Time Series: Project 1

André Luís, 98638 Carlos Sequeira, 87638

I. Introduction

II. Transformations

First, we aggregated the hourly data into daily data of PM10 particles in Avenida da Liberdade, from 2014 to 2019.

Then, we needed to deal with missing values, and we decided to perform linear interpolation on this values, using $na.approx$ function from zoo R package.

III. Exploratory Analysis

After the transformations the resulting data is displayed as follows in Figure 1, where in the vertical axis we have the PM10 particles in micrograms per cubic meter and in the horizontal axis we have the respective time.

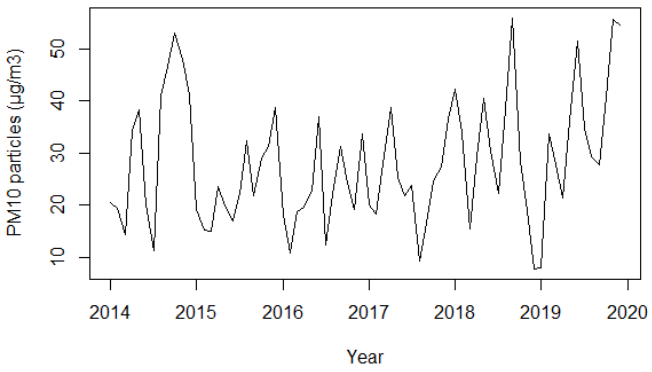

Fig. 1. 24 hours average levels of PM10 particles in Avenida da Liberdade

To perform a better analysis a STL procedure was performed over the data and is represented in Figure 2.

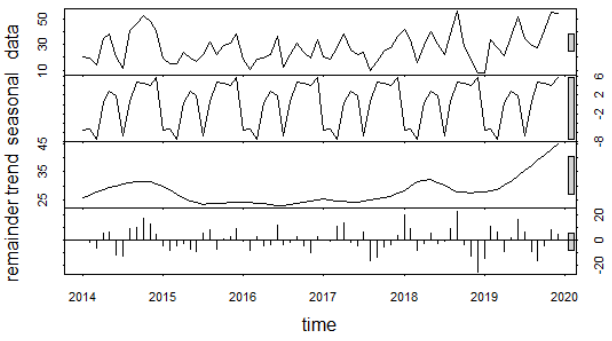

Fig. 2. Seasonal-trend decomposition procedure on the PM10 particles data

The resulting plot is divided in four parts which represent, from top to bottom, the original data, the seasonal component, the trend component and the remainder that we get after removing these two last components from the original data.

We can now analyze the two resulting components:

- Seasonal Component: PM10 particles have a higher concentration in May, September, October, November and December, and a lower concentration in January, February, March and July.
- Trend Component: PM10 particles concentration has been increasing since July of 2016, with a huge increase after March of 2019.

For the next step we analized both the original data and the residuals obtained from the STL procedure.

