Statistics for Data Science Project

Micluța-Câmpeanu Marius – 407

1. Time series

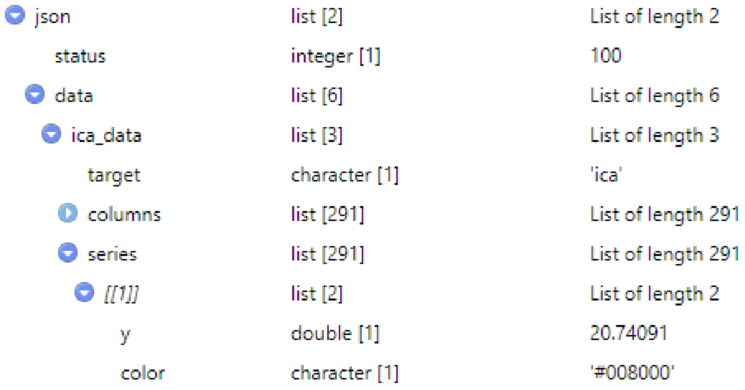
Dataset used

The dataset for this subject is the Air Quality Index (AQI) from Bucharest, recorded at University Square. Official data can be obtained from http://calitateaer.ro/public/monitoring-page/reports-reports-page/, but the process itself is tedious and time-consuming. Instead, I used another source which is more straightforward to use. The data can be viewed here: https://aerlive.ro/ica/universitate/ and it can be easily downloaded using the browser's developer tools and selecting the corresponding json response. Using cURL, it can be accessed using the following command in order to obtain more datapoints:

```
curl -k 'https://apps.roiot.ro/aerlive/api/cluster.php?q=history' \
--data-raw 'key=d09668ea-def5-44ea-8c77-ae32e9fa5572&s=2019-01-25&e=2020-06-22&cluster=8'\
> data_univ.json
```

Inspecting the two request parameters (`s` and `e`), we can determine that we sould have data spanning from 2019-01-25 until 2020-06-22. However, analyzing the data, we have 291 observations spanning from 2019-08-13 until 2020-06-03. The `key` parameter refers to the station (Universitate in this case).

The data has the following structure:



json	list [2]	List of length 2
status	integer [1]	100
data	list [6]	List of length 6
ica_data	list [3]	List of length 3
target	character [1]	'ica'
columns	list [291]	List of length 291
series	list [291]	List of length 291
[[1]]	list [2]	List of length 2
y	double [1]	20.74091
color	character [1]	'#008000'

The `columns` key represents the date of the measurement. The data is comprised of 6 time series:

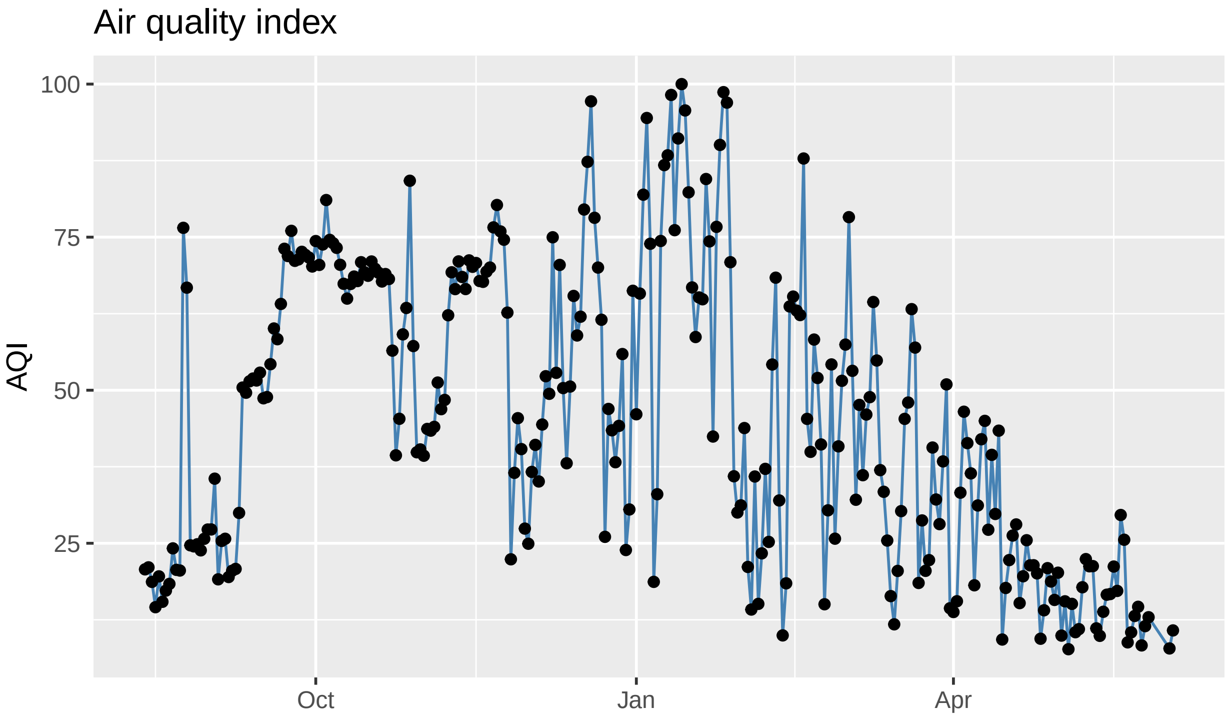
- ica_data
- pm10_data
- pm25_data
- co_data
- no2_data
- so2_data

I will limit this analysis to ICA (indicele calității aerului - AQI). From Wikipedia (https://en.wikipedia.org/wiki/Air_quality_index), the air quality index is used by governments to inform the general public about how polluted is the air or how polluted it will become.

According to the data source website (https://aerlive.ro/cum-masuram-poluarea/), since the project is new, only PM10 and PM2.5 are used to compute the AQI (ICA). The measurements for air pollutants CO, NO_2 and SO_2 are in testing phase ("1-3 months"). It is unspecified if these have been included by now, because the data spans almost ten months.

PM10 (Particulate matters) are particles with a diameter 2.5 and 10 micrometers; PM2.5 is for particles that have diameter 2.5um or less (https://en.wikipedia.org/wiki/Particulates).

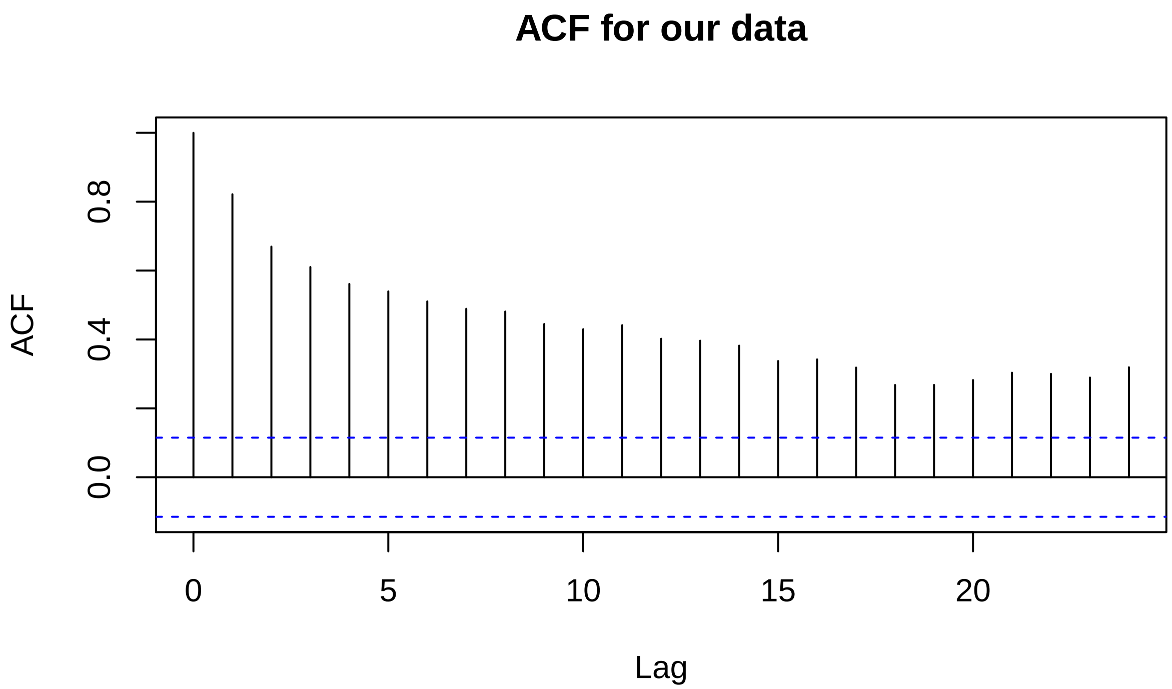
The data points are plotted below:

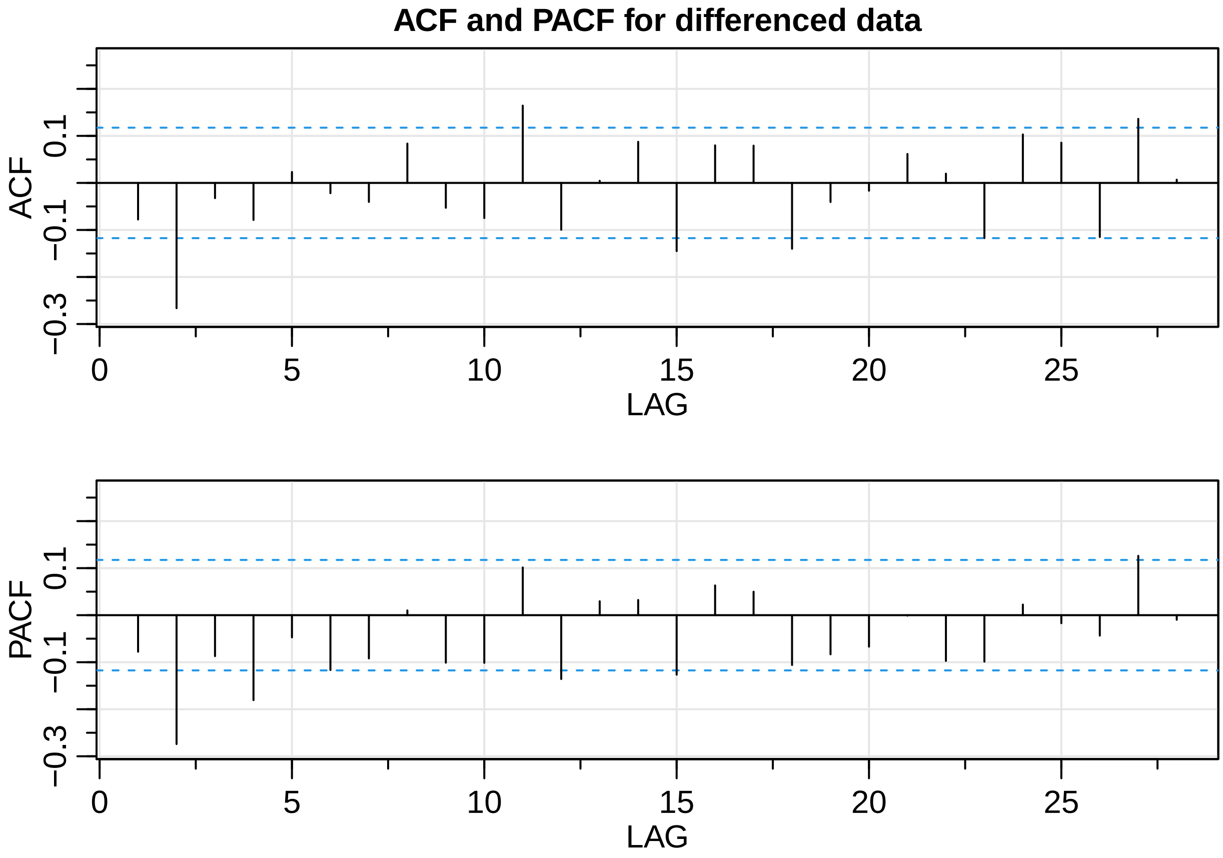


Building the model

From Lecture 6-1, a slowly decaying ACF (auto-correlation function) tells us if differencing is likely necessary. In order to verify this, first we plot the ACF.

