Approaches to Fuse Fixed and Mobile Air Quality Sensors

Gerhard Dünnebeil, Martina Marjanović, Ivana Podnar Žarko

Approaches to Fuse Fixed and Mobile Air Quality Sensors

Gerhard Dünnebeil[1], Martina Marjanović[2] and Ivana Podnar Žarko[2]

[1]AIT Austrian Institute of Technology GmbH, Vienna, Austria
gerhard.duennebeil@ait.ac.at
[2]University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
{martina.marjanovic, Ivana.podnar}@fer.hr

Abstract. Nowadays, air quality monitoring is identified as one of the key impacts in assessing the quality of life in urban areas. Traditional measuring procedures include expensive equipment in the fixed monitoring stations which is not suitable for urban areas because of the low spatio-temporal density of measurements. On the other hand, the technological development of small wearable sensor devices has created new opportunities for air pollution monitoring. Therefore, in this paper we discuss statistical approaches to fuse the data from fixed and mobile sensors for air quality monitoring.

Keywords: air quality monitoring, interpolation, kriging, mobile sensors

1 Introduction

Air pollution represents a serious threat in urban environments with a significant negative impact on human health. Therefore, the European Parliament and the Council of the European Union provided the Air Quality Directive to emphasize the importance of air quality monitoring in urban areas [1]. Also, scientists have proven that exposure to traffic-related air pollution can cause different respiratory problems [2]. Since heavy industry and vehicles are nowadays major producers of toxic gases, it is necessary to densely monitor air pollution in big cities, both in time and space, in order to identify contaminated areas promptly and devise appropriate actions.

Today Air Quality Monitoring is mostly done with stations that do long term monitoring at fixed location. Although the equipment often is mounted in containers which can be relocated (Figure 1 and Figure 2), there is a need to have undisturbed series of measurements over a long period of time that allow to exclude location based effects from the measurement campaign.

adfa, p. 1, 2011.

Figure 1 A typical AQ monitoring fixed station[1]

Figure 2 A set of analyzers as they are typically found in a fixed station

[1] All images from fixed stations are courtesy of Authorities federal state of Upper Austria.

The traditional air quality measurement infrastructure can therefore be extended to obtain a higher spatial resolution by using a larger number of mobile wearable sensor nodes for environmental monitoring (Figure 3).

Figure 3. Wearable sensors for air quality monitoring[2]

Although there is a significant discrepancy in the accuracy and sensitivity between the new mobile sensors and static meteorological stations, their affordability and simplicity have created the opportunity for wide usage of small and cheap sensor devices. Thus, we investigate how these two can coexist and benefit from each other. One of the major advantages of mobile sensors is the simplicity of taking samples on many locations. In this paper, we will focus on ideas how this wider coverage can be used to estimate the pollution at arbitrary points by exploiting the spatial and temporal coverage of mobile sensors in combination with the accuracy of fixed stations.

The rest of the paper is organized as follows: Section 2 introduces a model based sensor data interpolation with focus on the kriging method. In Section 3 we discuss the proposed model. We further introduce the problem of bogus sensor detection and determination of confidence factors in Section 4. Section 5 provides a brief overview of related work by addressing interpolation methods used in different areas of the environmental science. Finally, Section 6 concludes the paper and gives directions for future work.

---

[2] Wearable sensors and smartphone application for air quality monitoring developed at University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia.

2 Model-based Sensor Interpolation

The mobility of sensors leads to a dynamic sensing coverage of geographical areas, which can potentially result in certain areas being redundantly covered, while other areas may suffer from lack of available sensor measurements. Obviously, it is not possible to cover all geographical points in a certain area of interest by actual sensor readings. Therefore, we need to use a finite number of sensor readings and estimate the actual values in between. Mathematically this requires an interpolation approach.

Interpolation is a method of constructing new data points from a set of previously known values. Basically, it makes some assumptions about the values at locations that have to be estimated by using some kind of a model. Classical interpolation approaches, like polynomial or spline interpolation, completely ignore the fact that sensor readings always come with a certain inaccuracy. Hereafter, we present and discuss other techniques that can be used to estimate missing values.

