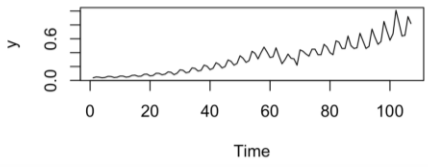
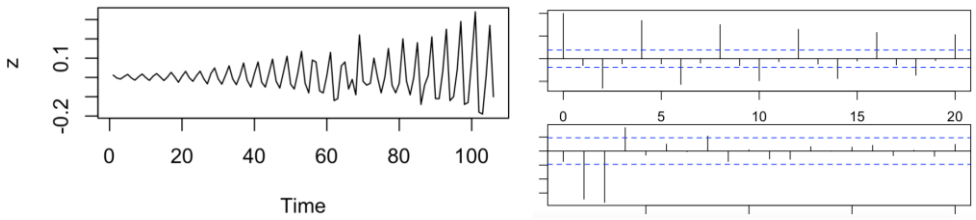
FORECASTING TIME SERIES

Assignment

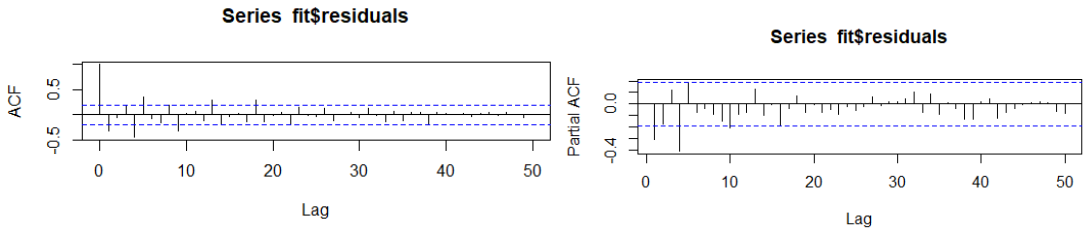
GROUP B PROJECT



By observing the plot, we can see that the data is not stationary in the mean or variance (due to the noticeable upward trend, since it is increasing systematically over time). Moreover, we can observe in Annex 1, the lags are decaying to 0. Finally, the ADF test was used for confirmation (with a 95% confidence level), and we calculated the differences needed as seen in Annex 2. After obtaining the results, we will apply one regular and one seasonal difference.



After the transformations, our data is visually stationary in the mean and variance, which suggest that we can use the data on a model. Now that our data is stationary, we are going to analyze the residuals of our model arima ((0,1,0) (0,1,0)4).



Finally, As seen in Annex 3 and 4, the residuals are stationary, suggesting that we can start improving the model to predict

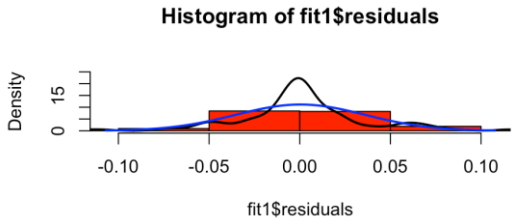
MODEL 1

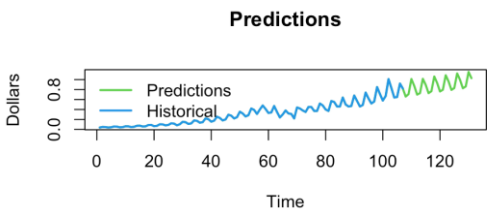
Arima (0,1,0) (1,1,0)4

The variables are relevant since we get a result above 1.96.

```
Coefficients:
         sar1
      -0.4842
s.e.   0.0899
```

- As seen in the residual PACF, we notice the seasonal lag 1 out of bounds, therefore we decided to use SAR (1) for our model.
- The Box Test shows a p-value higher than 0.05, meaning the data is White Noise.
- To analyze the existence of Strict White Noise, we input the squared Box Test which shows a p-value lower than 0.05, meaning the data is correlated so there is no Strict White Noise.
- The Schapiro test indicates that the p-value is less than 0.05 so the data is not normally distributed. For instance, the Shapiro test, together with the histogram confirm there is no Gaussian white noise.





Since our residuals are WN, we can use our model for prediction, as seen in the graph above.

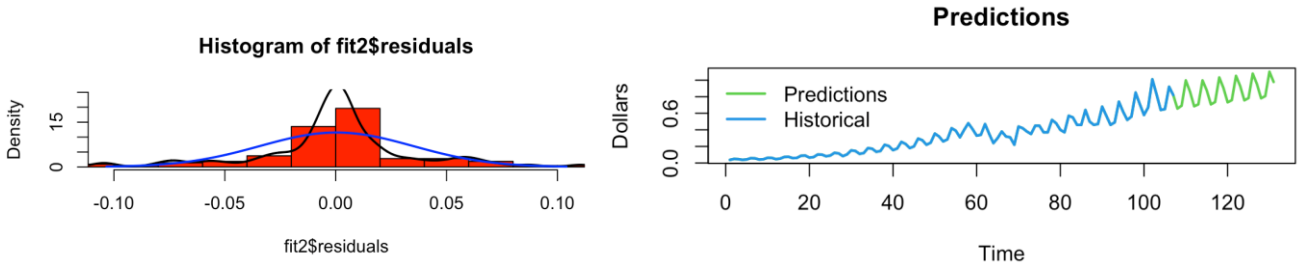
MODEL 2

Arima (2,1,0) (1,1,0)4

The variables are relevant since we get a result above 1.96.

```
Coefficients:
          ar1      ar2     sar1
      -0.1912  -0.2093  -0.4702
s.e.   0.1030   0.0996   0.1003
```

