Mini Research Project: Analyzing and Forecasting Air Quality

COHNDDS242F-010 –Time series analysis and forecasting

CW1

INTRODUCTION

OBJECTIVE - Analyze historical CO air quality data in Los Angeles,CA. and forecast future CO levels using univariate time series analysis.

RESEARCH QUESTION – 'How has CO air quality changed over the last 10 years (2014-2024) in Los Angeles, CA. and what are the forecasted CO levels for next 6 months?

LITERATURE REVIEW - In the atmosphere, CO gas plays major role as a "air pollution gas " , it reacts with (OH) radicals and it effecting to concentration of greenhouse gases like Ozone(O3)and Methane(CH4). Monitoring CO levels and forecasting its future may essentially helps to control air pollution and protect both human and environmental health.

Time series models like ARIMA , SARIMA ,Exponential smoothing was used by previous research due to the ability of capturing the trends and seasonal variations. It may helps to forecast and get better forecast of future air .

Previous researches which based on CO levels in urban areas represent that seasonality and meteoritical factors significantly influence pollutant levels. Research which based on CO, Los Angeles found notable decline in CO level due to emission control in vehicles and industry regulations. Some researches reported a significant sudden drop in CO levels in 2020-2021 period due to reduce of transportation and industry works during covid lockdown season.

RESEARCH GAPS –

Most of the previous research are focused on short term predictions.(during covid season research only two year include).

Many of research based on multivariate models (using multiple variables for forecast air quality like CO,NO,CH4….)

Previous researches which did during covid pendemic post 2021 trends remain as underexplored(sudden drops)

Data Collection & Pre-processing

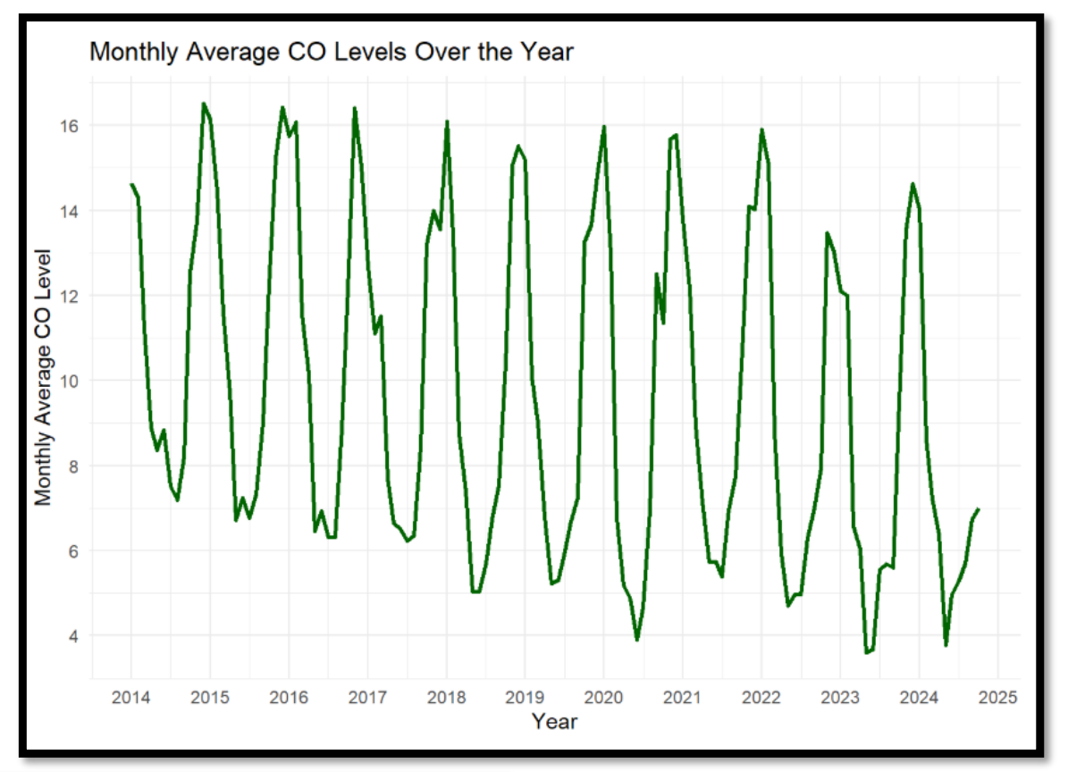
DATA COLLECTION - this dataset contain daily CO level of past ten years and 8 months (2014.01.01 -2024.10.01) in Los Angeles city in CA.(secondary dataset)