Interpolation can be done in space or in time or both. In this paper we implicitly restrict ourselves to interpolation in space. Our goal is to get subsequent maps of pollution for discrete points in time that incorporate as much information as possible to get the best accuracy. This includes not only information gained at or near the said points in time but also the knowledge about the involved sensors gained from the earlier maps.

2.1 Kriging approach

One well known approach to interpolate sensor readings is kriging (originated by Danie G. Krige in 1951[3]). The basic idea of kriging is to estimate a value at a specific location by computing a weighted average of the known values in the neighborhood of that location. In other words, kriging takes into account the nearby sensor readings to eliminate to a certain degree the random errors inherent in every reading. The mathematical meaning of the term nearby is defined by the so called co-variance function v(l $\vec{}$) which defines the significance that a certain reading r has at location l $\vec{}$.

The estimated value at location l $\vec{}$ can be calculated as

$$Z(\vec{l}) = \sum_{\alpha=1}^{n(\vec{l})} \lambda_\alpha \, Z(\overrightarrow{l_\alpha}). \quad (1)$$

The factors λ_α describe how much a reading is relevant for the interpolated value. This is determined by the co-variance function.

There is a certain degree of freedom in choosing this function. Indeed, this is the model behind the kriging approach.

Usually one chooses a function that will have a value of 1 at the exact location $\vec{l}$ and will monotonically decrease with the distance from $\vec{l}$. Either this function will be zero at a certain distance from $\vec{l}$ or it will converge to zero with the distance reaching infinity. So the co-variance function typically has the form of a coefficient between 0 and 1

which defines the statistical weight of a reading r at the location $\vec{l}$. Figure 4 shows some typical curve forms for such function.

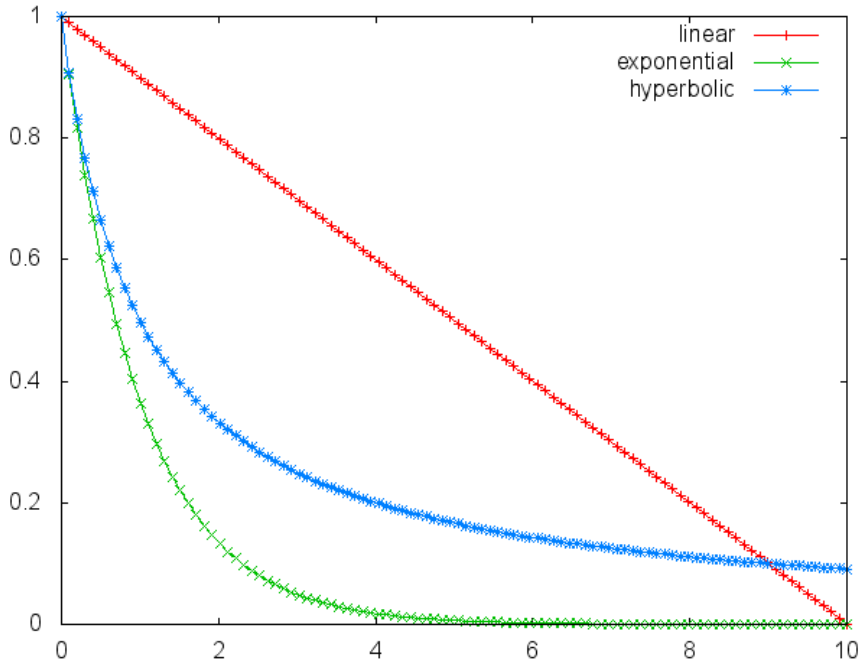

Figure 4 Examples of different co-variance functions

This co-variance function gives an individual value for each sensor that is expected at a location$\vec{l}$. All these values usually differ from each other. The interpolated value is now calculated such that the individual errors are minimized (least mean square).

2.2 Taking confidence into account

Nowadays, mobile sensors typically have a much larger error than fixed stations. When trying to fuse the readings between fixed stations and mobile sensors we have to take this into account. Even though a mobile sensor takes measurements exactly at a sampling location, usually the readings collected by different sensors at the same location and point in time will not be the same. This will be the case even if both sensors are calibrated.

