Group Assignment - Big Data & Artificial Intelligence in Operations Management

Group 5:

Guillermo Brun

Sami Boustani

Natalia Clark

Sasha Glatt

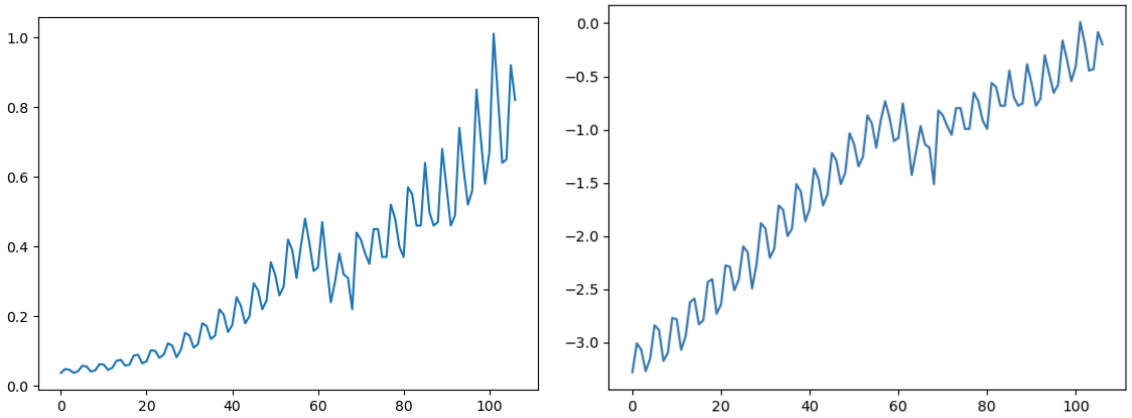
José Miguel Palencia

Date: 24/03/2024

<u>Series 1: Coca Cola Quarterly Earnings per Share from 1983-2009</u>

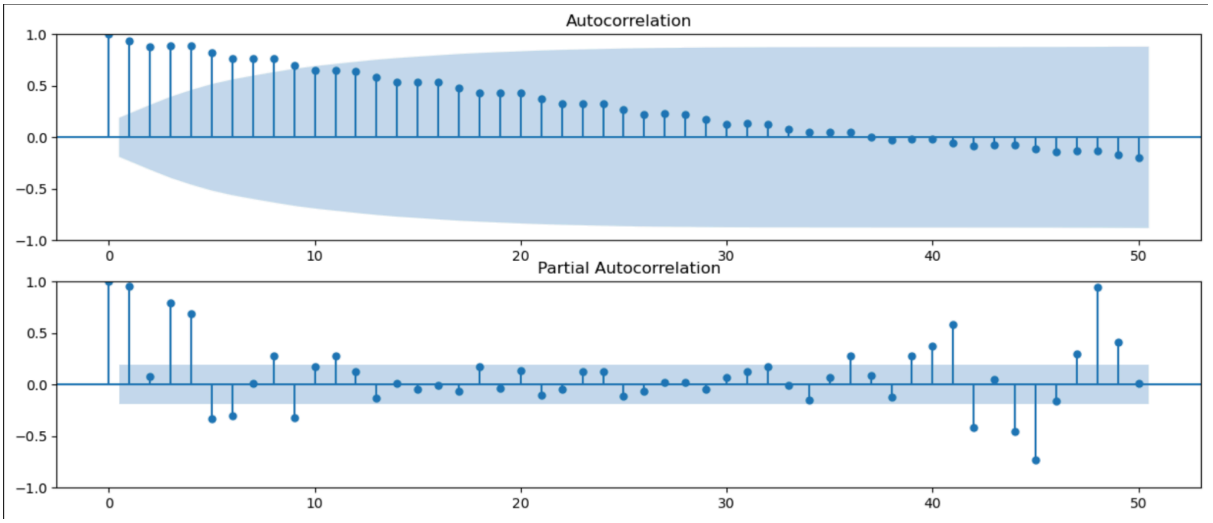
Transformations before the 2 models:

- Taking the log: (before and after taking the log):

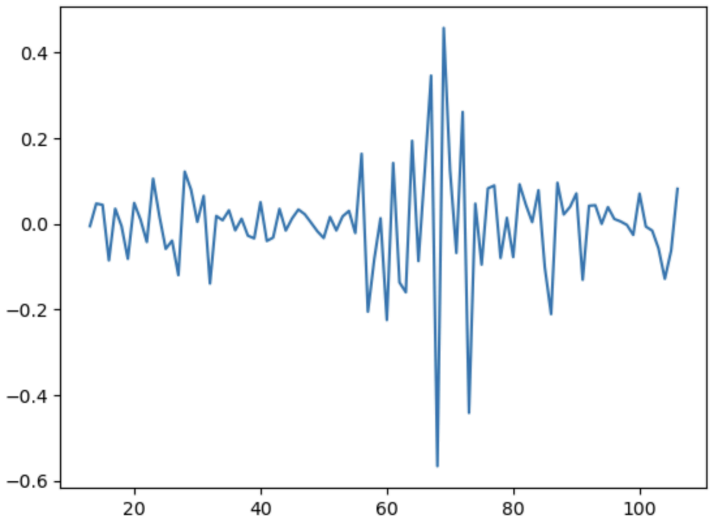


The first thing we noticed from this data is the non-stationarity in the variance (gradually increasing). From here, we knew to take the log to provide some stationarity in the variance to be able to analyze the time series from there (see above the transformed series from taking the log).

