Spatial Interpolation of Air Quality Data with Multidimensional Gaussian Processes

Paul Tremper[1], Till Riedel[1], Matthias Budde[1 2]

Abstract: The central question of this paper is whether interpolation techniques applied to a distributed sensor network can indeed provide more information than using the constant background of an urban reference station to measure air pollution. We compare different interpolation techniques based on temporal-spatial machine learning in terms of their applicability for correctly predicting personal exposure. Using a dataset of stationary low-cost sensors, we estimate exposure on a route through the city and compare it to mobile measurements. The results show that while different machine learning-based interpolation methods yield quite different results, validation of machine learning-based approaches is still challenging.

Keywords: Air Quality; Gaussian Process Regression; Validation

1 Introduction

New fine-grained measurement networks promise to better quantify individual exposure to pollutants such as particulate matter. While official measurements look at large spatial and temporal integrals (daily averages for entire cities), the urban foreground can vary, and individual risk can vary widely based on daily routines. Since it seems impractical for individuals to carry dosimeters and simulations rely on typically incomplete information, low-cost sensor networks are becoming increasingly popular and have even been investigated by government agencies.

However, even with cheap sensors, there are limits to the density of such networks, and even dense networks have limited readings, typically spaced a few hundred meters apart in space and minutes apart in time. Therefore, measurement information must be interpolated, which means that even measurement systems always rely on data-driven predictions.

In this paper, we discuss how to evaluate basic temporal-spatial prediction methods against real data to improve the quality of machine learning (ML) algorithms in this domain and better understand relevant metrics.

[1] Karlsruhe Institute of Technology (KIT), TECO / Pervasive Computing Systems, Vincenz-Prießnitz-Straße 1, 76131 Karlsruhe, Deutschland tremper@teco.edu

[2] Disy Informationsssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe, Deutschland matthias.budde@disy.net

To do this, we follow a simple methodology. The basis for the analysis is a large-scale open dataset collected in Augsburg over a three-year period. After calibrating a set of low-cost sensor readings against a monthly background station mean, we validate different interpolations of low-cost sensor readings with calibrated reference stations. We show how the different algorithms perform based on typical regression metrics. Our analysis shows that the difference in prediction is not covered by the metrics.

To represent a realistic scenario in exposure quantification, we then predict the expected exposure on a bicycle route through the city and compare it to actual mobile in situ measurements. While the bike route measurements are difficult to calibrate and are not accurate, we hypothesize that a better interpolation methodology will nonetheless reduce the mean square error against this imperfect ground truth and serve as a better metric for machine learning algorithms that reconstruct the structure of propagating emissions.

The paper is structured as follows: In section 2, we present some related work before describing the dataset and our calibration procedure in section 3. In section 4, we then describe our first analysis, in which we use data of the low cost network to predict ground truth values. In section 5, we describe our second analysis, in which we attempt to predict a scenario of personal exposure along a path through the city. The latter two sections are split into data, methods and results. Finally, we conclude our paper in section 6 with an overall discussion followed by a conclusion.

2 Related Work

The basic assumption underlying any spatial interpolation is succinctly expressed in *Tobler's First Law of Geography* [To70]: "*Everything is related to everything else, but near things are more related than distant things*". The most simple approach that incorporates this idea is *Inverse Distance Weighting* (IDW) [Sh68].

Kriging [Kr51] is another popular approach in geoinformatics. Its mathematical basis are Gaussian Processes (GPs) which have recently received much attention as a general framework to deal with predictions under uncertainty and their close relation to unbounded neural networks. Kilibarda et al. e.g. introduced an automated mapping framework for predictions of daily air temperatures (mean, min and max) using regression-kriging for a resolution of 1km [Ki14]. Pebesma et al. [PH16] use copulas to enable spatio-temporal kriging: They show the application and benefit of their approach with a prediction of daily mean PM_{10} concentration. In [Br18], the residual error of Gaussian Process has been reduced by coupling them with a calibration scheme for low cost sensors within a black box optimization approach.

Landuse regression (LUR) is another popular approach for modeling spatial variation in various domains, among them also air quality [Ho08]. Such a modelling approach has been successfully applied to the same dataset that was used in this study [Sh20] . LUR predicts

spatial variations in air pollution based on correlations with the measurement context instead of the geospatial one. LUR modeling requires air pollution measurements at multiple locations across the study area, stationary monitoring used by LUR is typically at 20 to 100 locations, spread over the study areas. To summarize the predictor variables used in the LUR models, frequently used data include: area of land-use, road network or traffic information, physical geography such as elevation and slope, and meteorological data. In this sense LUR has the same issues as simulation as it actually relies heavily on the availability of external information. In addition, machine-learning approaches such as ensemble regression methods have been utilized to handle complex and nonlinear relationships that exist within data and produce forecasting models with comparable performance in practice. Based on our review of papers [Yu16, ZLH13, Li17, Su16] from the domain of spatial data analysis, since the prediction accuracy follows algorithm design, the machine-learning algorithms are crucial for building air quality prediction models, whereas statistical models have not been heavily used recently. Moreover, the random forest based approach is a prominent technique in selecting variables and inferring air pollution values.

3 Dataset

3.1 Description

We selected our data from the SmartAQnet[3] [Bu17] dataset. We focus on PM_{10} (Particulate Matter of diameter $< 10\ \mu g/m^3$ as pollutant and chose the month of September because of data availability. The SDS011 Crowdsensor sub-dataset consists of 1.479.343 data points across 35 devices, the EDM80 Scientific Scout sub-dataset consists of 253.220 data points across 35 devices. As references, we chose the official measurement stations within Augsburg, which give 2.756 data points across 4 stations.

