**Analyzing air quality, providing low-pollution bicycle routing in Stuttgart, Germany**

Seminar: GIS Analyses with Free and Open-Source Software

Institute of Geography

Teacher: Christina Ludwig, M.Sc.

|  |  |
| --- | --- |
| **Amandus Butzer**  Matriculation number:  B.Sc. Geography  @stud.uni-heidelberg.de | **Julian Käflein**  Matriculation number: 2932816  B.Sc. Geography  j.kaeflein@stud.uni-heidelberg.de |

**0. Abstract**

Airborne particulate matter (PM) is in the focus of public health concern. In 2016, the city of Stuttgart, Germany introduced measures to reduce the PM levels within the city.

In this study, two sources of data to analyse the PM-concentration were used:

*Citizen-science* data from Luftdaten.info project using low-cost, self-built sensors; and official air quality sensor data by the State Office for the Environment, Measurements and Nature Conservation of the Federal State of Baden-Württemberg (LUBW).

The hourly data was automatically downloaded from the respective API, ingested into a database, filtered for outliers, then passed to a raster-interpolation process and a bicycle routing application.

An in-short analysis of comparability between the two datasets showed overall agreement in terms of general PM-concentration trends, yet considerable discrepancies (#how much?) on daily concentration levels.

A comparison of raster interpolation process showed (#was?)

Based on the resulting data, a bicycle routing application was developed, allowing users to avoid areas of high average PM-concentration.

**1. Introduction**

Since the introduction of a public alert for high particulate matter (PM) concentrations (“Feinstaubalarm”) in 2016, the issue of air quality in the city of Stuttgart, Germany, has been in the focus of media and scientific interest due to increased levels of particulate matter.

Atmospheric particulate matter (PM) is solid or liquid matter suspended in the atmosphere with diameters of under 10 µm (coarse-particle matter) or under 2.5 µm (fine-particle matter) (Fuzzi et al. 2015 pp. 8228). In the past years, there has been a high interest in particulate matter and its effects on human health (Fuzzi et al. 2015 p. 8217).

Short-term exposure to particulate matter can cause inflammation of the respiratory system, immune response and oxidative stress to the affected cells (Ristovski, Z. et al. 2005, p. 205).

Long term exposure to elevated levels of particulate matter can lead to increased mortiality due to nonallergic respiratory morbidity, allergic illness and symptoms (such as asthma), cardiovascular morbidity, cancer, and effects pregnancy, birth outcomes and male fertility (Heinrich et al 2005, pp. 132).

It is also the area of interest of an ongoing project by the German Federal Ministry of Transport and Digital Infrastructure (BMVI): The Satellite-based system for displaying, predicting and simulating air pollutants for sustainable urban and regional development ("Satellitenbasiertes System zur Anzeige, Prognose und Simulation von Luftschadstoffen für eine nachhaltige Stadt- und Regionalentwicklung - SAUBER").

At the same time, while cycling has no direct adverse effect on the air quality - cyclists themselves are directly subjected to air pollution within a city, as cycle paths are usually in immediate proximity to roads. Analyzing and mapping this concentration and distribution of particulate matter - and to map a 'least-polluted' route through particularly affected areas - is a step towards informing about - and protecting users from - air pollution in the city of Stuttgart.

It is also a step towards informing decision-making processes e.g. for guiding the traffic of combustion engine vehicles and possibly enforce restrictions on it in order to reduce air pollution.

It is vital for the analysis to combine official datasets with more extensive open source data, as a recent study on the impact of driving bans in Stuttgart states that "The sparsity [of the sensor network] introduces uncertainty which cannot be modeled correctly or can not be modeled at all"(Wolfmann et al. 2019, pp. 295)

The goal of this analysis is thus to provide a small-scale geographic analysis of the particulate matter concentration in the city of Stuttgart; comparing official and *citizen-science* data quality; making the resulting map available on a user-interface; and allowing users to combine the results with a routing service informed by the result of the analysis.