We could also see in the ACF a slow decline and the first bar of the PACF close to 1, meaning we would need at least one difference. We then use the ndiffs and nsdiffs functions as well as intuition (this data is not stationary) to tell us how many differences we need. We see the new data below after **1 regular difference and 1 seasonal difference**:



Series after taking differences:

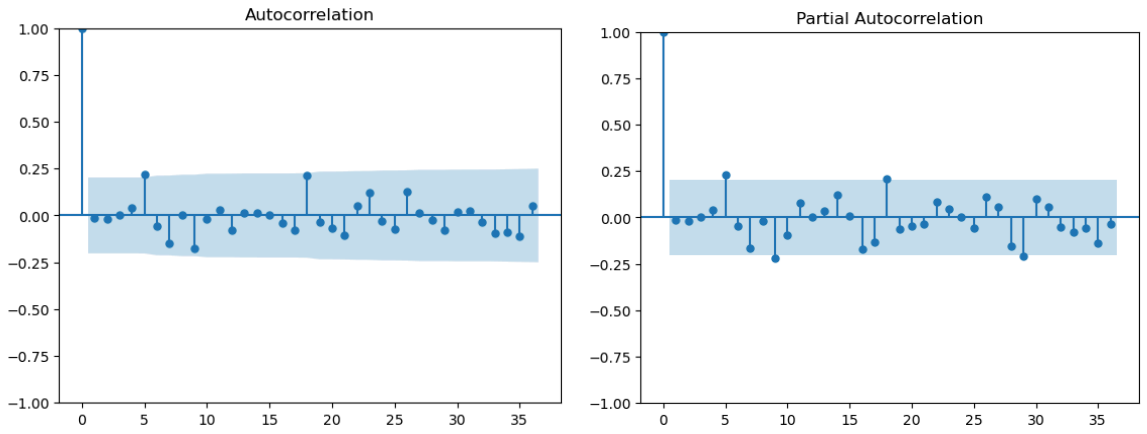


We will use the (p,d,q)x(P,D,Q)s=seasonal lags model to illustrate the models we have found for our data.

Answer 1: (0,1,1) x (0,1,1) s=4

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	−0.4096	0.058	−7.103	0.000	−0.523	−0.297
ma.S.L4	−0.8202	0.062	−13.221	0.000	−0.942	−0.699
sigma2	0.0072	0.001	12.597	0.000	0.006	0.008

As we see, all of the parameters are significant (different from 0) which makes this a good start for our model. We then needed to check the ACF and PACF to check if we think there is correlation in our residuals (along with the formal Box test).



Final analysis:

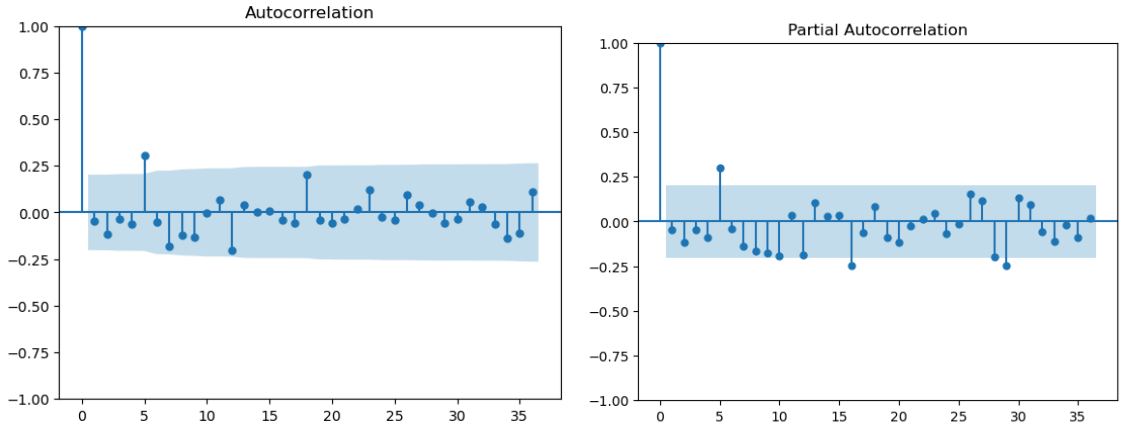
- *Stationary, 0 mean, uncorrelated, not Normal, not independent: just WN residuals*
- *We would need a non-linear model*

To us, this data looks pretty uncorrelated. The Box test confirms with a p-value of 0.5, meaning we could not reject the null hypothesis - this data is uncorrelated. This means our residuals are White Noise. We check the Shapiro test for the residuals to see if they are Guassian, and we reject the null hypothesis with a p-value of 0, meaning the residuals are not Gaussian. We check the Box test for the squared residuals, and it turns out that they are correlated (p-value of 0, reject the null hypothesis). This means that our residuals are not independent (confirmed by the Rank test), meaning that they are not Strict White Noise. It also means that we would want a non-linear model to model the residuals. As we were told to find a linear model, we stop our analysis here with the understanding that we would need the non-linear model for the residuals.

