Analysis of Spatio-Temporal Data – Seminar Project Report

Caro Niebl

March 7, 2025

Abstract

Particulate matter (PM) pollution in urban areas can be a health risk to citizens and commuters. It is especially in the interest of cyclists to reduce their exposure to PM pollution in order to prevent health issues in the long term. Sensors and monitoring stations allow for the monitoring of environmental PM pollution, and Open Data platforms allow public access to the resulting data. As most data are collected by stationary sensors, a temporally continuous but spatially unchanging data set is provided. Thus, an accurate view of PM-pollution at the location of the sensors may be given, but for locations without a sensor, interpolated data may be less accurate. This seminar project explores the use of mobile bike sensors to complement stationary sensors in the assessment of PM2.5 (PM data for particles below 2.5 micrometer diameter) pollution within the city of Münster, to create an information product that visualizes the mean PM pollution on cycleways throughout the city.

1 Introduction

OpenSenseMap is an open data platform that utilizes citizen science for data collection and distribution [5]. On it, users can register sensors and publish the data that they have collected. A public API allows for this data to be downloaded and used. The senseBox:bike project offers a bicycle-mounted sensor-box that makes use of various sensors, including a particulate matter sensor [1]. Data collected from this sensor box can be published via openSenseMap and publically accessed. For this project, PM2.5 data from a small amount of stationary sensors and several mobile bicycle sensors is sourced from openSenseMap.

The OpenSenseMap platform contains several senseBox:Bike trajectories throughout the year 2024, mostly created by researchers and participants of data-collection campaigns and studies.

This seminar project uses senseBox:Bike PM sensor-data as well as stationary PM data from openSenseMap to interpolate and estimate the mean PM pollution throughout the city area during the months november to january. From this, an information product is designed that gives an overview over the mean PM pollution of Münster cycleways within the aforementioned time frame. This information product, a static PDF document, has been submitted in parallel with this report.

This Project report starts by outlining the Personas, user stories, and mockups created for this project, and adjusts the initial design mockup into a design product that is more feasible for this project. Then it continues by explaining the methodology that was used to arrive at this information product and closes up with a discussion and future work chapter.

2 Personas and User Stories

In the planning phase of this seminar project, two personas with corresponding user stories were created.

2.1 Sabine

Sabine is in her early 40s and works at the city's planning office as an urban planner. She lives in a home on the outskirts of town with her wife and kids. She carpools to work with up two two other colleagues, she owns the car. Sabine cycles for leisure and errands on the weekends, weather permitting. At her work she is currently responsible for proposing planning decisions regarding the town's bicycle infrastructure. She has a decent amount of experience in GIS.

2.1.1 User Story

As a city planner with an interest in the communal bicycle infrastructure,
I need to know which factors impact the bikeability in what areas of the city,
so I can propose decisions to improve the communal bicycle infrastructure.

As a leisurely cyclist,
I need to know which areas in my surroundings provide a sufficiently safe, comfortable, and healthy cycling experience,
so I can plan my cycling routes accordingly.

2.2 Detleff

Detleff just turned 34 and works in accounting at the city's planning office. He lives around 5 km away from his workplace, which is to him still a viable cycling distance. Due to his job he has experience in evaluating graphs and tables and reaching a conclusion from them. But due to not studying data science (he did consider it before choosing office management), he gets overwhelmed easily by information that are too complex or abstract. He is very conscious of the environment and prefers to commute by bicycle. However, he does not enjoy cycling in harsh weather and prefers to take the bus on rainy days. He has to avoid fine dust particulate matter pollution due to his asthma.

2.2.1 User Story

As a cyclist with asthma,
I need to know which conditions (geographical or time of day) are predictors of strong particulate matter pollution,
so I can avoid cycling within these conditions and keep my health intact.

2.3 Notes on Personas

These personas needs have been tailored to have some experience with more complicated information products to allow for the resulting information product to require less work to make it accessible. This was done because I estimated the data processing part of the project to take up a large amount of time available. However, I am planning to involve parts of the resulting work in the ongoing research on bikeability indices in the Situated Computing Lab, which will involve the creation of a more accessible information product in the form of an interactive dashboard.

3 Design of the Information Product

3.1 Mockup

The initial mockup of the resulting information product was a dashboard that allowed the viewing of PM pollution of the road network, alongside a slider to select time of day and time of year and weather conditions, with the acknowledgment that these latter options were only really possible to implement with sufficient information available.

3.2 Adjustments to the Information Product

During the development some issues became apparent that put further restraints on the design of the final information product:

1. During the time of development, there were only sufficient data available for the time from November to August 2024. For this reason it was not possible to select for seasons or time of year.