Since we want to model different confidence levels, we introduce a confidence coefficient cc which lies between 0 (not reliable at all) and 1 (completely reliable). Still we have to determine a method how to gain such confidence coefficients

(which we show in the section below), but once they are available they simply attenuate the existing co-variance function v($\vec{l}$) by defining a new one

$$v_1(\vec{l}) = v(\vec{l}) * cc\ . \tag{2}$$

This will result in a domination of a nearby fixed station in the interpolation process. But if multiple mobile sensors in the vicinity share the same reading, they can eventually dominate a fixed station.

2.3 Some Definitions

To keep the other chapters free from complex and repeated mathematics and definitions as much as possible, we introduce a few phrases here. We use the term physical sensor readings r_{phys} which indicates sensor readings without further corrections. We are aware though, that this concept is a bit problematic as this concept is not taking into account that real sensors suffer from cross sensitivity and similar effects. Certain compensations for these effects must still be applied to the raw sensor readings before they can be used in this context.

The most important function is the interpolation function I which gives us the interpolated value. I is a function of the location $\vec{l}$. An index of ”F” will indicate that the interpolation is done only over the fixed stations, while an index of ”M” will indicate the same only for mobile stations (i.e. I_F and I_M).

A function LMS means we use the least mean square fit on the data. The result of an LMS is a set of coefficients. Due to the lack of better models we assume a linear mean square fit throughout this paper so the result is a pair of numbers, amplification a and offset b so that the correction function is

$$r_{Fcorr} = f(r_{phys}) = ar_{phys} + b. \tag{3}$$

3 Model discussion

The above described approach has a clear weakness as it assumes that only the distance between different sensors is important. This assumes an ideal situation where diffusion processes are not disturbed by wind or obstacles, neither natural nor artificial. The assumption does not hold in a typical urban environment where street canyons dominantly influence the spreading of pollutants. Further discussion of street canyon effects, their consequences and possible solutions is out of scope for this paper and it will be taken into account in future work.

3.1 Relating mobile sensors to fixed stations

Usually sensors tend to drift. Even though sensors are kept in a very controlled constant concentration of pollutants, the readings will still change over time. A lot of effort is taken to control and compensate the natural drift for fixed stations. The related analyzers are often calibrated and in many cases a so called function control procedure is performed once every 24 or 48 hours. During the function control phase the analyzers

are first exposed to gases with known concentrations, then the resulting error is recorded and all subsequent sensor readings are compensated numerically.

However, doing the same for mobile sensors would significantly devalue their advantages in mobility and costs. For that reason, it is necessary to find other ways of compensating the errors of mobile sensors. The first idea, to just compare readings between fixed and mobile sensors when a mobile sensor passes by a fixed station seems obvious but will not work. One reason is that mobile sensors typically measure one to four samples per minute as they have to operate on severely limited energy. If we assume that the sensor is moving with a bicycle or car, this means it will hardly measure at all within a distance close enough to a fixed station. Another reason is that mobile sensors typically measure at heights of 1 to 1.5 meters above ground level while fixed stations have their air inlet at least 4 meters high.

To compensate this errors, we first interpolate using only the fixed stations. This can usually be done easily as all fixed stations give their results for the same time. The result is a map that shows approximately the situation over the area for a given point in time.

The resulting interpolation function $I_F(\vec{l})$ is then used to calculate the deviation of each sensor reading to the related interpolated value. Of course, doing this for just one reading per sensor is still significantly influenced by different errors, some of them potentially huge. We need to gather enough of these (real reading, interpolated value) pairs to apply an LMS-function and get statistically significant coefficients for gain and offset to correct each sensor.

3.2 Relating mobile sensors to each other

The above described approach will only work when the probability of a mobile sensor to pass near a fixed station is high. However, the probability that two mobile sensors meet is much higher considering that only a certain (minimal) number of fixed sensors is in the field.