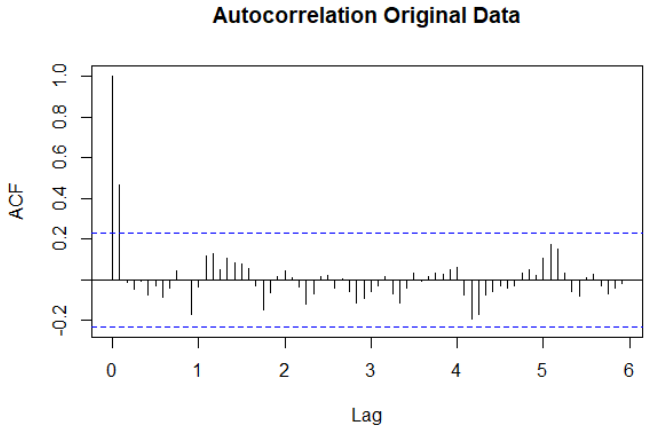

Fig. 3. Autocorrelation Function on the Original Data

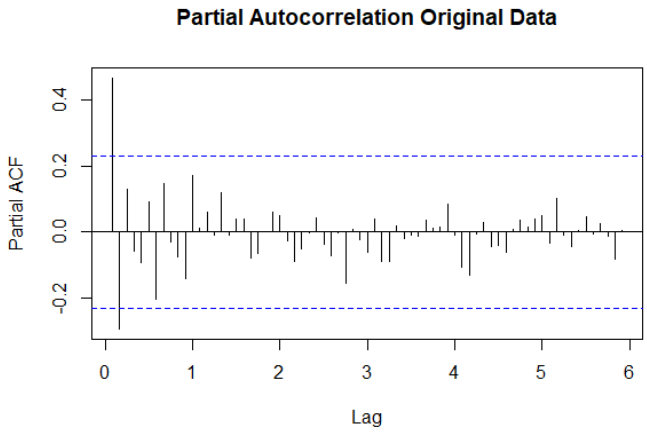

Fig. 4. Partial Autocorrelation Function on the Original Data

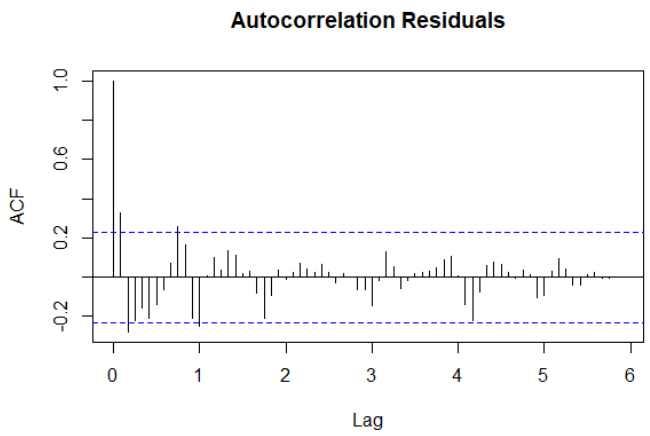

Fig. 5. Autocorrelation Function on the STL Residuals

It is possible to verify that, in both Figure 3 and Figure 5, there is very low correlation between two different time steps as most data points fit between the lag-wise 95% confidence intervals in blue. This proves that both the original and the residuals are random data sets.

IV. Dependence orders and degrees of differencing

Before fitting a model to our data, one needs to assess if our time series is stationary, and, if yes, how many degrees of differencing are needed, since it removes changes in the level of a time series and, therefore, reduces or eliminates trend and seasonality.

By visual inspection of the ACF and PACF plots, we can suspect that our time series is stationary, since they drop to zero relatively quickly. However, to confirm our visual analysis, we tested stationarity with three different hypothesis tests:

- Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test for level or trend stationarity: Unit root test to assess if a time series is stationary. The null hyphotesis is that the data is stationary, so small p-values suggest that differencing is required.
- Augmented Dickey-Fuller Test: Test if a unit root is present in a time series sample. The null hypothesis for this test is that the data is non-stationary.
- Ljung-Box test for independence: Examines whether there is significant evidence for non-zero correlations at given lags. We tested for 25 lags. The null hypothesis represents independence in a given time series, which suggests stationarity.

TABLE I
Stationarity tests

Test	Decision	Conclusion
KPSS Test	Do not reject H0	Data is Stationary
ADF Test	Reject H0	Data is Stationary
Ljung-Box Test	Do not reject H0	Independence between different lags

Thus, we conclude that our data is stationary, and we can continue our analysis without the need to difference it.

V. Parameter estimation

Seasonal Auto-regressive Integrated Moving Average (SARIMA) Models can be defined as:

$$SARIMA \underbrace{(p,d,q)}_{Non-seasonal} \underbrace{(P,D,Q)_s}_{Seasonal}$$

Where the non-seasonal parameters correspond to the auto-regressive order, degree of differencing and moving average order of the model, respectivelly. The seasonal parameters correspond to the modelling of the seasonal period, with s=12.

Our data shows a seasonal pattern, with some months having higher observations of PM10 particles every year, as we observed in the STL decomoposition. Thus, an ARIMA model cannot cope with this seasonal component, since this model assumes that data is either not seasonal or has the seasonal component removed. We need to fit a SARIMA type model to our data.

In the previous section, we showed that our data is stationary, so there is no need to difference it. However, when dealing with the seasonal component, we will use D=1 to calculate the differencing at a lag equal to the number of seasons (s) to remove additive seasonal effects.