2. During data analysis, a consistent daily or weekly periodicity of PM pollution could not be found, so a time of day slider could not be implemented.

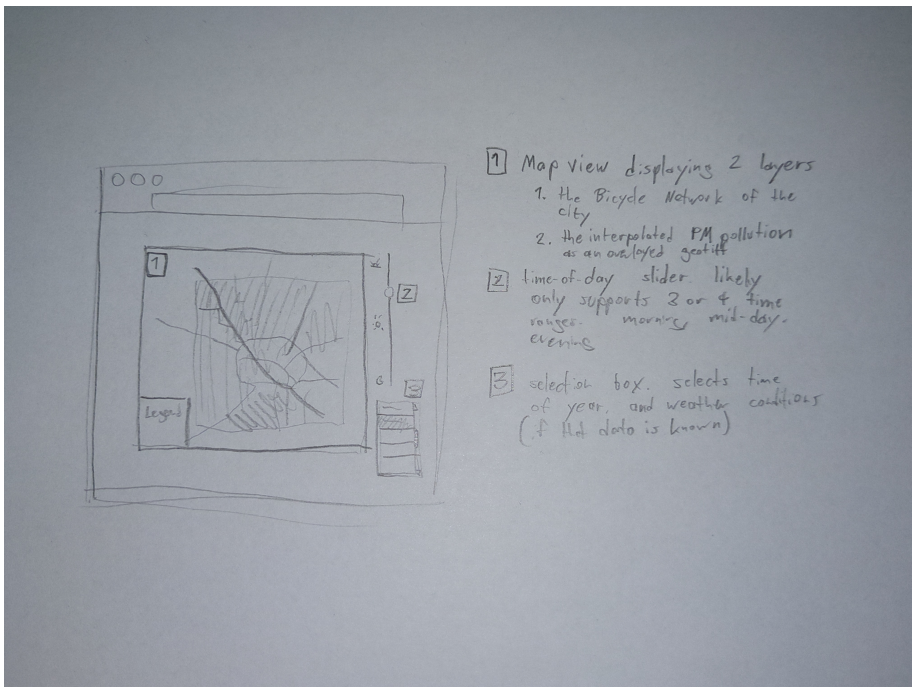

Figure 1: Initial sketch of the dashboard mockup

3. Because the development of basic functions took priority, the analysis of weather conditions became a stretch goal, that ultimately could not be met due to time constraints.

Because a season and time-of-day selection was not possible, and likewise the view of different weather conditions could not be implemented in time, the need for interactivity was strongly reduced. For this reason the design goal of the information product was shifted towards a static pdf document that displays a map of the city of Münster with visualizations of the normalized mean PM pollution of the cycleway network for the months of November to january.

4 Methodology

4.1 Data Description

The data used for the project describes the amount of PM2.5 particulate in µg/m^3, meaning particles suspended in the air below 2.5 micrometer diameter. The data consists of 9 continually recording stationary sensors and trajectories of 54 different mobile sensors from the time window from November 1st 2024 to January 31st 2025. All sensor time series were sourced from the OpenSenseMap open data platform [5]. The mobile sensor in question are bike-mounted sensebox:Bike [1] sensor boxes that contain a particulate matter sensor. While the stationary sensors record continually without interruption, the bicycle sensors' measurement periods are temporally constrained to the time of the bicycle rides in which they were utilized. Both stationary and mobile sensors record in one-second intervals.

It is to be noted that while stationary sensors are usually mounted in a more elevated location with some distance to roads, the bike mounted sensors tend to be recording at street level and may thusly have a bias towards higher PM measurements. Additionally, because the data is sourced from OpenSenseMap, it is not guaranteed that the stationary sensors are set up in accordance to regulations laid out in the 39. BImSchV [2]. For this reason the final product avoid making statements about µg/m^3 values, as discrepancies with regulation abiding sensor sources are otherwise possible.

4.2 Data Preparation

4.2.1 Adjusting for bias between measurement sources.

After removing outliers beyond 100 µg/m^3, a summary comparison (table 1) of the mobile and stationary dataset confirms the previously stated bias introduced by the bike sensor's closer proximity

to roads. Because both datasets will be used for spatiotemporal kriging for interpolation, quantile mapping (qmap) is used to transform the bicycle dataset to match quantiles with the stationary dataset. The resulting adjusted bike-sensor data is used henceforth.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
stationary sensors	0.00	2.10	5.10	8.159	11.40	100.00
bike sensors	0.32	4.60	11.65	15.11	20.50	99.73
qmap-adjusted bike sensors	0.00	2.10	5.068	8.095	11.372	100.00

Table 1: summary results of the stationary and mobile sensor data compared

4.2.2 Spatial and Temporal Data Aggregation

Data Aggregation Constraints. With around 100.000 observations from the mobile sensors and close to 900.000 observations from the stationary sensors, it becomes necessary to aggregate the data in order to decrease the expense of computing the empirical variogram and the subsequent kriging.