Answer 2: (1,1,0) x (2,1,0) s=4

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3371	0.048	-6.990	0.000	-0.432	-0.243
ar.S.L4	-0.5932	0.087	-6.854	0.000	-0.763	-0.424
ar.S.L8	-0.2260	0.124	-1.815	0.069	-0.470	0.018
sigma2	0.0086	0.001	11.198	0.000	0.007	0.010

As we see, all of the parameters are significant (different from 0) which makes this a good start for our model. We then needed to check the ACF and PACF to check if we think there is correlation in our residuals (along with the formal Box test).



Final analysis:

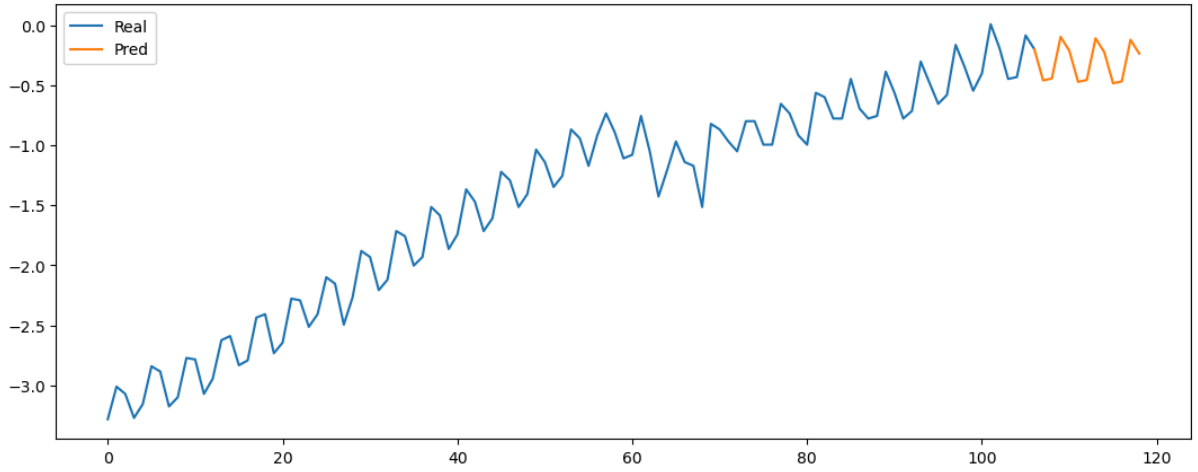
- *Stationary, 0 mean, uncorrelated, not Normal, not independent: just WN residuals*
- *We would need a non-linear model*

As we see from the PACF and ACF, the data looks uncorrelated (except for the one lag 5, but we find this lag to be outside of the interval no matter what model we propose). The Box test returns a p-value of 0.15, meaning we fail to reject the null hypothesis and the data is uncorrelated. We have White Noise residuals.
We then check if we have Gaussian White Noise from the Shapiro test, which returns a p-value of 0, meaning we reject the null hypothesis. We do not have GWN.
We check the Box test for the squared residuals to check if we need a linear model, and we find that the squared residuals are correlated. We check the Box test for the squared residuals, and it turns out that they are correlated (p-value of 0, reject the null hypothesis).

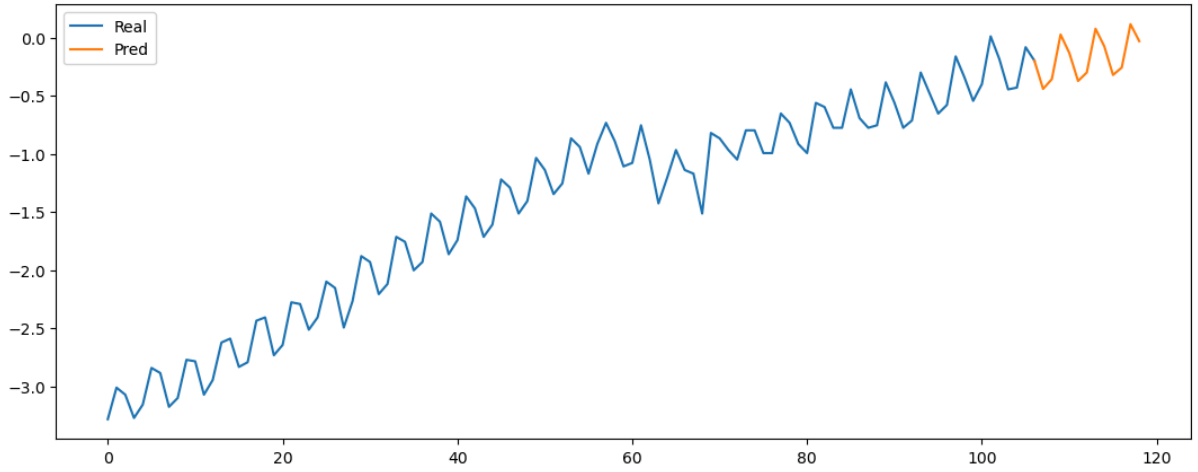
This means that our residuals are not independent (confirmed by the Rank test), meaning that they are not Strict White Noise. It also means that we would want a non-linear model to model the residuals. As we were told to find a linear model, we stop our analysis here with the understanding that we would need the non-linear model for the residuals.

