a different form is the Empirical Cumulative Distribution Function which would represent the probability of each ramp.

The Empirical Cumulative Distribution Function (ECDF) provides a way to represent the probability distribution of observed ramp values in a dataset. Unlike a standard Cumulative Distribution Function (CDF), which assumes a known probability distribution (e.g., normal), the ECDF is non-parametric and is directly constructed from the data. The ECDF is mathematically defined as:

$$\mathrm{ECDF}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \le x} \tag{3}$$

where n is the sample size, $1_{x_i \le x}$ is an indicator function that is equal to 1 if $x_i \le x$ and 0 otherwise. This function counts the proportion of data points that are less than or equal to a given value x. The ECDF is a essentialy a step function that increases by $\frac{1}{\mathrm{n}}$ at each data point. It is useful for analyzing the probability of wind power ramp fluctuations, as it transforms time series fluctuations into a probability distribution. Unlike parametric approaches, the ECDF does not assume an underlying distribution but instead treats the dataset as its own empirical distribution. Additionally, one can compare the ECDF to a theoretical CDF, such as a normal distribution, to assess whether the observed data follows a known statistical pattern. This way, the fluctuations can be modeled by getting the samples' means μ and standard deviations σ.

Additionaly, as a final caviat before fitting any distributions, the ramps must be split between positive and negative categories because increasing and decreasing ramps have different impacts on power systems and require different operational consequences.