The main constraint of the degree of spatial and temporal aggregation of the PM data emerges from the differences in spatiotemporal structure of the bike sensor data and the stationary sensor data (see fig. 2). While the stationary sensors are stationary in location, they are recording data continuously. In contrast, the bike sensors record data in short trajectories at a time, with these trajectories traversing geographic space in a way that most measurements do not share the same point with another. It can be said that for the stationary sensors, variance in measurements stems from its spread in the temporal dimension while for the mobile sensor, measurement variance results from its spread in the geographical dimension.

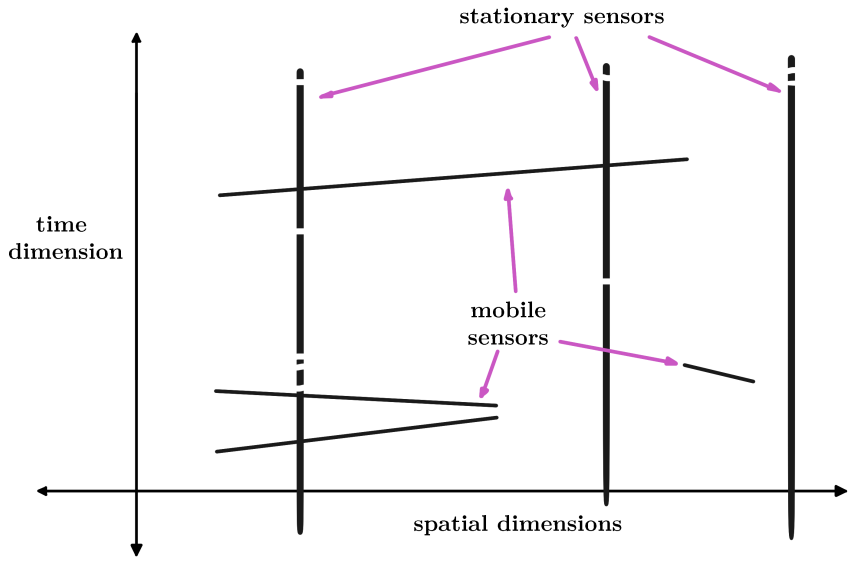

Figure 2: Temporal and geographical spread of bike sensor and stationary sensor data. The stationary data is spread temporally while being spatially constant. The sensor data is spread spatially while strongly resitricted to small temporal windows.

For this reason the scale at which to aggregate the data, spatially and temporally, needs to be chosen carefully. While the size of spatial cells needs to be chosen in a way that aggregates sufficient data points to even out strong local outliers, time windows too large can cause the aggregated values of cells containing stationary sensors to differ too strongly from those containing only mobile sensors.

initial parameter selection. The initial aggregation parameters were spatial grid cells of 500 meter side length and time slices containing a day each. For each spatial grid cell, the mean of every measurement that is located within it at a given time slice is taken, and assigned to the

centroid of this grid cell. Grid cells with no measurements contained within them in a given time slice are discarded. The result is a collection of means distributed in space that is stored as an STIDF object to be used to compute an empirical variogram.

Observing entire days at a time did not result in usable spatiotemporal variograms; while the grid-cell means aggregated from stationary sensors were calculated from an entire day of PM measurements, those that were aggregated from mobile sensors were generally aggregated from only a few minutes of data, resulting in the grid cells aggregated from stationary data to deviate strongly from the mobile sensor data.

observing peak times for parameter selection. For the selection of better time ranges for data aggregation, a histogram of the amount of measurements by time of day is observed (fig. 3). Because the rate of measurement for stationary sensors is largely uniform, peaks in the distribution indicate times of day with increased bike sensor usage. Based on the histogram, two traffic peak times were chosen a day.

1. morning: 07:00-09:00
2. afternoon: 15:00-17:00

Each of these peak times within the observed time frame from November to February is assigned a time slice for which the cell means are computed like previously stated.

