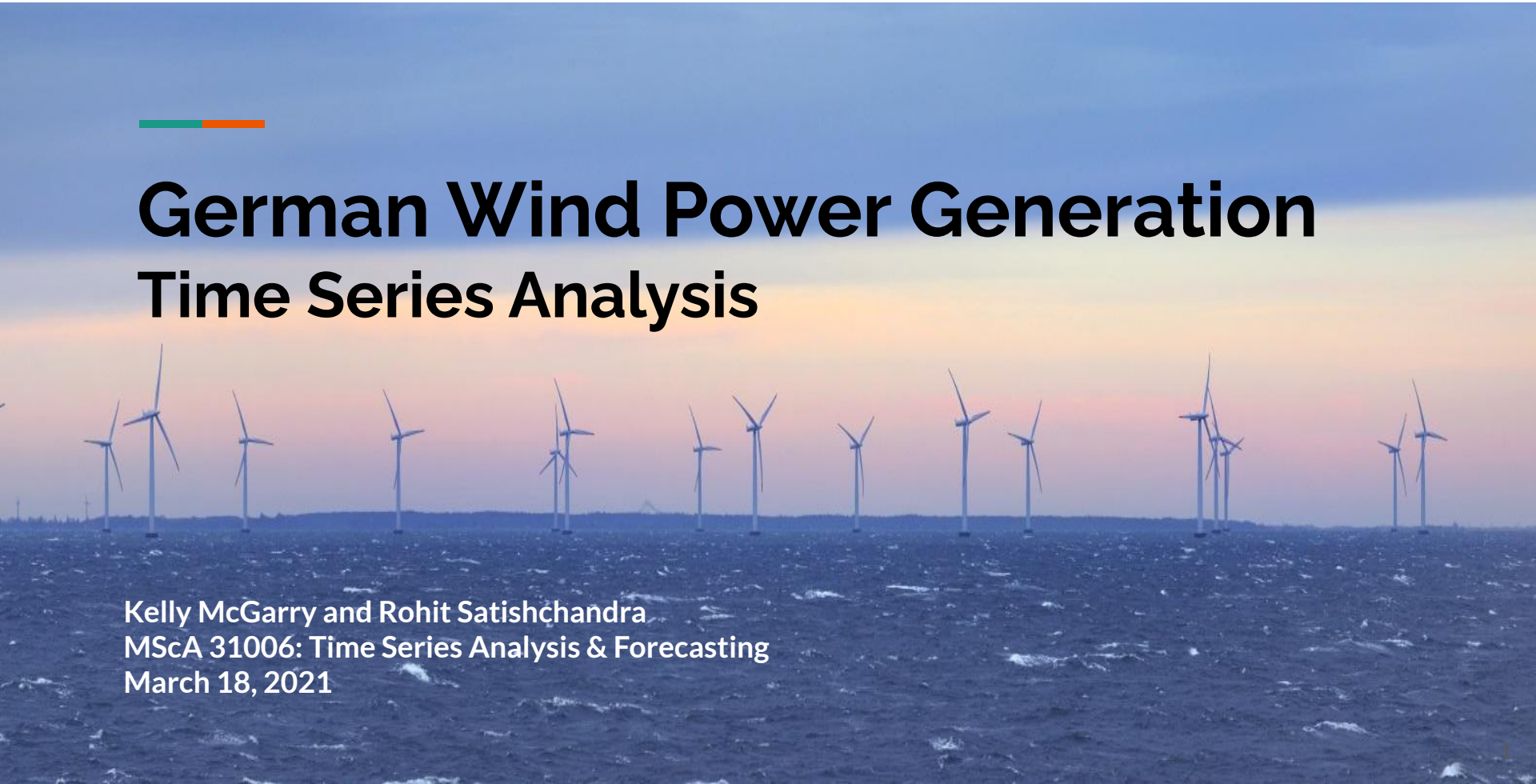


German Wind Power Generation
Time Series Analysis

Kelly McGarry and Rohit Satishchandra
MScA 31006: Time Series Analysis & Forecasting
March 18, 2021

Table of Contents

Introduction

Problem Statement

An essential component of any ***climate change mitigation*** strategy is the efficient transition from fossil fuel-based energy sources to ***renewable sources, such as wind and solar*** power. In this project, we will attempt to use time series data to ***forecast hourly and daily wind power generation in Germany***, as part of a broader effort to assess the future energy landscape and plan for appropriate resource allocation.

Problems to solve

1. **Explore time series properties.** Assess independence of the four separate time series, examine trends, inspect seasonal patterns, review outliers.

2. **Apply necessary transformations to satisfy model assumptions.** If necessary, de-couple mean and variance, smooth data with moving average, and/or apply differencing. Confirm stationarity of series.

3. **Train and compare fits of different models.** Dynamic Harmonic Regression, TBATS, Seasonal ARIMA.

4. **Hourly: Forecast 48 hours (2 days) of wind power generation for one power plant (50Hertz).** Visually inspect and compare model fits and forecast plots.

5. **Daily: Forecast 20 days of wind power generation for one power plant (50Hertz).** Visually inspect and compare model fits and forecast plots.

Wind Power Generation Data

An Overview

Data Overview

The data contains **wind power generation** in Germany for an approximate one year period. The datasets come **from four German energy companies**, known as the four German Transmission System Operators (TSOs).

The four German energy companies:

- 50 Hertz
- Amprion
- TenneT TSO
- TransnetBW

The full data consists of **four files**, each from one of the four TSOs.

Each file has **one row per date** (397 rows per file) and 96 columns for power recordings (non-normalized, **terawatt hours**) at **15-minute intervals** throughout the day.

The data is from **08/23/2019 to 09/22/2020** (397 days, 9528 hours).

Map of the four German energy companies

*image credit: https://www.cleanenergywire.org

Data Re-Formatting

Date	00:00:00	00:15:00	00:30:00	00:45:00	...	23:30:00	23:45:00
2019-08-23	9.68	10.16	10.94	11.39	...	69.97	69.58
2019-08-24	67.94	67.52	64.48	64.78	...	94.75	91.73

Example of Raw Data (15-minute sampling rate)

time	hourly_total
2019-08-23 00:00:00	42.17
2019-08-23 01:00:00	52.09
...	...
2019-08-24 00:00:00	264.72

Example of Re-Formatted Data Hourly Total (sum every 4 columns)

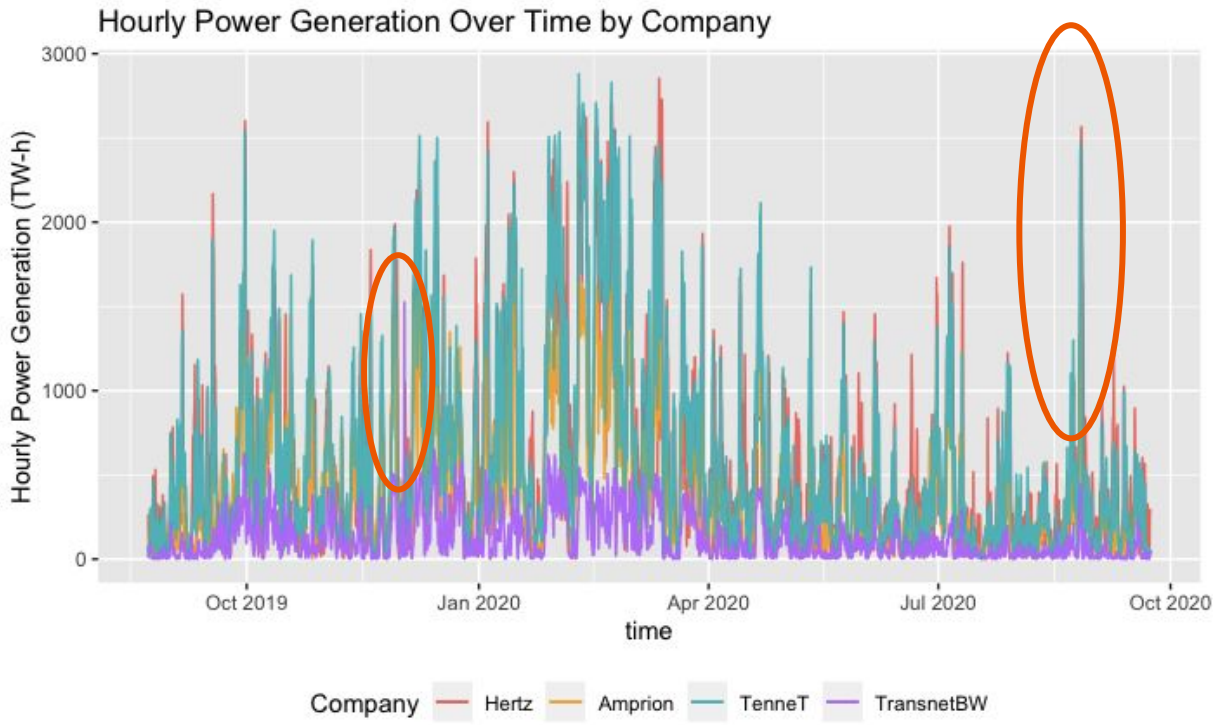
date	daily_total
2019-08-23	2655.76
2019-08-24	6567.22

Example of Re-Formatted Data Hourly Total (sum every 4 columns)

Data Exploration & Analysis

Hourly

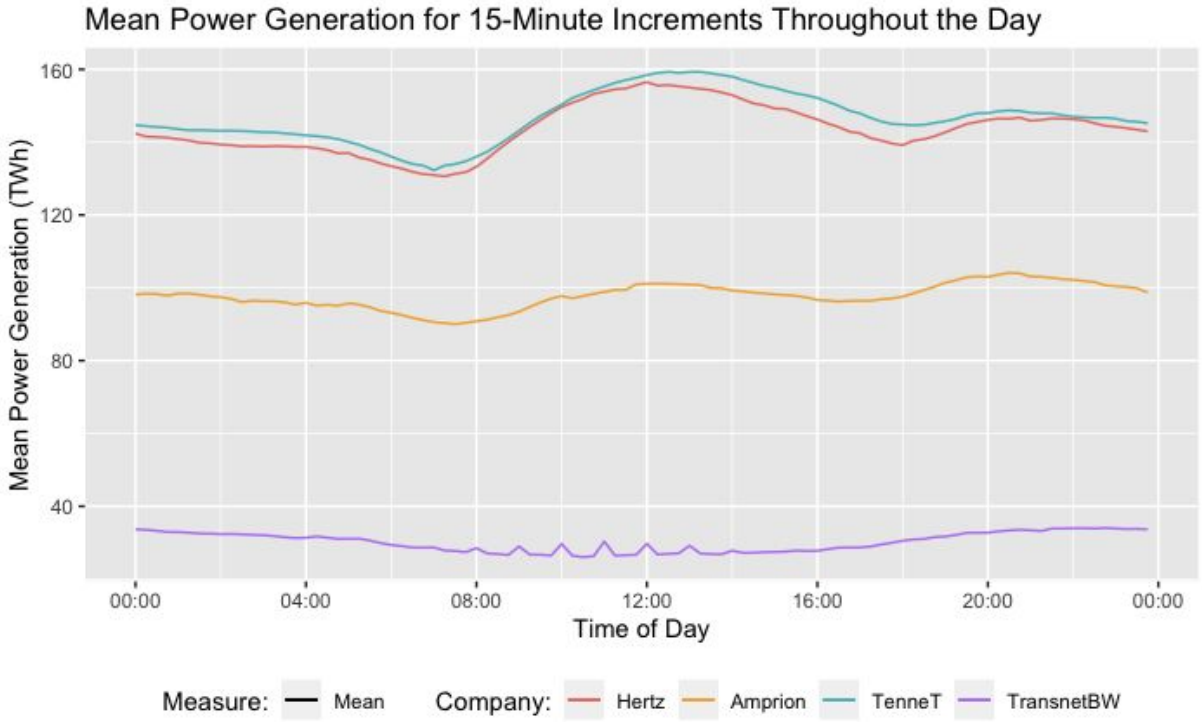
Hourly Time Series Plot



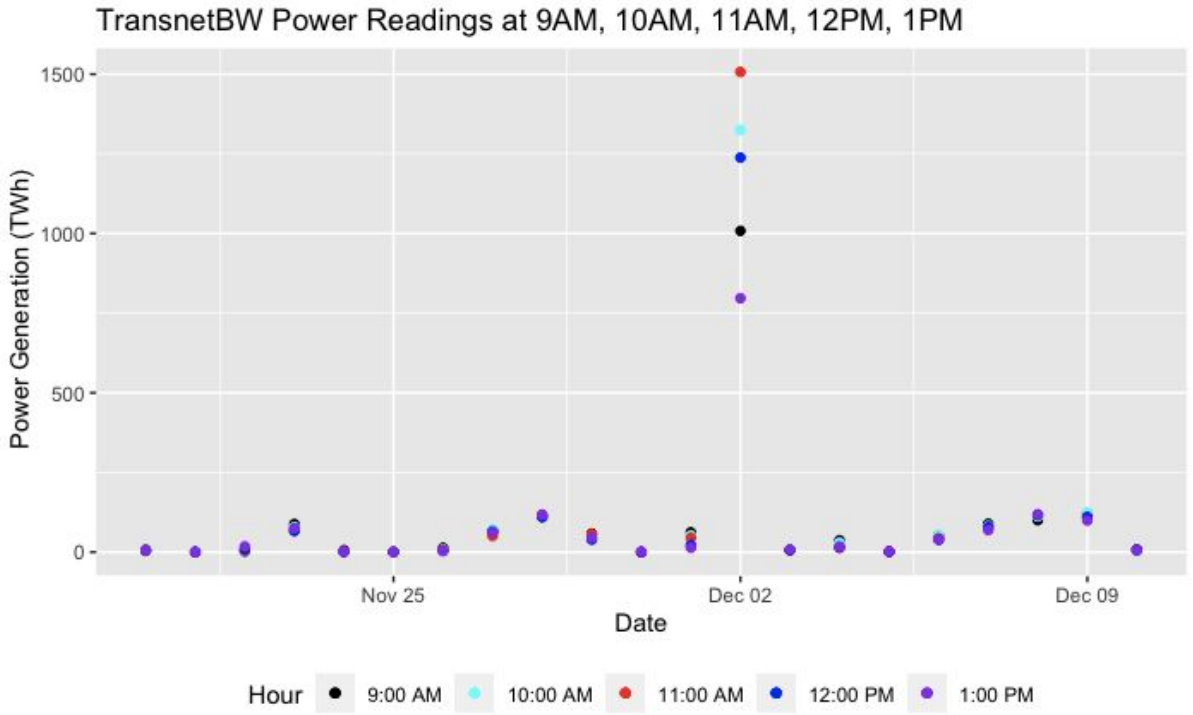
Hourly Series Summary

	TenneTTSO	50Hertz	Amprion	TransnetBW
Min	4.0	2.0	0.0	0.0
1st Quartile	169.0	174.0	94.0	22.0
Median	373.0	383.0	247.0	66.0
Mean	586.8	574.4	391.1	120.6
3rd Quartile	825.2	794.0	574.2	175.0
Max	2880.0	2853.0	1834.0	1524.0

