# Machine Learning - Final Project

**Rodrigo Lallana** - https://gitlab.cs.ttu.ee/rolall/iti8586_Rodrigo_Lallana_project

## 1. DATA ANALYSIS AND PREPARATION

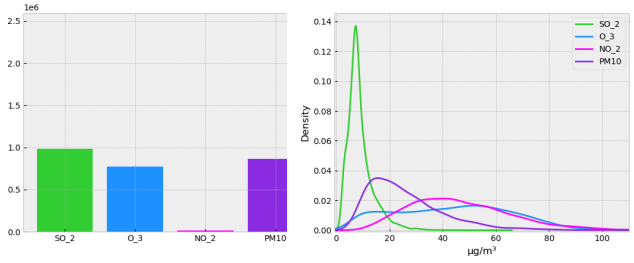
### *1.1 Data cleaning*

The total number of rows of the dataset for each pollutant is **2592864**. The dataset contains a lot of hourly missing values for three out of the four primary pollutants. The primary pollutants that we want to predict are:

- Sulphur dioxide (SO2)
- Ozone (O3)
- Nitrogen dioxide (NO2)
- Particulate matter (PM10)

The data sets contains a lot of NaN values due to each of the 24 differents stations measure different types of pollutants.

To make the dataset easier to work with, we made the next changes: for each primary pollutant we delete the missing rows and downscale the data from hourly data to daily data. The daily data is calculated with the mean of the 24 hours data. And the data of each station is merged.




(a) Missing values of the primary pollutants (b) Distribution of the data after the changes.

### *1.2 Data shape*

The daily data for the NO2 and PM10 pollutants is very noisy, also the PM10 pollutant has a lot of outliers among the others. The data as it was expected has a strong seasonality every 12 months. The horizontal lines indicates the Air Quality Index limit set by the European Environmental Agency.

### *1.3 Objective*

For each model we are going to predict the pollutant concentration mean for one day ahead and two day ahead. Also we will use the monthly mean of the maximum daily values to predict the next month maximum pollutant concentration. We are going to train the model for the data between the year 2001 to 2015 and evaluate from the year 2016 to 2018. The scores to evaluate the model will be the MSE, MAE and R square.

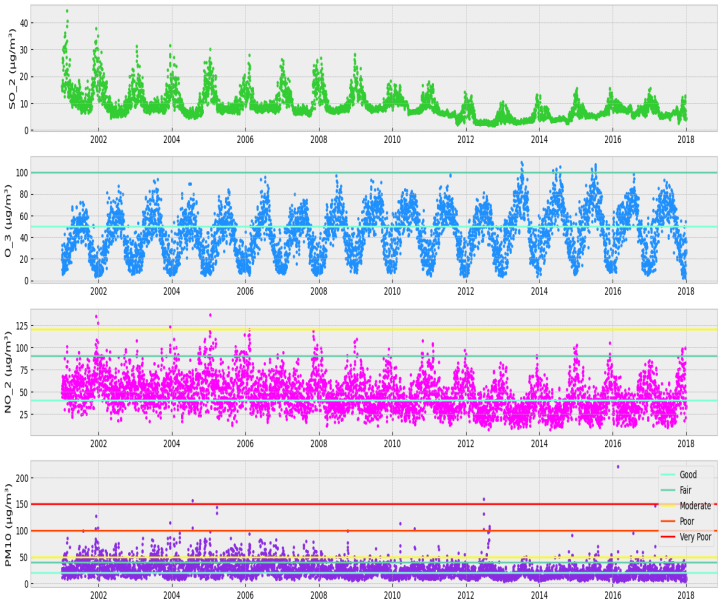



Fig. 2. Daily mean.


Fig. 3. Monthly mean of the daily maximum values.

### *1.4 Measure of the performance*

As we can see in the graph and tables each pollutant has different ranges of values so the evaluation of the MAE will change depending of the standard deviation of each pollutant. If the MAE is lower than the 25% standard deviation of the pollutant we can say the error has a low enough acceptable error.

Daily prediction threshold: SO2<1.18, O3<5.40, NO2<4.70, PM10<3.80

Monthly prediction threshold: SO2<4.55, O3<7.21, NO2<6.69, PM10<9.11

| Summary | SO_2 | O_3 | NO_2 | PM10 |
|---|---|---|---|---|
| mean | 8.826106 | 43.443976 | 46.306180 | 25.918347 |
| std | 4.742453 | 21.623826 | 18.827037 | 15.221613 |
| max_std | 18.334078 | 28.875684 | 26.760205 | 36.447830 |
| min | 1.675000 | 2.607143 | 7.732639 | 2.961538 |
| max | 44.569201 | 109.453731 | 137.028874 | 220.979167 |
| 25% | 5.884483 | 25.562063 | 32.207900 | 15.100491 |
| 75% | 10.503891 | 59.633631 | 58.008276 | 32.803467 |