Next, we plot the auto-correlation and partial auto-correlation functions for the differenced data.





The ACF plot suggests a possible MA(3) or MA(5) model. The PACF plot suggests a possible AR(3). The plots might also suggest larger models, but the risk of overfitting is greater.

Unfortunately, there is not enough data to observe the seasonality, and the observed trend might be a bit misleading.

Next, we plot the residuals in order to see if the model fits our data well. Below, I only show the results for the best model that I found (based on the indications from the lectures/laboratory), because plots take up a lot of space. They are included in the notebook code, but they are simply not rendered in the output document. I have experimented with MA(3), MA(5) and AR(3).

AR(3) fails the Ljung-Box test.

MA(5) has slightly higher AIC and BIC scores, so I only included results for MA(3).

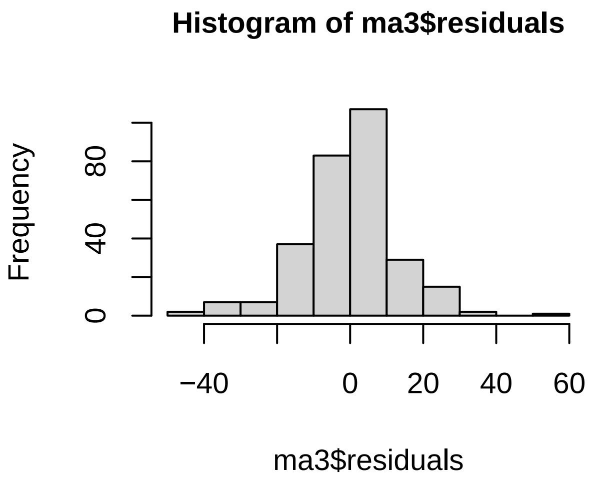
Analyzing the model, diagnostics

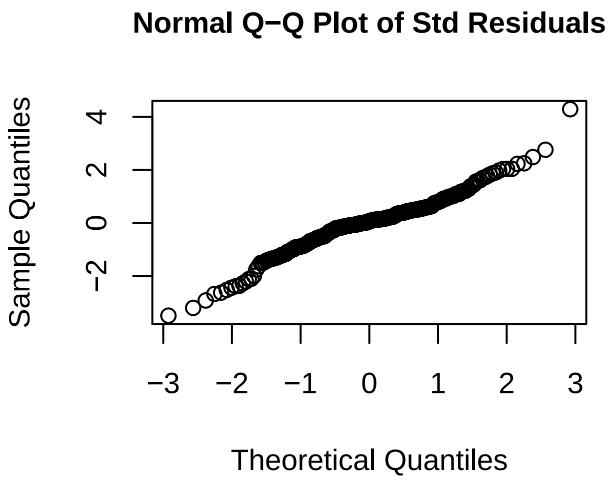
The histogram of residuals and the Q-Q plot are shown below. The Q-Q plot seems approximately linear. However, I have used the Shapiro-Wilk normality test and it strongly rejects the null hypothesis, which means that residuals do not come from a normal distribution (from the histogram, it is indeed skewed).

```
## [1] "AIC: "            "2322.22127623599"
```

```
## [1] "BIC: "            "2340.57068085089"
```