- As seen in the residual PACF, we notice the regular lag 2 and seasonal lag 1 out of bounds, therefore we decided to use AR (2) and SAR (1) for our model
- The Box Test shows a p-value higher than 0.05, meaning the data is White Noise.
- To analyze the existence of Strict White Noise, we input the squared Box Test which shows a p-value lower than 0.05, meaning the data is correlated so there is no Strict White Noise.
- The Schapiro test indicates that the p-value is less than 0.05 so the data is not normally distributed. For instance, the Shapiro test, together with the histogram confirm there is no Gaussian white noise.



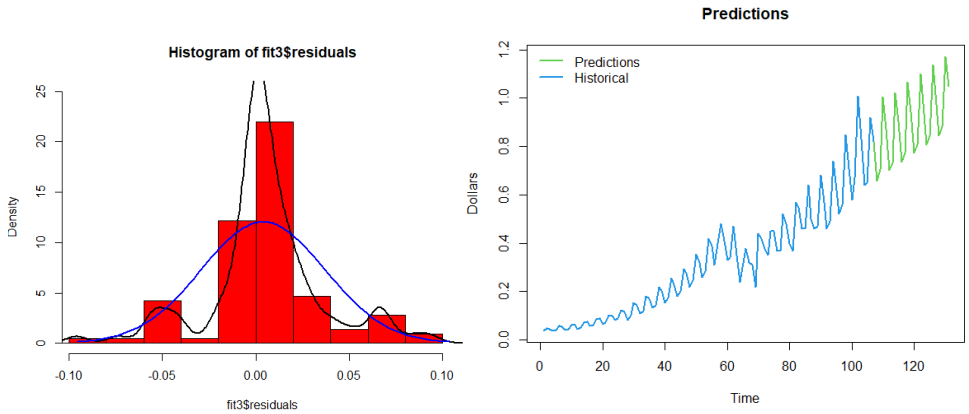
Since our residuals are WN, we can use our model for prediction, as seen in the graph above.

MODEL 3

Arima (1,1,1) (1,1,0)4

The variables are relevant since we get a result above 1.96.

In order to fully understand how the arima model worked, we decided to investigate more about the concept. As a result, we came across the auto_arima function in python, as seen in Annex 5, it calculates the Akaike information criterion (AIC) of all the possible models and chooses the one with the best score. The AIC refers to an estimator of out-of-sample prediction error



As seen in the model summary in Annex X, the variables are relevant since we get a result above 1.96. Furthermore, both Box-tests in Annex X and x, the data is White noise since the p-value is over 0.05 and suggest that there isn't Strict White Noise since the p-value is under 0.05, however we cannot be certain whether it´s SWN unless we know it's GWN. Moreover, the histogram and the Schapiro test suggest that there isn't Gaussian White Noise. Since our residuals are WN, we can use our model for prediction, as seen in the graph above.

RESULTS & CONCLUSION

To analyze which model is the most optimum, we have decided

- Use the Mean squared error, Root Mean Square Error, Mean Absolute Error and Mean Absolute percentage error metrics to compare the models.
- Remove the last 24 real values to compare all the models in terms of forecasting

Models	MSE	RSME	MAE	MAPE
1	0.00543044	0.07369152	0.05816588	9.725125
2	0.004472508	0.06687681	0.05356396	8.70509
3	0.004686575	0.06845857	0.05476783	8.962264

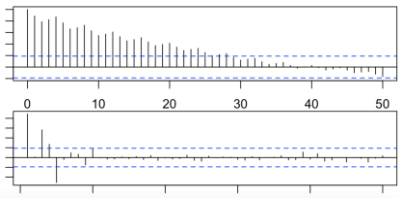
Best	
Average	
Worst	

In conclusion, as seen in the table by comparing the models with the metrics, model 2 has the best scores overall. Therefore, we conclude that the model sarima (2, 1, 2) (1, 1, 0)4 is our best identified model.

However, there are various ways of calculating the accuracy of the model, such as the auto_arima function in python that uses the AIC metric, where our model 3 had the best score. Nevertheless, since the metrics used in the table above were the ones studied in our program, we decided to choose model 2 as our best model.

ANNEX

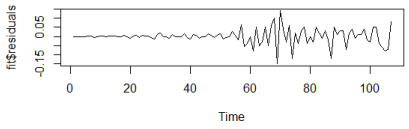
Annex 1



Annex 2

```
> s=4       # seasonal parameter FOR THIS DATA SET
> ndiffs(y, alpha=0.05, test=c("adf")) # regular differences?
[1] 1
>
> nsdiffs(y,m=s,test=c("ocsb"))  # seasonal differences?
[1] 1
```

Annex 3



Annex 4

```
> ndiffs(fit$residuals, alpha=0.05, test=c("adf")) # regular differences?
[1] 0
```

Annex 5