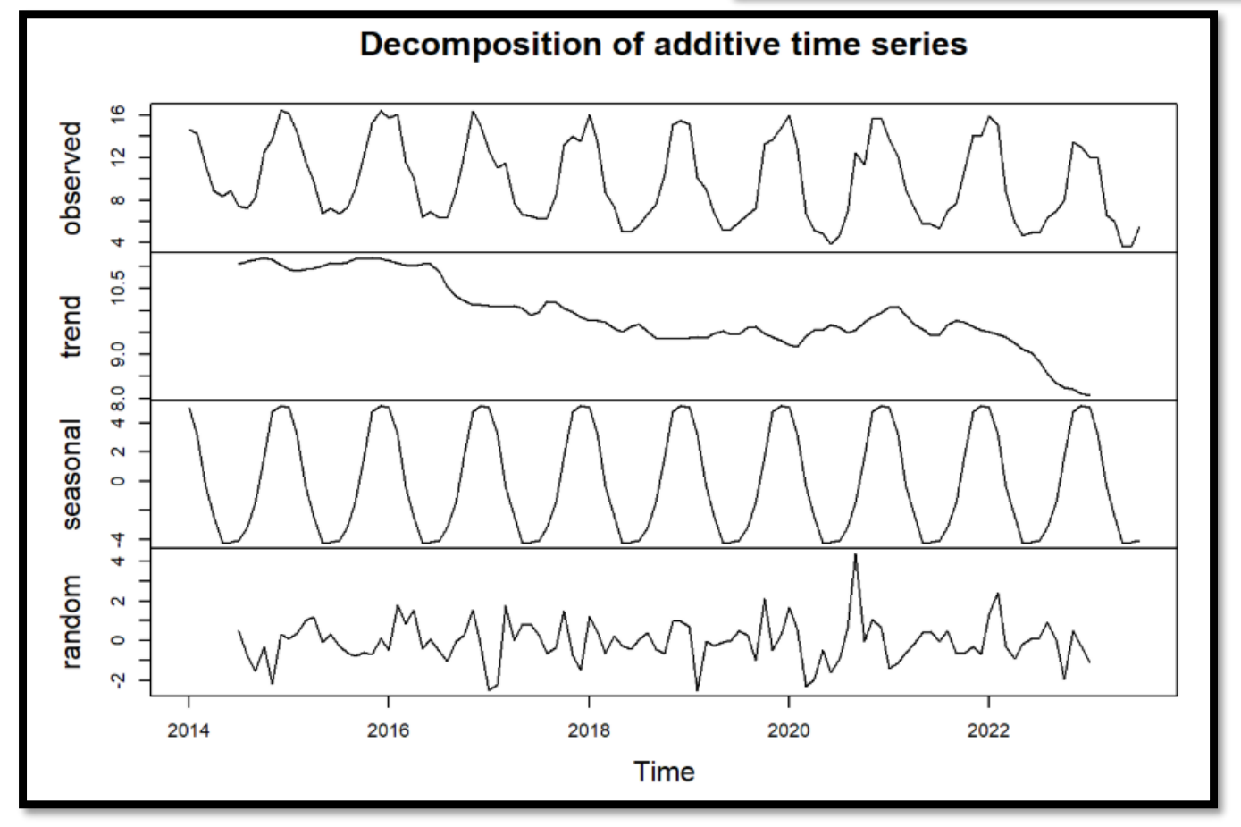
PRE- PROCESSING –

Checking missing values ,Check for the outliers and treat them by using IQR smoothing technique ,Aggregate daily data into monthly level and split the data set into train data and test data, total no. of months is 130 and split it into 115 months for train data and 15 test data.