## 2. LINEAR REGRESSION

### *2.1 One day ahead prediction*

We will train four different models, one for each pollutant: SO2, O3, NO2 and PM10. First we covert the data to a supervised training model. For each different pollutant we create a new column which is the value of that pollutant the next day. That column will be our target, y. And the feature set X, the daily mean of the pollutant. The next table shows the evaluation values after training the four different models. We will use the sckit Linear Regression function.

| Simple Linear Regression | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | 0.7861 | 0.8559 | 0.6782 |
| O_3 | 0.6716 | 159.135 | 9.911 |
| NO_2 | 0.4779 | 151.9496 | 9.7054 |
| PM10 | 0.2526 | 130.4145 | 6.5084 |

The model for the SO2 fit the data very well with a R square of 0.7861 and a low error: MSE=0.8559 and MAE=0.6782. The model for the O3 has a decent R square but the MSE and MAE too are high. For the other two models, NO2 and PM10, the results are poor.

Next we are going to test how using multiple variables to predict the pollutants affect the Linear Regression models. First we calculate the correlation matrix of the four pollutants.

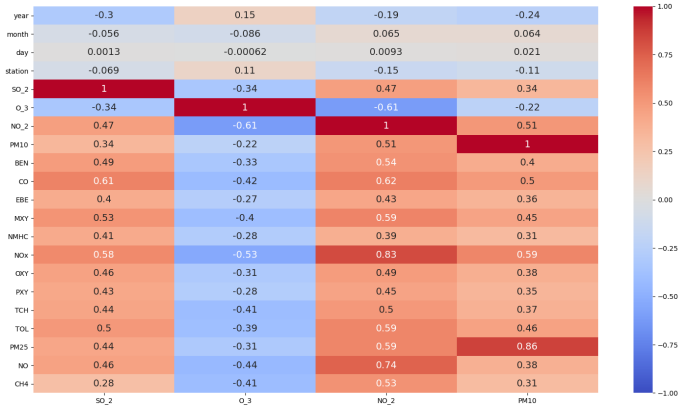



Fig. 4. The correlation of the dataset

We choose the features with correlation higher than 0.5 for each pollutant. The features for each model are:

SO2 : ['CO', 'MXY', 'NOx', 'TOL']

O3 : [(None)]

NO2 : ['PM10', 'BEN', 'CO', 'MXY', 'NOx', 'TCH', 'TOL', 'PM25', 'NO', 'CH4']

PM10 : ['NO2', 'NOx', 'PM2.5']

And we get the next results

| MLR (corr>=0.5) | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | 0.7817 | 0.8735 | 0.7046 |
| O_3 | 0.6716 | 159.135 | 9.911 |
| NO_2 | 0.5026 | 144.753 | 9.4552 |
| PM10 | 0.2956 | 122.9009 | 6.2372 |

There isn't an improvement at all. To improve the model we are going to add past values also called 'lag' values for all the features this means for example adding pollutants value of the past seven days as columns. For all the models we add the features of the past 15 days as columns.

| Multiple Linear Regression (lag=15) | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | 0.7903 | 0.8393 | 0.6618 |
| O_3 | 0.718 | 136.623 | 9.2352 |
| NO_2 | 0.5213 | 139.3104 | 9.213 |
| PM10 | 0.3268 | 117.4696 | 5.9164 |

From the first Simple Regression models to the last Multiple Linear Regression models we manage to increase slightly the R square values and lower the MAE and MSE. The final SO2 model manage to predict very well the mean of the next day with a very low MAE=0.6618 and a good R2=0.7903, the model for the O3 is also relative good.

| Comparison | SO_2 | O_3 | NO_2 | PM10 |
|---|---|---|---|---|
| R2 (First) | 0.7861 | 0.6716 | 0.4779 | 0.2526 |
| R2 (Last) | 0.7903 | 0.718 | 0.5213 | 0.3268 |
| MSE (First) | 0.7861 | 159.135 | 151.9496 | 130.4145 |
| MSE (Last) | 0.8393 | 136.623 | 139.3104 | 117.4696 |
| MAE (First) | 0.6782 | 9.911 | 9.7054 | 6.5084 |
| MAE (Last) | 0.6618 | 9.2352 | 9.213 | 5.9164 |

### *2.2 Two day ahead prediction*

| MLR (lag=15) | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | 0.6194 | 1.5247 | 0.8737 |
| O_3 | 0.6219 | 183.4285 | 10.9549 |
| NO_2 | 0.2739 | 211.3756 | 11.628 |
| PM10 | 0.0636 | 163.4619 | 7.8128 |

Discard completely using MLR to predict NO2 and PM10 pollutants meanwhile model for SO2 is lower than the 1.18 threshold and O3 model gets a decent R2 but is still higher than the 5.40 threshold.

