Data Analysis and Visualization in R (IN2339)

Case Study: There is Something in the Wind - Has Barcelona's Air Quality Changed Over the Past Four Years?

Jennifer Schlindwein, Philippe Thome, Jan-Steffen Ruck, Anastasiia Okonnikova

21 Januar, 2022

Introduction

Greener - healthier - more sustainable: modern cities compete along various dimensions to attract and retain both inhabitants and organizations. A central factor to providing a livable environment within densely populated metropolitan areas is air quality. Although invisible to the human eye, traces of various pollutants can impair both public and individual health. To keep track of their current level and its development, the city of Barcelona installed a set of measure stations across most of its districts. Three commonly measured pollutants are NO2, O3 and PM10. With Barcelona's endeavors to revolutionize their intra-city mobility network, NO2, mostly emitted by traffic, becomes of particular interest. We therefore studied the development of the NO2 concentration within several districts over the time span 2018 - 2021.

Dataset and Preparation

In addition to the initially provided data set that contains pollution data for November 2018, we found a similar continuation of measurements that span from 2019 - 2021. Those measurements (aka "external data") were published by the Direcció General de Qualitat Ambiental i Canvi Climàtic de la Generalitat de Catalunya as part of Barcelona's Open Data Movement and are managed by the Department of Statistics and Data Dissemination of the Municipal Data Office. It can be accesses with the following link: https://opendata-ajuntament.barcelona.cat/data/en/dataset/qualitat-aire-detall-bcn. We added text files detailing the sources as well as English translations for feature names/descriptions to the external dataset.

Both sets contain hourly measurements of the concentration for each of the three pollutants described above in µg/mˆ3. Each observation is complemented by a position, a station name or identifier and a time stamp. The following paragraphs describe individual characteristics for each of the two data sources.

The original dataset of the case study was labeled as containing data for November 2017, yet all internal time stamps refer to November 2018. We chose to assume that the time stamps are correct.
The external dataset contains data for all but the following months of 2019: January, February, March (missing values).

Although both the case study and the external dataset are in the documented format of the Open Data material, this format was updated, taking effect from May 2019 onward. To accommodate this change, we prepared the two data sources separately (coincidentally, all of the non-empty external data set files were found to adhere to the new format).

Both, the case study and the external dataset, were accompanied with a second table that provides further information regarding the measurement stations. For the external dataset, we ran a manual check and found that the number and feature of relevant stations did not change when comparing the two available versions from 2018 and 2021. We therefore chose to work with the 2021 version.

For the 2018 data, we expanded the provided 'date of generation' time stamp into separate features for hour, month, day and year. To locate the measurements and ensure consistency with external data, we further matched the information regarding the measurement's position with the available measurement station data via coordinates and station identifiers.