(EDA) , Stationarity and Transformation

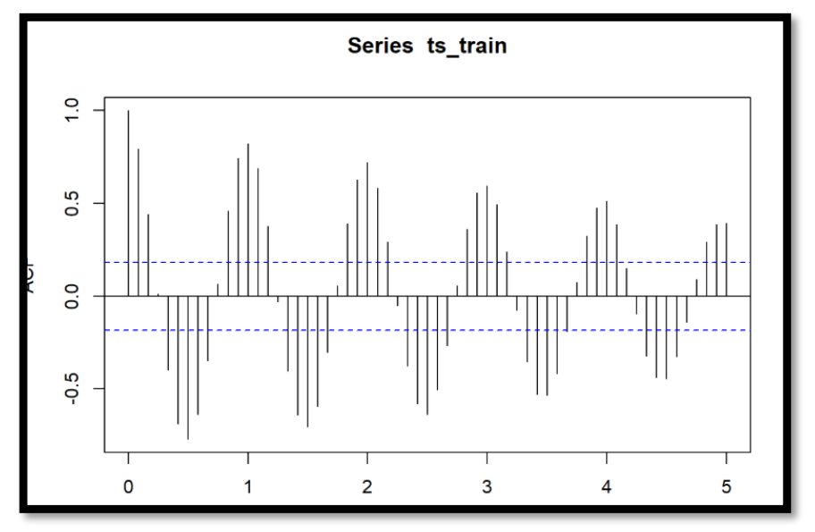
This plot shows the monthly average CO levels change over the past 10 years(2014-2024) over the months.

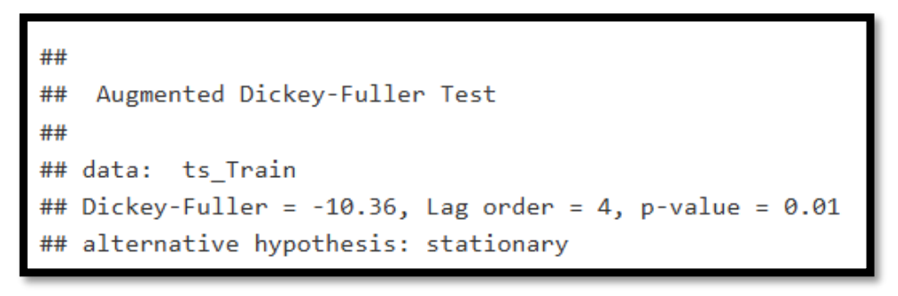




Convert train dataset into timeseries and decompose the time series using classical decomposition. In this decomposition time displayed as 2 by 2 yrs and the decomposition plot shows a downward trend and a seasonality over the time.

When applying ACF plot to check the auto correlation we can see , ACF shows strong spikes at fixed lags it indicates ,the seasonality continue over the time period we study.