Doing an interpolation on the mobile sensors themselves and comparing individual sensor readings to the interpolated values can provide individual deviations for each sensor. When having sufficient deviations, a best-fit-line which will compensate individual sensors with respect to their companions can be calculated. When a network of mobile sensors is compensated, it is also possible to calculate compensation factors of certain sensors to the fixed stations. With this approach it is possible to adjust the complete set of sensors to the readings of fixed stations.

For this approach, we calculate the interpolation function for the mobile sensors, $I_M(\vec{l})$. Theoretically this is done for all sensor readings at a particular point in time. In praxis, it not possible to have this one point in time, instead we use a short time interval.

Subsequently all physical readings are paired with the interpolated readings at the same location. As singular readings are not sufficient to compute LMS statistics, these pairs are stored for later evaluation. A sufficiently large set of these pairs is then used to compute coefficients that correct sensors to be harmonized with the other sensors.

This list of pairs must contain enough data to calculate a meaningful statistic. On the other hand, as the sensors tend to age and drift, it must not contain values too long. In practice, we found out that having a list length of a few hours is more than sufficient.

Next, for each fixed station n another pair ($I_M(\overrightarrow{l_n})$, station reading) is calculated. This dataset is subsequently used to compute an LMS. The resulting coefficients (a_F, b_F) are merged with the individual coefficients for each sensor.

$$r_{M,corr} = (a_F * (a_M * r_{phys} + b_M) + b_F = a_F a_M r_{phys} + a_F b_M + b_F \quad (4)$$

1. This can also be formulated as an algorithm:
2. Calculate the interpolation function IM for all mobile sensors.
3. Gather all pairs ($I_M(\vec{l})$,r_{phys}) for all locations where sensor readings are available and add those to the set of existing pairs for each sensor.
4. Calculate correction factors for each sensor.
5. Calculate the interpolation function IF for all fixed stations.
6. Gather all pairs ($I_F(\vec{l}), I_M(\vec{l})$) for all locations of the fixed stations.
7. Calculate an LMS for all pairs and obtain the corrective coefficients a_F and b_F.
8. Apply both, individual and global, factors to each sensor reading to obtain harmonized readings.

3.3 Results and discussion

To verify the proposed algorithm, we have used a real-world dataset acquired from the air quality measurement campaign "SenseZGAir" performed in the City of Zagreb, Croatia, in early July 2014 as part of the Smart City Zagreb initiative. The "SenseZGAir" dataset contains 151,000 data points, including temperature, humidity, pressure, concentrations of carbon monoxide (CO), and either nitrogen dioxide (NO2) or sulfur dioxide (SO2), obtained at 13,000 unique locations in Zagreb (according to GPS coordinates) in 3 days that the campaign lasted [4]. To evaluate our model, we have used CO measurements from mobile sensors and official gas concentrations from the Croatian Ministry of Environment and Energy on July 7, 2014.

Figure 5 shows individual paths which 8 of our sensors did during the field trial in the time span from 9:00 to 10:00 a.m, while the "zoom" part shows the sensors "near" one of the fixed monitoring stations in Zagreb (marked as red cross) which we have chosen for our experiments.