SDS011 Crowdsensors are ~ 30€ sensor nodes, consisting of a Nova Fitness SDS011 fine dust sensor, an esp32 or esp8266 microcontroller and a BME280 or DHT22 humidity/temperature/pressure sensor, depending on the build[4]. In the context of this analysis, these differences in build are considered equivalent. These sensor nodes are considered in the ultra low cost range and well suited for use in large numbers and thus in various builds popular in citizen science contexts. While they have weaknesses in measuring precision, they perform reasonably well in measuring relative changes in air quality, which makes them an interesting component in large sensor networks.

EDM80 Scientific Scouts are considered low to mid cost sensors of $O(1000€)$, and come in two variants which we consider equivalent in the context of this analysis. The older

[3] https://www.smartaq.net

[4] A construction manual can be found, e.g., here: https://www.smartaq.net/en/participate/

variant (EDM80 NEPH) contains a nephelometer measuring cell, while the newer variant (EDM80 OPC) contains an Alphasense OPC (optical particle counter) measuring cell as fine dust sensor component. They have been developed, maintained and calibrated in project SmartAQnet by GRIMM[5], a company that produces and maintains air quality sensors. With regard to the SDS011 Crowdsensors, the scientific scouts are considered as a higher tier in quality.

Reference Stations are the official measuring stations of the Bayrisches Landesamt für Umwelt (Bavarian state agency for environment). There are four such stations located in Augsburg[6] which collect hourly values of various air quality parameters, including PM_{10}. In the context of this analysis, these stations are considered the ground truth.

3.2 Calibration Procedure

Off-the-shelf, low-cost light-scattering PM sensors often exhibit systematic deviation from ambient particle mass concentrations [Bu18] and therefore need to be calibrated.

First we took the monthly median of the official urban background station at Augsburg Bourges-Platz (15.0 $\mu g/m^3$) as our baseline to calibrate the sensors on. This choice is strictly speaking only valid in case of sensors which are also located in urban background environments. In the case of traffic environments, this will certainly under-calibrate the relevant sensors. Since we need to establish some kind of calibration on the sensors, we still proceed in this way and compare our interpolation results with the non-calibrated predictions later on to get information about the validity of the calibration. We note that for this reason, we do expect the calibration to be not 100% and thus leave an offset in the predictions. This method could be improved by taking into account the respective environment of each sensor and calibrating it against a reference station which is classified as the same environment.

Second we computed the monthly median for each single sensor (Crowdsensors and Scouts) and used the difference to the aforementioned baseline to shift the individual measurements. The effects of the calibration can be seen in subsection 4.3. For a discussion on why we used the median instead of the arithmetic mean, please see the discussion section.

[5] `https://grimm-aerosol.com`

[6] for a description of the official stations, see `https://www.lfu.bayern.de/luft/immissionsmessungen/dokumentation/index.htm` (in German)

4 Validation of Sensor Network Interpolation

4.1 Data

To validate the sensor network interpolation, we predicted the values of the four reference stations. Since the reference stations only give hourly values, we first had to aggregate the data to hourly values to be able to make predictions. We used the median instead of the arithmetic mean for the aggregation (again, see the discussion section). We then took data from the 21st and 22nd of September (48 hours) since the bike route, which we want to predict in the next section, falls into this interval. In that time, we have 31-33 active SDS011 Crowdsensors with a total of 1404 hourly values after aggregation in an area of 4.5 km x 7.9 km [lat x lon] (36 km^2) and 6-27 active EDM80 OPC Scientific Scouts with a total of 839 hourly values in an area of 10.2 km x 9.2 km [lat x lon] (94 km^2).

4.2 Methods

For the validation of the sensor network interpolation, we used two methods:

- One is the naive inverse distance weighted interpolation (IDW), which gives a prediction value v_p for each point by $v_p = N \sum_i \frac{v_i}{d_i^2}$ with the normalization constant N given by the condition $N \sum_i \frac{1}{d_i^2} = 1$. The d_i and v_i are the distances to each sensor and their values.
- The other method is Gaussian Process Regression (GP), which uses multidimensional Gaussian distributions to interpolate. We used an RBF Kernel (Radial Basis Function, $K(x, x') = \sigma^2 \exp\left(-(x - x')^2/2\ell^2\right)$) for the Gaussian Process Regression in every dimension (two spatial, one temporal). This kernel has one relevant parameter, the length scale ℓ, which is gauges the scale on which correlations between points are measured.

In case of the IDW, which is oblivious of the time coordinate, we performed an independent spatial interpolation for each set of hourly values. The GP, however, was trained on the full data of the two days. For that time frame, the GP finds RBF length scales of [0.00582, 0.00309, 1.38] (SDS011 Crowdsensor dataset) and [0.00785, 0.0149, 1.06] (Scientific Scouts dataset) in units of [lat,lon,hours]. To have better comparability, we adjust the length scales slightly and fix them at [0.00450,0.00680,1], which corresponds to ~500m in lat and lon dimensions, as well as one hour in the time dimension.

While we coded the inverse distance weighted interpolation ourselves (in python) using the formula above, we used the scikit-learn [Pe11] implementation of the Gaussian Process Regression.