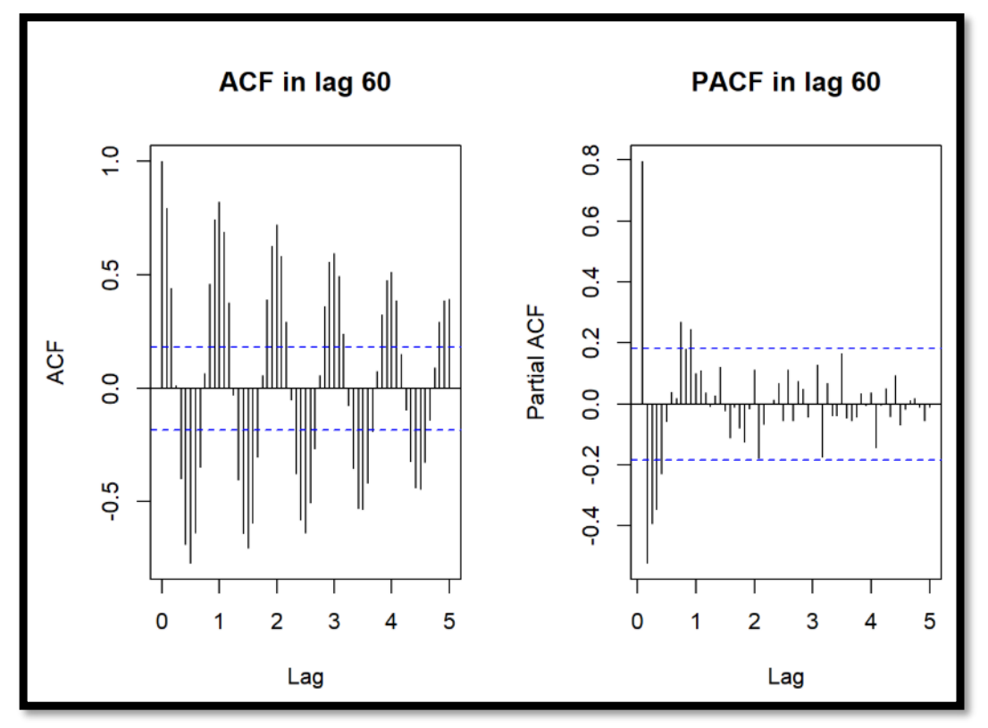




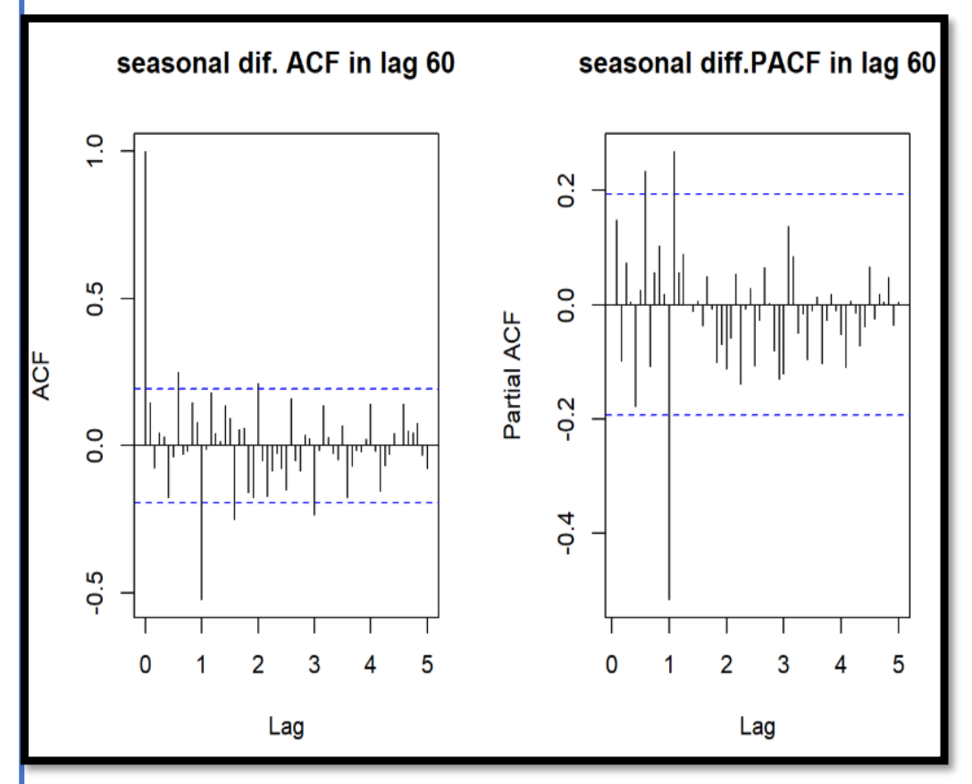
```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_Train
## Dickey-Fuller = -10.36, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

When comes to ACF and PACF plots we can see there is a seasonality over the time , as I described in ACF plot., we can check our seasonality is a real seasonality or noice by adding more lags to our ACF plot, when comes to that ACF plot we can see our seasonality is continued over the all lags we applied . So , when apply modeling for forecast future we can apply seasonal ARIMA (SARIMA) model and ETS.

When applying ADF test to check stationarity , we can see our train data set is stationary with 0.01 p-value so, we don't need to apply differencing in this step.

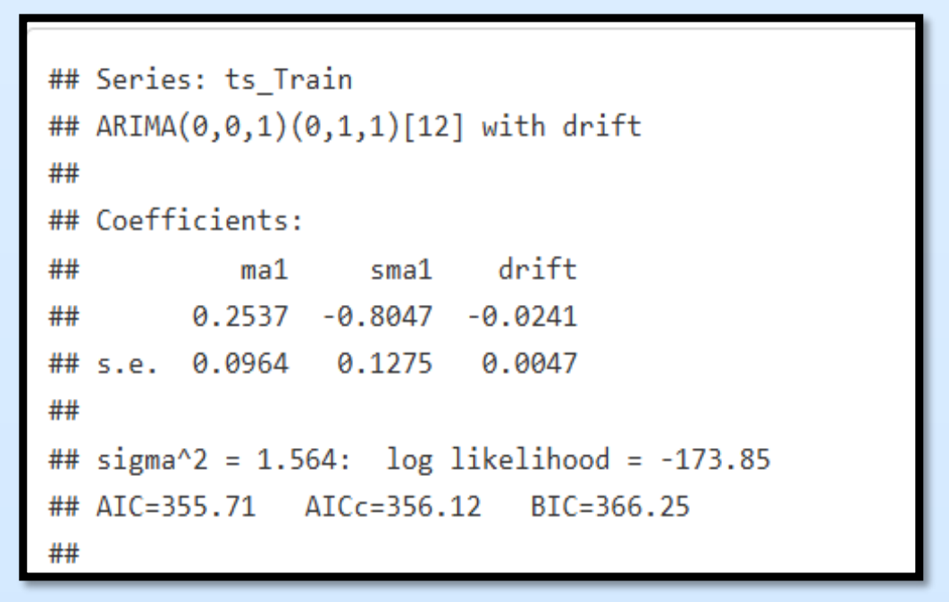


Before applying SARIMA model manually , we need to do seasonal transformation to remove our seasonality , because in manual SARIMA model (D) was predicted by seasonal differenced aplyed time . So , we can apply seasonal difference in 1 time and check over seasonality is removed if not we need to re apply seasonal differencing 2nd time . From the given plot we applied seasonal differenced 1 and check the seasonality , we can see the seasonality was removed from 1st seasonal differencing .



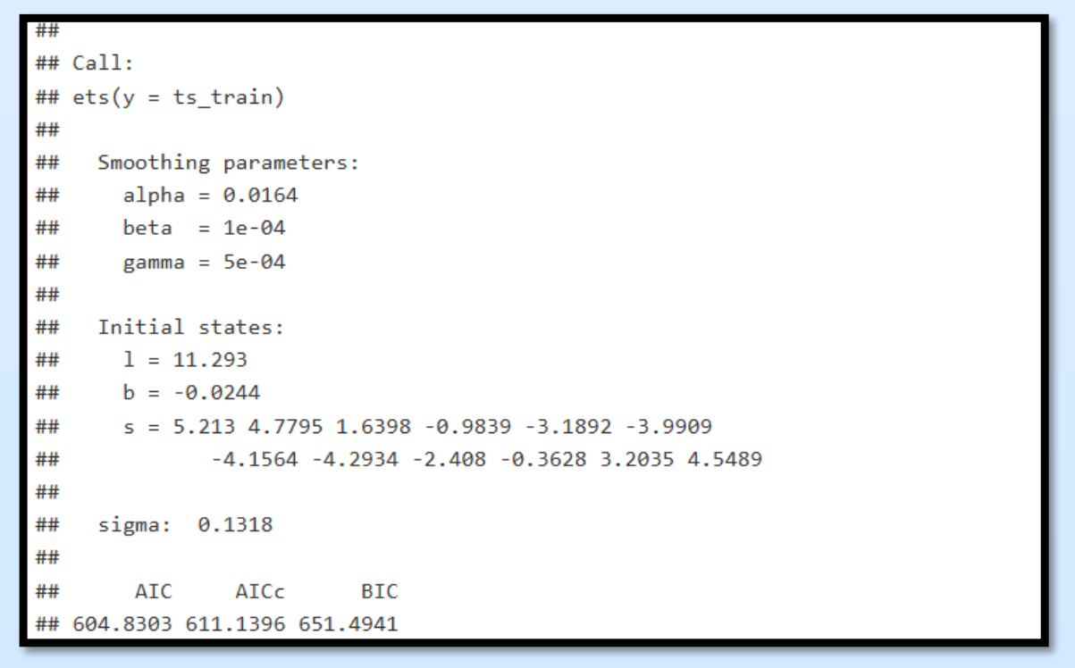
Modeling

Due to the seasonality in train data, we can apply Seasonal ARIMA (SARIMA) model and ETS model to check the best fitting model, by comparing the AIC and BIC values, we can get the best fitting model by checking the model with lowest AIC and BIC values .