4.2 Results

4.2.1 Negative Ramps

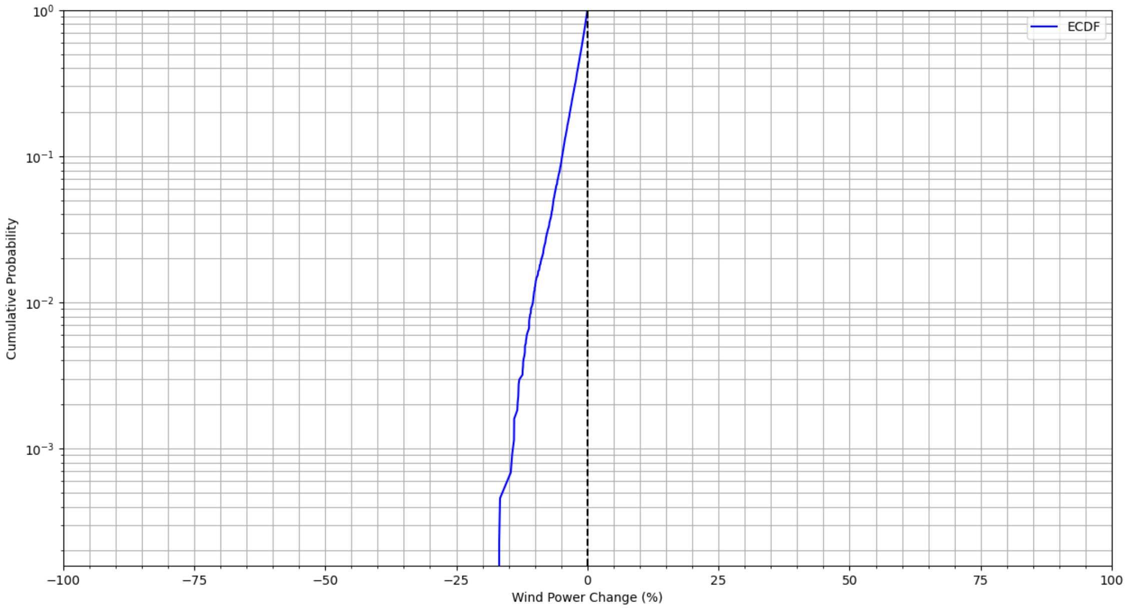

Figure 13. Empirical Cumulative Probability of Negative Ramp Changes

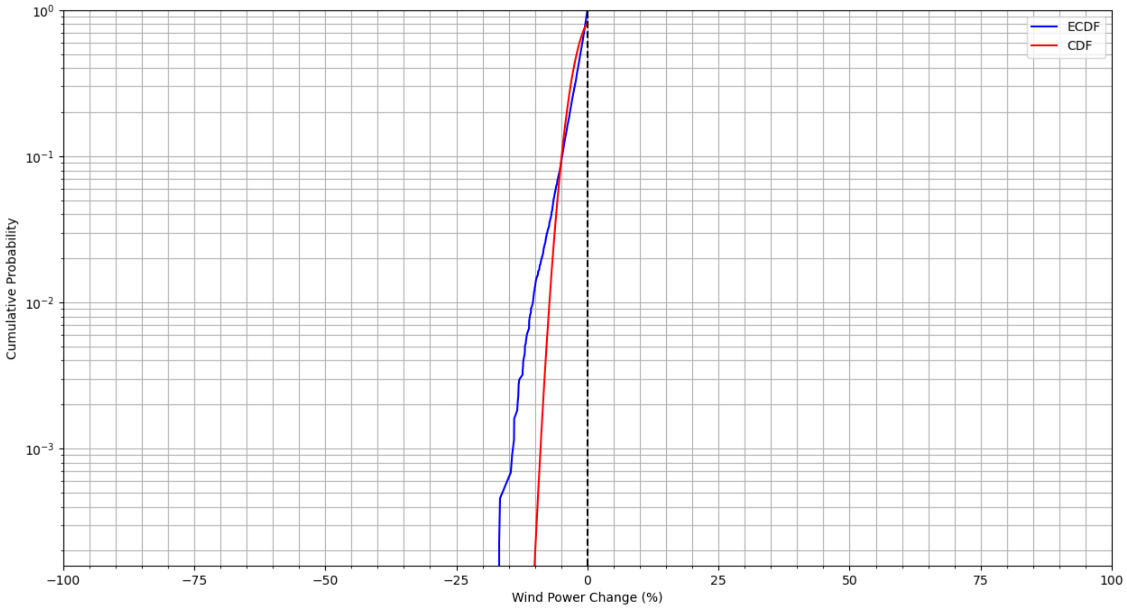

Figure 14. Empirical Cumulative Probability of Negative Ramp Changes Fitted to a Normal Distribution. $\mu \approx 2.06$, $\sigma \approx 2.23$