Coca Cola predictions:

First predictions (0,1,1) x (0,1,1) s=4:



Second predictions (1,1,0) x (2,1,0) s=4:



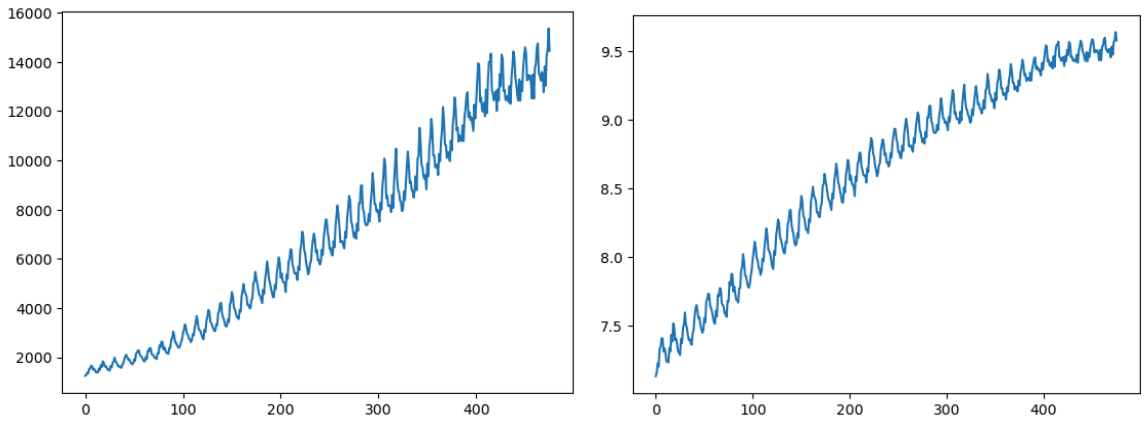
We see slight differences in the outputs of the time series models for this data. The first model, while slightly simpler, shows that it will give a more horizontal output for the following values. As we see the upward trend, we would expect the predictions to look more like the second model.

Therefore, though slightly more complex, we suggest the second model as the better model because it seems its predictions are more accurate for our data and time series information.

Series 2: Monthly Australian Electricity Demand from 1956-1995

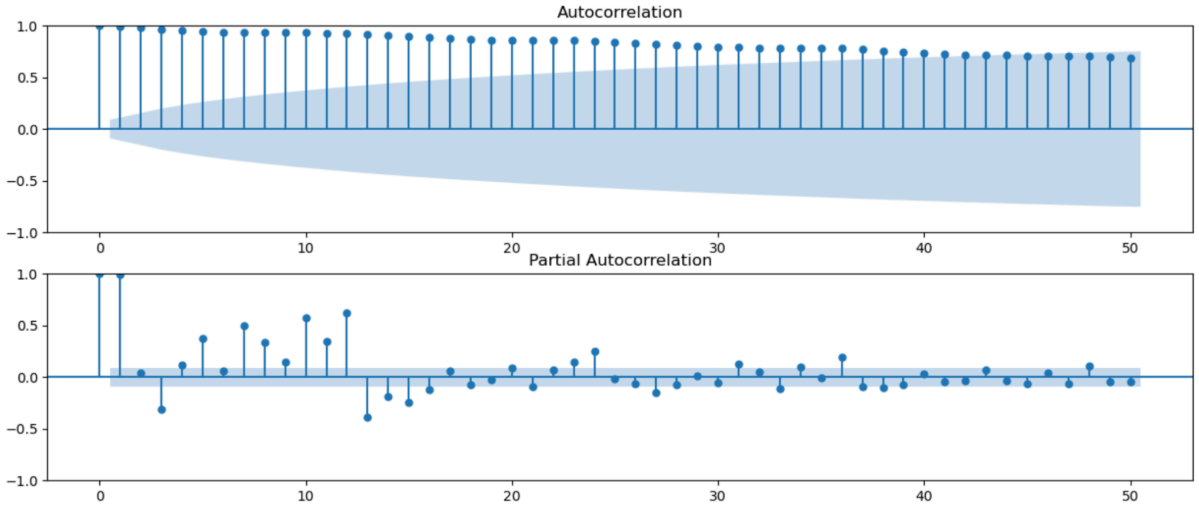
Transformations before the 1st model:

- Taking the log: (before and after taking the log):