### *2.3 Next month maximum value*

| MLR (1Month) | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | 0.0899 | 6.5912 | 1.9876 |
| O_3 | 0.5561 | 90.8013 | 6.5824 |
| NO_2 | -0.9439 | 558.0215 | 18.9683 |
| PM10 | -0.4499 | 2541.9331 | 29.4591 |

SO2 model performs very well and the O3 model MAE is lower than 7.21 threshold.

## 3. TRIPLE EXPONENTIAL SMOOTHING

The triple exponential smoothing, aka Holt-Winters method, is an algorithm that works best when there is a seasonal trend behavior in the data. Due to the strong seasonality every year in all the four pollutants we choose this model. We are going to use the implementation of this algorithm from the python library: statsmodels.

There are two variations of the model depending of the type of seasonality additive or multiplicative. When the values changes through each season it means the seasonality is multiplicative, when the values stays in a similar range each season is means that the seasonality additive. Because of the daily values are very different from each other we use the seasonality parameter as multiplicative for the daily model. While for the the monthly maximum model the values remain the same through the year, so we will use the additive seasonality.

To evaluate how this Exponential Smoothing model perform we train with the daily values through the years 2001 to 2016 (5478 values) set the seasonality of the data 365 (1 year) and use the seasonality multiplicative approach.

### *3.1 One day ahead*

| | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | 0.7475 | 1.1232 | 0.734 |
| O_3 | 0.615 | 187.956 | 10.7532 |
| NO_2 | 0.3976 | 198.0508 | 11.0745 |
| PM10 | 0.2311 | 99.0221 | 6.3547 |

Worse performance than the MLR model for all the pollutants.

### *3.2 Two day ahead*

| | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | 0.5244 | 2.1134 | 1.0039 |
| O_3 | 0.4636 | 261.1271 | 12.6261 |
| NO_2 | -0.0541 | 347.1425 | 15.0093 |
| PM10 | -0.3903 | 179.2795 | 8.601 |

Worse performance than the MLR model for all the pollutants.

### *3.3 Next month maximum value prediction*

| | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | 0.5553 | 3.1686 | 1.3886 |
| O_3 | 0.6159 | 76.8015 | 6.4655 |
| NO_2 | 0.6497 | 97.1052 | 7.3798 |
| PM10 | -0.0767 | 1827.8448 | 23.9431 |

The monthly prediction of this model for the SO2, O3 and NO2 pollutants perform very well. While the PM10 keeps having a very low R2 most likely due to the peak of PM10 during one of the months of 2016.

### *3.4 Evaluation*

Exponential Smoothing for the next month maximum value performs very well, except for the PM10 pollutant. Meanwhile for the one day and two day ahead predictions the MLR model performs better.

## 4. RECURRENT NEURAL NETWORK - LONG SHORT-TERM MEMORY (LSTM)

We choose this type of neural network because it is widely use to solve problems related to prediction of temporal data. This is thanks to the inside memory which remember a set of past inputs that makes the model very powerful for any sequenced data. We are going to use the Keras LSMT implementation.

### *4.1 Model architecture*

As we said the LSTM neural network has a cell that remembers values over arbitrary time intervals, the number of time intervals the neural network will be able to remember is 30 (past 30 days). We will only use one feature (multivariate high-correlated features added an insignificant improvement in our case). The training data will be divided in batches of 32 days. The input shape of the LSTM is a 3D tensor with the shape of [32, 30, 1] as [batch size, time steps, Features].The loss function used in the model is the MAE over MSE due to it's more sensible to slights errors.

The layout of the model is:

- Input: 3D tensor (batch-size=32, 30, 1)
- First layer: LSTM (activation=hyperbolic tangent, output=32)
- Second layer: LSTM (activation=relu, output=16)
- Last layer: Dense layer (output = 1)
- Loss function: MAE

Before the training we normalized the values to speed up the learning. The fit of the network lasted for 50 epochs. The last 33% of the training data was used as validation. So the size of the training data was 3670 and the validation size 1809. The steps per epoch were 114. The training was done on the CPU and time training for each model was about 1 to 2 minutes. As before we test the data for the years 2016 to may 2018 (851 values)

### *4.2 One day ahead prediction evaluation*

| LSTM (lag=30) | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | 0.5627 | 0.2907 | 0.4598 |
| O_3 | 0.581 | 67.6704 | 6.4914 |
| NO_2 | 0.39 | 176.535 | 10.6989 |
| PM10 | 0.2986 | 30.716 | 2.9513 |

The biggest improvement was with the O3 model which managed to decrease from the 136.62 MSE with MLP to 67.6704 and the PM10 model from a MSE of 117.46 with MLP to 30.71.