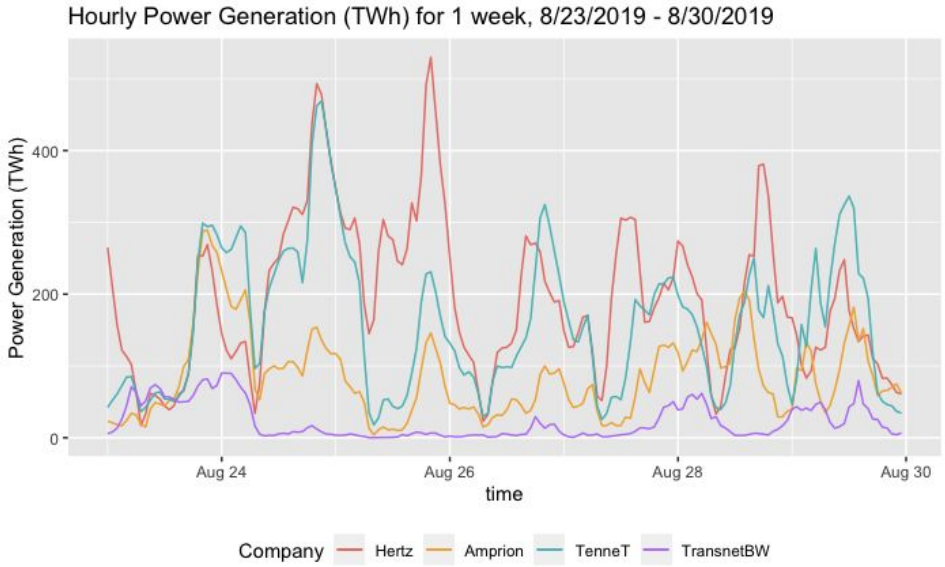
15-Minute Column Means

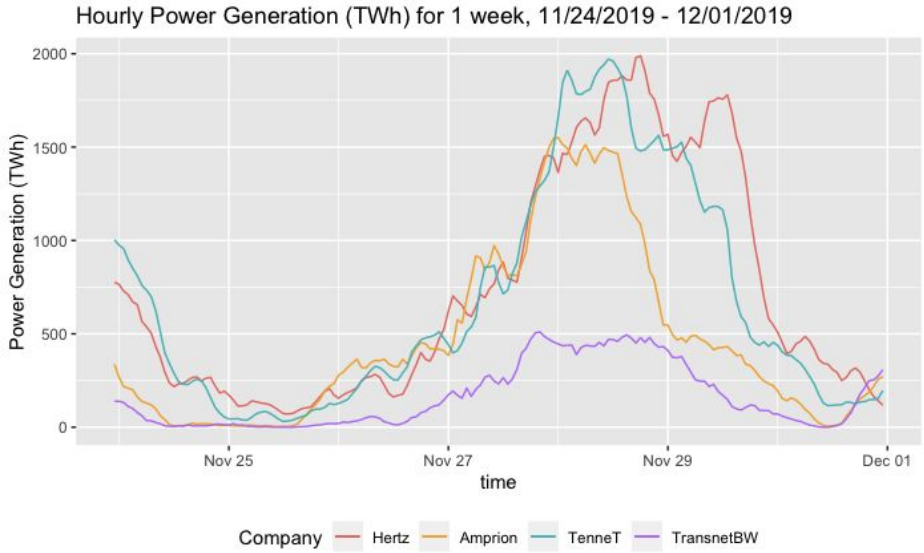


TransnetBW Outliers

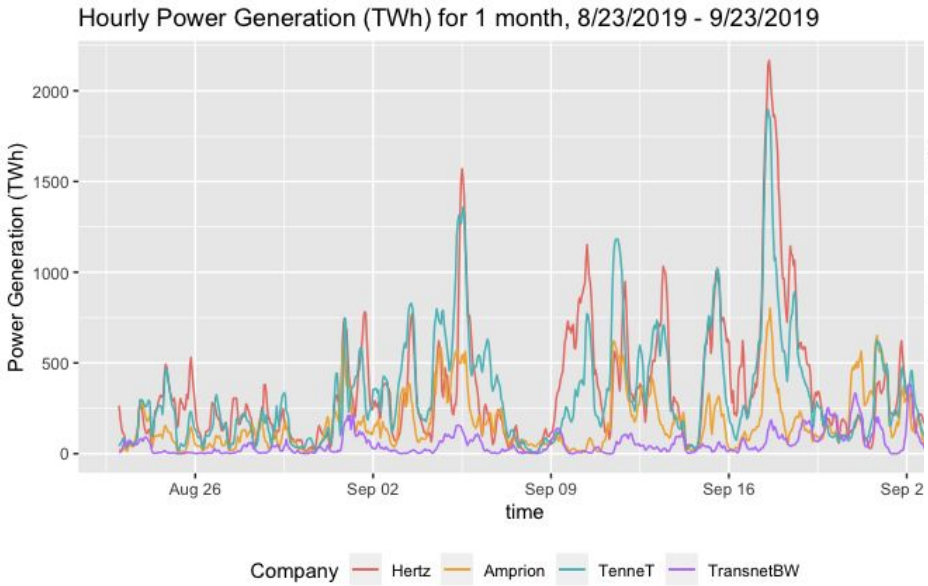


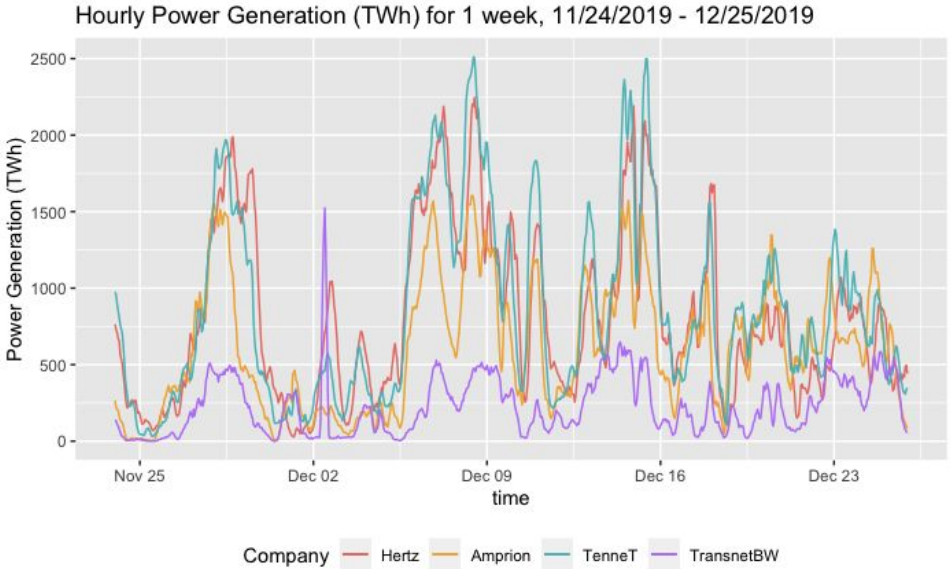
Complex Seasonality (Week)





Complex Seasonality (Month)





Hourly Power Generation Analysis

Is the Box-Cox transformation necessary?

- lambda=1 for all four -> no transformation is necessary

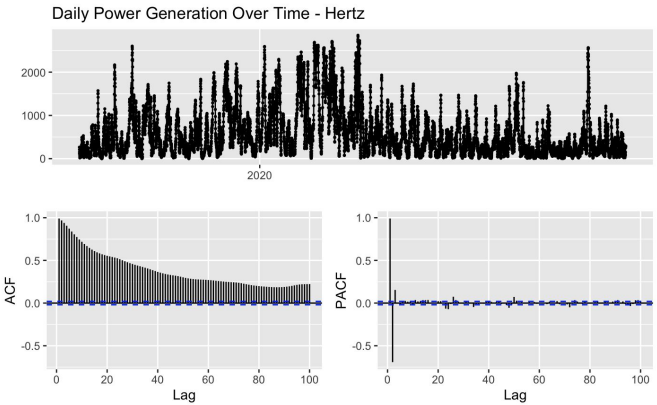
Is our data stationary?

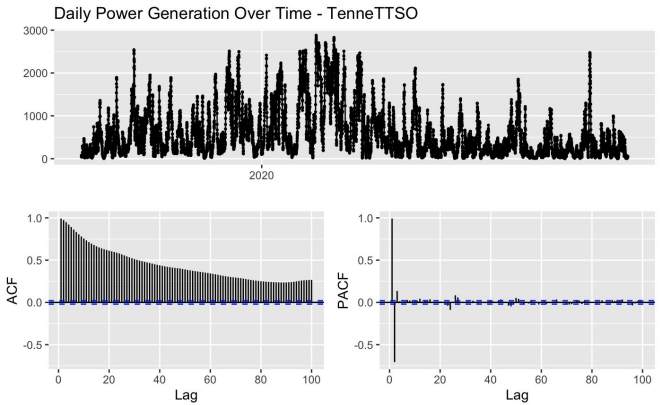
- ADF test p-value = 0.01 -> statistically significant, we reject the null hypothesis that the data is nonstationary (**data is stationary**)
- KPSS test p-value = 0.01 -> statistically significant, we reject the null hypothesis that the data is stationary (**data is non-stationary**)
- KPSS indicates non-stationarity and ADF indicates stationarity - **The series is difference stationary.** Differencing is to be used to make series stationary. The differenced series is checked for stationarity.

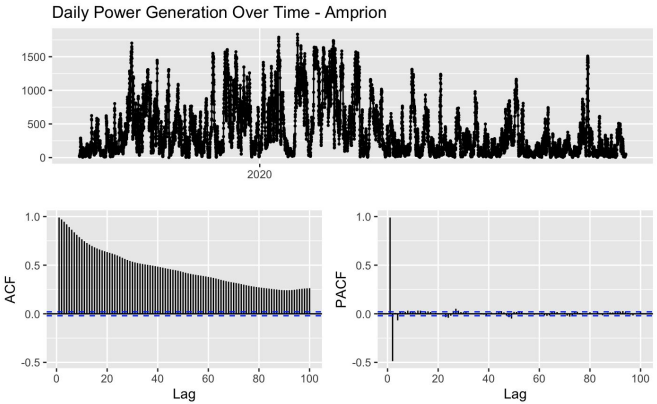
Differencing

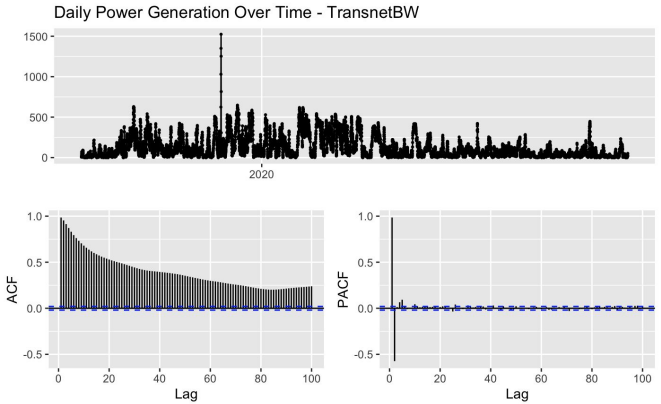
- Number of differences required for a stationary series = 1
- ADF and KPPS both indicate stationarity
    - **ADF p-value < 0.01 → reject null hypothesis → stationary**
    - **KPSS p-value > 0.1 → accept null hypothesis → stationary**