4.2.2 Positive Ramps

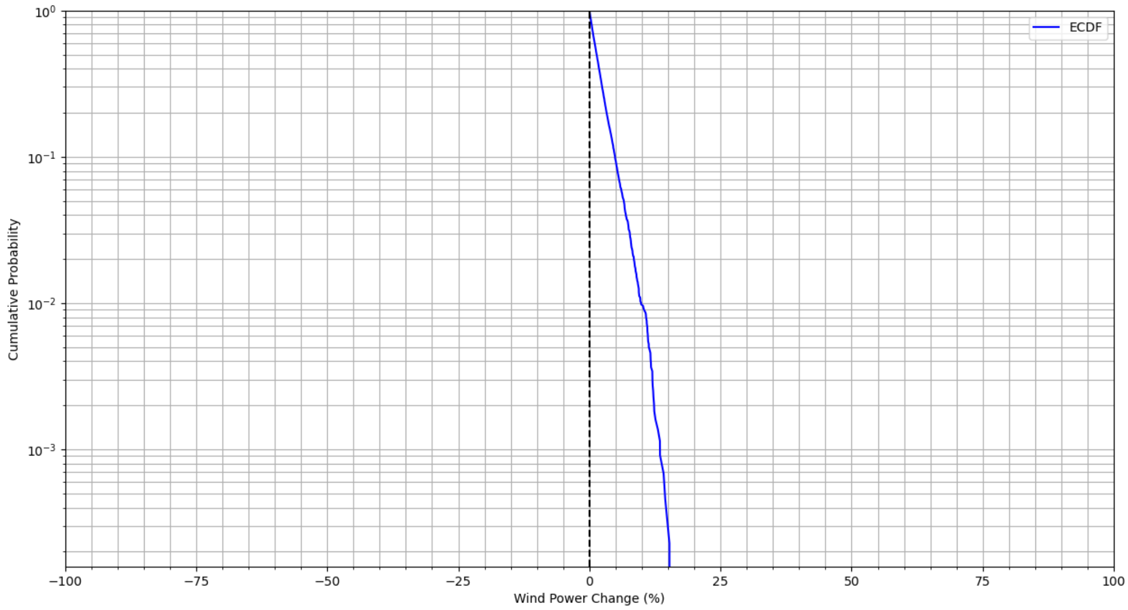

Figure 15. Empirical Cumulative Probability of Positive Ramp Changes

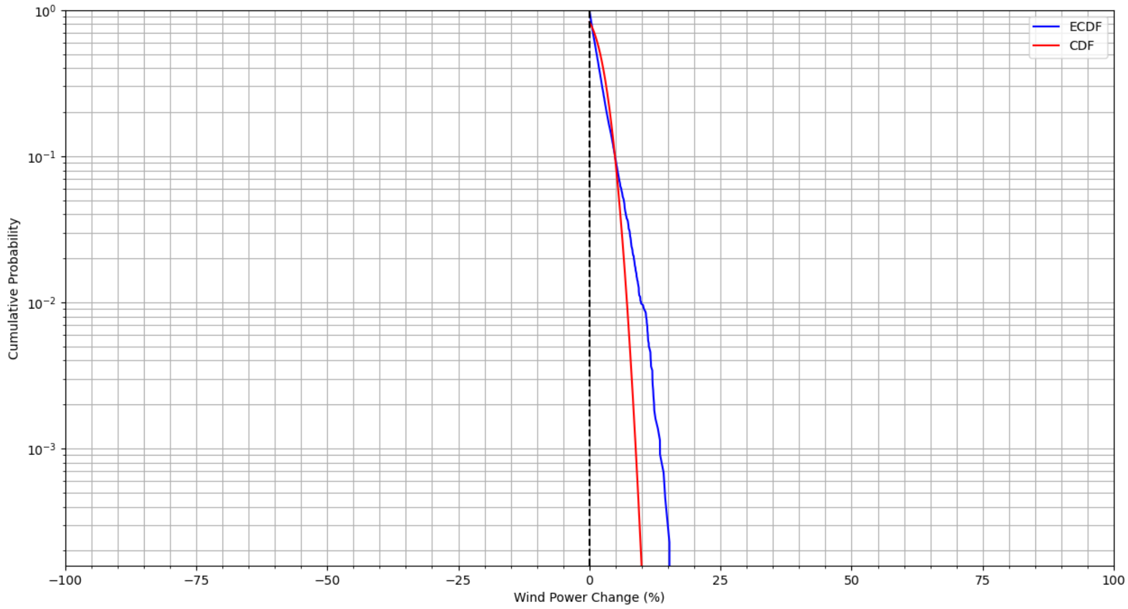

Figure 16. Empirical Cumulative Probability of Positive Ramp Changes Fitted to a Normal Distribution. $\mu \approx 2.06$, $\sigma \approx 2.19$

4.3 Inference

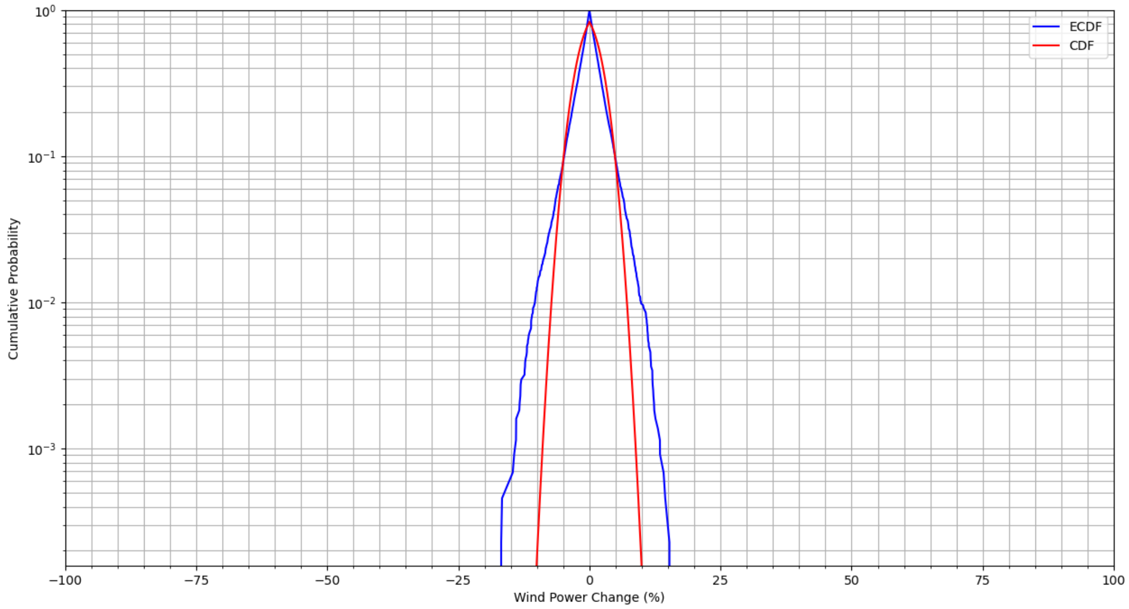

Figure 17. Empirical Cumulative Probability of Positive and Negative Ramp Changes Fitted to a Normal Distribution

First, one can argue that both positive and negative ramps have almost a symmetrical shape (but are not really symmetrical) in this case with the negative ramps going as far as -17% whereas the positive ramps going as far as 15%.

Now, using a normal distribution to predict extremes in wind power does not seem to be the most optimal model as seen in Figure (17). The CDF is light-tailed whereas the ECDF of the ramp data is heavy-tailed.

5 Question 4

5.1 Steps

Now, power system operators are tasked with the challenge of balancing supply and demand. So, they need to understand the variability in wind generation over different timescales. In other words, when predicting into the future, what change should they expect to see in wind power generation? One can investigate that phenomenon by looking at different ramp values for a different d. For example, a power system operator would like to know what to expect in 5 hours– will wind generation decrease or increase, and more importantly, how likely it is to happen.

Then, each $r(t, d)$ can be computed for $d\epsilon[0, 24]$ to find the ramp value for each hour (up to a day).

To calculate the percentile, it is simply the

$$P = \frac{n}{N} \times 100 \tag{4}$$

where n is the the number of data points below the data point of interest and N is the sample size. The ramps are also split between positive and negative for this case.