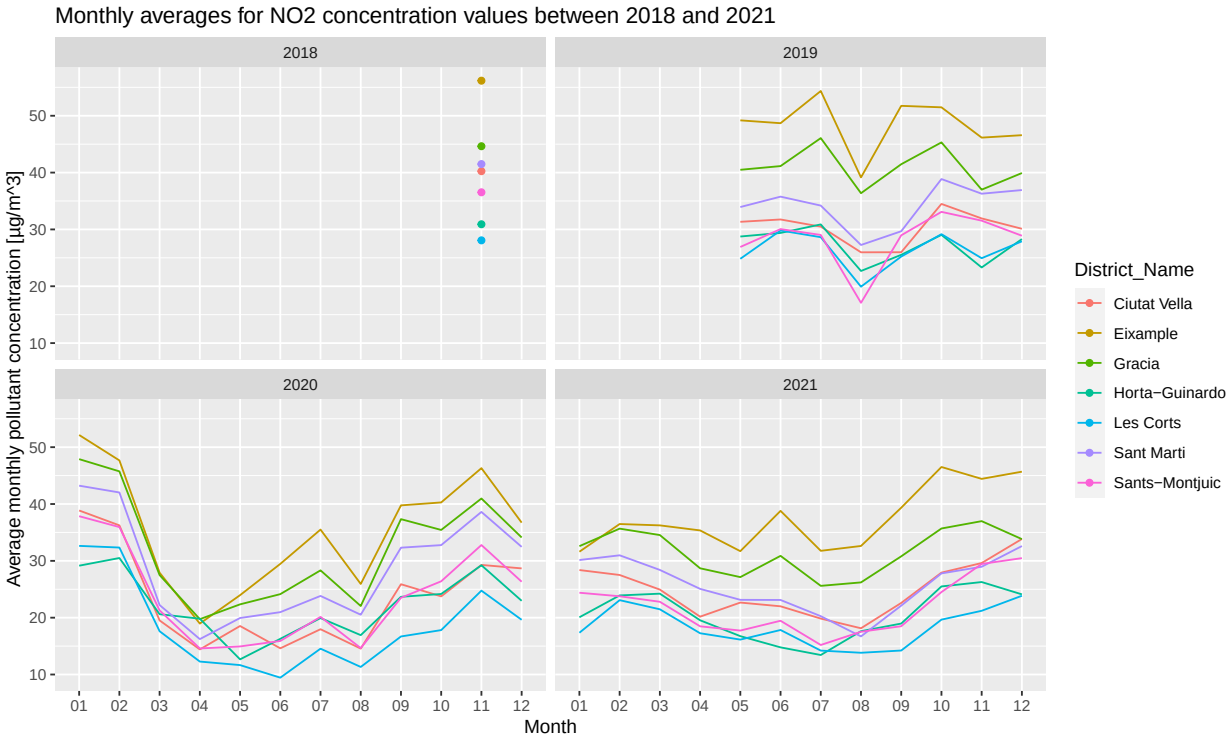
Besides the empty March 2019 file, the external dataset was published in a wide-table format with columns for each hour of the day which had to be converted to our target long-table format. Time stamp conversions were similar to 2018. The observed pollutant was provided as a code that had to be matched with its respective pollutant using a separate file (qualitat_aire_contaminants_pollutants.csv).

Combining the pre-processed entries from 2018 and the external dataset returns the 'airquality' dataframe, a long-list that features one measurement per row and contains features regarding Longitude, Latitude, Station, Hour, Value, Day, Month, Year, District Name, Neighborhood Name, and 'Gas' (which refers to the observed pollutant) in its columns.

An overview over the prepared data, aggregated by month and district, is provided in the plot below:

```
airquality_agg <- aggregate(Value ~ District_Name + Month + Year, data = airquality[Gas ==
    "NO2", ], FUN = mean)
airquality_agg$District_Name <- as.factor(airquality_agg$District_Name)

plt <- ggplot(data = airquality_agg, aes(x = Month, y = Value, color = District_Name,
    group = District_Name))
plt <- plt + geom_line(data = subset(airquality_agg, Year > 2018)) + facet_wrap(~Year) +
    geom_point(data = subset(airquality_agg, Year == 2018))
plt <- plt + labs(x = "Month", y = "Average monthly pollutant concentration [µg/m^3]",
    title = "Monthly averages for NO2 concentration values between 2018 and 2021")
plt <- plt + scale_fill_discrete(name = "District Name")
plt
```



One notable feature is that NO2 concentration is not equally distributed across the city. There seems to be a relative ranking of districts, which is rather stable over months and years.