Hourly Power Generation Plots









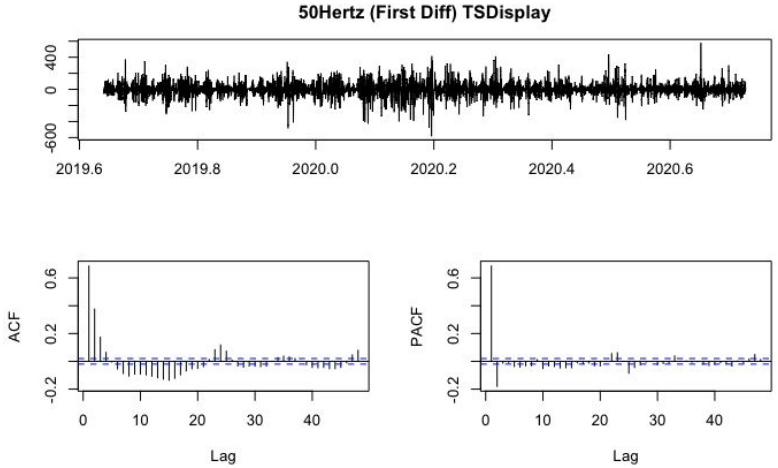
ACF Plots

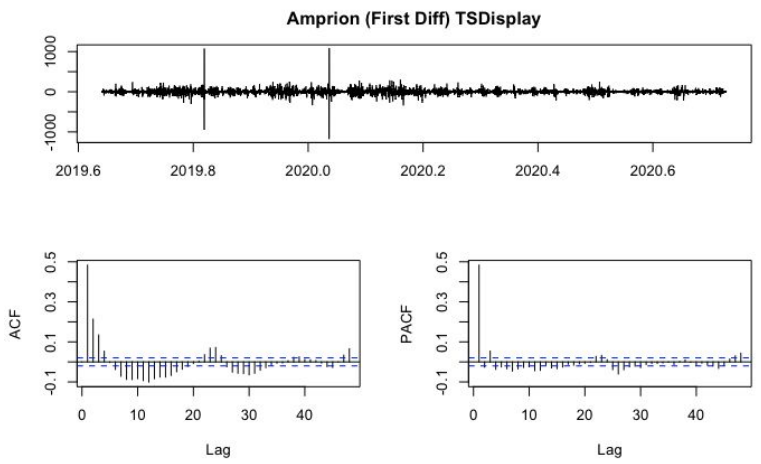
Trails off/decay -> Autoregressive (AR) process

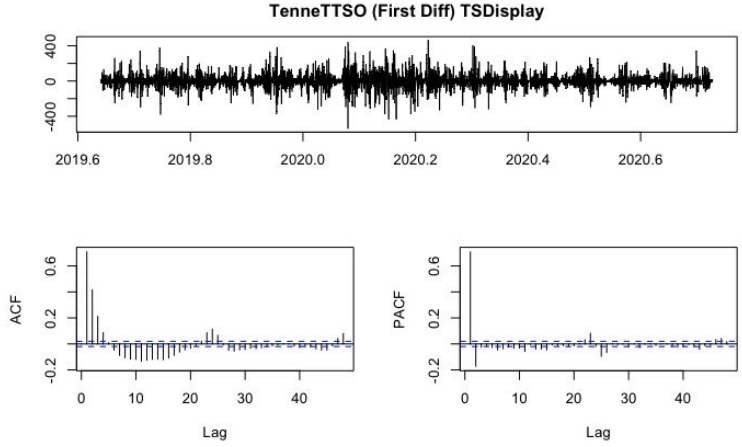
PACF Plots

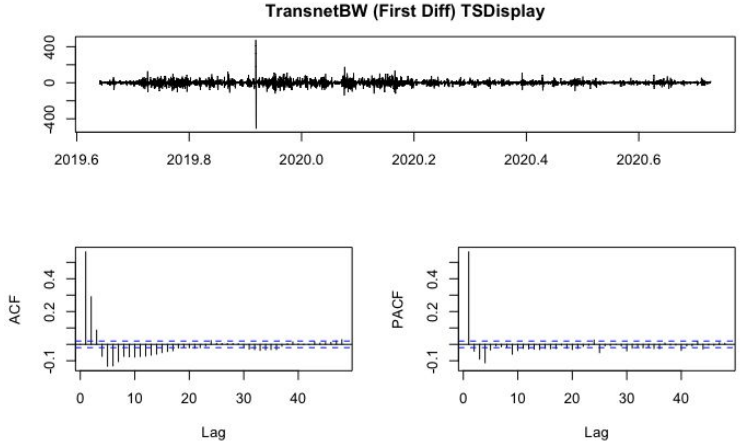
Hard cut-off/drop = Autoregressive (AR) process

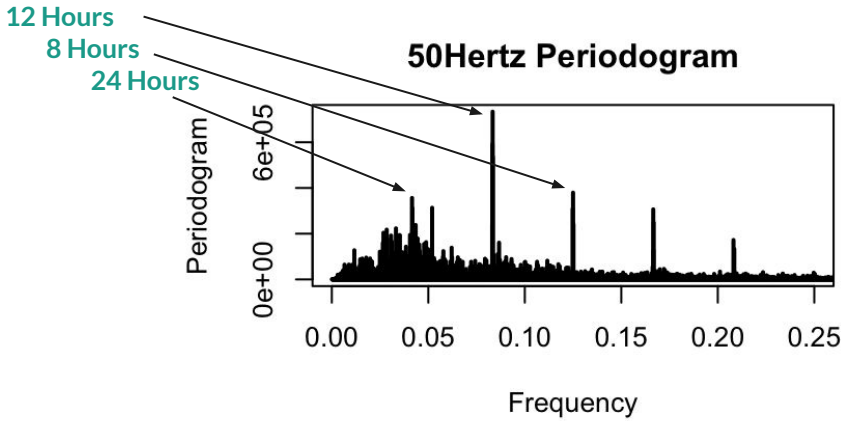
This data represents an ARIMA(p,d,0). The model is AR because the ACF trails off after a lag and has a hard cut-off in the PACF after a lag. This lag is taken as the value for p.

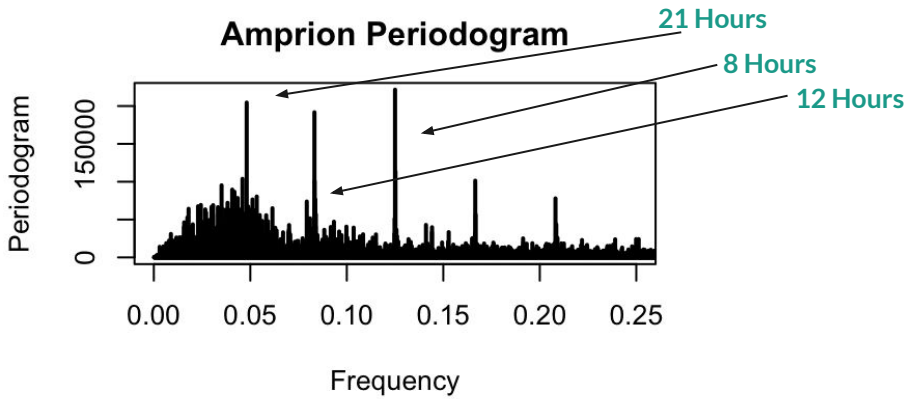

50Hertz (First Diff) TSDisplay
2019.6
2019.8
2020.0
2020.2
2020.4
2020.6
ACF
PACF
Lag

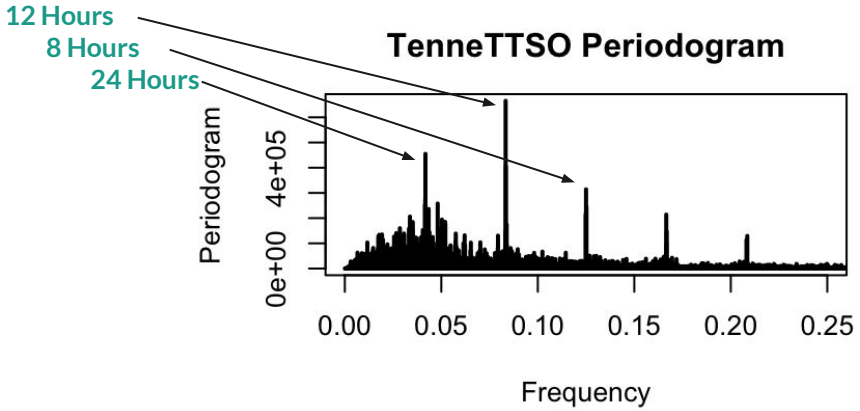

Amprion (First Diff) TSDisplay
2019.6
2019.8
2020.0
2020.2
2020.4
2020.6
ACF
PACF
Lag

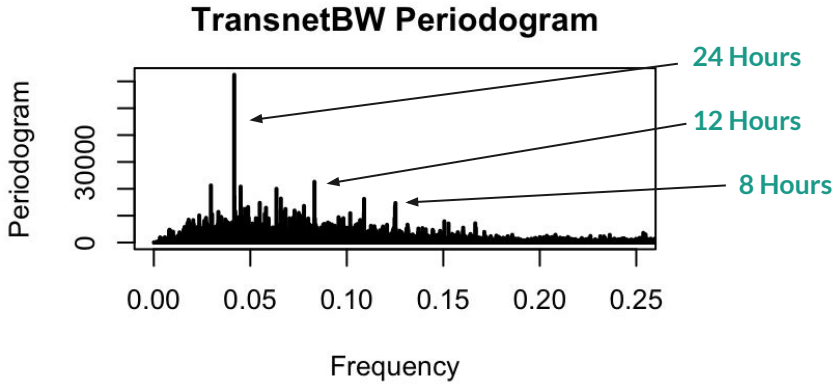

TenneTTSO (First Diff) TSDisplay
2019.6
2019.8
2020.0
2020.2
2020.4
2020.6
ACF
PACF
Lag


TransnetBW (First Diff) TSDisplay
2019.6
2019.8
2020.0
2020.2
2020.4
2020.6
ACF
PACF
Lag


12 Hours
8 Hours
24 Hours
50Hertz Periodogram
Periodogram
0e+00
6e+05
0.00
0.05
0.10
0.15
0.20
0.25
Frequency


21 Hours
8 Hours
12 Hours
Amprion Periodogram
Periodogram
0
150000
0.00
0.05
0.10
0.15
0.20
0.25
Frequency

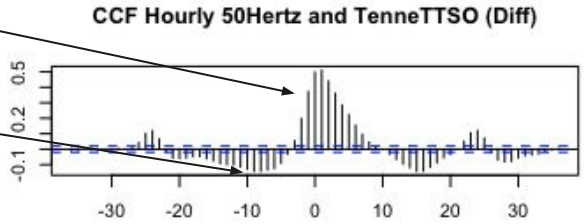

12 Hours
8 Hours
24 Hours
TenneTTSO Periodogram
Periodogram
0e+00
4e+05
0.00
0.05
0.10
0.15
0.20
0.25
Frequency

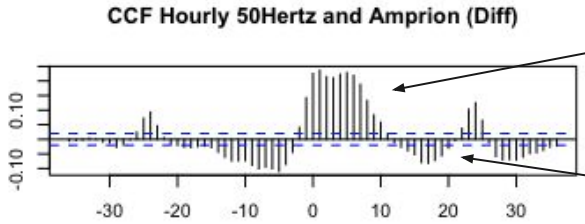

24 Hours
12 Hours
8 Hours
TransnetBW Periodogram
Periodogram
0
30000
0.00
0.05
0.10
0.15
0.20
0.25
Frequency

Cross Correlation Functions

50Hertz “lags” TenneTTSO

50Hertz “leads” TenneTTSO



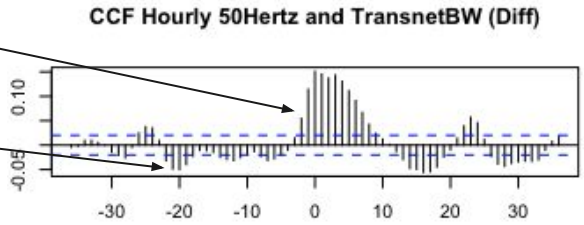


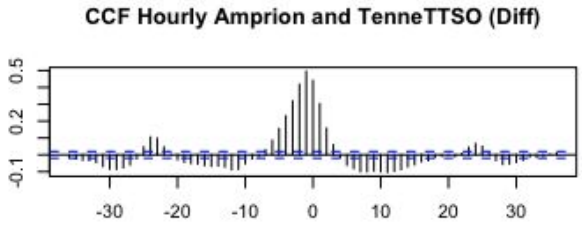
High values of both series associated with each other

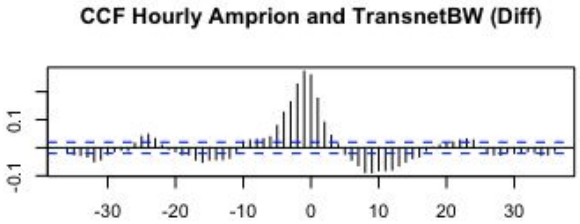
High values of one series associated with negative values of other series

