FACULTY OF SCIENCE

KU LEUVEN

Modern Data Analytics (G0Z39a)

Assignment Report - Kosovo

Ngoc-Thien Nguyen (r0774196)
Wenting Jiang (r0824739)
Frida Trepca (r0730756)
Christian Daniel Sepulveda Arzuza (r0821283)
Rebecca Gösker (r0923036)

June 4th, 2023

Contents

1 Introduction

As urban spaces grow, policy makers have been increasingly confronted with managing noise propagation in cities, as they aim to balance quality of private and public life. It is therefore imperative to solve issues around noise management. In the age of big data, how can we make use of the vast amount of information available to us to enact better city planning that reduces and limits noise pollution? And why are some locations that are geographically very close still noisier than others? This report aims to answer these questions by utilising comprehensive data on Leuven as a case study, to predict noise levels and find possible solutions to mitigate the noise problem.

The city of Leuven is trying to tackle the noise issues in the Naamsestraat neighborhood, a main street of the city, that combines the hustle and bustle of the city by day, as well as a very active nightlife. The city does not want residents to be bothered by unnecessary noise levels, so, via the help of sound monitors and behavioral interventions, we can help mitigate this issue. Many studies have studied the effect of different urban characteristics (man-made) while holding weather conditions stable, so it may be necessary to look at weather as controls, while exploring the issue.

2 Datasets and pre-processing

Two datasets made available for analysis during the course of the project were used: one containing noise levels during the year 2022, and one containing Meteo data during the same period. The weather data was combined with noise using relative distance, and noise aggregated on hour-date basis. Two locations with almost no data were removed.

Additionally, since many other factors contribute to urban noise levels, external data was gathered for a more comprehensive analysis. Firstly, data festivals/concerts held throughout 2022 was collected, including the FIFA World Cup dates, as the events of e.g. public viewings and parties could have significantly increased noise levels. We also downloaded the locations of the bars/pubs close to these locations of interest, and the number of bars/pubs in a radius of 15m, 30m and 50m is aggregated to serve as an additional feature. Another source of external data is the transportation networks, such as bus and cycling routes for bus. Both of these were available on OpenStreetMap. As such, the file in use (noisemeta) was created using google maps, in order to get the exact coordinates for the noise locations. Finally, since Leuven has a considerable student population, data on the semester dates, as well as significant dates/holidays within the semester, was obtained.

3 Modelling

First, the data was used to build a linear regression model.However, in order to understand the factors that are associated with noise, a machine learning model was developed. Different modelling techniques were experimented, and we chose Random Forest Regression due to its performance on the cross-validated test set.

As this model is not interpretable, we used an explanation technique, namely, SHAP. The reason for using SHAP instead of other explanation methods, is to provide a local explanation for a single observation. This means, for a particular noise location at a particular minute, the model will output a predicted noise level based on the available features, and we visualize the relative importance of these features via a bar plot output by SHAP.

4 App development and deployment

An app has been created to display visually the distribution of max noise levels in Leuven. The app has two main features. First, at the top of the page, a heatmap that is updated based on the user input. The heatmap displays the levels of noise throughout the different locations of the measurements in Leuven. Second, the app visualizes a real-time graph of the maximum noise of each location by date. This is interesting to see, as its uses can be furthered on by live data, which would be very useful for a user to see: It could help in pinpointing the noisiest location of the neighborhood, and in the long-term evaluation of the street noise. Also, for residents of the city of Leuven, this could be a great way of visualizing the noise as well as the neighborhoods that they might prefer to avoid at certain times.

The deployment was done in AWS, specifically using AWS Elastic Beanstalk, which is PaaS (Platform as a service). An environment was created here, which is a collection of resources set up and managed by AWS. An EC2 instance of type t3.micro was created to serve as the host for the application, for this a role also had to be created for this instance. A database is not necessary for this deployment since the processed data is under 2 GB, which makes it available in the EC2 instance that acts as a host for the application. Finally, the application was served on Flask to provide default compatibility with this deployment option. This means that while Dash takes care of the front-end of the application, it is Flask that is handling the back-end as the server.

5 Tech stack

The model was built using Python. For data processing, we used polars, pandas and numpy. For machine learning, we used the standard libraries including sklearn and statsmodels. For app development, we used Dash, Dash-bootstrap components, Flask, folium, and plotly.

6 Results

Overall, the model provided a good prediction of noise levels using the features, given the scale of the data. Based on the SHAP explanation, the most important features were temperature, humidity, and hour, all three of which score above 1.0. As temperature and humidity increase, noise levels tend to be higher. These two features are theoretically related to each other, as hotter air is able to hold more water, propagating noise levels more effectively. Furthermore, a higher temperature might also make people more people go out, as it makes the environment more comfortable for humans, increasing the noise level in the street. The hour feature shows that noise levels are higher during the day, which makes sense given the lesser traffic, fewer public transport routes, and the lesser number of businesses opened at night.

Some features had surprisingly low SHAP values, such as transportation and events. This implies that within the time frame of the model, the traffic and public transportation routes did not have a strong predicting effect on noise levels. Given this, the most relevant factors for the noise result to be variables that cannot be controlled. However, this can help with planning, as now it is cleared that hot more humid days will be more critical and it would be worth to implement in those days additional noise control strategies. Furthermore, the build dash allows to clearly visualize the evolution in time and the distribution of noise in the locations.

7 Github repository and deployed app

Project's repository

Deployed application