5.2 Results

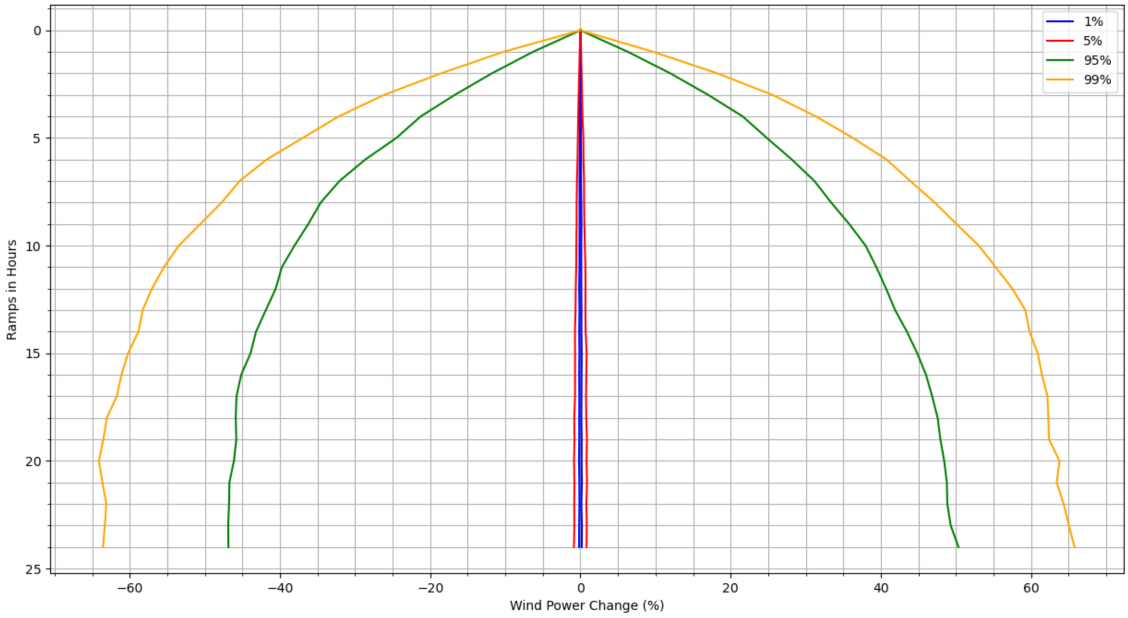

Figure 18. 1^{st}, 5^{th}, 95^{th}, and 99^{th} percentiles of wind power ramps

5.3 Inference

The main thing to take away from the plots in Figure (18) is that the spread of the curves widens as the ramp duration increases. This means that wind power fluctuations become more extreme over longer timescales making it harder to predict what happens because the range of possibilities increase and the probability of an extreme event gets higher.

6 Question 5

6.1 Steps

The Autocorrelation Function measures how values at time t in a time series are correlated with previous values. In other words, how data is correlated with its past data which would indicate an inherent seasonality attribute. For this case, the autocorrelation is being plotted for the *actual* *wind generation*, meaning that generation values in the future will be correlated with generation values in the past, which from previous sections was evident that it might not be the most optimal setup to identify seasonality. In any case,

10-day lags in the data is $10\,\text{days} \times \frac{24\,\text{hours}}{1\,\text{days}} \times \frac{1\,\text{data point}}{1\,\text{hour}} = 240$ data points.

6.2 Results

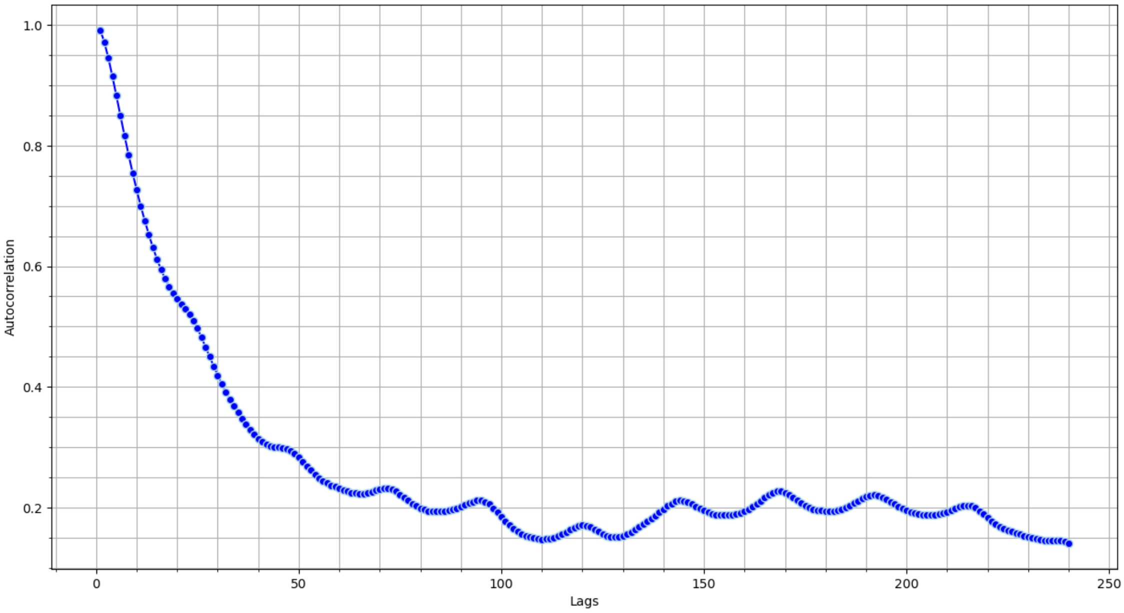

Figure 19. Autocorrelation graph of Wind Generation in MWh over 10 days or 240 hours.