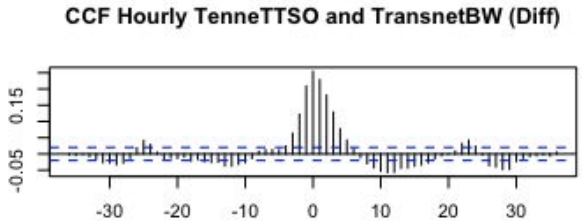
50Hertz “lags” TransnetBW

50Hertz “leads” TransnetBW







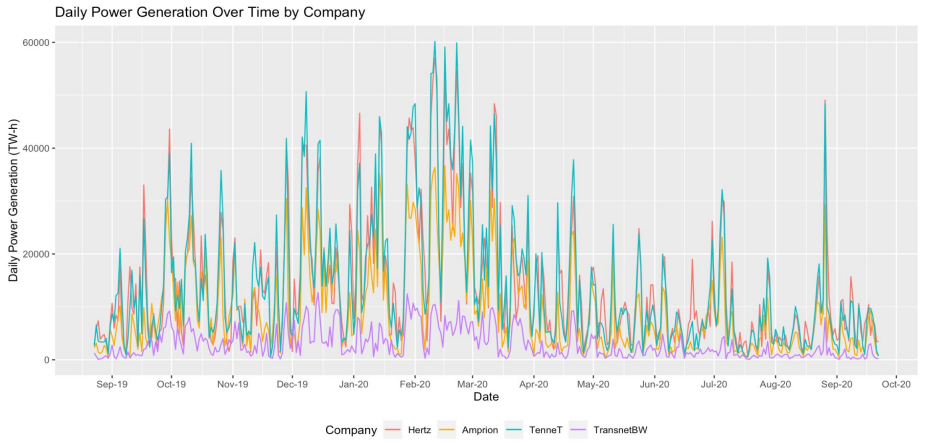


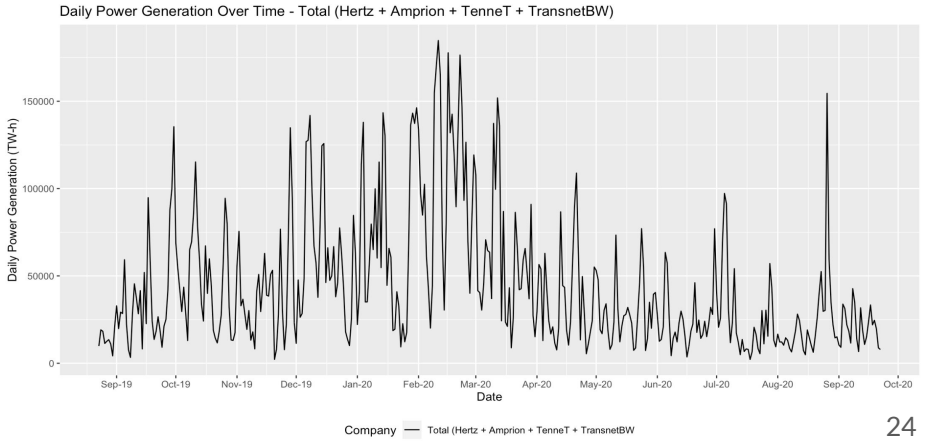
Daily

Daily Power Generation Overview

The daily power generation plots are easier to look at to see the overall trend.

- The data looks to show some seasonality behaviour.
- There are prominent daily increases about every 7 and 12-14 days, suggesting that there might be weekly and bi-weekly seasonality.





Total Daily Power Generation Analysis

Is the Box-Cox transformation necessary?

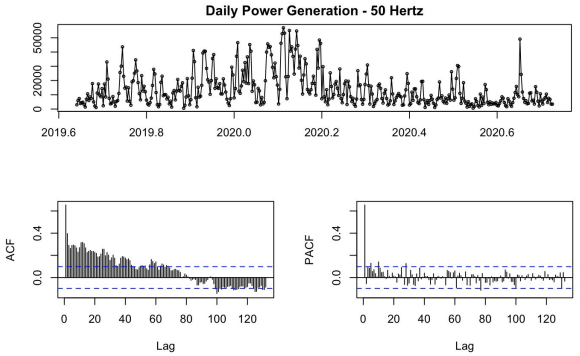
- lambda=1 -> no transformation is necessary

Is our data stationary?

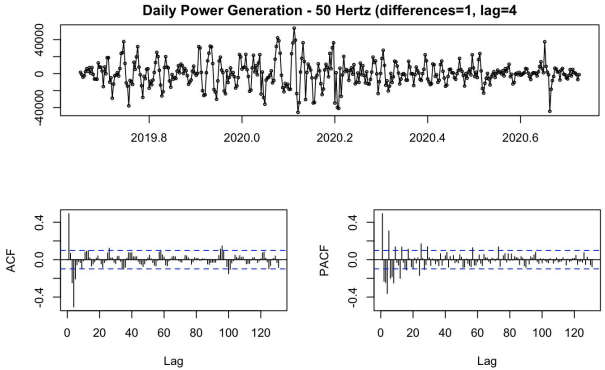
- ADP test p-value = 0.01 -> statistically significant, we reject the null hypothesis that the data is nonstationary (**data is stationary**)
- KPSS test p-value = 0.01 -> statistically significant, we reject the null hypothesis that the data is stationary (**data is non-stationary**)
- KPSS indicates non-stationarity and ADF indicates stationarity - **The series is difference stationary.** Differencing is to be used to make series stationary. The differenced series is checked for stationarity.

Differencing

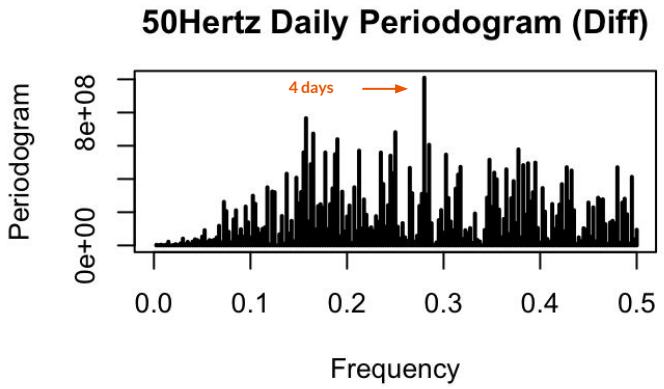
- Number of differences required for a stationary series = 1
- ADP and KPPS both indicate stationarity

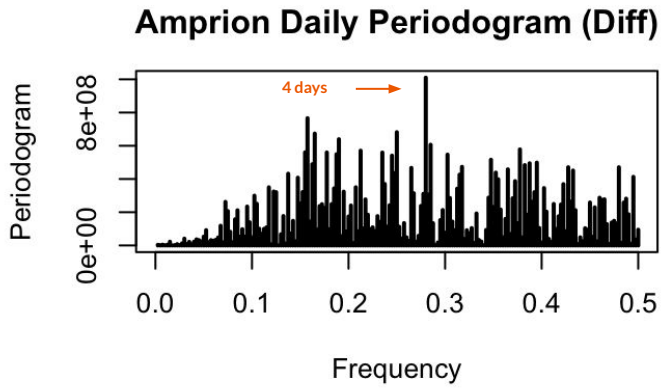


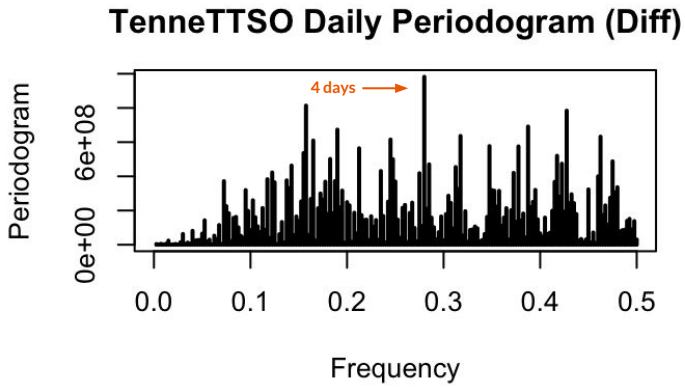
This data represents an ARIMA(p,d,0). The model is AR because the ACF trails off after a lag and has a hard cut-off in the PACF after a lag. This lag is taken as the value for p.

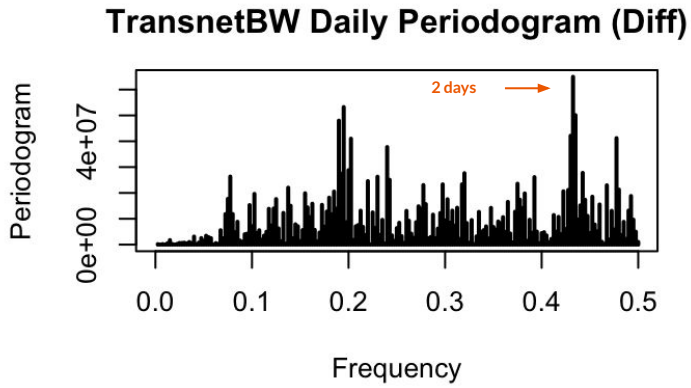


The periodogram graphs a measure of the relative importance/ strengths of possible frequency values that might explain the oscillation pattern of the observed data.









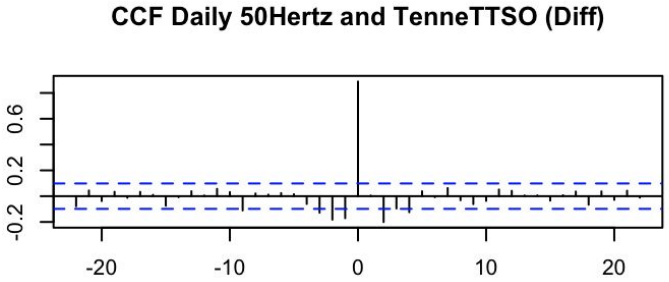
The dominant periods (or frequencies) of a time series.

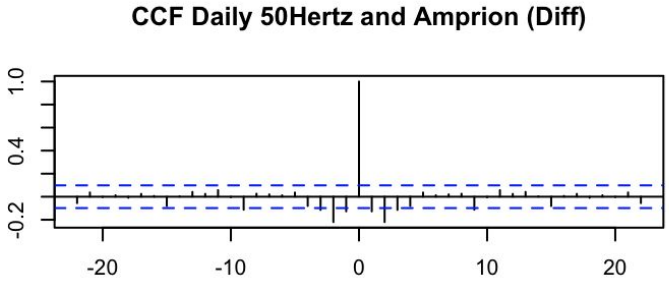
50 Hertz and Amprion
- 6 days and 4 days

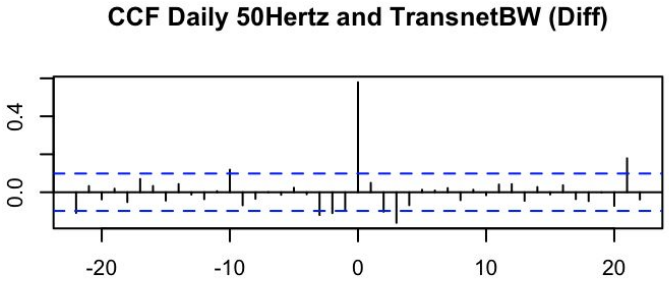
Tenne TSSO
- 6 days, 4 days and 2 days

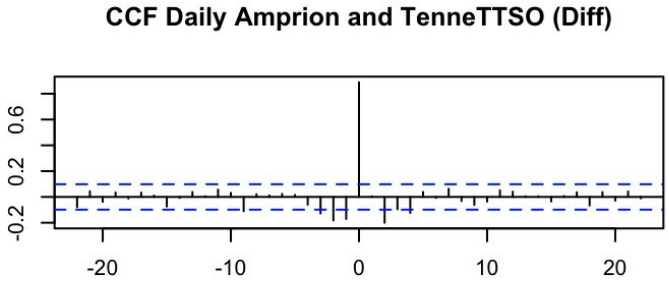
TransnetBW
- 5 days and 2 days

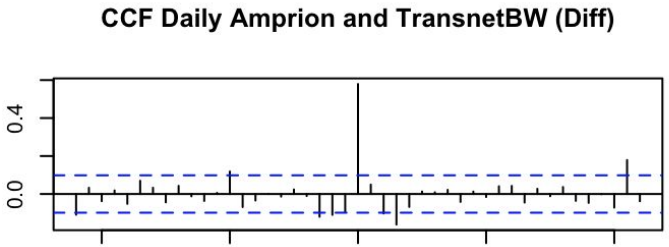
Cross Correlation Functions and Lagged Regressions | Are there any lags of the x-variable (company A) that might be useful predictors of the y-variable (company B)?

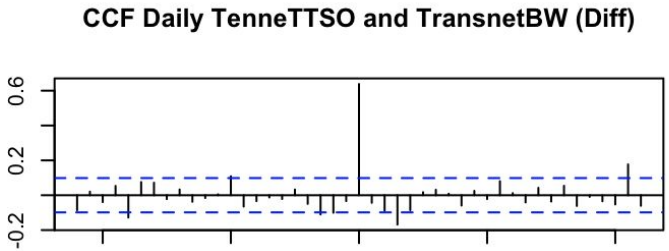












Key Takeaways
High values of 50 Hertz are associated with high values of Tenne TTSO

The peaks in the middle of the plots (lag=0) indicate that the companies are most strongly correlated at the same time.