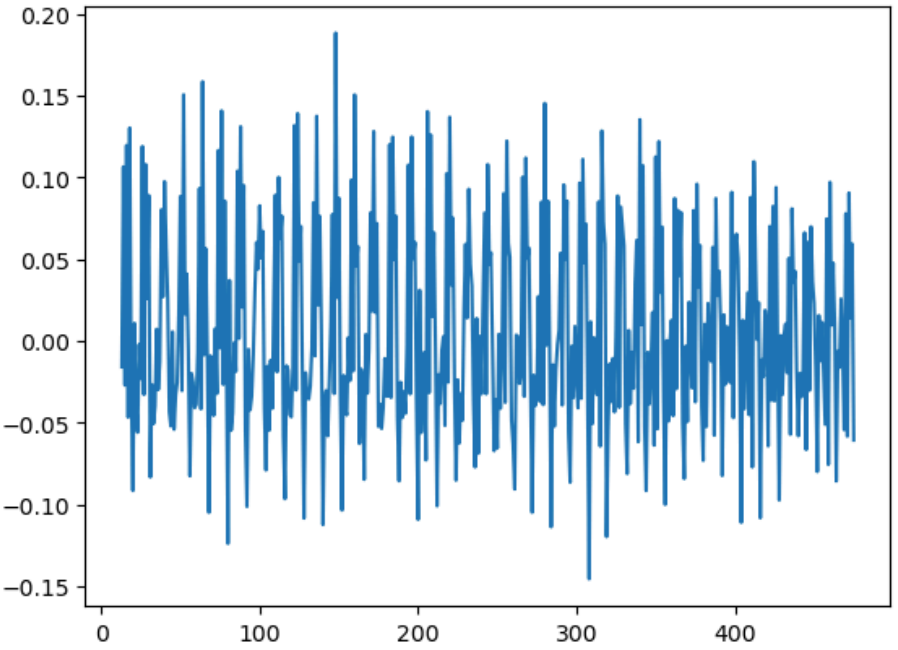


The first thing we noticed from this data is the non-stationarity in the variance (gradually increasing). From here, we knew to take the log to provide some stationarity in the variance to be able to analyze the time series from there (see above the transformed series from taking the log).

We could also see in the ACF a slow decline and the first bar of the PACF close to 1, meaning we would need at least one difference. We then use the ndiffs and nsdiffs functions as well as intuition (this data is not stationary) to tell us how many differences we need. We see the new data below after **1 regular difference and 0 seasonal differences**:



Here is the series after taking the one regular difference:

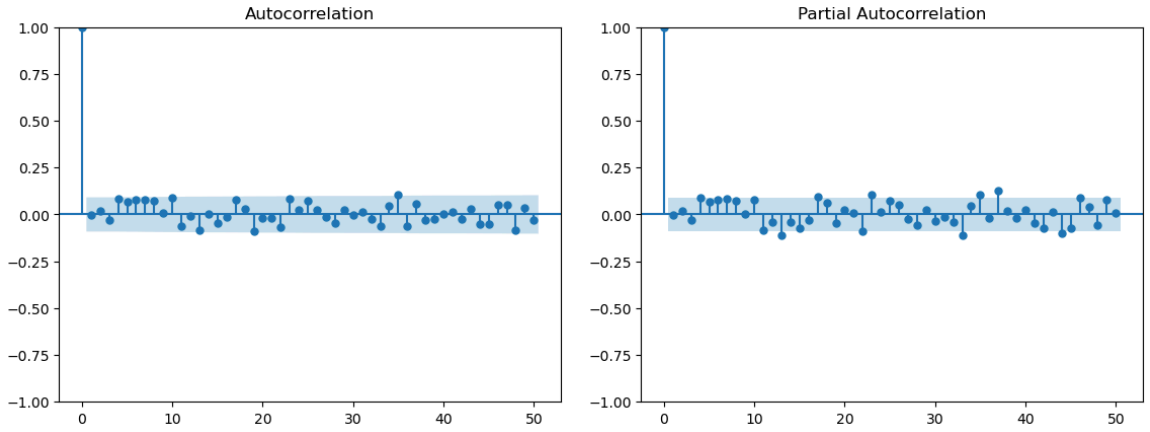


We will use the (p,d,q)x(P,D,Q)s=seasonal lags model to illustrate the models we have found for our data.

Answer 1: (10,1,0) x (3,0,1) s=12

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	−0.6362	0.050	−12.714	0.000	−0.734	−0.538
ar.L2	−0.3929	0.058	−6.812	0.000	−0.506	−0.280
ar.L3	−0.2092	0.063	−3.335	0.001	−0.332	−0.086
ar.L4	−0.2414	0.061	−3.990	0.000	−0.360	−0.123
ar.L5	−0.2964	0.055	−5.344	0.000	−0.405	−0.188
ar.L6	−0.3617	0.052	−6.984	0.000	−0.463	−0.260
ar.L7	−0.4100	0.049	−8.360	0.000	−0.506	−0.314
ar.L8	−0.3714	0.052	−7.078	0.000	−0.474	−0.269
ar.L9	−0.2384	0.058	−4.130	0.000	−0.352	−0.125
ar.L10	−0.2400	0.046	−5.241	0.000	−0.330	−0.150
ar.S.L12	0.8707	0.080	10.869	0.000	0.714	1.028
ar.S.L24	−0.1664	0.071	−2.340	0.019	−0.306	−0.027
ar.S.L36	0.2598	0.063	4.125	0.000	0.136	0.383
ma.S.L12	−0.5854	0.080	−7.334	0.000	−0.742	−0.429
sigma2	0.0005	2.89e−05	16.023	0.000	0.000	0.001

As we see, all of the parameters are significant (different from 0) which makes this a good start for our model. We then needed to check the ACF and PACF to check if we think there is correlation in our residuals (along with the formal Box test).