```
## Series: ts_Train
## ARIMA(0,0,1)(0,1,1)[12] with drift
##
## Coefficients:
##          ma1     sma1    drift
##       0.2537  -0.8047  -0.0241
## s.e.  0.0964   0.1275   0.0047
##
## sigma^2 = 1.564:  log likelihood = -173.85
## AIC=355.71   AICc=356.12   BIC=366.25
##
```

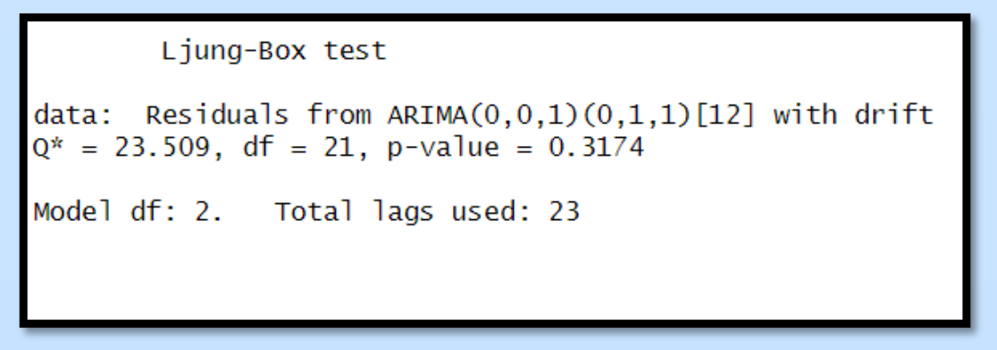
SARIMA



