Peaks and the Influence of Weather, Traffic, and Events on Particulate Pollution

Stefan Hagedorn[1], Kai-Uwe Sattler[1]

The task of the Data Science Challenge as part of the BTW 2019 conference is to analyze air quality data collected by the *luftdaten*[2] project. This project provides sensor measurements recorded from volunteers around the world. With do-it-yourself setups people can deploy their own sensors and report various environmental values to the project's servers, where they are made available as open data for further analyses. Thus, data is available only in regions where volunteers decided to participate in the project. Since in our city, Ilmenau, as well as in the state Thuringia only very few sensors are present, we decided to shift our focus to a broader area around Thuringia.

1 Frameworks, Technology & Preparation

Used Frameworks & Technology To investigate the schemata and contents of the many measurement files, we preferred a notebook system that let us easily explore the files. We use Apache Spark (and SparkSQL) in combination with our STARK[3] framework for spatial and temporal analyses, preprocessing as well as visualization. We additionally utilize Spark ML, e. g. for finding rules in the measurements. As programming languages we chose Scala, Python, and R.

Data Cleaning First analyses of downloaded data reveiled that alone for a single day, 2018-02-01, six different schemata exist. Some sensors report temperature and humidity only (`DHT22`), while others measure the particulate matter (`SDS011`). In addition to these two we found further six sensor types that measure e. g. air pressure, altitude or another level of particle concentration. We used the per-sensor measurement files provided in the *luftdaten* archive and integrated them using our Piglet[4] as well as Python scripts. In addition to the integration, a challenging task is to find invalid measurements, e. g. due to extreme weather condiditons, missing values, etc.

[1] Technische Universität Ilmenau, Databases & Information Systems, Ilmenau, first.last@tu-ilmenau.de
[2] `https://luftdaten.info/`
[3] `https://github.com/dbis-ilm/stark/`
[4] `https://github.com/dbis-ilm/piglet`

doi:10.18420/btw2019-ws-33

2 Analyses Goals

Particle Concentration Peaks Our first goal is to find peaks in the particle measurements for regions. Peaks are measurements exceeding the average particle value for a short period of time. Such peaks occur e. g. on New Year's Eve or during commuting hours. This is achieved by performing a spatial join with the sensors and a data set containing the regions of interest using STARK. Then, we calculate the average value per such region and subsequently filter the input measurements for values greater than that average. The result is visualized on a map, showing a timeline per region when and by how much the average particle concentration was exceeded. This information can be used to learn how long such extreme air pollution persists. Since this also depends on the weather conditions, we use values from weather data sets provided by the National Oceanic and Athmospheric Administration[5].

Weather Influence The weather conditions have an impact on the particle concentration measured by the sensors. Our hypothesis is that with rainy weather more people use their car instead of walking or taking the bike. However, since the rain will clean the air, we expect not many peaks. In order to support this hypothesis, we plan to additionally integrate open traffic data that can be found on various open data portals[6][7]. Besides the case for rainy weather, we also look at the opposite weather condition: Do people use the car when it gets too hot?

The overall idea is to find weather conditions when many people (or more people than usual) will use their private vehicles and thus produce more particulates. The results are temperature and precipitation values (lowest and highest) where the particle concentration rises above average.

Particulate Matter Polution Prediction & Correlation We employ a frequent itemset mining approach to find rules to predict the particle concentration and limit exceedings from the weather conditions and current volume of traffic. For this, we use the traffic data as well as "historic" weather data sets to categorize the particulate pollution. An additional correlation is to be tested between "events" (such as the begin of New Year's Eve, begin/end of school vacations, or concerts, football matches etc.) in cities and the air pollution. Since the *luftdaten* project continuously collects new values, the set of existing values is used as training data and future values are used for validation.

The result of this analysis is a set of rules that can be used to predict the range in which the particle concentration will be under forecasted weather conditions and traffic volume.

[5] https://www.ncdc.noaa.gov/cdo-web/datasets#GHCND
[6] http://opentraffic.io/
[7] http://govdata.de