6.3 Inference

Although this graph does not show the stems but rather only shows the correlation of the data with its previous values, an evident pattern is that there is some kind of seasonality with some kind of weak correlation at around 20-25%. The correlation drops from 100% in about 60 lags to reach about 25% and then exhibits some kind of seasonality, but it is not very clear.

7 Question 6

7.1 Steps

Now, for this case, the *change in wind generation* is being used to find autocorrelation, and as evident in previous sections, this was the more optimal setup that yielded graphs and insights into potential seasonality in the data. To graph the statistical significance of $p = 0.05$ on the graph, the equation is that of confidence interval as follows

$$\text{CI} = \frac{1.96}{\sqrt{N}} \tag{5}$$

where N is the sample size.

7.2 Results

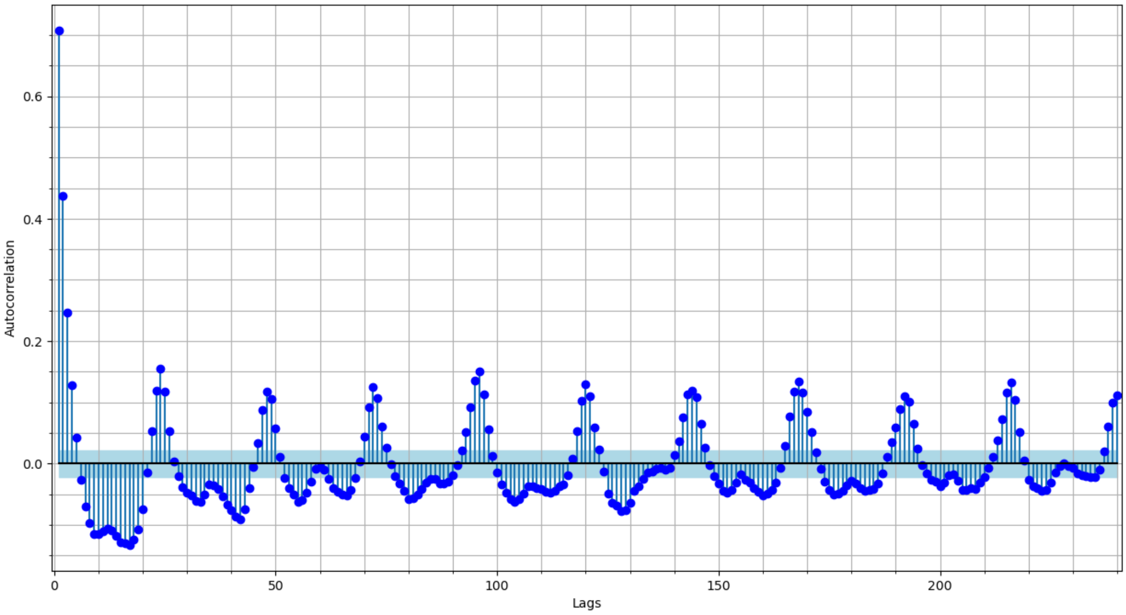

Figure 20. Autocorrelation graph of Change in Wind Generation in MWh over 10 days or 240 hours with a confidence interval of $p = 0.05$.

7.3 Inference

First, for points to be statistically significant on the plot, they must fall outside of the shaded-lightblue area. All seasonal datapoints fall within the statistically significant area, which means that they do not occurr by luck. Additionally, seasonality now is very clear and evident and seems to be around 22-23 lags (almost 1 day). When compared with Figure (19), modeling *change in wind generation* seems is much more appropriate than modeling *actual wind generation*.

8 Question 7

8.1 Steps

The autocorrelation graph suggests that the seasonality is statistically significant. However, for further investigation, a Variance Ratio Test should be performed to check if the wind generation time series follows a random walk or whether it doesn't. Hence, a hypothesis test can be performed as follows

H_0: The wind generation time series follows a random walk

H_1: The wind generation time series does not follow a random walk

If the null hypothesis H_0 is to be rejected, the p-value for the Variance Ratio Test must be less than 0.05. Now, if the p-value is less than 0.05, then there will be some evidence of mean-reversion or mean-aversion. If the VRT has a VR statistic < 1, the series is mean-reverting or has a tendency of returning to an average value. If the VR statistic > 1, then the series is mean-averting or suggests a trend away from the mean.

8.2 Results

Rejected H_1	**pvalue**	**statistic**
2 Lags	0.75	-0.32
3 Lags	0.26	-1.14
4 Lags	0.06	-1.88

Table 1. Lags that rejected H_1

All the other lags from $[5, 240]$ accepted the alternaticve hypthesis H_1 and all lags had a VR statistic < 1.