```
##
## Call:
## ets(y = ts_train)
##
##   Smoothing parameters:
##     alpha = 0.0164
##     beta  = 1e-04
##     gamma = 5e-04
##
##   Initial states:
##     l = 11.293
##     b = -0.0244
##     s = 5.213 4.7795 1.6398 -0.9839 -3.1892 -3.9909
##            -4.1564 -4.2934 -2.408 -0.3628 3.2035 4.5489
##
##   sigma:  0.1318
##
##      AIC     AICc      BIC
## 604.8303 611.1396 651.4941
```

ETS

based on the comparison of auto ARIMA(SARIMA)test and ETS model - the SARIMA model is the beter fit model for the our time series data.because, SARIMA model has lower AIC ,BIC values (AIC=355.71 BIC=366.25)comparing to ETS model AIC and BIC results(AIC=604.8303 BIC= 651.4941) , so we can take SARIMA as the best fitting model.

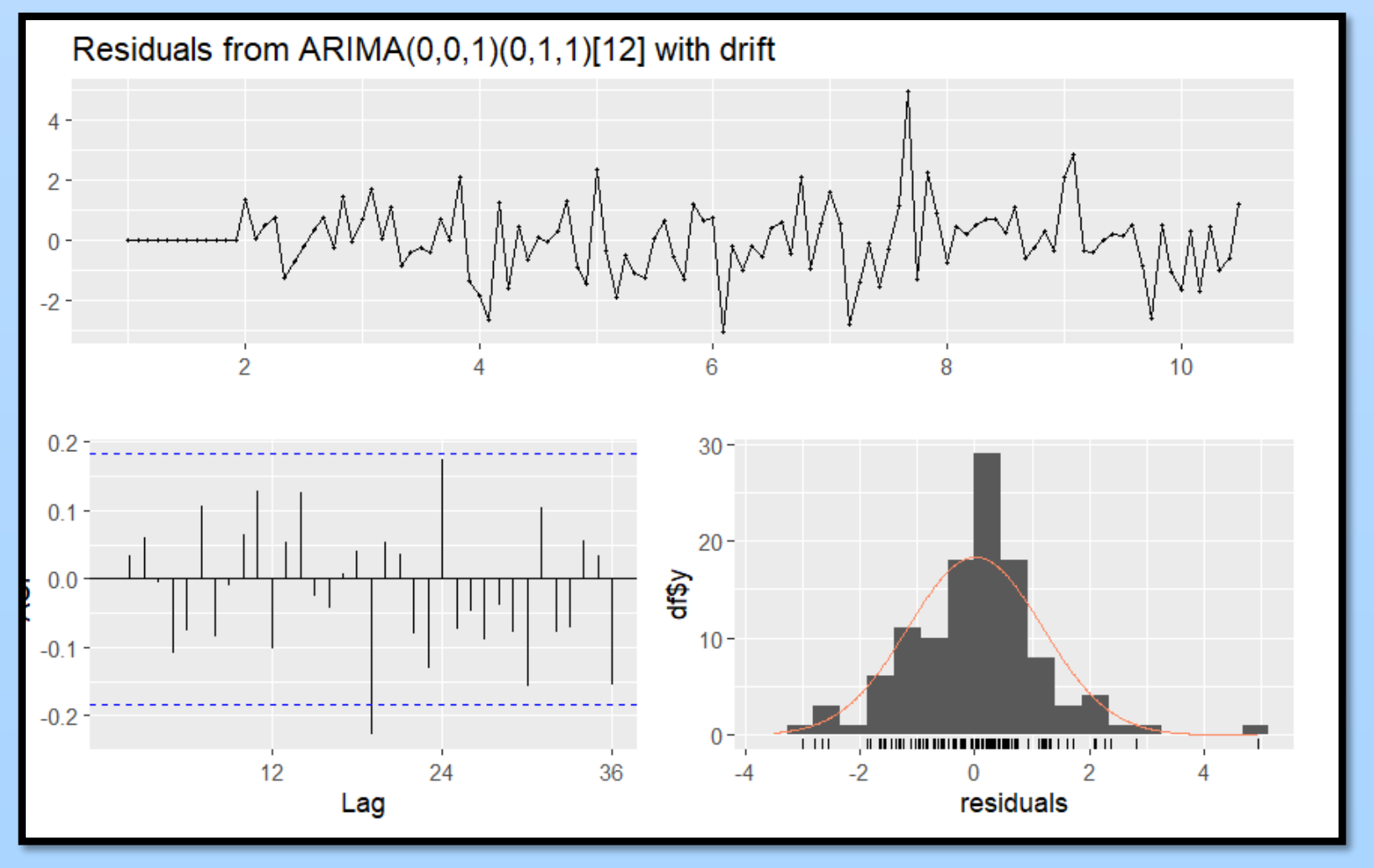