**2. Literature research**

Using PM monitoring site data, LI ET AL. (2017) offer a literature review of interpolation methods (LI ET AL. 2017, pp. 9103) and identify the Inverse Distance Weight (IDW) method as most widely used in epidemiological studies on air pollution (LI ET AL. 2017, p. 9104).

In an early approach of mapping air pollution using data from low-cost sensors, Briggs et al. mention Kriging as standard method of geographic interpolation; but developed a multiple regression model fitted for different cities individually (Briggs et al. 1997, p. 701). The use using weighted land cover instead of simple distance weight (Briggs et al. 1997, p. 707) appears useful for inner-city mapping. For modelling air pollution, Matějíček et al. (2006) used a simple Inverse Distance Weighting method and ordinary Kriging with estimates of variability (Matějíček et al. 2006, p. 266).

The Office for the Environment, Measurements and Nature Conservation of the Federal State of Baden-Württemberg (LUBW) has conducted research into the comparability of the Nova Fitness SDS011 sensor used by Luftdaten.info (see 3.1), and a calibrated sensor by Grimm GmbH. The SDS011 sensor performs comparable to sensors used by the LUBW in conditions of 50-70% air humidity and below 20µg/m3 PM concentration. Above these thresholds, or under changing atmospheric conditions, the SDS011 data significantly higher values (LUBW 2017, pp.17). The tests were conducted under ‘laboratory’ conditions and direct comparison between one LUBW measuring station and one nearby SDS011 (LUBW 2017, pp. 17).

**3. Data**

**3.1 Particulate matter sensor data**

Air quality measurements in Stuttgart and the State of Baden-Württemberg is conducted by the State Office for the Environment, Measurements and Nature Conservation of the Federal State of Baden-Württemberg (LUBW). The LUBW operates 44 measuring stations within the State of Baden-Württemberg, with 5 stations located within Stuttgart itself. Upon request, direct access to the measuring station data was not provided by the LUBW. Thus, the first goal of the project will be to access and automatically download the output of the LUBW's publicly available data at <https://www.lubw.baden-wuerttemberg.de/luft/messwerte-immissionswerte>. The site provides an hourly average of Ozone (O3) and Nitrogen dioxide (NO2); and a moving 24-hour average of particulate matter (PM2.5, PM10). The data is provided in the GeoJSON format.

To supplement these measurements, the Stuttgart-based project Luftdaten.info was brought into life. As a 'Citizen Science'-project, it offers shopping and building instructions for inexpensive particulate matter sensors (mainly the Nova Fitness Co., Ltd. SDS011), and multiple APIs at <https://github.com/opendata-stuttgart/meta/wiki/APIs> to access the recorded data, which are queryable for e.g. geographic bounding boxes, and with a temporal resolution of down to 5 minutes. The Luftdaten.info-data is provided in JSON formatting.

The strength of the project is the relatively very high density of sensors - especially within the city of Stuttgart. A simple bounding-box query of all active sensor around in the greater area of Stuttgart has yielded 1350 sensors. Apart from the non-standarized installation of the sensor, the accuracy of measurements of the SDS011 sensor was investigated by the LUBW (2017) and discussed in Blon (2017).

Direct comparisons of the accuracy of the PM-measuring devices used in the Luftdaten.info project has shown that while the SDS011 perform measurements comparable to the calibrated, high-accuracy sensors used by the LUBW in conditions of 50-70% air humidity and below 20µg/m3 PM concentration. Above these thresholds, or under changing atmospheric conditions respectively, the reported concentrations by the SDS011 show significant discrepancies (LUBW 2017, p.5).

Since the smallest temporal resolution offered by the LUBW is one hour, and the accuracy of the SDS011 data seems to benefit from using aggregate data, the project's aim is to use hourly data from the Luftdaten.info project. Additionally, subsequent analyses based on this project can utilize different queries on Luftdaten.info APIs in terms of geographical area, temporal resolution, etc. with little effort.