4.3 Results

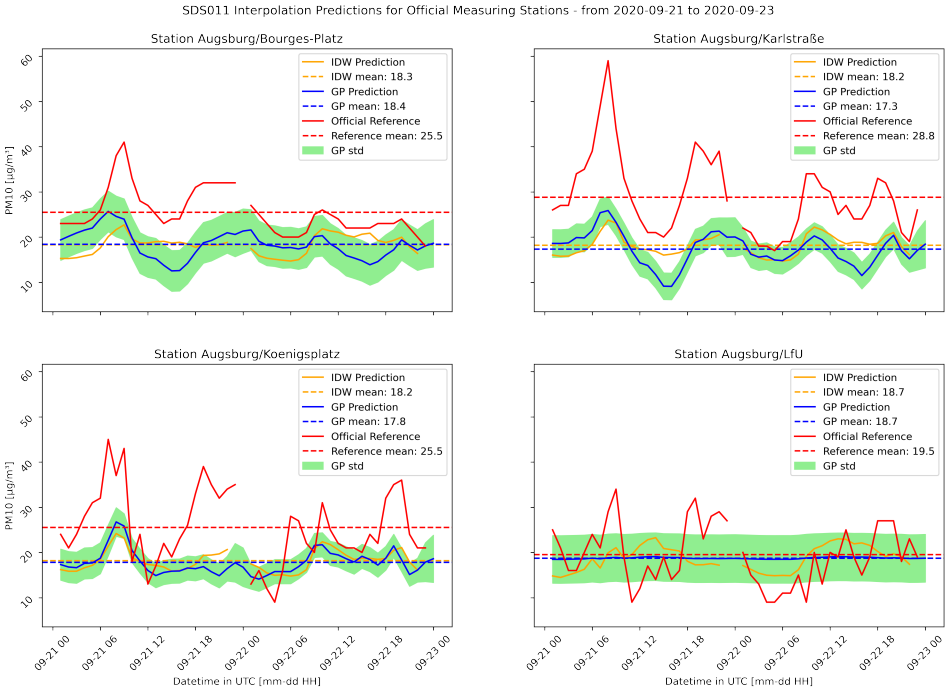

Fig. 1: SDS011 Crowdsensor dataset PM_{10} predictions for each of the four reference stations. The inverse distance weighted interpolation is in orange, the (2+1)D Gaussian Process in blue, the measurements of the official reference station in red, the standard deviation of the Gaussian Process Regression in green.

The results of the prediction of the four reference stations are show in Figure 1 (Figure 2) for the Crowdsensors (Scouts). We can see that in both cases, Crowdsensors and Scouts, the interpolations were able to capture the changes very well in case of Bourges-Platz and Karlstraße. The varying offsets of the prediction graphs to the ground truth graphs are systematic errors we will discuss later.

In case of the Crowdsensors (Figure 1), this is encouraging since the closest sensors to these stations are located 589m (Bourges-Platz) and 490m (Karlstraße) away. The length scale of the GP was set into that range to be sensitive to correlations of that length scale (ℓ = 500m, see above). We thus interpret the similar shapes of the prediction and the ground truth graphs as a sign that:

a) on the hourly level there are relevant correlations at the O(500m) scale in the data.

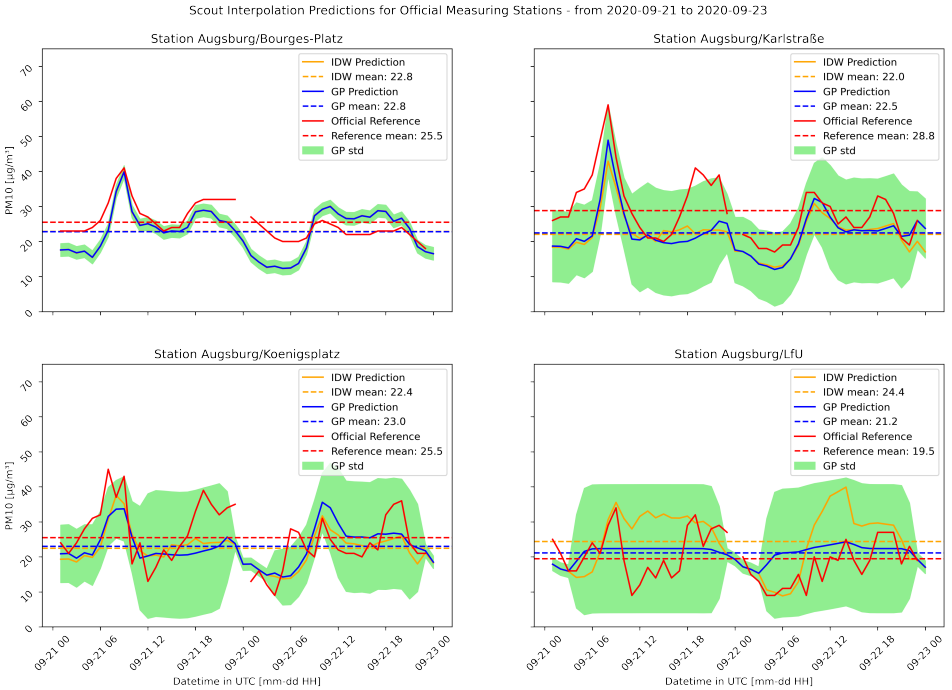

Fig. 2: Scientific Scout dataset PM_{10} predictions for each of the four reference stations. The inverse distance weighted interpolation is in orange, the (2+1)D Gaussian Process in blue, the measurements of the official reference station in red, the standard deviation of the Gaussian Process Regression in green.

b) a network of low cost sensors, like the SDS011 Sensors at hand, can produce useful predictions, if supported by interpolation methods. That is, if it is meshed tightly enough to be sensitive to that length scale.

c) The GP, as a representative of an AI based method that makes use of the temporal information, performs a lot better in tracing the graph of the ground truth than the naive IDW.

The Scientific Scouts are harder to interprete here. There is a Scientific Scout located closely to each of the reference stations (<10m). The results in Figure 2 show, however, that at least those at Karlstraße, Königsplatz and LfU seem not to produce any data. This is most notable at the size of the standard deviation band of the GP (in green), which represents a measure of confidence of the method in its own prediction. While the result at Bourges-Platz seems plausible for a sensor located rather close the the actual station (the IDW lying on top of it seems to back that), the predictions at the other three stations exhibit such large uncertainties, that it seems unlikely that the prediction is supported by a sensor very close.