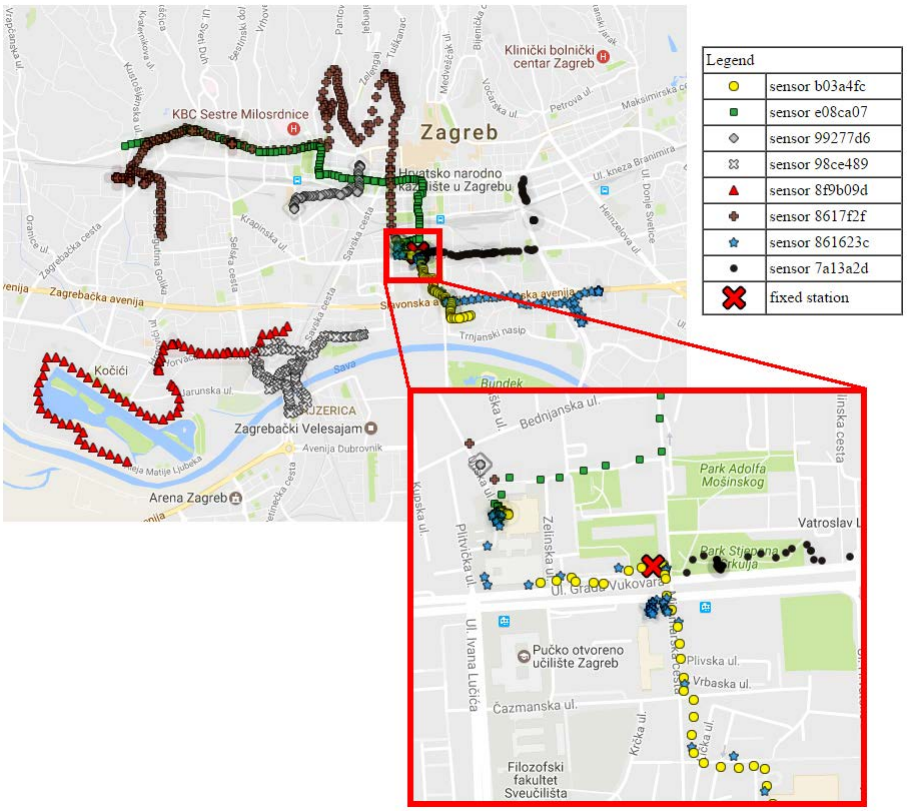

Figure 5 Mobile sensor measurements in the City of Zagreb for July 7, 2014, from 9:00 to 10:00 a.m.[3]

We have visualized individual sensor measurements together with the fixed station concentrations during the chosen time span, as shown on Figure 6. It can be seen that CO values differ a lot between sensors which is obviously not only due to the measurement errors but also to local effects. Also, the figure shows the fixed station measurement value (hourly mean) for the same period.

[3] Visualized by the CopyPasteMap tool available at http://www.copypastemap.com/

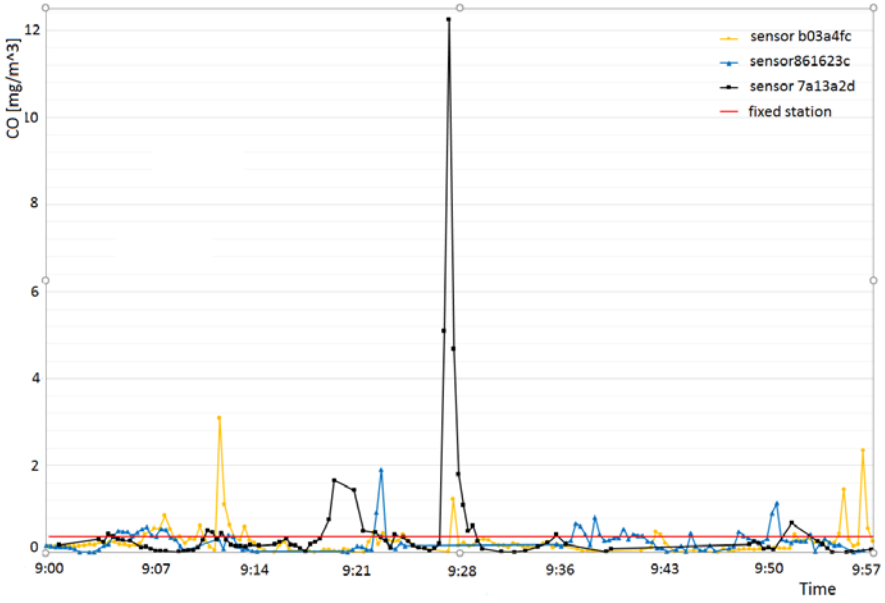

Figure 6 All sensor measurements between 9:00 and 10:00 a.m. on July 7, 2014

Our initial experiments with real-world data have shown that the proposed procedure suffers from the following:

- Even if you restrict yourself to sensor measurements near fixed stations the mobile sensors show a lot of variance in the measurements which is most probably caused by the real changes in the local gas concentration.
- Fixed station data on the other hand is usually only published in form of aggregated (mean) values over a significant period of time.