The hourly data can be categorized by using the *Common Air Quality Index* (CAQI) which was standardized for yearly, daily and hourly measures ([3])

**3.2 Road network**

#Kurz zum OSM Project, etc.

Automatic download from OpenStreetMap

**4. Methodology**

**4.1 Data model**

In order to organize the stream of relevant input sensor data, a time-series data model for low redundancy and disk usage was developed:

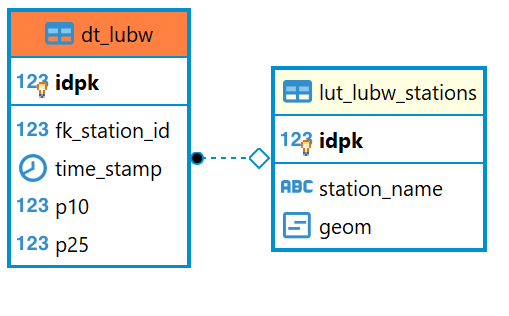


Figure : Entity relationship data model

The E-R diagram shows the data model. The actual sensor data is stored in a data table containing each hourly data point as one row. The columns consist of

* idpk: an individual primary key / ID
* fk\_station\_id: the station ID, which is assigned during parsing
* time\_stamp: the time stamp of the measurement
* p10: the PM10 value
* p25: the PM25 value

To prevent having to store non-changing values such as the name and location of the associated sensor on disk, the data tables for both API data contain a column “fk\_station\_id” to reference this information once for each sensor in a lookup-table as a foreign key.

**4.2 Data download**

To store the expectedly numerous datapoints, a database, including a fitting data-model was developed. PostGIS, an extension to the Free and Open Source relational Database Management System PostgreSQL, offers support of geographic data types and simple geographical analyses.

In order to obtain a complete, 24-hour dataset from the respective APIs, automated download scripts were developed to access, download and ingest the data, which were run by the task scheduler *cron.* For these hourly values, both APIs did not offer any filter on retrieval. Thus, the “global” datasets from each API for each hour were downloaded.

This was realized via two *shell-*scripts (#see Github repository). The scripts download the data using the *cURL*-library. In case of a failed download due to temporary connectivity problems, the scripts will reattempt downloading the dataset every 30 seconds.

On success, the files are stored on the host system under given a unique name based on the time of download.

As the LUBW API already provided valid GeoJSON data, the *ogr2ogr* command of the GDAL library was used to insert each file into a temporary database table. Afterwards, a parser function is invoked by the script. Since the input data contains the name of each sensor location, the function finds all values matching the name “Stuttgart”, then distributes the values according to the data model described in #4.1.

The Luftdaten.info data is similarly ingested into a temporary database table. Due to the structure of the data, first the geographical location of each sensor is built from lat/lon values, then a geographical filter is applied by intersecting the values with a reference polygon for the area of interest. Lastly, due to apparent non-valid values for the PM-concentration, a preliminary filter discarding values higher than 200µg/m³ is applied before the values are ingested into the persistent data- and lookup-tables.

#OSM download

**4.3 Data** **pre-processing**

Access to the stored data was realized within the database via stored SQL queries (“views”).

While LUBW data showed no obvious outliers, Luftdaten.api data was filtered for extreme PM-values during ingestion into the database (see 4.2). The remaining dataset still contained outlying measurements of several hundred percent above the city-wide mean value of the same timeframe.

In order to provide additional filtering for outliers, the database views on Luftdaten.info data calculates the 99th percentile of PM-concentration for the requested timeframe, then discards any value above it. Trials with this method showed that for timeframes without extreme values, the 99th percentile will fall directly on the highest recorded value, which in consequence is not discarded from the query.

For the bicycle routing application, a set of 24 views for each hour of the day were created via a batch script. Each view combines all filtered Luftdaten.info and LUBW sensors with their respective, average PM-concentration of the queried hour.