Nevertheless, the Scouts were able to capture the shape of the ground truth very well in some parts for Karlstraße and Königsplatz.

For the LfU station, the behaviour of the GP for both, Crowdsensors and Scouts, shows that there is no sensor close enough to make a reliable prediction. This can be seen on the one hand by the large band of uncertainty, but most notably by the GP prediction dropping to the prior most of the time. And indeed, the closest Crowdsensor to the LfU station is located 1.163m away, which is more than twice the length scale of the GP. For the Scouts, the closest sensor is within 10 meters of the station, but the results suggests, that this Scout did not produce any measurements for most of the relevant timeframe.

The aforementioned offsets of most of the predictions are closely related to the calibration of the sensors. We indicated in subsection 3.2, that we expect the calibration to leave an offset because it underestimates foreground sensors. Therefore, a part of the systematic error that produces the remaining offset can be attributed to the under-calibrated foreground sensors in the Crowdsensor and Scout datasets.

In case of the Crowdsensors (Figure 1) this seems a plausible explanation as they underestimate Bourges-Platz, Karlstraße and Königsplatz by a somewhat comparable amount. LfU, being located far away in an suburban background, seems to match the means of the rather crude predictions by pure coincidence.

In case of the Scouts (Figure 2), the same argument applies and the pattern seems to be the same as for the Crowdsensors. The offset is considerably less, but that may well be due to them being higher quality sensors and individually calibrated during the duration of SmartAQnet a year ago. Also, if the ratio of background to foreground sensors is different in the dataset, this will also produce offsets of different sizes in the predictions.

Ref. Name	Ref. mean	IDW (nc)	IDW (c)	(2+1)D GP (nc)	(2+1)D GP (c)
Bourges-Platz	25.5	14.2	18.3	11.5	18.4
Karlstraße	28.8	12.8	18.2	5.6	17.3
Koenigsplatz	25.5	13.0	18.2	9.5	17.8
LfU	19.5	16.7	18.7	15.9	18.7

Tab. 1: Prediction means of the SDS011 Crowdsensor dataset for the reference stations. We give calibrated (c) and not-calibrated (nc) values.

Ref. Name	Ref. mean	IDW (nc)	IDW (c)	(2+1)D GP (nc)	(2+1)D GP (c)
Bourges-Platz	25.5	18.9	22.8	18.9	22.8
Karlstraße	28.8	18.8	22.0	19.6	22.5
Koenigsplatz	25.5	17.4	22.4	18.0	23.0
LfU	19.5	20.1	24.4	16.4	21.2

Tab. 2: Prediction means of the Scientific Scout dataset for the reference stations. We give calibrated (c) and not-calibrated (nc) values.

Table 1 (Table 2) shows the prediction means of each interpolation method for the SDS011 Crowdsensors (Scientific Scouts) at the locations of the reference stations. We see that in most cases, the calibration shift is substantial. We may safely disregard the reference at LfU, since it is a suburban background station, which we expect to be lower than both, the Crowdsensor and the Scout dataset, which are both located mostly in urban environment. We also see that (excluding LfU), not-calibrated (calibrated) predictions range between 19.4% - 55.7% (60.1% - 72.2%) for the Crowdsensor dataset and 65.3% - 74.1% (76.4% - 90.2%) for the Scientific Scouts dataset. We can see that the calibration, while not adjusting the sensors 100% accurate, substantially improved the predictions. We also observe, that the Crowdsensors improve more than the Scouts, which matches our expectation, since the scouts have been previously calibrated in project SmartAQnet.

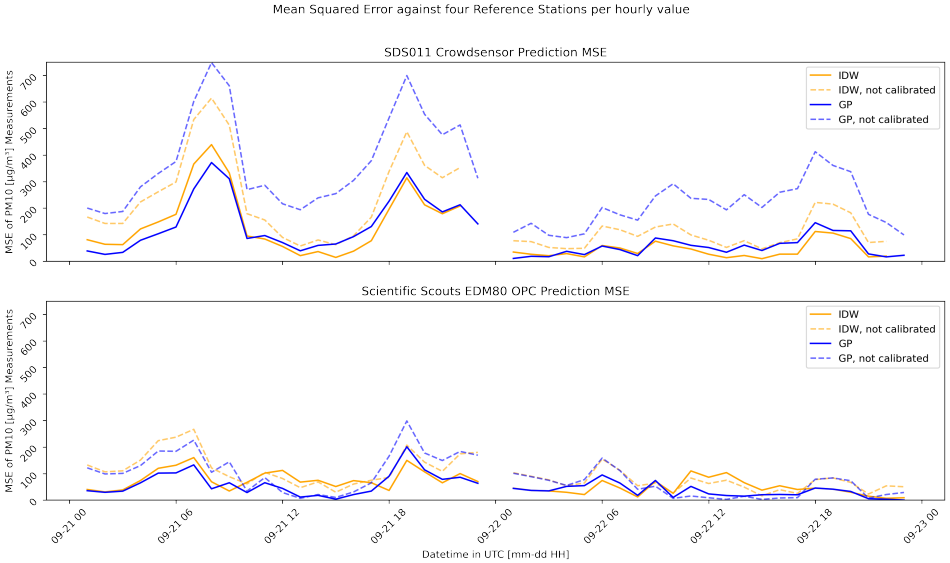

Fig. 3: MSE without (dashed) and with (solid) Calibration. Top (bottom) is the prediction of the SDS011 Crowdsensor (EDM80 Scientific Scout) dataset. We predicted each hourly value for 21st and 22nd of September 2020 (48hrs) for each reference station. Squaring the difference to the measurement and averaging over all four reference stations produced the shown plot. The hole at midnight of the 22nd is caused by missing data. In blue is the (2+1)D Gaussian Process Regression; in orange the inverse distance weighted interpolation.