### *4.3 Two day ahead prediction evaluation*

| LSTM(lag = 30) | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | 0.5414 | 0.3048 | 0.425 |
| O_3 | 0.4412 | 90.2942 | 7.4544 |
| NO_2 | 0.2169 | 226.8861 | 12.424 |
| PM10 | 0.0583 | 41.2825 | 3.8583 |

Great improvement for the SO2 model over MLP and also decrease by half of the MSE for the O3 and PM10 models.

### *4.4 Next month maximum value prediction*

For the monthly data the number of lag features that the network receive is 12 (1 year) and the batch size set to 1, so the input is [1, 12, 1]. The training data was split into training and validation also, each epoch had 108 steps the total training data was 180, the 72 last months were used for validation.

| LSTM(lag = 12) | R2 | MSE | MAE |
|---|---|---|---|
| SO_2 | -3.7422 | 41.0353 | 6.0285 |
| O_3 | 0.4965 | 106.5104 | 8.3576 |
| NO_2 | 0.6075 | 134.1464 | 9.3314 |
| PM10 | 0.0562 | 763.0872 | 15.8071 |

SO2 model perform very bad. Values achieves with the triple exponential smoothing are better for the SO2, O3 and NO2 pollutants while the LSTM happen to predict better the PM10 pollutant respect the exponential smoothing model, 15.61 over 23.94.

### *4.5 Evaluation*

During the training the loss and validation accuracy never converged for any of the 4 models

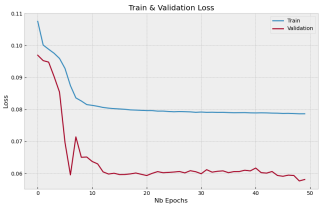



Fig. 5. Training evaluation of the PM10 one day ahead model

## 5. FINAL RESULTS

### *5.1 One day ahead models*

To predict the pollutant concentration for one day ahead and two day ahead the LSTM model performed the best achieving the next MAEs for the different pollutants:

$SO2 = 0.4598, O3 = 6.4914, NO2 = 10.6989, PM10 = 2.9513$

Except for the NO2 model were the MLP model got a lower MAE=9.213, but still either model achieve the NO2 threshold. The threshold was: SO2<1.18, O3<5.40, NO2<4.70, PM10<3.80, it was achieved for the SO2 and PM10 models, and the O3 was close en enough with 6.49 over 5.40. The only model that not represent well the pollutant concentration is the NO2 model.

### *5.2 Two day ahead models*

LSTM models worked the best achieving the next MAEs and surpassing the threshold for the models of the SO2 and PM10 pollutants.

$SO2 = 0.425, O3 = 7.4544, NO2 = 12.424, PM10 = 3.8583$

As before the MLR NO2 model is slightly better with a MAE of 11.628.

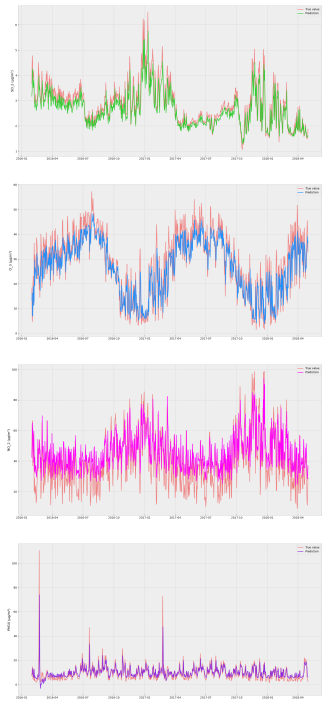


Fig. 6. One day ahead prediction for SO2, O3, PM10 with LSTM models and NO2 with MLP model

### *5.3 Comparison of the next month maximum models*

Triple exponential smoothing models were the best to predict the maximum pollutant values with monthly data. Except for the PM10 pollutant where the LSTM model achieve better results.

Exponential smoothing models MAE's: SO2=1.38, O3=6.46, NO2=7.37, PM10=23.94

LSTM models MAE's: SO2=6.02, O3=8.35, NO2=9.33, PM10=15.80

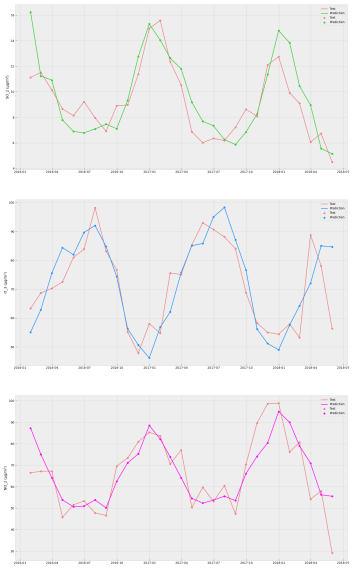


Fig. 7. Next month maximum value prediction for SO2, O3 and PM10 concentration with Triple Exponential Smoothing models