Model Selection - 50 Hertz

Hourly
Dynamic Harmonic Regression
TBATS

Daily
Dynamic Harmonic Regression
Seasonal ARIMA

Dynamic Harmonic Regression

Hourly

Dynamic Harmonic Regression

This is where periodic seasonality can be handled using pairs of Fourier terms. The assumption is that the seasonal pattern is unchanging.

Advantages

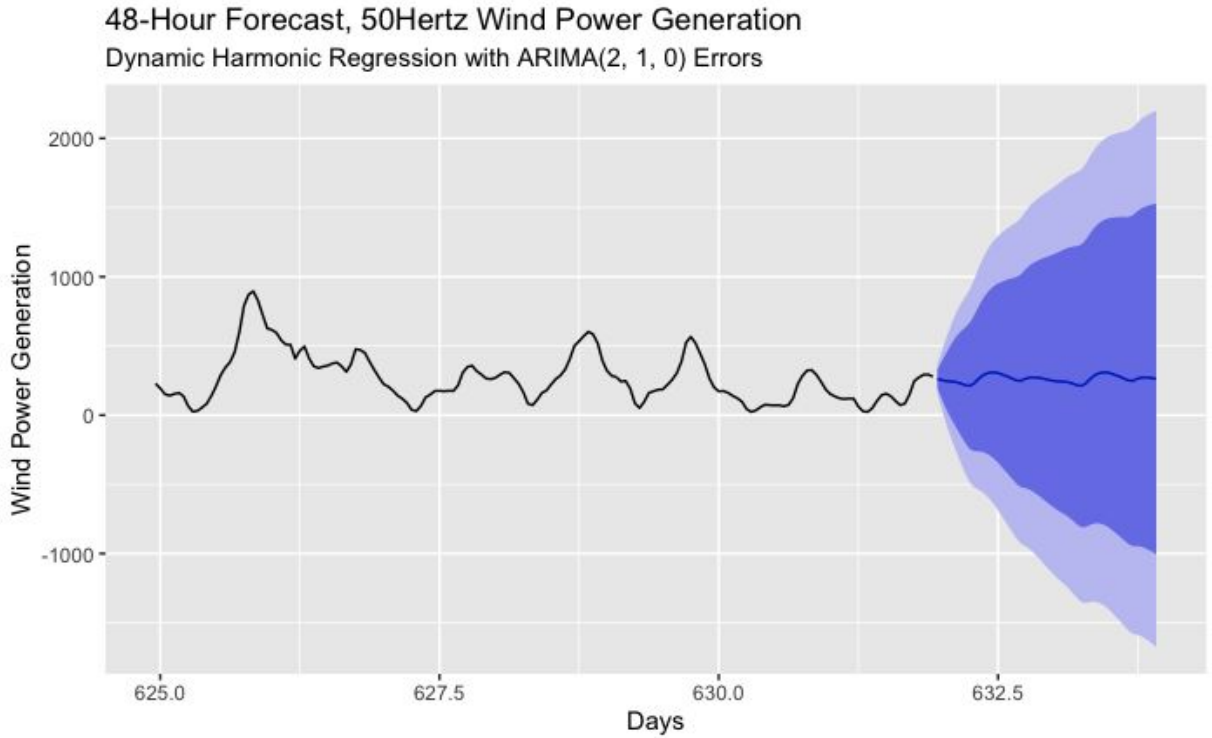
- It allows any length seasonality ;
- For data with more than one seasonal period , you can include Fourier terms of different frequencies;
- The seasonal pattern is smooth for small values of K(but more wiggly seasonality can be handled by increasing K);
- The short term dynamics are easily handled with a simple ARMA error

Disadvantages - seasonality is assumed to be fixed

Multi-Seasonal TS Object

```
hertz_hourly_msts <- msts(hertz_hourly[, 2],
                          # 8 hour, 12 hour, and 24 hour
                          seasonal.periods = c(24/3, 24/2, 24),
                          # natural frequency
                          ts.frequency = 24,
                          # 235th day of the year
                          start = c(235, 0))
```

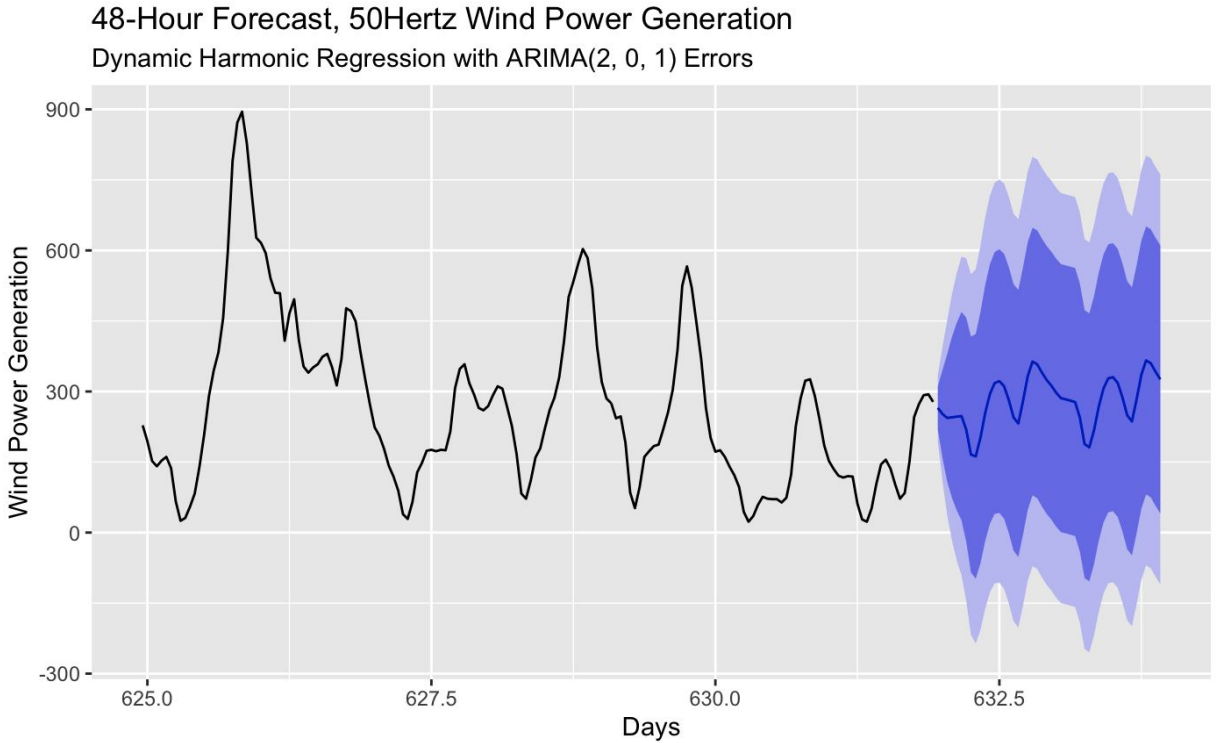
Dynamic Harmonic Regression



- Fit to **full multi-seasonal hourly series**
- Note large forecast intervals
- Some variance captured
- **AIC: 102972.7**

**** Plot only shows last 7 days + 48 Hour Forecast**

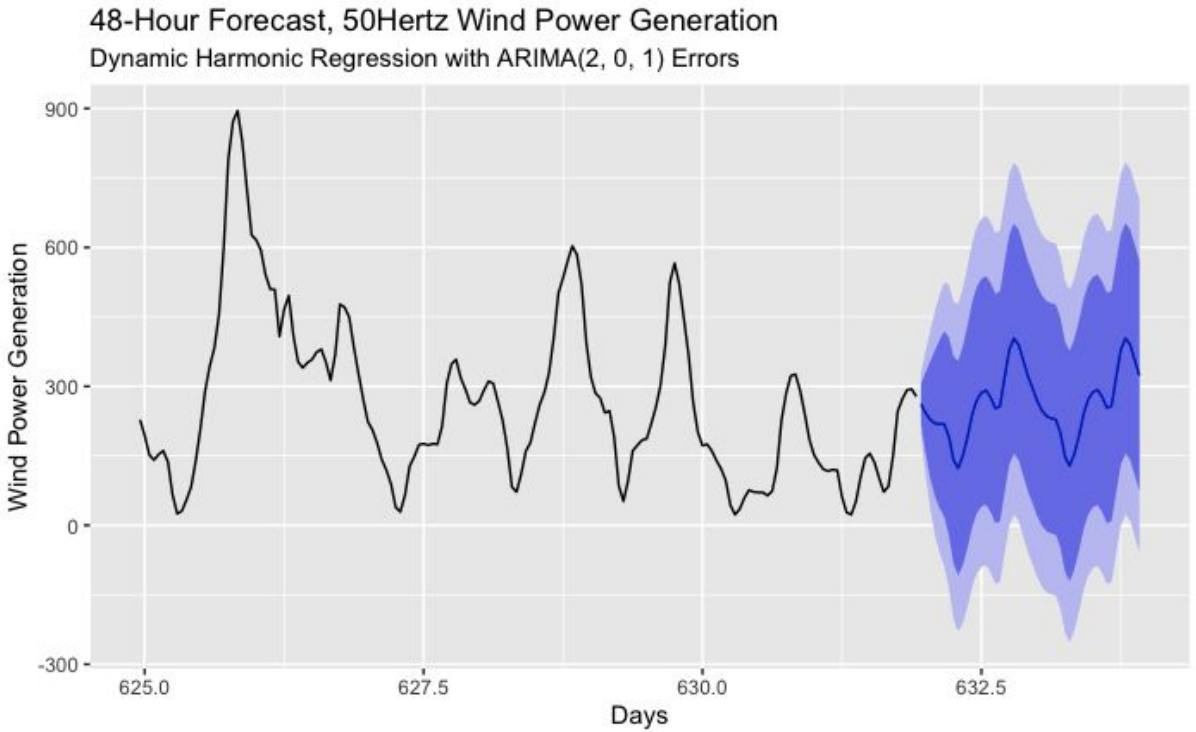
Dynamic Harmonic Regression



- Fit to **last 21 days worth of observations**
- Forecast intervals still large
- ARIMA (2, 0, 1) Errors
- **AIC: 5086.73**

**** Plot only shows last 7 days + 48 Hour Forecast**

Dynamic Harmonic Regression



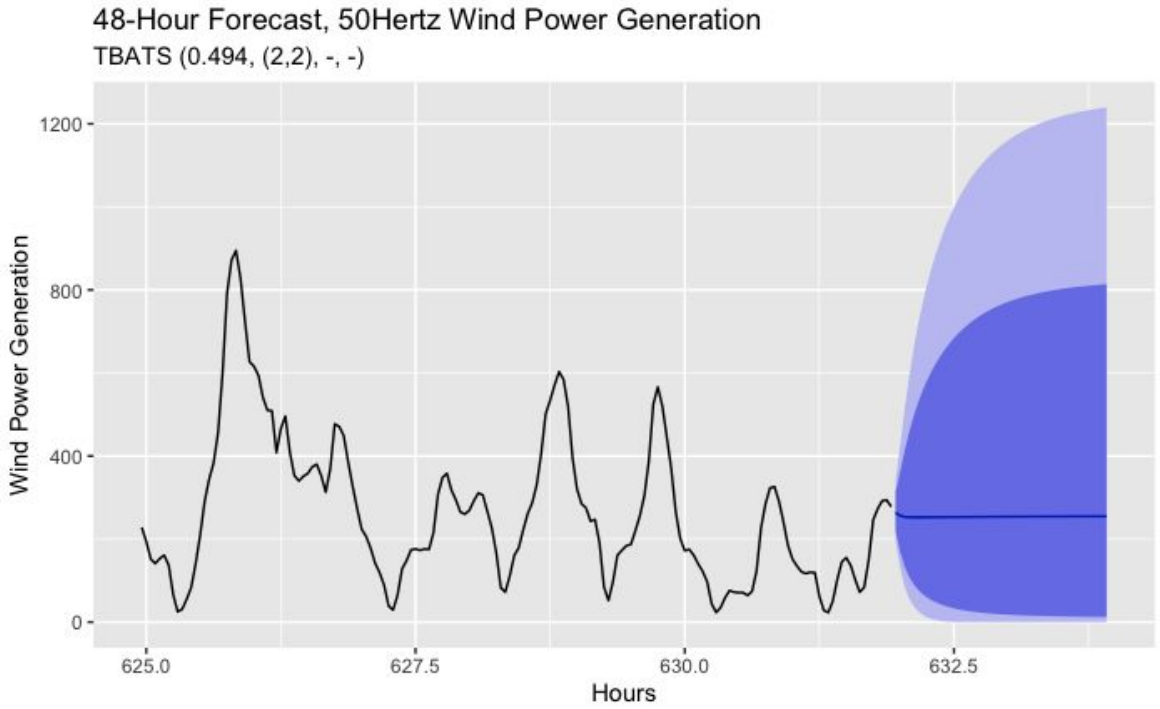
- Fit to **last 14 days worth of observations**
- Note large forecast intervals
- ARIMA (2, 0, 1) Errors
- **AIC: 3356.16**

** **Plot only shows last 7 days + 48 Hour Forecast**

TBATS - Trigonometric Exponential Smoothing State Space model with Box-Cox transformation, ARMA errors, Trend and Seasonal Components