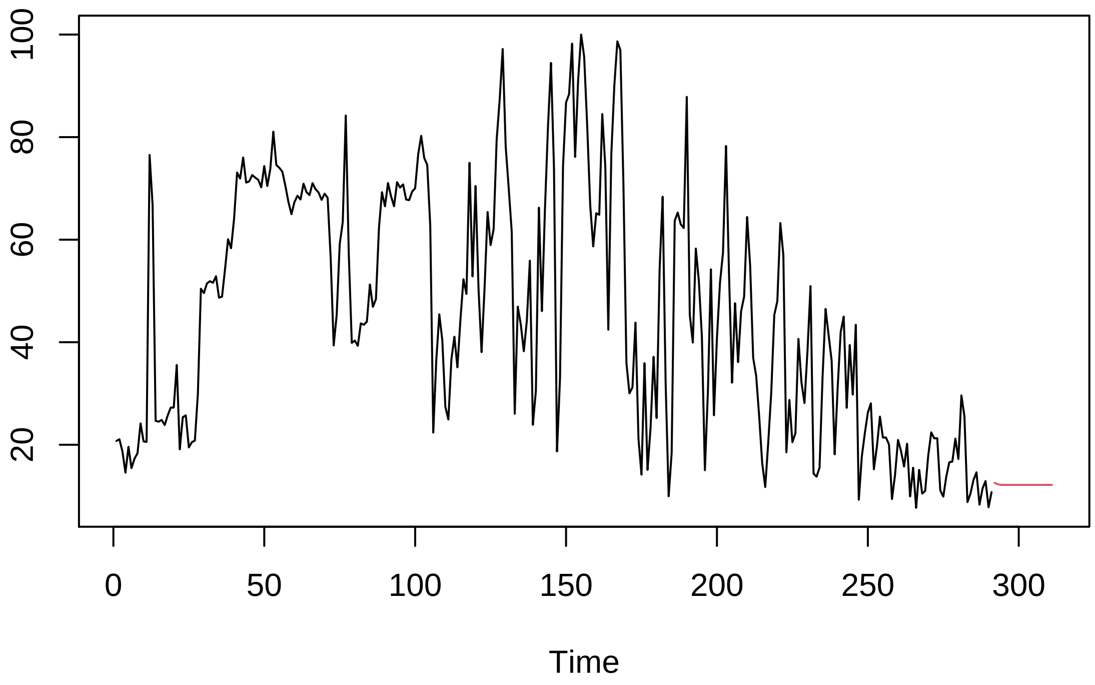
```
##
##  Shapiro-Wilk normality test
##
## data:  ma3$residuals
## W = 0.96304, p-value = 9.413e-07
```





Predicting new values

We differenced once, so `d=1` and the model would be ARMA(0, 1, 3). The next 20 predicted values are plotted below. Due to the lack of better data (based of what I have read besides the lectures - see references), it is not very surprising to obtain a flat prediction. From my understanding, this happens to be the case when (at least for this model) the data does not provide a better trend‘. We might also obtain better results with a seasonal model, since that is usually the case with weather-related phenomena. It might also be the case that further initial transformations of the data are required.



Two-way ANOVA

Dataset used

Results

Linear regression

Dataset used

Building the model

Results

References:

- the code is inspired from laboratories; I also tried to include below most of the resources I used
- https://www.r-graph-gallery.com/279-plotting-time-series-with-ggplot2
- http://www.cookbook-r.com/Graphs/Titles_(ggplot2)/
- https://www.earthdatascience.org/courses/earth-analytics/time-series-data/date-class-in-r/
- https://stackoverflow.com/questions/58907202/
- https://stats.stackexchange.com/questions/111010/

- https://stats.stackexchange.com/questions/143073/non-normal-residuals-in-arima
- https://stats.stackexchange.com/questions/79400/
- https://stackoverflow.com/questions/37115276/
- https://stackoverflow.com/questions/4241798/how-to-increase-font-size-in-a-plot-in-r
- https://stats.stackexchange.com/questions/286900/arima-forecast-straight-line
- https://stats.stackexchange.com/questions/124955/is-it-unusual-for-the-mean-to-outperform-arima/125016#125016
- https://rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf
- https://stackoverflow.com/questions/28032846/