Figure 3 shows the mean squared errors for Crowdsensors and Scouts with and without calibration. The Scouts perform overall better than the Crowdsensors, while the applied methods - inverse distance weighted interpolation and the (2+1)D Gaussian Process - perform similarly. This figure basically summarizes the essential information from this section.

5 Bike Route Prediction

5.1 Data

To tackle a prediction of a personal exposure scenario, we chose a bike route from 2020-09-22, which collected data from 6:48:20 to 7:10:56 over a length of 4.121 km for a total of 465 data points. The sensor used on the bike was a SDS011 sensor. The path is shown in Figure 4, along with the locations of the SDS011 Crowdsensors, EDM80 OPC Scientific Scouts and the official reference stations.

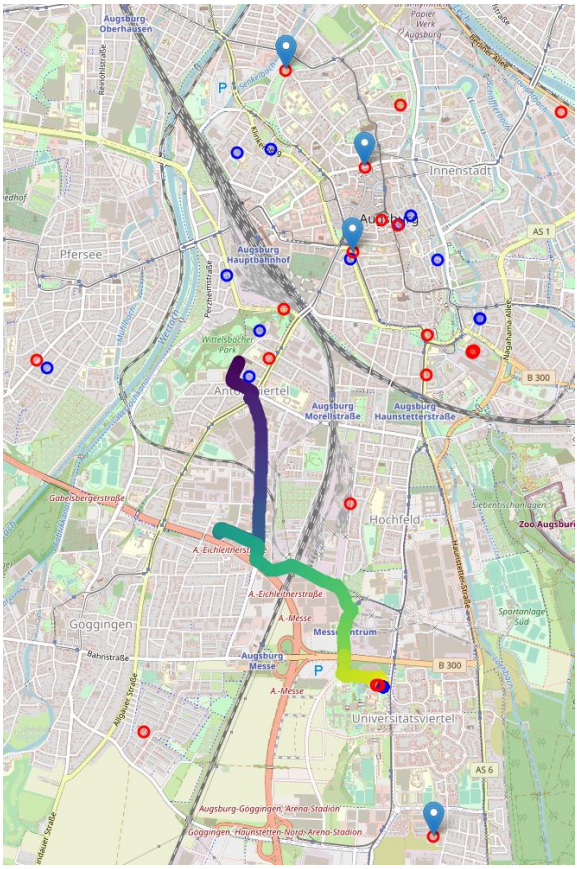

Fig. 4: The Location of the bike route within the city of Augsburg. Positions of the SDS011 Crowdsensors (Scientific Scouts) are marked in blue (red). The positions of the official reference stations are marked by the blue pins. The color coding of the bike path refers to the travel time with purple being the start and yellow the end. The map has been created using folium for python.

We used a total of 3454 (331) data points in the window from 6:00:00 to 7:15:00 to train the (2+1)D Gaussian Process for the Crowdsensors (Scouts). In contrast to the sample used in

the interpolation described in section 4, these were not further aggregated and correspond to 30 second to 1 minute values for the Crowdsensors and 5 minute values for the Scouts. We chose the RBF length scale as [0.00450,0.00680,10], where the spatial dimensions again refer to lat,lon and correspond to ~500m in each direction. The time direction is now encoded in seconds and fixed at 10 seconds to be sensitive to the frequency of the not aggregated sensor values.

5.2 Methods

We used the same two methods as in subsection 4.2, inverse distance weighted interpolation (IDW) and a (2+1)-dimensional Gaussian Process Regression ((2+1)D GP). Additionally, we performed a simple spatial Random Forest Regression (RF) using the scikit-learn implementation and a 2D (spatial) Gaussian Process Regression (2D GP) on the relevant time slice corresponding to each datapoint on the bike path.

To account for the timestamps of the bike path measurements and the sensor measurements not aligning, we used linear interpolation between values of the sensors to obtain values for the timestamps of each bike path datapoint for each sensor. The (2+1)D Gaussian Process was trained on the original data without interpolation and learned the interpolation.

We then subsequently aggregated the bike path measurements and calculated the predictions of each interpolation method for each aggregation interval.

For comparison, we calculate an IDW of the four reference stations. To that end, we split the values of the hours in question in the following way: Since the bike path ran from 6:48 to 7:10, we took 55% of the 6:00-7:00 hourly value and 45% of the 7:00-8:00 hourly value.

5.3 Results

Figure 5 shows the mean squared errors for each interpolation method as a plot over the aggregation time. There are two effects, which drew our attention: a) The interpolation methods all seem to perform vastly better than the background station and b) the (2+1)D GP performs way better than the other methods in case of the Scientific Scouts. However, in our subsequent analysis, we found that both effects stem from other sources and can not be attributed to any interpolation method performing any better than any other, nor the sensor network having an advantage over the reference stations, given the data at hand. We will first discuss the two effects that lead to Figure 5 and then discuss possible issues and improvements of our study.

a) The difference in MSE between the interpolation methods used on the Crowdsensors and Scouts and the prediction of the reference station depicted in Figure 5 is most likely purely