Hourly

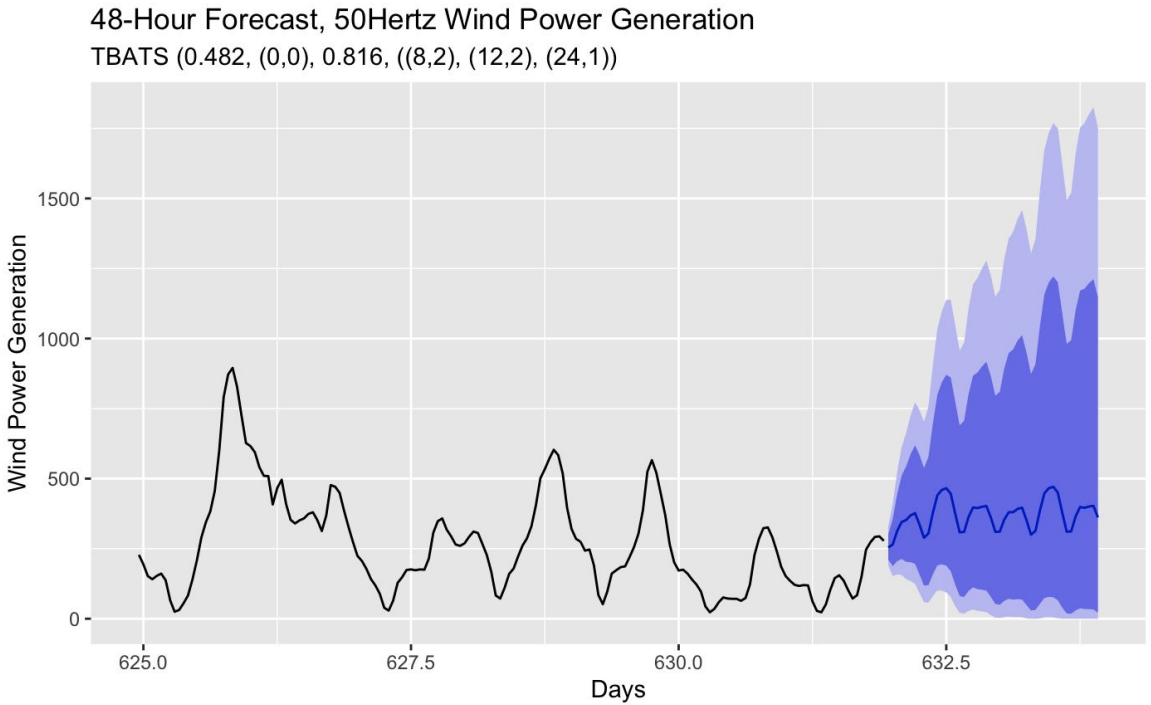
TBATS



- Fit to **full multi-seasonal hourly series**
- Note **even larger forecast intervals**
- 0 Fourier terms
- 0.49 Box-Cox
- ARMA (2,2) Errors
- **AIC: 158678.4**

**** Plot only shows last 7 days + 48 Hour Forecast**

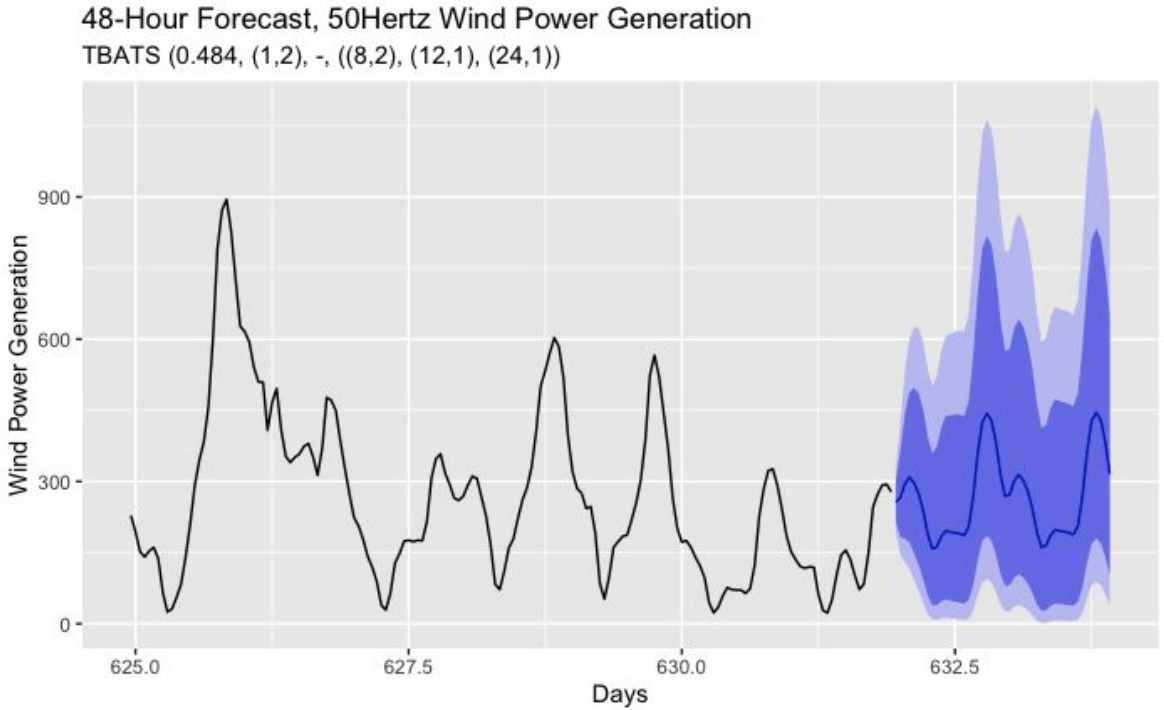
TBATS



- Fit to **last 21 days worth of observations**
- Note **even larger forecast intervals**
- 5 Fourier terms
- 0.48 Box-Cox
- No ARMA Errors
- **AIC: 6713.61**

**** Plot only shows last 7 days + 48 Hour Forecast**

TBATS



- Fit to **last 14 days worth of observations**
- Note **even larger forecast intervals**
- 4 Fourier terms
- 0.48 Box-Cox
- ARMA (1, 2) Errors
- **AIC: 4297.55**

**** Plot only shows last 7 days + 48 Hour Forecast**

Residuals Summary

- Among Dynamic Harmonic Regression models fit to the hourly series, only the **model that was fit using the last 14 days (336 hours) of data had residuals that passed Ljung-Box test for serial correlation and none pass the Shapiro-Wilk test for normality**

- Among TBATS models fit to the hourly series, **none of the models has independent residuals and only the model fit to 3 weeks recent data passed normality test**

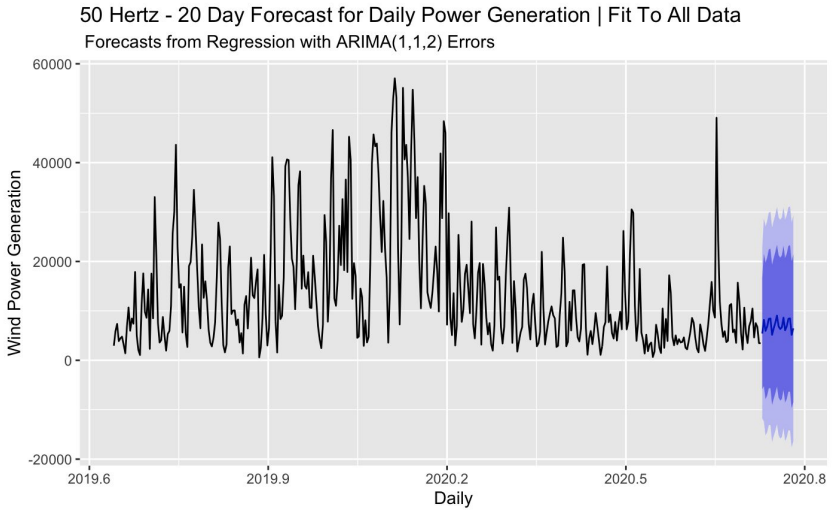
Dynamic Harmonic Regression
Daily

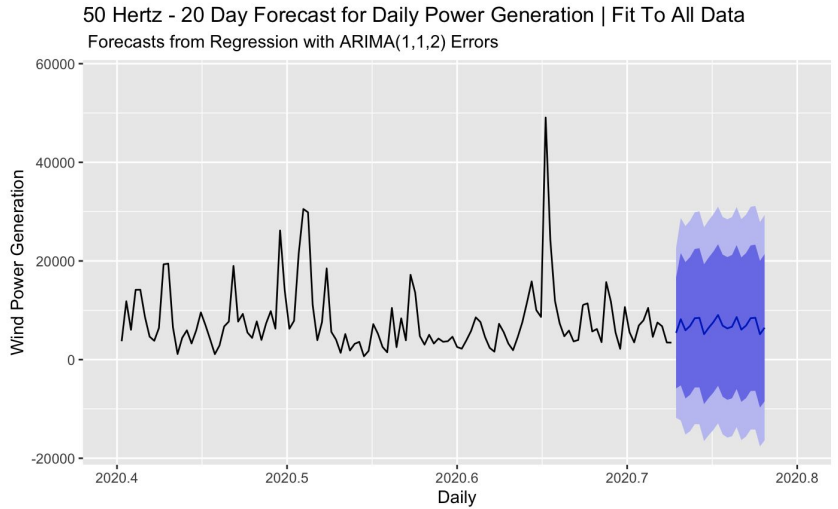
Combine Fourier terms with ARIMA errors

Dynamic Harmonic Regression - All Data

The chart on the left shows the full daily time series with 20 days of forecast. The plot on the right shows just the last few months (with same forecast). In both plots, we see large forecast intervals.

AIC = 8330.084

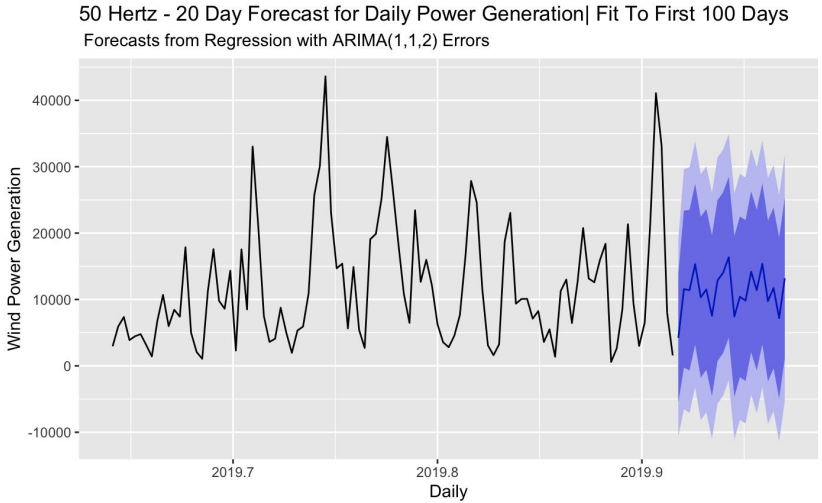


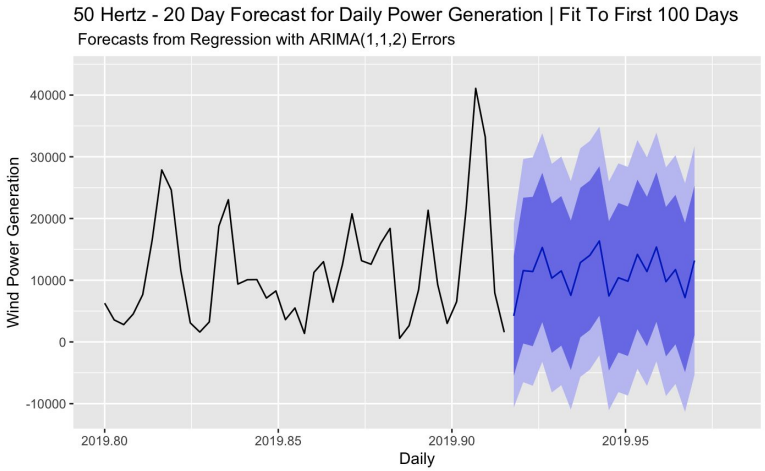


Dynamic Harmonic Regression - First 100 Days

Subset time series to just 100 days. The **plot on the left shows the full daily time series with 20 days of forecast**. The **plot on the right shows just the last few months (with same forecast)**. In both plots, we see large forecast intervals.

AIC = 2103.962

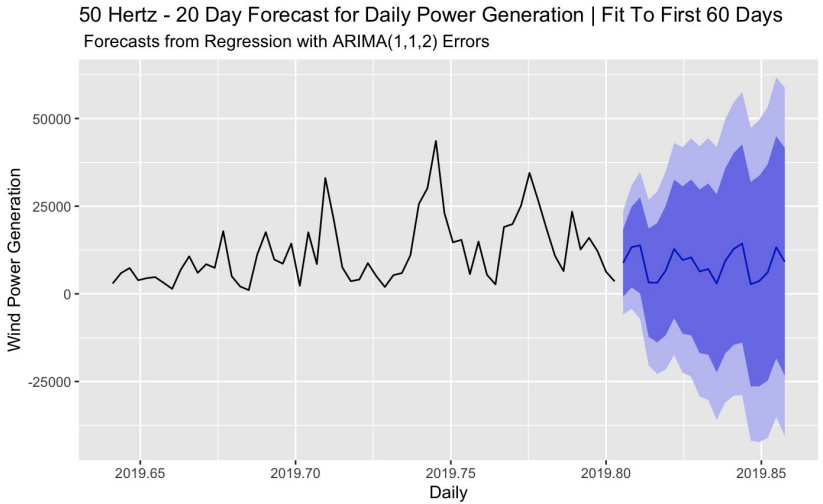




Dynamic Harmonic Regression - First 60 Days

Subset time series to just 60 days. The **plot shows the full daily time series with 20 days of forecast**. In the plots we see large forecast intervals.