```
	Ljung-Box test

data:  Residuals from ARIMA(0,0,1)(0,1,1)[12] with drift
Q* = 23.509, df = 21, p-value = 0.3174

Model df: 2.   Total lags used: 23
```

from the results of Ljung-Box test check the residuals of our SARIMA model are uncorrelated or not,

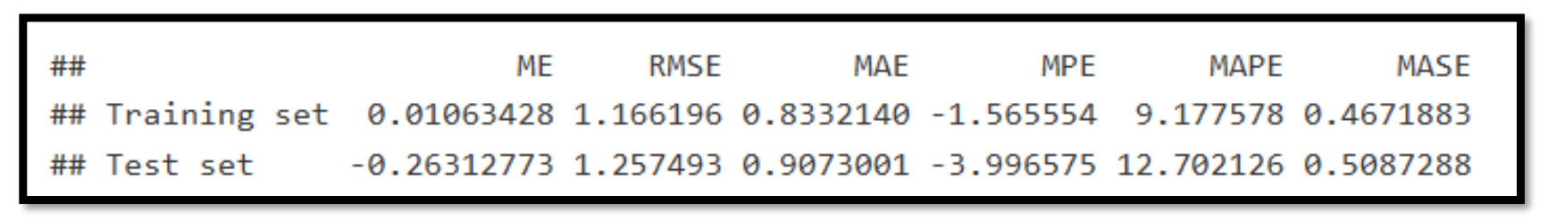
from the results of SARIMA test P-value is 0.3174 it is greater than 0.05 and we fail to reject null hypothesis (H0),

(H0 Null Hypothesis = Residuals are independent model is a good fit)it means the residuals of SARIMA model don't show autocorrelation and SARIMA model has captured the patterns in the data. As the residuals shows the P value(0.3174) is higher than 0.05 which suggest that this model does not have significant autocorrelation , that means the SARIMA model capture all the patterns and trends in this series.



As the final result of best model fitting we can choose , SARIMA is the best fitting model for this time series data.

Forecasting