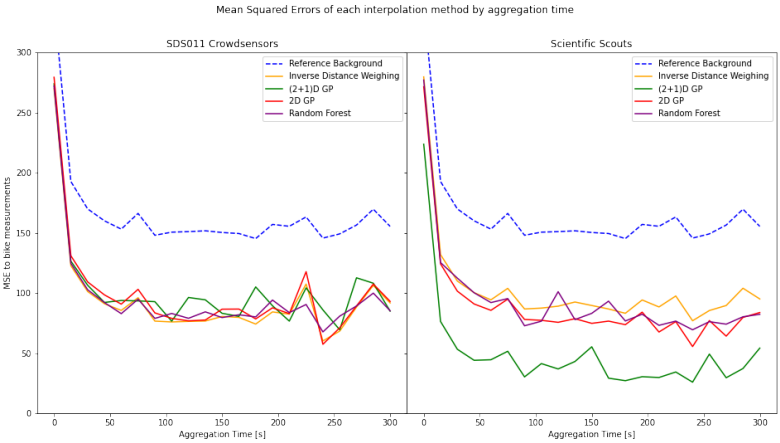

Fig. 5: Mean Squared Error for the SDS011 Crowdsensor dataset (left) and the Scientific Scout dataset (right) over aggregation time. The dashed blue line is the MSE against the mean of the reference stations for the time span of the bike path. MSE of inverse distance weighting, (2+1)D Gaussian Process, 2D Gaussian Process and 2D Random Forest in orange, green, red, and purple, respectively.

due to calibration. On the one hand, as stated earlier, our calibration of the Crowdsensors and Scouts improved their predictions a bit, but still left room for improvement. A proper calibration, where each single sensor is calibrated according the the specific environment it is deployed in would likely shift the interpolations of the Crowdsensors and Scouts in Figure 5 towards the prediction of the background stations. The large offset of the background station prediction MSE likely stems from the bike sensor not being well calibrated. The low mean of the bike sensor measurements, as shown in Figure 6, when compared to the Crowdsensors, Scouts and reference stations, supports that argument. Thus, any offset in Figure 5 likely stems from imperfect calibration of the relevant sensors and cannot be used to support our hypothesis.

b) The (2+1)D GP prediction MSE in the right side of Figure 5 seems to suggest, that the method outperforms the other methods. However, on further analysis, we found that this effect is just coincidence: When the temporal length scale is very low compared to the interval between measurements, the (2+1)D GP drops to the prior most of the time. This is because the GP interprets the temporal distances between the measurements as too large to be correlated. In case of the Scouts, which have a 5min measuring frequency, this is the case when training the GP with a temporal length scale of 10 seconds. The prior itself happens to be closer to the bike path measurements than the other interpolation methods by pure coincidence. This can be seen in Figure 6. We conducted further calculations with increasing temporal length scale (30s, 60s, 90s) and found that the (2+1)D GP starts to align

more with the shapes of the other methods and the MSE in Figure 5 also moves up to join the other methods. Therefore, the lower MSE of the (2+1)D GP also cannot support a claim of the method being superior in predicting personal exposure.

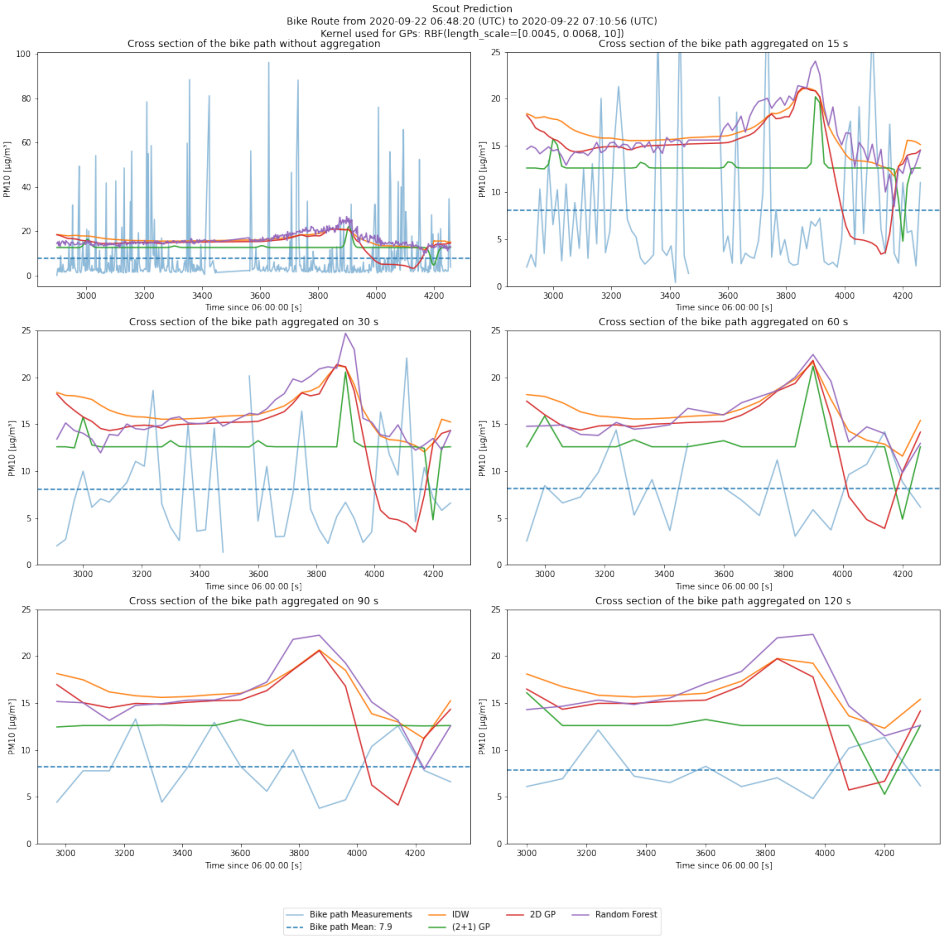

Fig. 6: Comparison of the various interpolation methods with the measurements along the bike path for different aggregation intervals.

We suggest, that the slope of a linear regression for each line in Figure 5 can show, that the corresponding method starts to pick up some predictive power as higher aggregation times wash out some of the fast fluctuations. In the case of the Crowdsensors, however, no reasonable slope is detectable. In case of the Scientific Scouts, the three AI supported methods (GPs and RF) seem to have a slope which, however, is still small compared to the fluctuations of the lines. This would be an interesting metric to follow for future work on a different dataset.