**4.4 Comparing official and crowd-sourced data**

Based on the findings concerning measurement accuracy of Luftdaten.info data in chapter 4.4, a simple comparison of both datasets regarding their comparability was performed.

Using an SQL-query, the PM-concentration recorded by the three LUBW stations was compared to the average of all Luftdaten.info sensors within a radius of 1000 meters for each timestep. The resulting line graph is shown in chapter #5.

**4.5 Interpolation**

Air quality and air pollution are spatio-temporal ('4-D') data. Meanwhile, the measurements of air pollution by official or distributed measuring stations are point-data.  
An analysis of the concentation of particulate matter within the city of Stuttgart needs to interpolate the measured concentration of particulate matter at one station with its surrounding environment.

#The Free and Open Source GIS *GRASS* wasused to perform the interpolation via…

**4.6 Categorization**

To facilitate the identification and graphical representation of polluted areas, the interpolated data is further categorized using the European Air Quality Index (EAQI).

The index is specifically developed for short-term air quality situations and is therefore the appropriate choice for the calculated hourly mean data set. The basis for the calculation of the EAQI are the concentration values of up to five key pollutants, namely particulate matter (PM10), fine particulate matter (PM2.5), ozone (O3), nitrogen dioxide (NO2) and sulphur dioxide (SO2).

The poorest level in any of the pollutants is decisive for the index level and it can therefore be used for single pollutants as well.

The driver of the index is the relative risks associated to short-term exposure to PM2.5 as defined by the Health Risk of Air Pollution in Europe project (HRAPIE project).

According to the report for an increase of 10yg/m3 in PM2.5, a higher daily mortality of up to 1.23% is estimated for different urban agglomerations in developed countries (WHO 2013).

The index levels are derived from this factor and the PM2.5 concentration values correspond to the following air quality standard:

- 0 - 10: Good

- 10 - 20: Fair

- 20 - 25: Moderate

- 25 - 50: Poor

- 50 - 75: Very Poor

- 75 - 800: Extremely Poor

These categories are applied to the interpolated concentration raster using

#blabla functions

which are then converted to vector geometries using TODO: GRASS functions.

At this point the edges follow the pixel borders. As we are working with interpolated values pixel-perfect precision is not required. Therefore the vector geometries need to be simplified which also reduces the file size significantly. The resulting polygons are subsequently exported as geojson and transformed to EPSG:4326 projection to enable "avoid routing" in the webapp.

**4.6 Routing**

Java Project

**5. Results**

Data quality

Figure : Comparison of Luftdaten.info and LUBW daily average PM10 values.

Figure 3 shows the average daily values of PM10 for all LUBW stations and Luftdaten.info sensors within 1000 meters distance of each station in µg/m³.

The timeframe starts at 2020-01-02 to account for the large influx of particulate matter from fireworks on New Year’s Eve. The timeframe ends on 2020-02-30 due to the Luftdaten.info API being unresponsive during several days in March 2020. The time axis is parted in weekly increments from the starting date.

The figure shows that the aggregated Luftdaten.info sensor values are higher for most days of the timeframe. Out of a total 59 days, the Luftdaten.info network returned higher values than the LUBW for 47 days.

The two datasets display the largest discrepancy around 2020-01-23, but appear generally in agreement with another. On average, the absolute difference of PM10 values within the timeframe was ~3.1 µg/m³.

Figure 4: Factor of Lufdaten.info vs LUBW daily average values.

Figure 3 shows the collected data during the same timeframe in a single graph as a factor of the Lufdaten.info daily average divided by the LUBW daily average.

While the previous figure shows the highest absolute discrepancies to lie around 2020-01-23, figure 3 shows the largest relative difference in the week between 2020-01-30 and 2020-02-06, where the Luftdaten.info network returned values up to 5.8 times higher than the LUBW stations. The average factorized difference between both datasets in the timeframe was appr. 2.8.