The choice of parameters to our model was based on the autocorrelation and partial autocorrelation plots of this time series:

- p - order of the autoregressive part was based on the PACF plot. Since we have 2 spikes on this function plot, that represent observations outside of the random confidence intervals, we set p=2
- d - degree of differencing. Equal to zero, as explained above. No differencing is required to model the non-seasonal part.
- q - order of the moving average part was based on the ACF plot. Since we have one spike in this plot, we set q=1

Three different performance metrics were used to assess the fit of the model to our data:

- Akaike's Information Criterion (AIC), computed as follows:

$$\text{AIC} = -2\log(L) + 2(p+q+k+1)$$

- Corrected AIC, computed as follows:

$$\text{AICc} = \text{AIC} + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2}$$

- Bayesian Information Criterion (BIC), computed as follows:

$$\text{BIC} = \text{AIC} + [\log(T) - 2](p+q+k+1).$$

Where L is the likelihood of the data, k=1 if c!=0 and k=0 if c=0.

These criteria only served to compare between different values of p and q. It did not influence the choice of d,

since differencing changes the data on which the likelihood is computed, making the values for these metrics not comparable.

The choice of seasonal parameters was based on the ones used in non-seasonal part of the model. The minimization of the previous performance metrics, specially the corrected AIC statistic, were our objective.

Final model is the following:

$$SARIMA(1,0,2)(1,1,1)_{12}$$

And it produced the parameter estimates of Figure 6, alongside with the performance metrics of Table II.

```
ARIMA(1,0,2)(1,1,1)[12]

Coefficients:
          ar1     ma1     ma2     sar1     sma1
      -0.9301  1.6776  0.7329  0.0115  -0.9999
s.e.   0.1026  0.1178  0.1062  0.1713   0.3935
```

Fig. 6. SARIMA parameter estimates

TABLE II
PERFORMANCE METRICS

AIC	470.75
AIC_C	472.33
BIC	483.31

VI. Residual Diagnostics

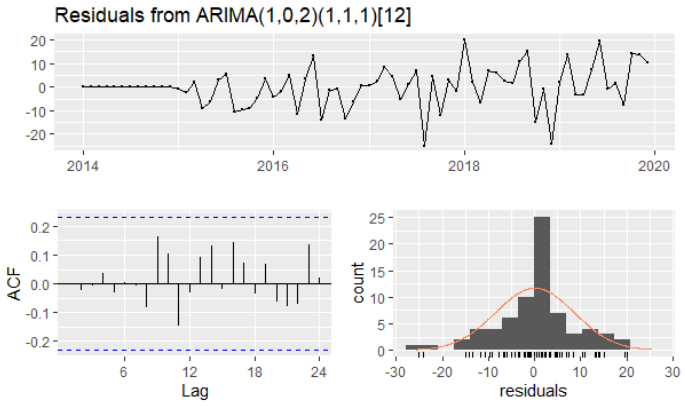

Fig. 7. Residual Analysis

Residuals ACF is shown on Figure 7 and there we can see that they are within the significance limits, so they appear to be white noise.

Besides that, we performed a Ljung-Box test on the residuals and it showed that there are no remaining autocorrelations.

VII. Results

So we can proceed to forecast the next 12 months using our model, since all the requirements were satisfied. The results can be seen in Figure 9.

```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,2)(1,1,1)[12]
Q* = 8.3468, df = 9, p-value = 0.4996

Model df: 5.   Total lags used: 14
```

Fig. 8. Ljung-Box test for the residuals

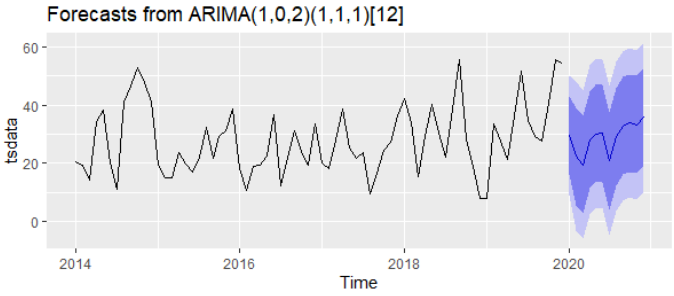

Fig. 9. SARIMA forecast

VIII. Conclusion

We can conclude that our data has some seasonality that cannot be modeled by ARIMA models, and thus we need to use SARIMA to cope with this.

In Figure 9, we can identify a similar seasonality comparing with our initial exploratory analysis, with peaks in the months of May, September, October, November and December, which gives us an intuition that our models approximates well the seasonality presented in previous years.

As future work, we would like to divide the initial dataset into train and test set, building the model on the training set and adjusting its parameters using the test set. We believe this could improve the quality of parameter estimation