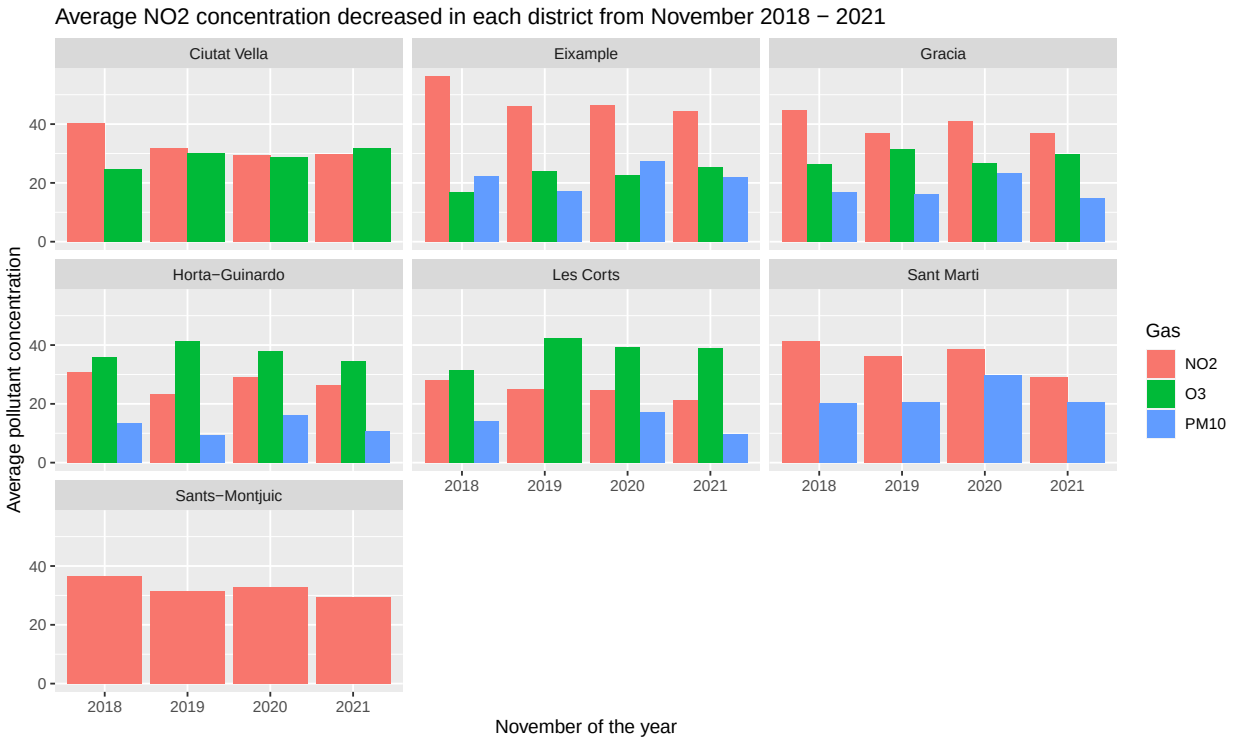
The plot further suggests seasonal fluctuations in the form of higher NO2 concentration values during winter months. As our individual datasets for each year do not cover the entire year, an aggregation of concentration values on a yearly basis might introduce seasonal distortions (e.g. 2018 averages would fully consist of November values, while 2019 ones would further include spring and summer months). Because our most limited dataset, 2018, covered the month of November, which is further present in all subsequent years, we narrowed our investigation down to comparing the NO2 concentration aggregated by month of each year's November.

Coming up With an Initial Hypothesis

The city of Barcelona launched and expanded a number of environmental initiatives during the observed time span (e.g. the 'Superilles', essentially banning vehicle traffic from entering certain intersections and streets, as well as measures related to fulfilling the UN's Agenda 2030 targets). Thus, we expected that over time, air quality had been enhanced - or respectively, in terms of available data, NO2 concentration had been decreased.
More precisely, our initial hypothesis was that there is a linear relationship between time and the mean NO2 concentration per month. For the purpose of testing this hypothesis, we defined H0 as the opposite: "There is no relationship between time and the mean NO2 concentration."

A first, visual approach to this hypothesis is shown in the plot below.

```
plt <- ggplot(airquality_November_dt, aes(Year, average_air_value, fill = Gas))
plt <- plt + geom_bar(position = "dodge", stat = "identity") + facet_wrap(~`District Name`)
plt <- plt + labs(x = "November of the year", y = "Average pollutant concentration",
    title = "Average NO2 concentration decreased in each district from November 2018 - 2021")
plt
```



This visualization shows the average air values in the city of Barcelona related to NO2, 03 & PM10 in November of the years 2018 - 2021. For NO2, one can observe a decreased value in November 2021 compared

to November 2018.