```
##                        ME     RMSE       MAE       MPE      MAPE      MASE
## Training set  0.01063428 1.166196 0.8332140 -1.565554  9.177578 0.4671883
## Test set     -0.26312773 1.257493 0.9073001 -3.996575 12.702126 0.5087288
```

When comes to selecting forecasting models we use the accuracy metrics for the training set and test set. training set metrics tells how well the model fits the training data. test set metrics tells how well the model forecasts match the actual test data to see how well it performed in predicting.

so for forecast accuracy we use test set metrics.

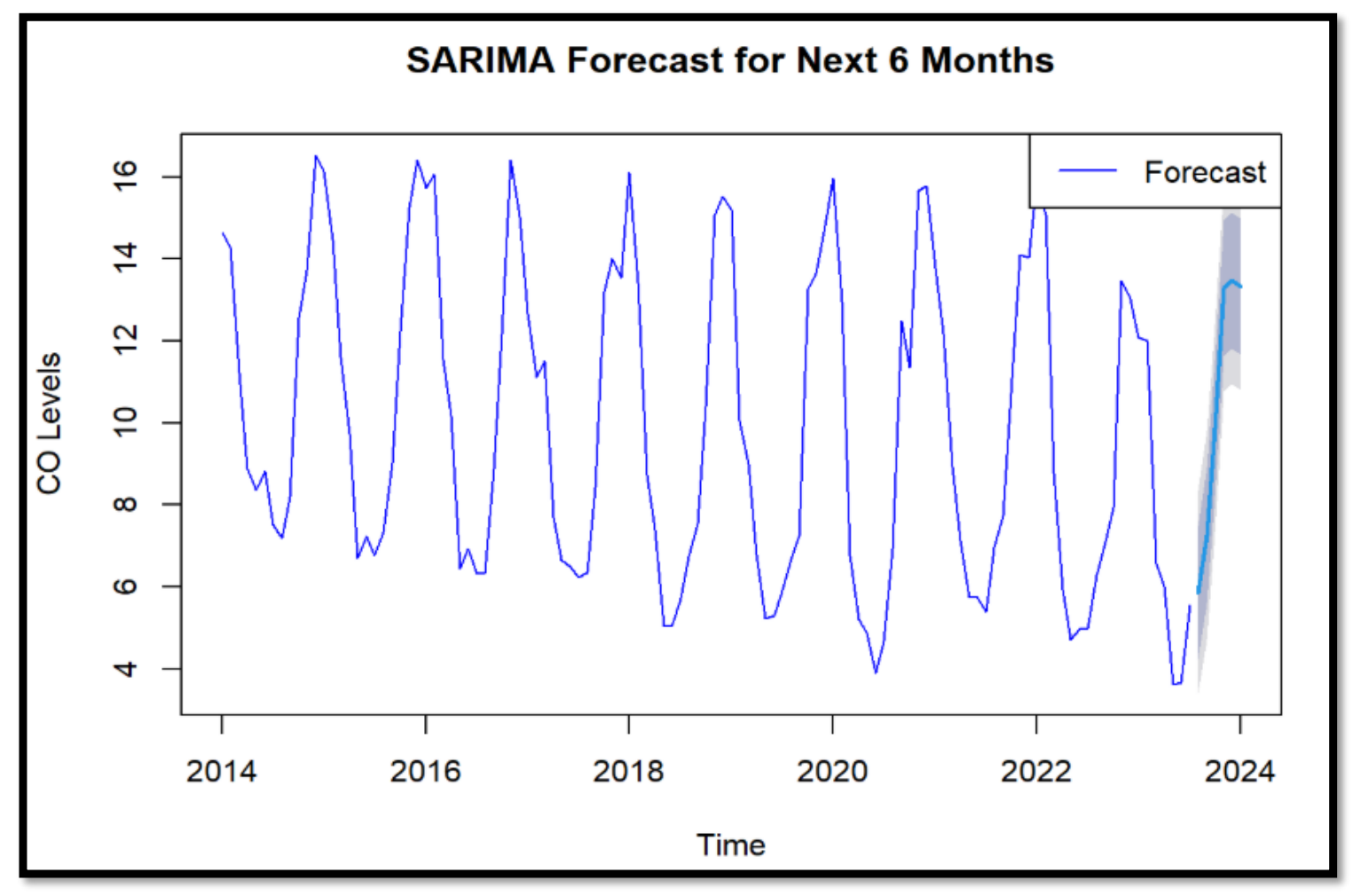
*MAE is 0.9 so on average the forecasted values deviate from the actual test values by 0.9 units.

*RMSE is 1.25 this is the average magnitude of forecast errors giving more weight to larger errors.

*MAPE is 12%. this is below 50% so the model is relatively acceptable.

the test set has higher MAE, RMSE and MAPE than training set which means model performs better on training set. this could be due to overfitting and capturing patterns too specific to the training data and does not generalize as well to new data.

Based on the calculated accuracy metrics SARIMA model's train error values are lower than the ETS model error values ,RMSE 1.16 (forecast errors are smaller for SARIMA),MAE 0.83 (SARIMA model's average errors are smaller value),MAPE 9.17 (SARIMA precentage error is smaller), based on the accuracy error metrics of the above two models we can select SARIMA model as the best forecasting model.we can use SARIMA model to forecast future ,



Results and Conclusion

The SARIMA model was used to forecast future (6 months) of CO levels and the Model fit best with series with lower AIC,BIC, MAPE , RMSE and MAE values.

Strength

This study covers all the aspects in time series forecasting and applies both the SARIMA and ETS models and validates them using cross-validation to select the best model.

Limitations

This study was limited to Los Angeles CO levels over an extended time period (2014-2024).The data set which used for this was univariate data set which only have data on CO level and the date.Data was collected daily , if the data was collected hourly it may give more clear and reliable forecast in future and If the data was with over 20 yrs we can do a better forecast.

Recommendations

Expand the data sources, like use other different sources that can get data and adopt multivariate data model and factors (more air pollution gases)for this study.