Final analysis:

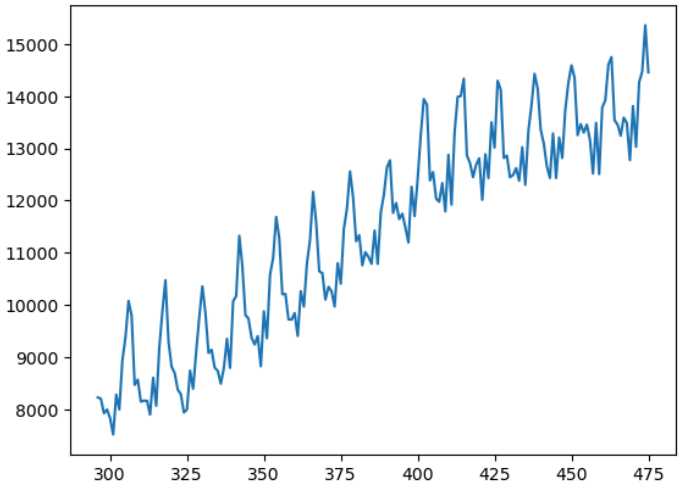
- *Stationary, 0 mean, uncorrelated, Normal, independent: WN, SWN, GWN*
- *We would NOT need a non-linear model*

We looked at the ACF and the PACF and found that these looked pretty uncorrelated, so we felt it could be a good model for us. We checked the Box test, and the p-value is around 0.03. This means that we would technically reject the null hypothesis, making the residuals correlated.

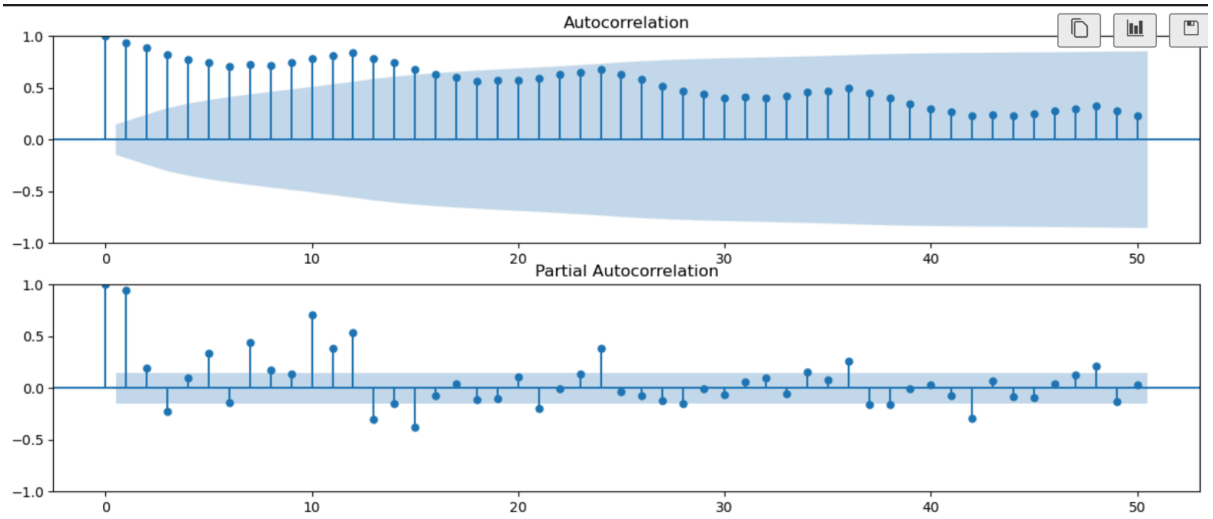
However, we have chosen to use our intuition with this model. The residuals seem very uncorrelated, so we will assume they are white noise. We then check the Shapiro test and see a p-value of 0.07, meaning that we fail to reject the null hypothesis, meaning we have Gaussian White Noise residuals. We then check for independence in the squared residuals, and we see they are uncorrelated in their ACF/PACF. Confirmed by the Rank test, the squared residuals are independent (meaning Strict White Noise). This means that we have found a model with WN, GWN, and SWN. Since the squared residuals are uncorrelated, this is our final model. We would not need a non-linear model for the variance.