Correlating these two datasets leads to the fact that many substantially different mobile measurements are compared to the same fixed station's measurement. This in return makes it impossible to calculate a least mean square fit.

In another experiment we tried to relate sensors to each other. For that we used the time slot from 9:00 to 9:10 from the same data set mentioned above.

We have visualized the sampling points in Figure 7. It can be seen that there are quite some places where sensors are located close to each other. We have used those sensors to find correlations between them, i.e. to show whether the sensors deviate or not.

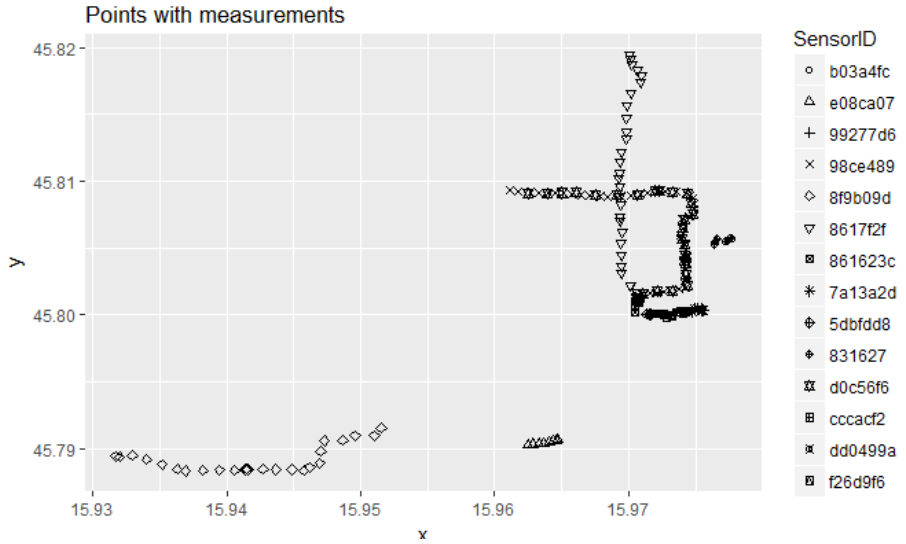

Figure 7 Geographical locations of mobile sensors from 9:00 to 9:10 a.m.

The interpolation showed the small variations between sensor measurements in the observed area. It also showed that the overall concentration of CO is mostly low, except on the crossing of two big roads with heavy traffic where we observed slightly lower air quality depicted as a red area in Figure 8.This is a well-known hot spot of lower air quality, so the public authorities have already placed the fixed station to continuously monitor the air quality as shown on Figure 5.

We have also compared the difference between the interpolated values and real sensor measurements as shown in Figure 9. The comparison gave an interesting result as it showed very few deviations between sensors and their interpolated values. Most of interpolated values differ less than 500 $\mu g/m^3$ from the measured value, while few of them show a difference between 1000 and 1500 $\mu g/m^3$. Those readings from a sensor have the value of zero and are not valid anyway. The few sensors that show real deviations do not allow to make any useful statistic.