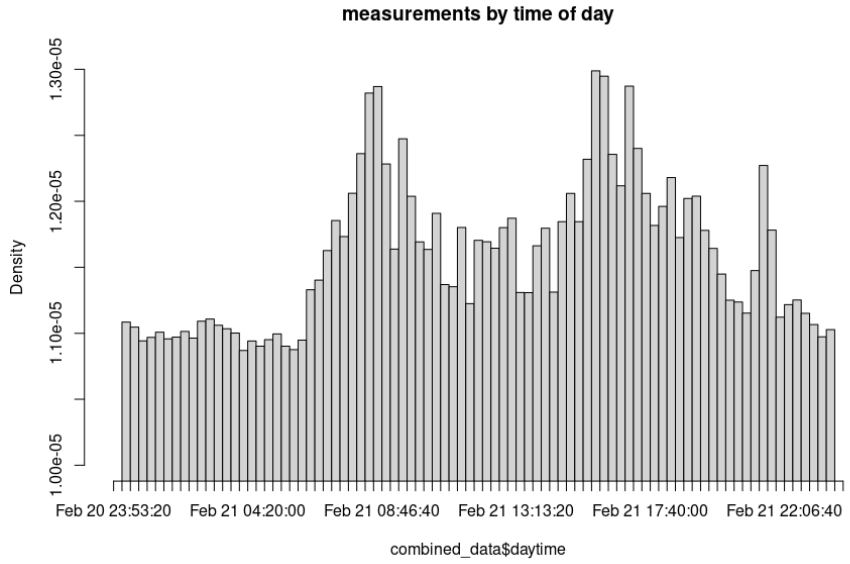

Figure 3: Amount of PM measuremets (mobile and stationary) by time of day

4.3 Computing and Fitting Spatiotemporal Variograms

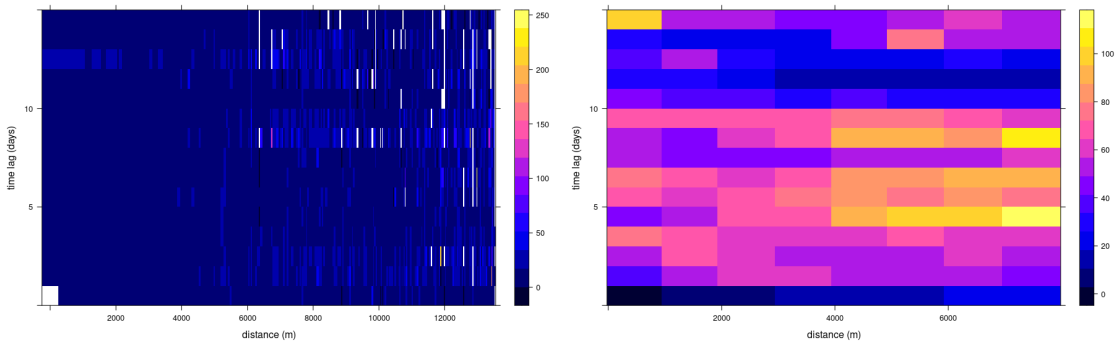

Figure 4: empirical variogram based on daily time-slices (left) compared to empirical variogram

4.3.1 computing empirical variograms

Computing the empirical variogram while tweaking the data aggregation parameters proved to be a time intensive step of trial and error that was slightly accelerated by making use of the jupyterhub instance provided by the university of Münster.

The first empirical variogram computed on daily aggregation of data resulted in a variogram with a barely changing variability (fig. 4, left). On a glance it appears to have a relative nugget very close to 1, which was indicative that the aggregation method needed to be refined.

After aggregating by peak times and learning how to correctly define an irregular time-lag for the variogramST function, the computed empirical variogram (fig. 4, right) is closer to one that a variogram model could be fitted on. On several attempts I observed a dip in variability after a lag of around 10 days, but for the time being could not determine why. The empirical variogram that I continue to work with is calculated with the time slices of November and December and a pre-set time lag of 20 slices (10 days).

4.3.2 fitting a variogram model.