Answer 2: (2,1,0) x (2,1,0) s=12

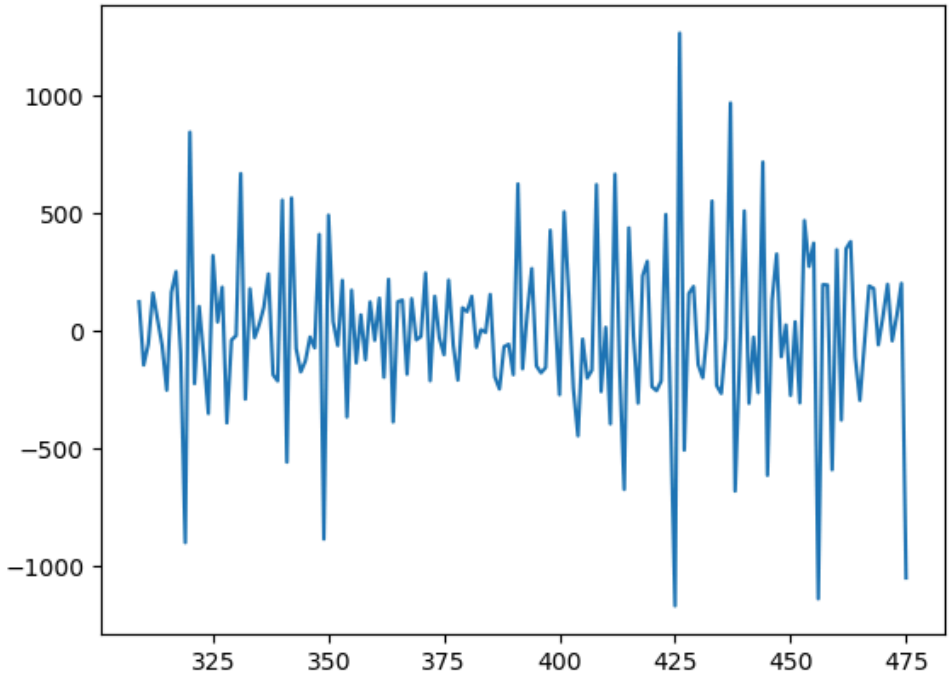
For this second model, we consider only the last 15 years of the data. The original data looks like this:



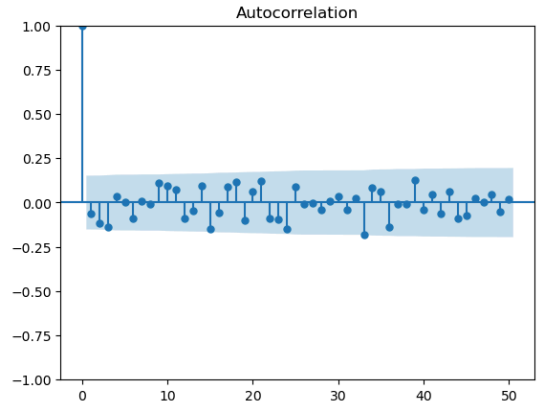
We see that we do not need the log transformation in this case, maybe just a difference. We decide to take 1 seasonal difference and 1 non-seasonal difference, as per the formal tests and also by the gradual decay of the ACF/first bar of the PACF being close to 1. We take these differences and this is our data:

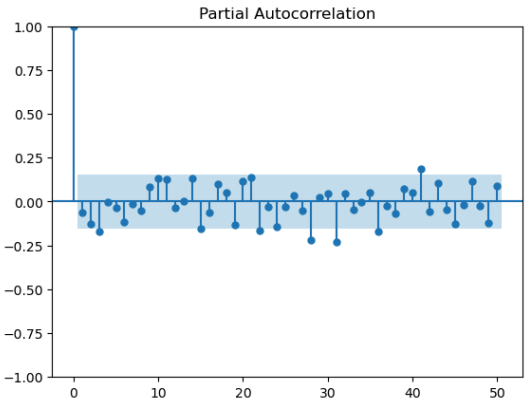


New Data:



From here, we play with the non-seasonal and seasonal parameters based on the new ACF and PACF.





	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	−0.6475	0.081	−8.008	0.000	−0.806	−0.489
ar.L2	−0.3180	0.081	−3.925	0.000	−0.477	−0.159
ar.S.L12	−0.6835	0.074	−9.194	0.000	−0.829	−0.538
ar.S.L24	−0.4148	0.086	−4.833	0.000	−0.583	−0.247
sigma2	5.343e+04	6058.358	8.819	0.000	4.16e+04	6.53e+04

Final analysis:

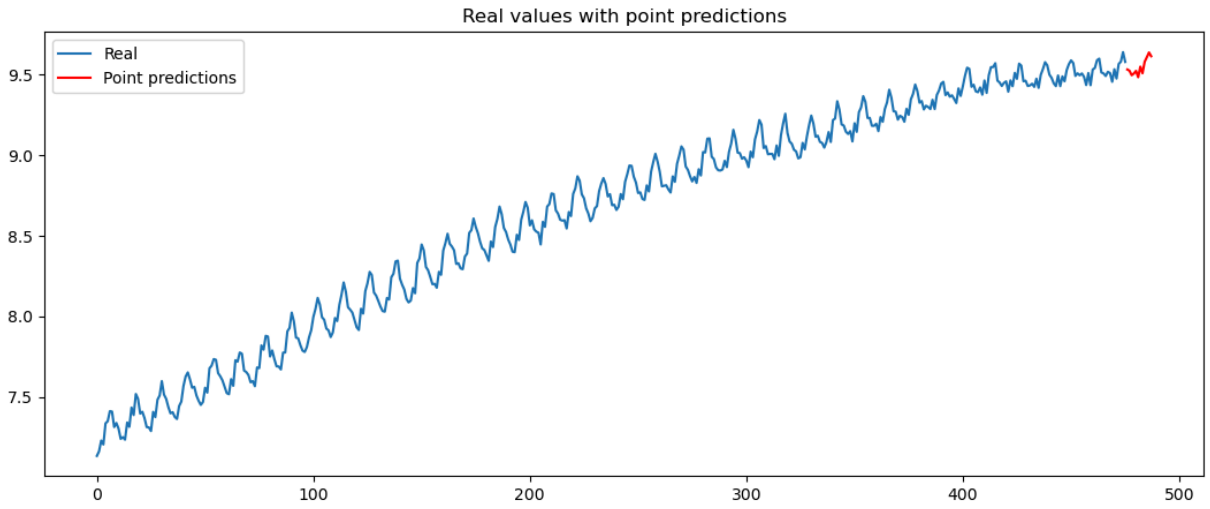
- *Stationary, 0 mean, uncorrelated, not Normal, independent: WN, SWN*
- *We would NOT need a non-linear model*

As we see, all of the parameters are significant (different from 0) which makes this a good start for our model. We then needed to check the ACF and PACF to check if we think there is correlation in our residuals (along with the formal Box test).

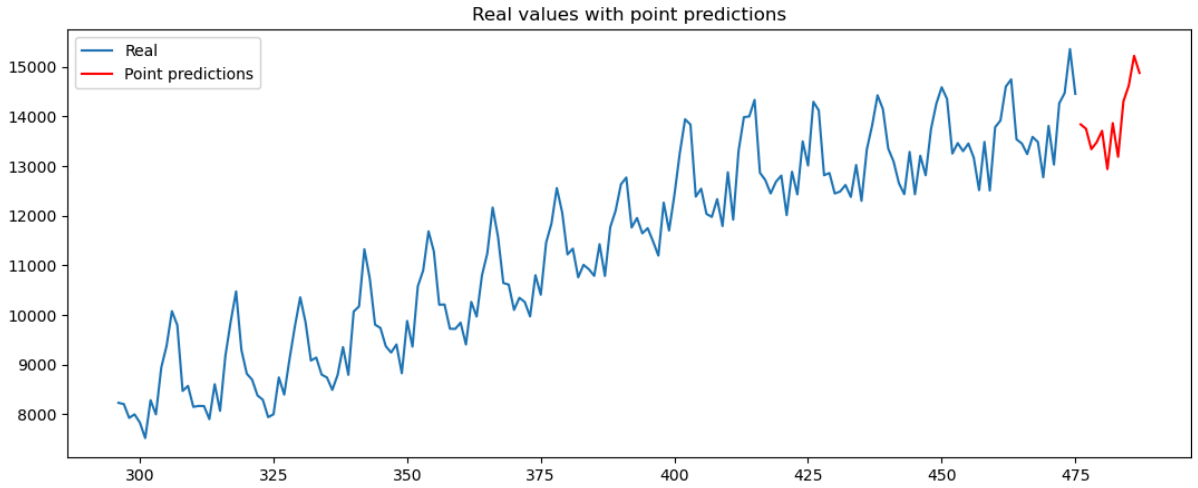
Our ACF and PACF seemed pretty uncorrelated, and our Box test returned values of 0.04. We can see in the PACF that there is maybe some correlation, but it starts at lag 30. We recognize that we do not need to complicate our model this much, and we choose to continue with this model. We reject the Shapiro test, meaning our residuals are not Gaussian. We fail to reject the Box for the squared residuals, which means that the squared residuals are uncorrelated. This implies independence, verified by the rank test. This means we have WN and Strict White Noise, but no Gaussian White Noise in this case. We will not need a non-linear model to model the variance, so we are done with this model.

Australian Electricity predictions:

Model 1 (10,1,0) x (3,0,1) s=12:



Model 2 (2,1,0) x (2,1,0) s=12



We see slight differences in the outputs of the time series models for this data. The first model, while much more complex, is using all of the data to make this prediction, which is what the assignment detailed. THe second model only uses the last 15 years, meaning we are studying the seasonality only on the last 15 years instead of all of the data. However, the model 2 is noticeably simpler. It makes sense, in this case, to just study the last 15 years and it provides a proper analysis for the seasonality of the last 15 years going forward, which should be sufficient.

Therefore, though only considering the last 15 years, we suggest the second model as the better model because it seems its predictions are more accurate for our data and time series information given the considerations of time.