8.3 Inference

This means that for the most part the time series does not follow a random walk as the null hypothesis H_0 was rejected by all lags from $[5, 240]$. Moreover, all the VR statistics were less than 1, meaning that the series is mean-reverting or has a tendency of returning to an average value. This answer is coherent with the previous autocorrelation graph where most datapoints fell into the statistically significant category while also exhibiting strong seasonality evidence.

9 Question 8

9.1 Steps

A Simple Moving Average (SMA) is a smoothing technique used to reduce short-term fluctuations in a time series. In this analysis, an SMA model is used to predict wind generation by averaging the past n values, where n represents the window size in hours. The goal is to determine the optimal window size that minimizes forecasting error.

The Mean Absolute Error (MAE) is used as the evaluation metric, measuring how far the SMA predictions deviate from actual wind generation. A lower MAE indicates better predictive accuracy.

To benchmark the SMA model, it is compared against a Persistence Forecast, which assumes that future wind generation will remain the same as the most recent observed value. This is used a common baseline in forecasting models.

1. Vary the SMA window size from $n = 1$ to $n = 24$ hours.

2. Compute the SMA for each n using rolling averages.

3. Calculate the MAE between the SMA predictions and the actual wind generation.

4. Compare the MAE of SMA models to the Persistence Benchmark to determine if SMA provides an improvement.

5. Identify the optimal SMA window that minimizes MAE and assess whether it performs better than persistence.

9.2 Results

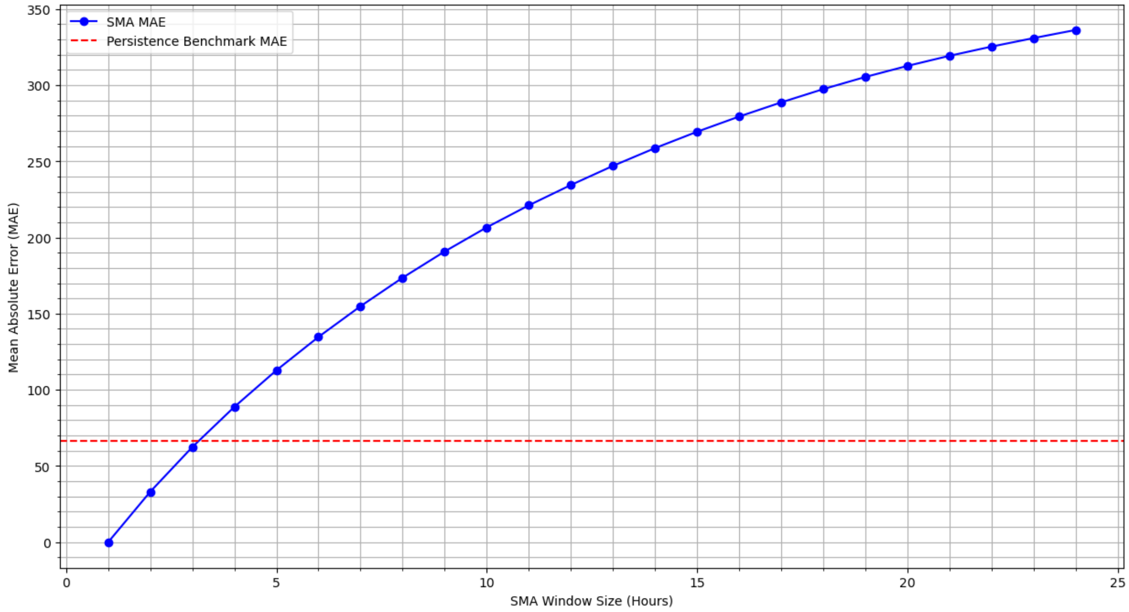

Figure 21. Optimal Window Size for SMA Based on MAE

9.3 Inference

From the results in Figure (21), the Mean Absolute Error (MAE) of the Simple Moving Average (SMA) model increases as the window size grows, indicating that shorter window sizes are better for forecasting wind generation. The persistence benchmark, represented by the red dashed line, trivially remains constant across different window sizes.

The lowest MAE is observed for SMA windows of 1 to 3 hours, which means that short-term averaging can capture wind generation trends without excessive lag. As the SMA window size increases beyond 3-4 hours, MAE steadily rises. The persistence model has a consistently lower MAE than SMA for window sizes greater than 4 hours. The persistence model remains superior to SMA at most window sizes, except for very short-term predictions.

In conclusion, for short-term forecasting (1-3 hours), a small SMA window is useful, whereas for longer horizons, persistence is a better predictor than SMA.