6 Discussion

We conclude, that for the given data, we were not able to show that a AI based interpolation of the network of Scientific Scouts or SDS011 Crowdsensors provides a significantly better prediction of personal exposure than an IDW of the reference stations. One reason why this part of our analysis led to no result is likely the geographical location of the route. We see from Figure 4, that the route most of the time is far away from both types of sensors, making a prediction by pure interpolation very unstable. An interesting question for future work would therefore be repeating the same analysis on a better suited dataset.

We were able to show in section 4, however, that accurate predictions are within reach of sensor networks comparable to those of the Crowdsensors and Scouts, and depend less on the sensor quality, as one might naively think. Our results suggest, that an individual calibration of the sensors based on the classification of their environment is the most important ingredient in obtaining accurate predictions (c.f. Figure 1). At the same time, the spatio-temporal (2+1)D Gaussian Process performs better than the non-AI based naive inverse distance weighted interpolation. This can be further bolstered by increasing the amount of training data and/or modifying the kernel function, where we used the most basic choice of an RBF kernel.

6.1 Limitations

The biggest limitations on the analyses we presented are:

a) The location of the bike path relative to the locations of the sensors as well as the overall density of the sensor network with $O(1)$ sensor per km^2

Denser networks promise to better resolve the fast fluctuations seen on the bike path measurements, thereby allowing interpolation methods to capture these patterns. These high fluctuations likely stem from the bike sensor being in traffic and thereby occasionally very closely exposed to car emissions. We suggest that, while it may not be practicable to fully resolve the second-by-second fluctuations of the measurements, a denser network with sufficient time resolution might be able to resolve the changes after minimal aggregation. This translates into the question which spatial and temporal network densities lead to accurate predictions at which aggregation times. While it is clear, that it would require unrealistically high temporal and spatial resolutions to resolve the most highly localized, second-by-second fluctuations, the interesting question is which temporal accuracy of a prediction would be realistic to target and what network density would be required to achieve this. This is a task for future work to map out.

b) Computing power to train spatio-temporal machine learning models on more data.

In our analysis, we restricted ourselves to two days in September, which we considered the relevant time frame for the bike route. For the prediction of the reference stations, we also did one run on which we trained the (2+1)D Gaussian Process using all September data (hourly aggregates, ~18.000 data points for the Crowdsensors and ~15.000 data points for Scouts) and predicted the four positions of the reference stations for each hour for the whole of September, which amounts to 4x24x30 ~3000 predictions for each method. Predicting both, calibrated and not calibrated, for the evaluation makes roughly 12.000 predictions. On a local machine, using jupyter notebooks and scikit-learn as library for the GP, the process of training and predicting took more than two hours. Using several days of the unaggregated dataset described in section 3, would exceed the capabilities of a local machine. This run already excluded the most time consuming step: finding the optimal hyper-parameters for the Gaussian process. We argue, that the effects a network with an average distance of 1km between stations can estimate are in the ballpark of 500m correlation length. We confirmed that the GP can find lengths of that order with much smaller datasets (max. 2000 data points), and then fixed the GP for training with larger datasets on those 500m in each spatial direction. Increasing the input for finding the optimal hyper-parameters significantly increases the computing time needed, thus using the entire dataset of September (even if aggregated) for a training that tries to find the hyper-parameters is out of question on a local machine, let alone working with the raw, not-aggregated data. Yet, there may be phenomena and patterns (e.g. daily periods), which escape us when working with a restricted or pre-aggregated dataset.

6.2 On the use of the median in calibration and aggregation

The Crowdsensors occasionally have corrupt values which spike as high as nearly 1.000.000. These values extremely distort any calculation if the arithmetic mean is used and these values have not been discarded. This leaves us with picking the threshold by hand which induces a degree of arbitrariness. We chose to resolve this more elegantly by taking the median instead of the arithmetic mean for the calibration. Note that for the treatment of these spikes, it is secondary if they map real effects or are pure electronic errors within the device. Since high spikes can only occur in one direction, they expose the Poisson nature of the data to a degree, that any approximation as a normal distribution breaks down. The degree to which the dataset loses its normal shape, the arithmetic mean becomes erroneous as a measure of centricity. When calibrating on a monthly mean where huge spikes are present, the spikes will pull the arithmetic mean so far up, that they dominate the calibration procedure and render the result useless, e.g. by leading to a correction that pushes 95% of the sensors values into the negative range. This could be avoided by introducing a cutoff. The value of the cutoff, however, needs to be justified since it unavoidably excludes some high but likely real, physical measurements. Its purpose is only to shape the data to an approximate normal distribution so that the arithmetic mean can be used as a measure of centricity again. We chose to use the median as a measure of centricity to avoid having to pick an arbitrary threshold. While possibly the choice of the median for a given dataset can

be mapped on a mean with a fixed cutoff, the choice to us seemed less biased than choosing a threshold to reshape the data until we consider it sufficient for our methods.

Interestingly, in case of the Scientific Scouts, where no such high spikes were present, it did not matter whether we picked the mean or the median and Figure 2 and the respective parts of Figure 3 are nearly identical when using the median or the arithmetic mean for calibration and aggregation. This leads to the question how the scouts avoid such spikes. While it is possible, that unwanted effects like pure electronic errors or coarse dust entering the device may be prevented by appropriate measures in higher quality sensors, there is still the possibility for real high exposure, e.g. through a person carrying a cigarette walking by, which the sensor should correctly detect. As mentioned, it does not matter whether a true or false measurement destroys the approximate normal distribution, any spike will do. This poses a challenge when considering low cost sensors in urban areas and scenarios like mobile sensors, where high exposure is expected and the use of high-cost-high-quality sensors may be restricted.