AIC = 1231.183



Autoregressive Integrated Moving Average (ARIMA)

Daily

Models for Non-Stationary Time Series

Multi-Seasonal Time Series Object

```{r}
hertz_daily_msts_4 <- msts(Hertz_Daily_TS,
                           # 6 days and 4 days
                           seasonal.periods = c(365/60.83, 365/91.25),
                           ts.frequency = 365,
                           start = decimal_date(as.Date("2019-08-23")))
```

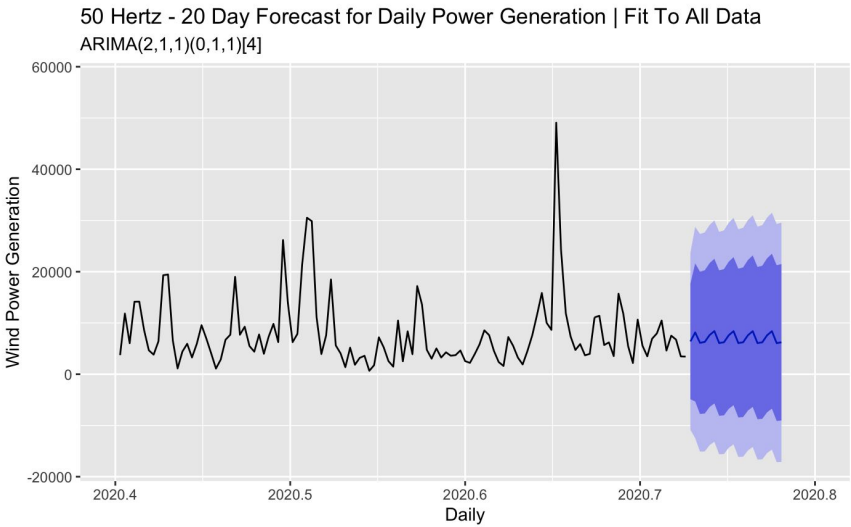
SARIMA - All Data

The dominant periods (or frequencies) of the 50 Hertz daily time series are 6 days and 4 days, with 4 days frequency being more dominant. Therefore, we wanted to try and capture this 4 day seasonal trend with a seasonal ARIMA.

Fit the data to an SARIMA model with the 4 day seasonal pattern and predict 20 days.

ARIMA(2,1,1)(0,1,1)[4]

- AIC = 8257.77
- Residuals resemble white noise and are independently distributed-> p-value =p-value = 0.9189 (> 0.05)
- RMSE = 8659.692
- Forecast looks to be capturing some seasonality and looks better than a basic ARIMA, but still looks like a naive forecast

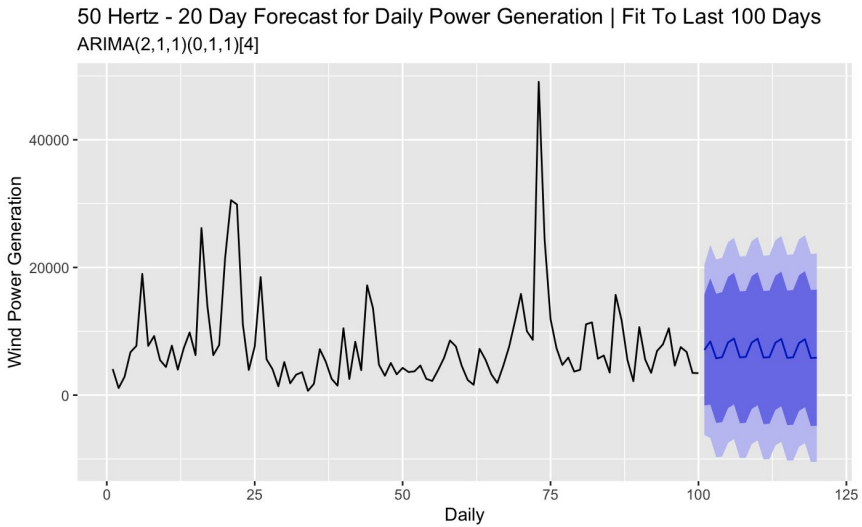


SARIMA - Last 100 Days

Fit the last 100 days of daily data to an SARIMA model with the 4 day seasonal pattern and predict 20 days.

ARIMA(2,1,1)(0,1,1)[4]

- AIC =1964.03
- Residuals resemble white noise and are independently distributed-> p-value =p-value = 0.3477(> 0.05)
- RMSE = 6398.356
- Forecast looks to be capturing some more seasonality than the first SARIMA and we see more volatility.

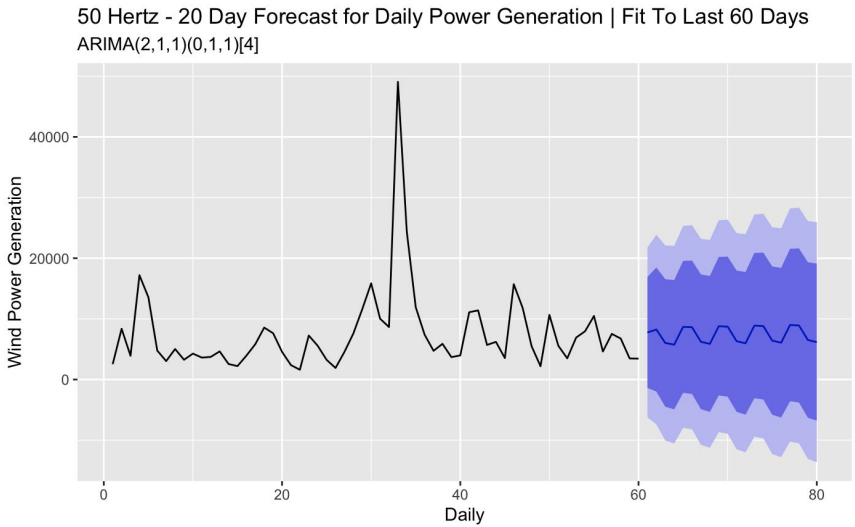


SARIMA - Last 60 Days

Fit the last 60 days of daily data to an SARIMA model with the 4 day seasonal pattern and predict 20 days.

ARIMA(2,1,1)(0,1,1)[4]

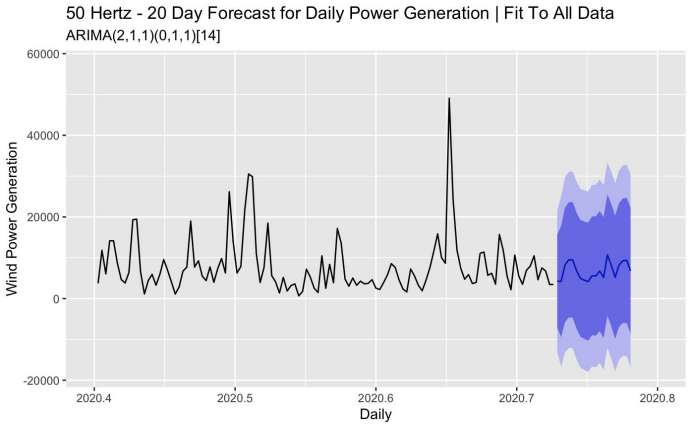
- AIC =1147.58
- Residuals resemble white noise and are independently distributed-> p-value =p-value = 0.9233(> 0.05)
- RMSE = 6334.997
- Forecast looks to be capturing some more seasonality than the first SARIMA and we see more volatility.



SARIMA - All Data (Other Seasonal Patterns?)

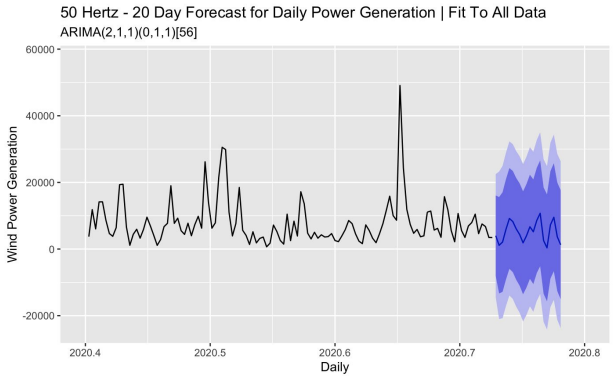
In reviewing the daily time series plots, it looked like there was a bi-weekly pattern every 14 days so i wanted to check out the period=14.

AIC=8070.54



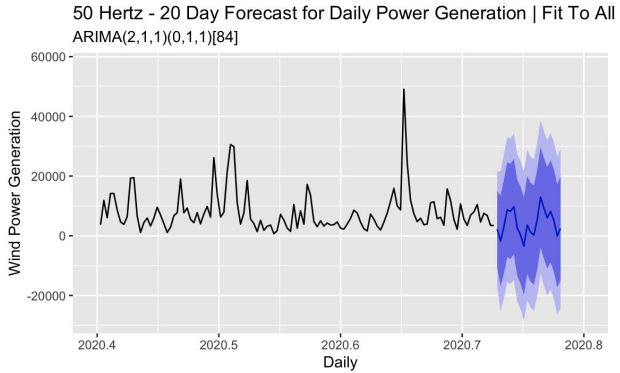
What does period=56 (every 4 weeks) look like? -> Better!

AIC=7259.052



What does period=84 (every 12 weeks) look like? -> Even better!

AIC=6694.482



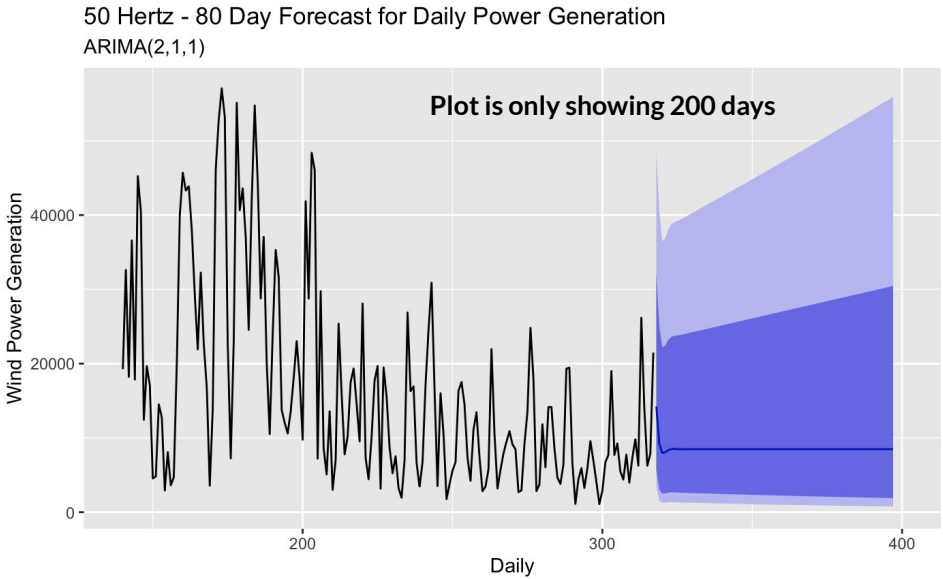
ARIMA - Bad Model for Reference

We split the daily data into test and train by 80/20 split. Then fit the data to an ARIMA model and predicted 80 days.

```
# ARIMA model - auto.arima() - basic model
daily_arima <- auto.arima(daily_train,
           d=1,
           seasonal=TRUE,
           stepwise = FALSE,
           approximation = FALSE,
           allowdrift = TRUE,
           lambda = "auto")
```

ARIMA(2,1,1)

- AIC= 1340.16
- Residuals resemble white noise and are independently distributed-> p-value = 0.3767. (>0.05)
- RMSE for training set is larger than test set (+1,961)
- Forecast is “OK” for first few days then levels off around 8487 (**naive forecast**)
- ARIMA is not a good forecast for this data



	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	2336.6536	9741.058	7133.879	-33.94375	73.06746	0.9112875	0.1378943
Test set	-975.0511	7780.009	6173.156	-181.50754	203.83281	0.7885640	NA

Conclusion and Future Work

Conclusion

Hourly Forecasts: the dynamic harmonic regression forecast (fit to the last 14 days worth of observations proved to be the best forecast model with AIC of 3356.16. Within TBATS, the forecast fit to the last 21 days worth of observations with 5 fourier terms and 0.48 box-cox was the best TBATS model (AIC=6713.61).

Daily Forecasts: The SARIMA model fit to the entire 50 Hertz time series dataset was the most promising and captured the seasonality the best - ARIMA(2,1,1)(0,1,1)[91].

Future Work

German Wind Power Generation

Future Analysis

- Further analysis on 15 minute or 30 minute, 4 hour, weekly intervals

Future Modeling Work

- Time series models for the other three German energy companies: Amprion, Tenne TTSO, TransnetBW
- Fine-tune 50 Hertz seasonal ARIMA model
- Introduce exogenous predictors (wind, temperature, other plants' series, etc)
- Time Series Train/Test and/or Cross-Validation

Thank you!