10 Question 9

10.1 Steps

The Persistence Benchmark Forecast assumes that the future wind generation value is equal to a past observed value from n hours ago:

$$X_{\text{predicted}}(t) = X(t-n) \tag{6}$$

where n is the forecast horizon, ranging from 1 hour to 24 hours. The objective is to evaluate the performance of the persistence model over different forecast horizons by calculating the Mean Absolute Error (MAE) and expressing it as a percentage of the maximum wind generation

First, drop NaN values. Then, compare the predicted values $X_{\text{predicted}}$ to the actual wind generation values $X(t)$. Compute $\text{MAE} = \frac{1}{N}\sum |X_{\text{actual}} - X_{\text{predicted}}|$, and then normalize it to ensure that it is being computed in terms of the maximum wind generation for consistency ($\text{MAE}_{\text{norm}} = \frac{\text{MAE}}{X_{\max}} \times 100$).

10.2 Results

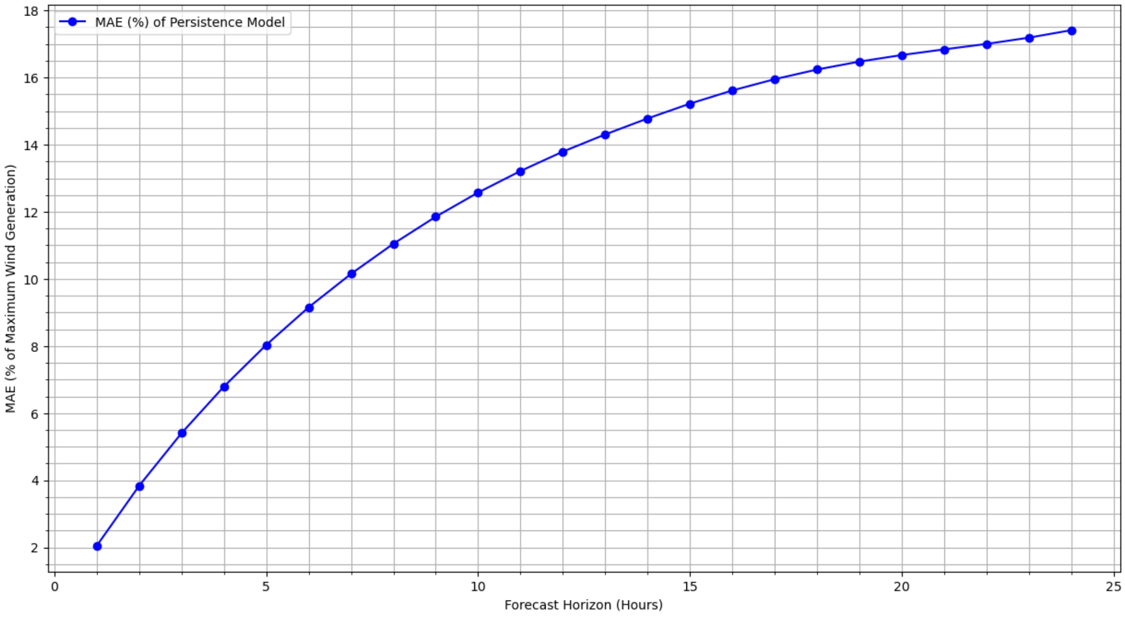

Figure 22. Persistence Benchmark MAE vs Forecast Horizon as a percent of Maximum Wind Generation

10.3 Inference

First, the most obvious improvement is when it is compared with SMA in Figure (21), where the MAE is significantly lower, achieving a maximum of 18% at 24 hours compared to 340% at 24 hours for the SMA model. Important to note is that MAE is low for short-term forecasts (1-3 hours) -which makes sense- but steadily increases as the forecast horizon grows. The curve follows an exponential increase initially, but beyond 15 hours, the MAE saturates and grows at a slower rate. The persistence model is effective for short-term wind forecasting, $\leq$6 hours, where the MAE remains relatively low. However, somewhere around 12-15 hours, the MAE becomes much largers.

11 Question 10

11.1 Steps

The AutoRegressive Integrated Moving Average (ARIMA) model is a time series forecasting method that combines *Autoregression (AR(p))*, which uses past values to predict future values, *Differencing (I(d))*, which ensures stationarity by removing trends, and *Moving Average (MA(q))*, which captures dependencies in past forecast errors.

The objective is to find the optimal ARIMA(p,1,q) model for wind generation by minimizing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

First, loop through the different combinations of p, $d=1$, and q, where $p, q\epsilon[1,4]$. Then, train each ARIMA model for each parameter combination and extract their BIC and AIC values. Then, evaluate the model performance based on AIC and BIC (lower AIC means better model fit and lower BIC means more parsimonious or peanlizes complexity). Then, to compare to two previous models, compute the MAE of the most optimal model and compare it with them.