In the case of aggregation, the same argument applies to a lesser degree. Here, one might argue that we are not primarily after a measure of centricity but a division of an integrated exposure. This would want to include real, valid measurements in the calculation. But since the mean in the presence of high spikes is dominated by them, it becomes crucial that the sensor in question still measures precisely in that high concentration range and that we are sure that the measured effect is indeed a real effect. Otherwise, errors will dominate our calculation. This sensitivity to the spikes being well measured and true effects thus effectively poses the same problem for low cost sensors, since they won't be able to guarantee the precision and truthfulness of a single measurement.

7 Conclusion

We conclude, that interpolation techniques applied on a distributed sensor network offer a lot of potential. While there are cases, in which they can provide additional value over the traditional, high-cost-high-quality measuring stations, the applications and their validation remains very challenging. We argue, that, in general, AI based interpolation - like any interpolation method - work better on more sensors than on less, and thus go hand in hand with sensor networks as an alternative approach to the traditional, high-cost-high-quality measuring stations. Especially when trying to tackle the intricate problem of accurately predicting real, personal exposure. We thus argue for both, sensor networks and AI based interpolation methods, as they can be seen as integrated system - neither can develop their full potential without the other. In the scope of future smart city applications, our analysis touches questions of urban decision makers, where the scenario is finding the best solution for monitoring air quality given a finite resources. The choice is basically to invest in very few, very accurate measuring stations – which is the traditional and standard way – or to invest into more, yet individually less accurate sensors. With machine learning methods

improving and starting to tap the potential of distributed sensor networks, they become an integral part of future smart cities.

8 Acknowledgement

This work was partially funded by the German Federal Ministry of Transport and Digital Infrastructure (BMVI) as part of project *SmartAQnet* [Bu17] (funding no. 19F2003B) and by the German Federal Ministry for Research as part of a *SDI-C* microproject.

References

[Br18] Bruns, Julian; Riesterer, Johannes; Wang, Bowen; Riedel, Till; Beigl, Micheal: Automated quality assessment of (citizen) weather stations. arXiv preprint arXiv:1802.06018, 2018.

[Bu17] Budde, Matthias; Riedel, Till; Beigl, Michael; Schäfer, Klaus; Emeis, Stefan; Cyrys, Josef; Schnelle-Kreis, Jürgen; Philipp, Andreas; Ziegler, Volker; Grimm, Hans et al.: SmartAQnet: remote and in-situ sensing of urban air quality. In: Remote Sensing of Clouds and the Atmosphere XXII. volume 10424. International Society for Optics and Photonics, p. 104240C, 2017.

[Bu18] Budde, Matthias; Schwarz, Almuth D; Müller, Thomas; Laquai, Bernd; Streibl, Norbert; Schindler, Gregor; Köpke, Marcel; Riedel, Till; Dittler, Achim; Beigl, Michael et al.: Potential and limitations of the low-cost SDS011 particle sensor for monitoring urban air quality. ProScience, 5:6–12, 2018.

[Ho08] Hoek, Gerard; Beelen, Rob; De Hoogh, Kees; Vienneau, Danielle; Gulliver, John; Fischer, Paul; Briggs, David: A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric environment, 42(33):7561–7578, 2008.

[Ki14] Kilibarda, Milan; Hengl, Tomislav; Heuvelink, Gerard BM; Gräler, Benedikt; Pebesma, Edzer; Perčec Tadić, Melita; Bajat, Branislav: Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. Journal of Geophysical Research: Atmospheres, 119(5):2294–2313, 2014.

[Kr51] Krige, Daniel G: A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Southern African Institute of Mining and Metallurgy, 52(6):119–139, 1951.

[Li17] Lin, Yijun; Chiang, Yao-Yi; Pan, Fan; Stripelis, Dimitrios; Ambite, José Luis; Eckel, Sandrah P; Habre, Rima: Mining public datasets for modeling intra-city PM2. 5 concentrations at a fine spatial resolution. In: Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems. pp. 1–10, 2017.

[Pe11] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[PH16] Pebesma, Edzer; Heuvelink, Gerard: Spatio-temporal interpolation using gstat. RFID Journal, 8(1):204–218, 2016.

[Sh68] Shepard, Donald: A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM national conference. pp. 517–524, 1968.

[Sh20] Shen, Yao; Lehmler, Stephan; Murshed, Syed Monjur; Riedel, Till: Characterizing Air Quality in Urban Areas with Mobile Measurement and High Resolution Open Spatial Data: Comparison of Different Machine-Learning Approaches Using a Visual Interface. In (Santos, Henrique; Pereira, Gabriela Viale; Budde, Matthias; Lopes, Sérgio F.; Nikolic, Predrag, eds): Science and Technologies for Smart Cities. Springer International Publishing, Cham, pp. 115–126, 2020.

[Su16] Sun, L; Wei, J; Duan, DH; Guo, YM; Yang, DX; Jia, C; Mi, XT: Impact of Land-Use and Land-Cover Change on urban air quality in representative cities of China. Journal of Atmospheric and Solar-Terrestrial Physics, 142:43–54, 2016.

[To70] Tobler, Waldo R: A computer movie simulating urban growth in the Detroit region. Economic geography, 46(sup1):234–240, 1970.

[Yu16] Yu, Ruiyun; Yang, Yu; Yang, Leyou; Han, Guangjie; Move, Oguti Ann: RAQ–A random forest approach for predicting air quality in urban sensing systems. Sensors, 16(1):86, 2016.

[ZLH13] Zheng, Yu; Liu, Furui; Hsieh, Hsun-Ping: U-air: When urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1436–1444, 2013.