Noise in Leuven

Modern Data Analytics [GOZ39a]

Michalopoulos Tryfon [r0829272]

Introduction

The purpose of this project is to analyze any existing underlying patterns involved with the constantly growing issue of noise in Leuven. There was an attempt to give an answer mainly to two scientific questions:

1. Can we identify the factors (latent or not) that are correlated with or contribute to the issue of noise in Leuven?
2. What is the projected trend for the future? Will the noise problem continue to worsen?

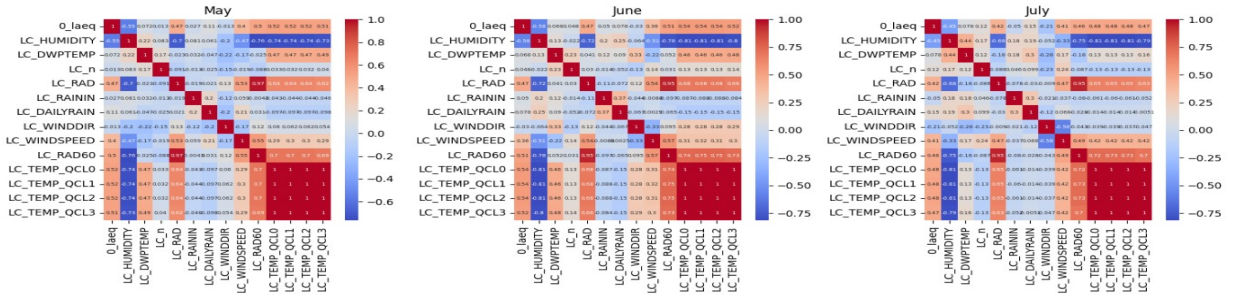
Data extraction and preprocessing

The basis of this analysis was made upon two datasets coming from completely different sources; the primary one that consists of data gathered by researchers at KU Leuven of enormous volume (data that has frequency of 1 sec) capturing 11 features from 9 different locations within Leuven, while the secondary one, including records capturing historical meteorological information, has been collected and is distributed by Leuven.cool network. Shortly after opening up the noise-related dataset, we encounter the issue of memory overflow, which leads us to no other option but to downsample the dataset. In particular, the dataset is resampled and aggregated over windows with intervals of 30 minutes (as opposed to the 1sec intervals that it initially had) and some basic imputation for the missing values takes place, by linear interpolation over (timewise) nearby values. The exact same transformation/aggregation is applied to the secondary dataset also, and thus, conveniently both datasets can use the timestamp values, which align perfectly, as the index, and the key for merging the tables together, if needed. Linear interpolation, as mentioned before, can be applied to impute missing values, but for parts of the noise-dataset there are huge intervals where no data was captured. For those large data-missing intervals, there was a thought (which actually was initially implemented) that the imputation should took place horizontally, meaning that missing values should be computed by a weighted average based on the distances of the locations of interest. However, this approach was later proven to lead to catastrophic results by introducing both collinearity and noise to the dataset and since, as already discussed, the magnitude of the data was way larger than required for our needs, this method was dropped out. A representative part of the dataset after all the preprocessing steps have taken place is shown in the figure below.

The Analysis

Early on, exploratory analysis reveals to us that most (if not all) features included within the Noise-dataset are highly correlated; this has to do with the fact that some of these features can be derived as a linear/non-linear combination or aggregation over a window of the others, meaning that we can as well just focus on one variable, since any insights can be extended along to other features also; for

this reason our focus is on the Equivalent sound level (average sound level over a specific time period), 'laeq' variable. For demonstration purposes the correlation of the 'laeq' variable against other variables from the meteo- dataset is presented on the figure below. Although only three months are presented here, we expect that there should be a seasonal pattern at a monthly level so we could effectively extrapolate our knowledge on other months also. Clearly the variables 'humidity', 'solar radiation', 'windspeed' and 'temperature' seem to correlate heavily with the variable of interest. For confirmation we proceed with a granger causality test between 'laeq' and all four variables mentioned above, which solidifies the fact that there is some sort of causation/correlation between those.



Given the strong correlation between the variable of interest and the rest of the meteorological variables we might consider that these variables could carry information and thus a multivariate time series model could be fitted so that we could predict the 'laeq' variable and consequently build intuition on how the phenomenon is evolving in Leuven (e.g. is it increasing/decreasing, what seasonal patterns does it show, which variables affect it the most etc.). But first we fit two univariate time series models, a so-called 'naïve' model and a sarimax model, for which the optimal parameters are found to be: AR component = 5, the order of differencing = 1 and the moving average component equal to 5. These two models yield an MSE equal to ~40 and ~39 respectively, and basically act as a benchmark for the model that we fit next , a LSTM, which takes in as an input all 5 variables that seemed to have an impact on 'laeq' and regresses on 'laeq' itself, which yields an MSE equal to ~37.

Final thoughts

Although a vast number of different configurations were tried out for each of the different locations (each location is affiliated with a different Neural network), none of them managed to outperform the baseline classifiers by a satisfactory margin; thus, although there clearly is some relationship between the aforementioned variables of the two datasets, we cannot confidently claim that there clearly exists a strong connection between them. For demonstration purposes, the prediction for the 'parkstraat' sensor is shown in the figure below (only showing a subset of the predicted interval).

Overall, this analysis provides insights into the noise issue in Leuven and lays the foundation for further investigation and improvement of predictive models.