Several methods and parameters were tried, as informed by a blogpost by Veronesi [6] and a Journal article by Gräler et al. [3]. The best fitting variogram model was one that used a simple sum metric method, but a mean square error lower than 211 could not be reached

4.4 Spatiotemporal Interpolation and subsequent use in information product

4.4.1 Kriging

Once the variogram model was fit, all conditions for kriging were met. Since for the final information product the mean PM pollution is to be determined, a spatiotemporal interpolation for each of the time slices is performed, and subsequently combined using the mean . The process of interpolating each time slice is time intensive but requires little oversight. The final result (fig. 5) is then upscaled bilinearily and cut with the mask of a 20x20 meter cell raster of the cycleways of Münster that was extracted from OpenStreetMap. This results in a raster of the Münster cycle network that contains the values of the interpolated mean PM pollution in each location. Finally the road raster with mean PM values is normalized and used in the final information product.

this is too clunky. do again

Figure 5: Mean of all interpolated time slices from 2024-11-01 to 2025-01-31

5 Discussion and Future Work

5.1 The Information Product.

The Information product being a static PDF document leaves room for improvement. Due to insufficient data it was not possible to facilitate the interactive components as originally planned, time constraints were an additional factor that motivated me to create a pdf instead. However, parts of the developed system are intended to be used in an interactive dashboard for visualising a bikeability indicator in urban areas as part of my work in the Situated Computing Lab at ifgi.

While the personas and user stories were tailored to fit the eventuality that I could not invest sufficient time into making the information product accessible to users without much prior GIS experience, it is in my opinion debatable whether Detleff's persona would be capable of understanding the information product. It would be dependent on whether or not Detleff has a sufficient grasp of means and normalization, which is possible but not guaranteed within accounting jobs.

5.2 Periodicity

while I could not satisfyingly determine a daily periodicity of observed PM data, possibly occurring daily periodicity may have already been captured by the peak-time aggregation. What I was able to determine was a difference in means between weekend and weekday data, with weekends having a higher mean of PM pollution than weekdays. This difference was found to be statistically significant using a permutation test. It appears counter-intuitive that the weekend has a higher PM pollution than weekdays, as one might assume workday commuting traffic to be responsible for a large amount of particulate matter pollution. Due to time constraints and not having a hypothesis I could not explore this phenomenon any further, but future work may be able to determine whether this difference is truly statistically significant, and if so explore what causes this difference between weekday- and weekend- PM values.

5.3 Correctness.

The correctness of the interpolation results has not been evaluated due to time reasons. Leaving out a random sample of trajectories and evaluating how well the interpolation fills in the missing data with the help of the remaining data would have been the method I would have used but this would only have given a view on the reliability of the interpolation method, not the data itself.

Because OpenSenseMap is an open data platform in which each user is responsible for their own sensor-setup, it can't be guaranteed that the sensor setups are up to the standards laid out in the 39. BImSchV [2], nor can it be guaranteed that the sensor data can otherwise be reflecting of the ground truth due to errors, hardware failures, environmental factors, etc.

comparison to official sources of PM2.5 data would be vital to ensure correctness of the developed method.

5.4 Efficiency.

Two steps of this method are computationally highly intensive: the computation of the empirical variogram, and the kriging of each time slice to later compute a mean of. While the empirical variogram only needs to be computed once, the kriging of the time slices is constantly required as new data comes in. The exploration of alternative spatiotemporal interpolation methods could be viable future work. The hybrid ordinary kriging inverse-distance-weighing approach outlined in Middya and Roy 2021 [4] would be a good candidate to accelerate the interpolation of each time slice.

5.5 Conclusion.

While a working method for interpolation was developed, a lot of room for improvement is still left. If done properly, the scope would far outweigh that of a 5 credit points seminar project. Working on this project has helped me gain a deeper understanding of spatiotemporal kriging methods and workflows and has equipped me with the needed knowledge to further refine the resulting work.

References

[1] Thomas Bartoschek, Verena Witte, Eric Thieme-Garmann, David Weigend, and Sergey Mukhametov. Citizen science on bikes in museums and schools: being part of mobility change research with the senseBox:bike. In *ARPHA Proceedings*, volume 1, pages 135–140. Pensoft Publishers. ISSN: 2683-0183.

[2] Bundesministerium der Justiz. 39. BImSchV - neununddreißigste verordnung zur durchführung des bundes-immissionsschutzgesetzes. https://www.gesetze-im-internet.de/bimschv$_3$9/$BJNR106510010.html$.

[3] Benedikt Gräler, Edzer Pebesma, and Gerard Heuvelink. Spatio-temporal interpolation using gstat. 8(1):204.

[4] Asif Iqbal Middya and Sarbani Roy. Spatial interpolation techniques on participatory sensing data. 7(3):1–32.

[5] Matthias Pfeil, Thomas Bartoschek, and Jan Alexander Wirwahn. OPENSENSEMAP - a citizen science platform for publishing and exploring sensor data as open data.

[6] Fabio Veronesi. R tutorial for spatial statistics: Spatio-temporal kriging in r. https://r-video-tutorial.blogspot.com/2015/08/spatio-temporal-kriging-in-r.html.