11.2 Results

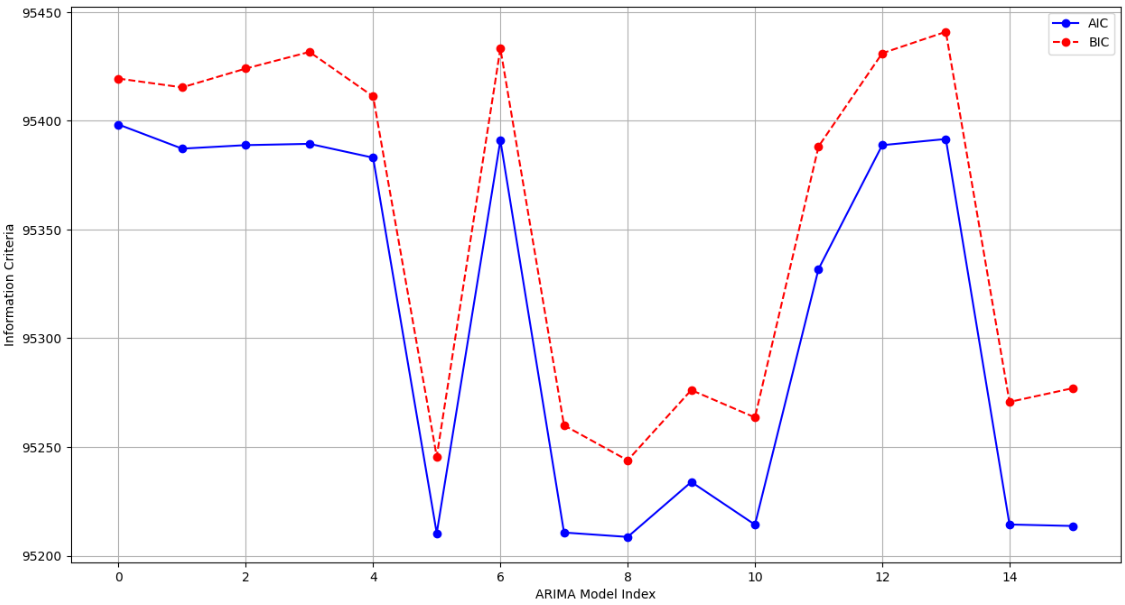

Figure 23. ARIMA model evaluation using AIC and BIC

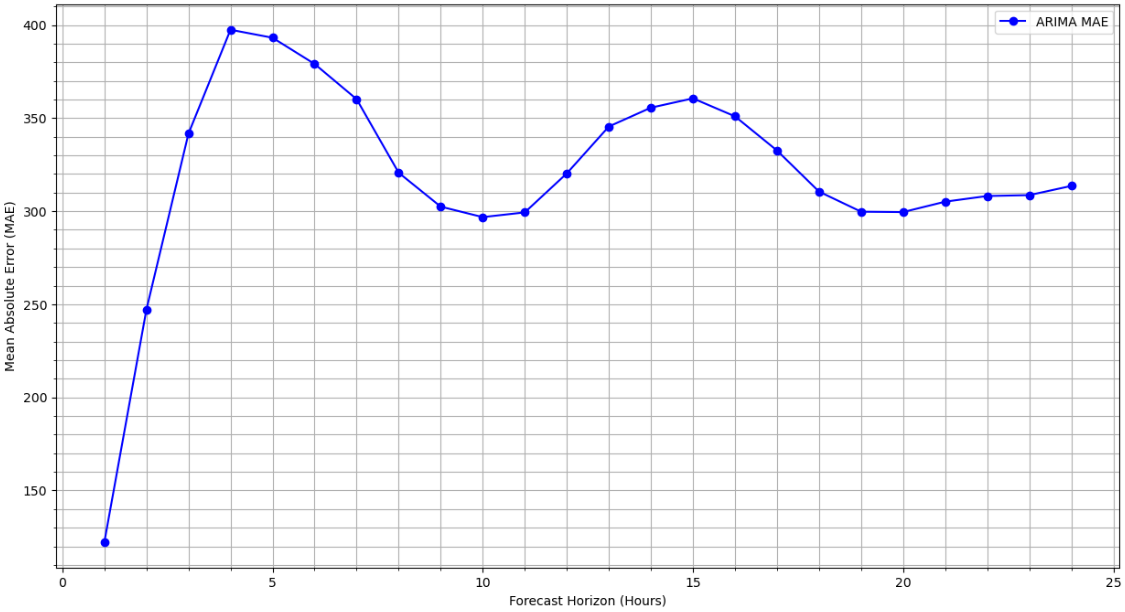

Figure 24. ARIMA Model Performance using MAE

11.3 Inference

The ARIMA model evaluation using AIC and BIC shows clear variations in model performance across different parameter choices. The lowest AIC and BIC values indicate the optimal model, suggesting that certain parameter combinations fit the wind generation data better than others. However, there are fluctuations, meaning some models may overfit or fail to capture the underlying patterns effectively.

Looking at the MAE performance plot, the ARIMA model's accuracy declines as the forecast horizon increases. Initially, the MAE grows rapidly, which is expected since short-term predictions are typically more reliable. However, after reaching a peak around the 10-hour mark, the error stabilizes and even slightly decreases. This suggests that while ARIMA struggles with mid-range forecasts, it eventually adapts to longer-term trends, though not necessarily improving beyond a certain point.

Overall, while ARIMA provides a structured approach to forecasting wind generation, its performance deteriorates with longer forecast horizons, but still more stable than SMA and persistence models.