**#Mapping air quality in Stuttgart**

1-2 Karten + ganz kurze Erklärung

**#Routing**

1-2 Screenshots + Erklärung

**6. Discussion**

The following chapter critically discusses the results of this study shown in the previous chapter. The reliability of the analysis depends on two main factors: Data quality, and data processing.

The analysis of the relationship between LUBW and Luftdaten.info sensors has shown that by using *citizen-science* data, the amount and density of measurements of air quality can be greatly extended; both of which are crucial for attempts to model the air pollution of an area.

On the other hand, the comparison of both sensor networks, coinciding with results from field-laboratory tests in the literature, has shown that there is a noticeable, non-systemic discrepancy of Luftdaten.info data towards the reference LUBW data.

Using simple statistical methods, obvious outliers from the Ludtdaten.info dataset could already be excluded from the analysis and data quality significantly improved.

By incorporating known factors influencing the deviation of Lufdaten.info sensors and reference networks, such as temperature, wind speed and humidity, future studies using this sensor data should attempt to model and reduce the discrepancies between both networks; and enrich the resulting models with these additional atmospheric factors.

In summary, while the Lufdaten.info sensor network could not measure PM-levels in Stuttgart with enough accuracy to allow for the prediction of exact pollution levels, the network was capable of capturing the overall trend of PM-pollution.

In accordance with the conclusions of the data quality analysis that the combined sensor network is able to return the general trend of PM-pollution in Stuttgart, the resulting interpolated maps were categorized according to WHO air quality classification. Within this classification schema, the particulate matter pollution in Stuttgart can be modeled with comparable certainty, as opposed to giving the exact predicted PM-values for each location – creating a false sense of accuracy.

The comparison of different interpolation methods yielded the (# welcher? Gibt es einen?) algorithm as best suited for the study.

**7. Conclusion**

The aim of the project was to automatically download and store both official and citizen *citizen-science* data, perform a comparison of usability of both datasets, create maps showing areas of high particulate matter pollution and develop a routing app to avoid high-concentration areas.

The automatic download and ingestion were realized successfully. Database views were developed to filter outliers from the Luftdaten.info dataset. The comparison of both datasets shows discrepancies both in absolute and relative values, yet overall comparability over time.

Based on the results of the data-preprocessing, the one-dimensional point data was interpolated into classified raster maps in order to model the two-dimensional distribution of air pollution.

Finally, a Java application was developed, providing cyclists with the ability to find a route avoiding areas of high particulate matter concentration.

**8. Sources**

Heinrich, J. et al. (2005): Studies on health effects of transport-related air pollution. In: Krzyzanowski, M. et al.: Health effects of transport-related air pollution. WHO Regional Office Europe.

Li, L.; Losser, T.; Yorke, C.; Piltner, R. (2014): Fast Inverse Distance Weighting-Based Spatiotemporal Interpolation: A Web-Based Application of Interpolating Daily Fine Particulate Matter PM2.5 in the Contiguous U.S. Using Parallel Programming and k-d Tree. Int. J. Environ. Res. Public Health 2014, 11, 9101-9141.

Li, L. et al. (2014): Fast Inverse Distance Weighting-Based Spatiotemporal Interpolation: A Web-Based Application of Interpolating Daily Fine Particulate Matter PM2.5 in the Contiguous U.S. Using Parallel Programming and k-d Tree. In: International Journal of Environmental Research and Public Health 2014, 11 (9), 9101-9141.

Ristovski, Z. et al. (2012): Respiratory health effects of diesel particulate matter. In: Respirology (2012) 17, 201-212.

Woltmann, L. et al. (2019): Assessing the Impact of Driving Bans with Data Analysis. In: H. Meyer et al. (Hrsg.): BTW 2019 — Workshopband, Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 287. Online: https://dl.gi.de/bitstream/handle/20.500.12116/21819/G2-1.pdf?sequence=1