Appendix

Transformation and Differencing

Transformation | Is the Box-Cox transformation necessary for this data?

We should use a mathematical transformation if the data show variation that increases or decreases with the level of the series. It will attempt to balance the seasonal fluctuations and random variation across the series.

Differencing | Is our data stationary?

ADF - Augmented Dickey Fuller Test

- The null hypothesis = there is a unit root and the data is nonstationary.
- The alternate hypothesis differs slightly according to which equation you're using. The basic alternate is that the time series is stationary (or trend-stationary).

KPSS - Kwiatkowski-Phillips-Schmidt-Shin Test

- The null hypothesis = the data are stationary
- The alternate hypothesis = the data is not stationary.

Autocorrelation Function (ACF) Plot

The ACF can be used to identify the possible structure of time series data. It describes how well the present value of the series is related with its past values.

- The ACF tells us whether a substantial linear relation exists between the series and its own lagged values.
- The ACF gives a profile of the linear correlation at all possible lags and shows which values of $k$lead to the best predictability.

ACF Plot

- Trails off/decay = Autoregressive (AR) process
- Hard cut-off /drop= Moving average (MA) process

Partial Autocorrelation Function (PACF) Plot

The PACF Measure relationship between Yt and $Yt+k$, when the effects of other time lags (i.e., $1;2;3;\cdots;k-1$) are removed.

Instead of finding correlations of present with lags like ACF, it finds correlation of the residuals (which remains after removing the effects which are already explained by the earlier lag(s)) with the next lag value hence 'partial' and not 'complete' as we remove already found variations before we find the next correlation.

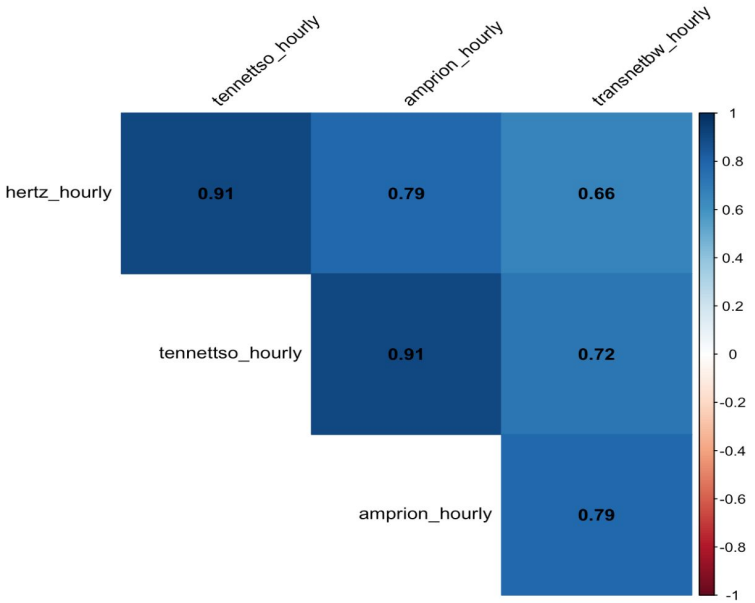
PACF Plot

- Trails off/decay = Moving average (MA) process
- Hard cut-off/drop = Autoregressive (AR) process

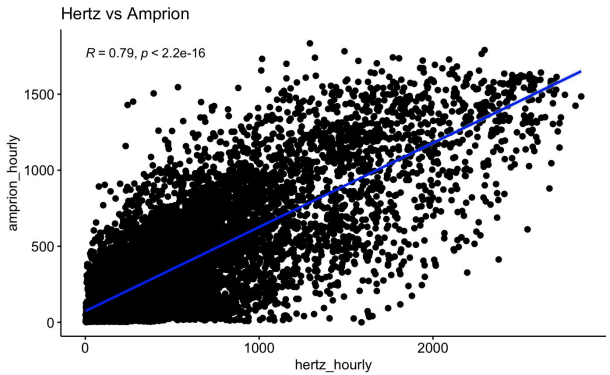
Correlation | Wind power generation from four German energy companies

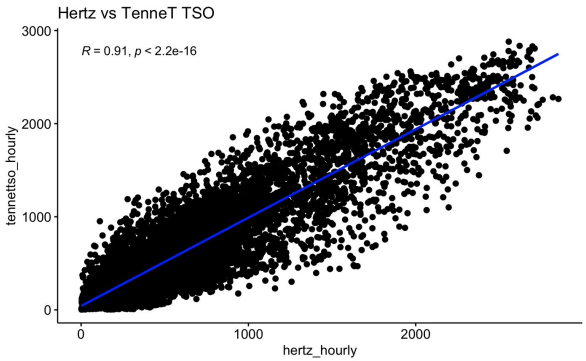
Hertz and TenneT TSO are very correlation with correlation coefficient = 0.9092
Amprion and TenneT TSO are very correlation with correlation coefficient = 0.9084

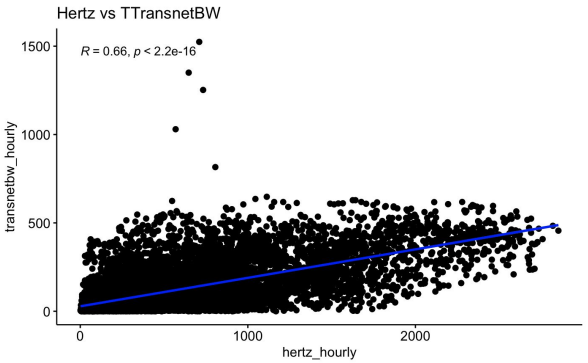
x (Company 1)	y (Company 2)	correlation coefficient	p-value
hertz hourly	amprion hourly	0.7856393	< 2.2e-16
hertz hourly	tennettso hourly	0.9092241	< 2.2e-16
hertz hourly	transnetbw hourly	0.6604917	< 2.2e-16
amprion hourly	tennettso hourly	0.9084243	< 2.2e-16
amprion hourly	transnetbw hourly	0.7898667	< 2.2e-16
tennettso hourly	transnetbw hourly	0.720573	< 2.2e-16

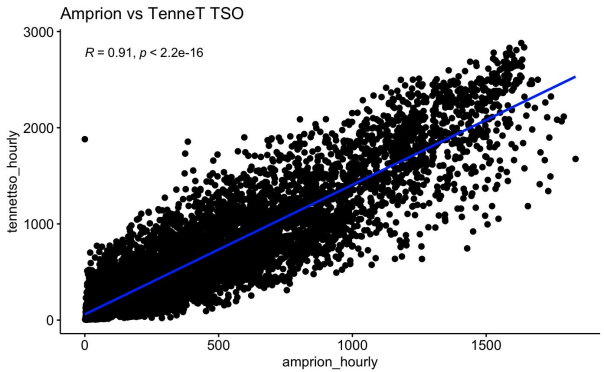


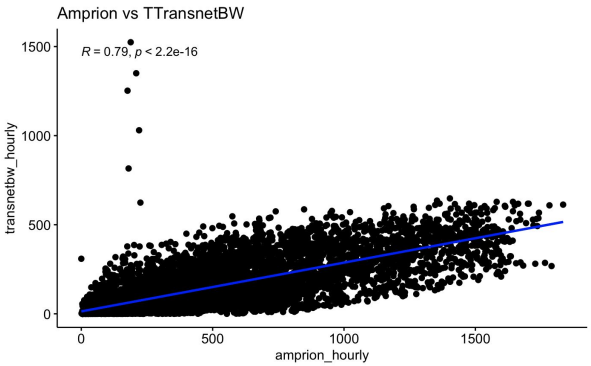
Correlation Scatter Plots

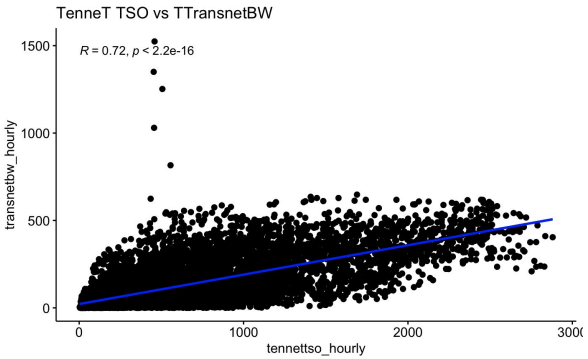












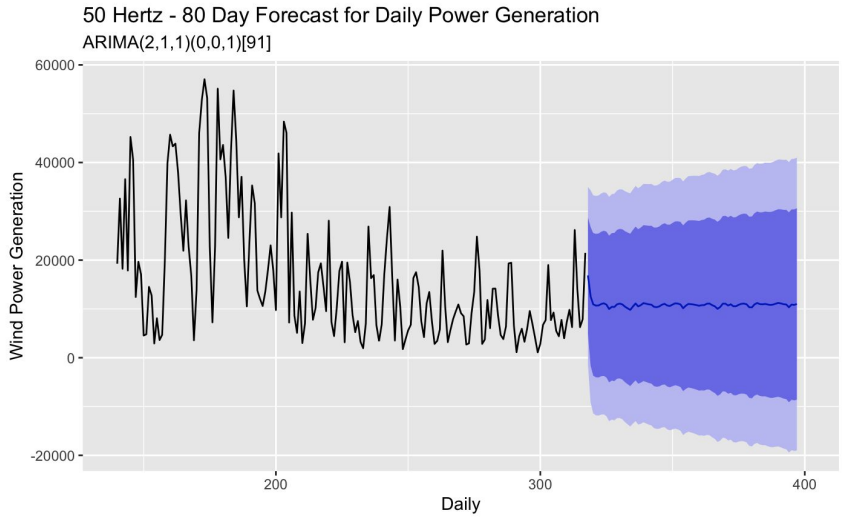
SARIMA - Train/Test

The dominant periods (or frequencies) of the 50 Hertz daily time series are 6 days and 4 days, with 4 days frequency being more dominant. Therefore, we wanted to try and capture this 4 day seasonal trend with a seasonal ARIMA.

We split the data into test and train. Then fit the data to an SARIMA model with the 4 day seasonal (91 per year) and predicted 80 days (amount of test data).

ARIMA(2,1,1)(0,0,1)[91]

- AIC = 6676.393
- Residuals resemble white noise and are independently distributed-> p-value =p-value = 0.477(> 0.05)
- RMSE for training set is much larger than the test set (+1025.126)
- Forecast looks to be capturing some seasonality and looks better than the ARIMA, but still looks like a naive forecast

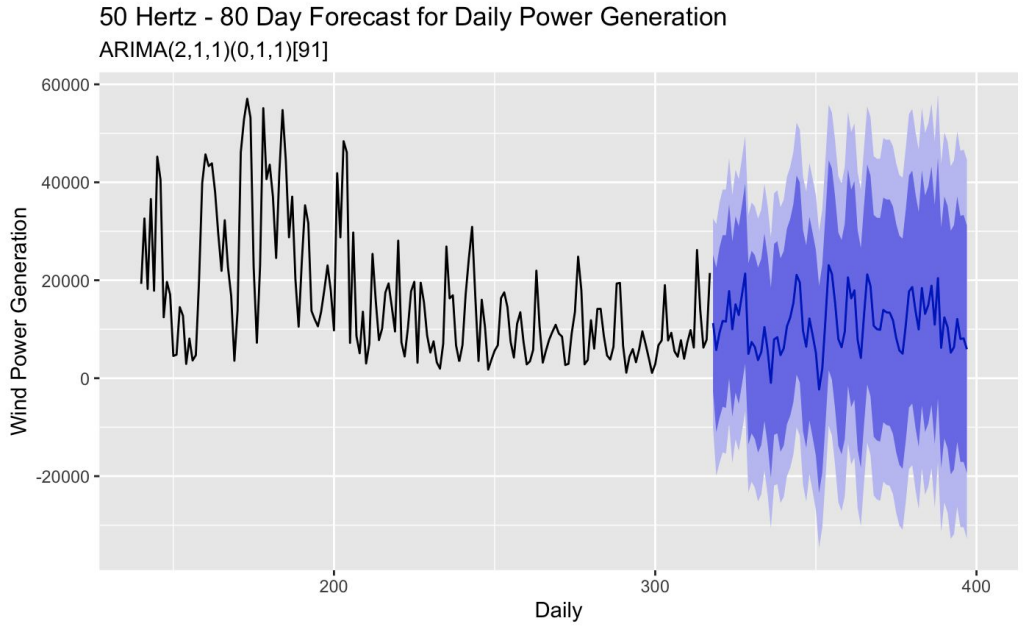


	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	149.0381	9183.979	7007.872	-55.30585	84.05031	0.8951913	-0.006576967
Test set	-3264.7942	8158.853	7123.754	-254.61290	269.05314	0.9099942	NA

SARIMA - Train/Test

ARIMA(2,1,1)(0,1,1)[91]

- AIC = 4876.356
- Residuals resemble white noise and are independently distributed-> p-value =p-value = 0.3613 (> 0.05)
- RMSE for training set is much larger than the test set (+1948.989), the difference is bigger than first SARIMA model
- Forecast looks better than first SARIMA model with D=0.
- Large forecast intervals



	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-421.0312	7877.411	5114.082	-32.51905	56.70588	0.653277	-0.005745144
Test set	-1937.8140	9826.430	7206.068	-184.16355	221.84010	0.920509	NA