Teststatistic

In the following, we investigated whether this relationship was significant.

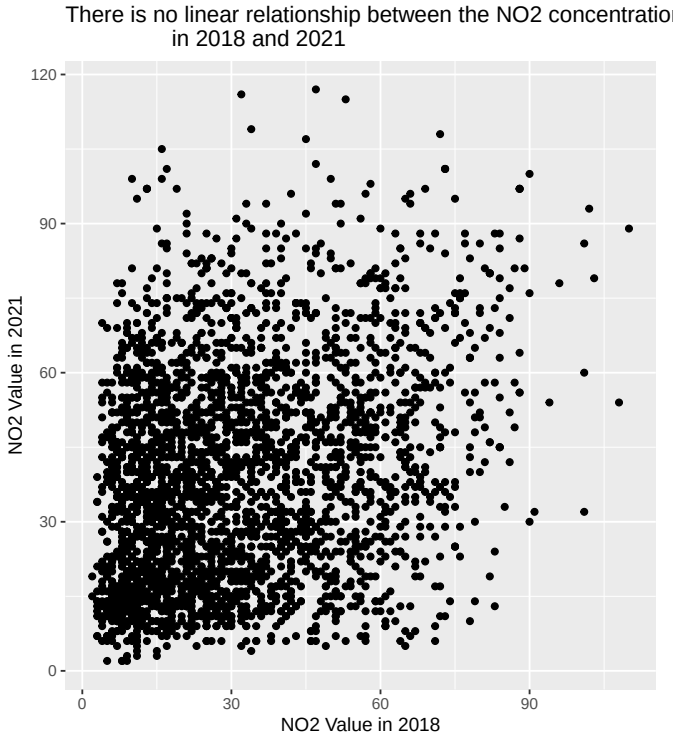
```
airquality_2021 <- airquality[Year == 2021, ]
airquality_2018 <- airquality[Year == 2018, ]
airquality_stat <- merge(airquality_2021, airquality_2018, by = c("Station", "Hour",
    "Day"))

idx <- sample(seq(nrow(airquality)), size = as.integer(0.8 * nrow(airquality)), replace = FALSE)
train <- airquality[idx, ]
test <- airquality[-idx, ]

cor.test(airquality_stat[Year.y == 2018, Value.y], airquality_stat[Year.x == 2021,
    Value.x])
```

```
##
##  Pearson's product-moment correlation
##
## data:  airquality_stat[Year.y == 2018, Value.y] and airquality_stat[Year.x == 2021, Value.x]
## t = 16.743, df = 2949, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2613213 0.3272294
## sample estimates:
##       cor
## 0.2946257
```

```
plt <- ggplot(airquality_stat, ) + geom_point(aes(x = Value.x, y = Value.y))
plt <- plt + labs(title = "There is no linear relationship between the NO2 concentration
                  in 2018 and 2021",
    x = "NO2 Value in 2018", y = "NO2 Value in 2021")
plt <- plt + coord_fixed()
plt
```



Since the p- value of the model is very small, there is at least some relationship between the dependent and independent variables explained. Since at least one factor of each variable is highly significant the chosen variables can stay in the model. The model also passed a validation test, since the MSE on a randomly sampled test set is roughly the same as the MSE of the train set. We therefore reject H0 and accept that there seems to be a time-related trend in the pollutant concentration in the observed districts for the time span of 2018 - 2021: The estimates for each factor of the year variable indicates, that the November NO2 concentration is decreasing as the years progress.

Conclusion

Our case study suggests an optimistic outlook for Barcelona's population: There seems to be a reduction in the NO2 concentration in the month of November over the past four years. This relationship is also found to be valid across districts.

Further investigation might now isolate and identify effects of the above-mentioned environmental initiatives or study how of external shocks such as corona-related measures impacted the city's air quality.