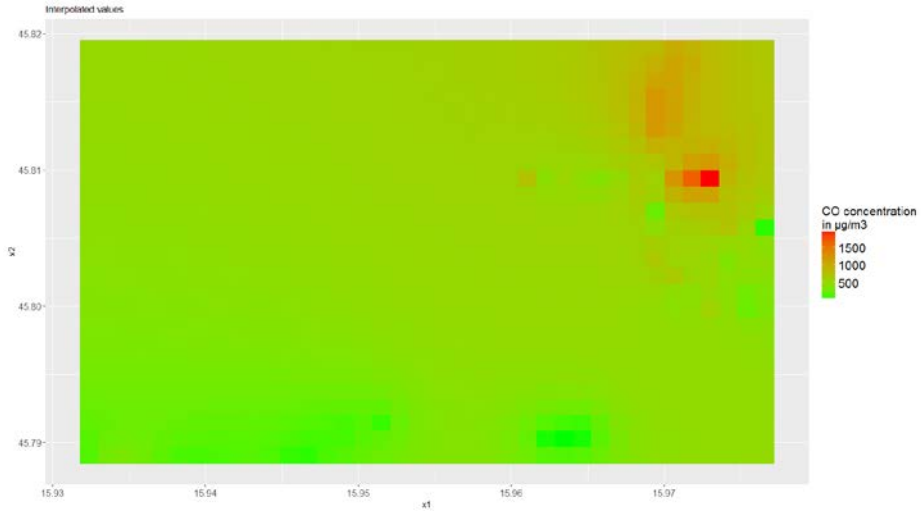

Figure 8 Interpolated sensor measurements

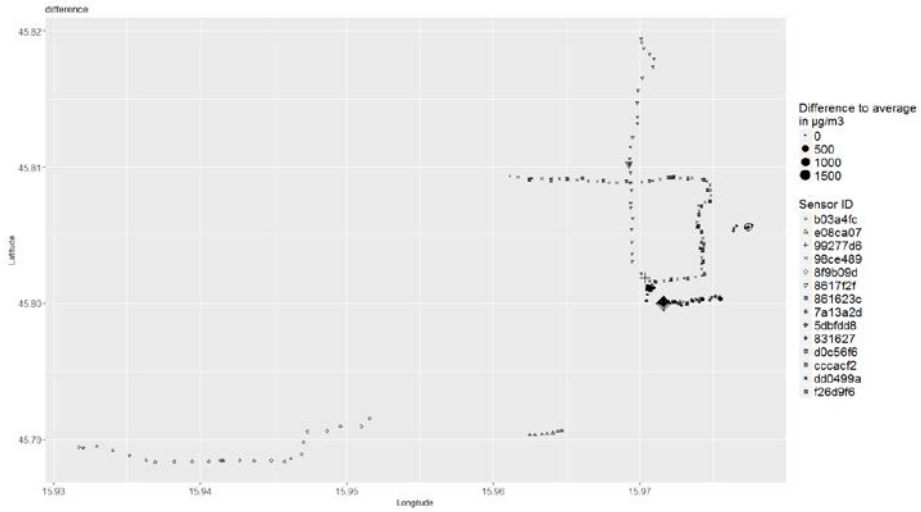

Figure 9 Difference between real sensor measurements and their interpolated values

4 Bogus sensor detection and determination of confidence factors

Mobile sensors tend to age and get unreliable over time. Detecting this effect is crucial for the evaluation of individual sensors. The procedures described above determine a best-fit-line for each sensor. Calculating the correlation factor for this best-fit-line provides an estimate of the quality of a sensor response to a given pollutant exposure.

Sensors that occasionally give bad correlations most probably have severely changed their environment during the time window currently in focus, e.g. parts of the measurements are indoor while others are near a road.

A more sophisticated analysis may be able to classify these individual measurements and treat them as separate classes. We simply propose to use this correlation factor as the confidence factor, maybe scaled down with an additional fixed factor that describes the overall confidence in mobile sensor. Sensors that occasionally show a bad fit are not necessarily defective. There can be many reasons for small sets of measurements being "off". Therefore, we propose to keep track of these correlation factors. If they show a long-term degrading trend, the sensor will lose permanently influence to the complete system. A sensor that has less influence is not important anymore and can be removed from the network.

5 Related Work

There are several papers considering data interpolation in different areas of the environmental science. Gummadi [5] gives a short overview of conventional interpolation techniques and neural network approaches commonly used for modelling and estimation of radon concentrations in Ohio. Kravchenko et al. [6] evaluate different interpolation principles to determine the optimal method for mapping soil properties, similar as Li et al. [7] who compare the accuracy of spatial interpolation techniques to identify the best prediction method to illustrate the spatial variability of the studied soil properties.

Spatial interpolation is widely used for creating continuous data where estimation at any unobserved location is within the data boundary and it is spatially dependent [8, 9]. Vuran et al. have developed the theoretical framework for the spatio-temporal correlation in wireless sensor networks (WSN) and showed that correlation can be exploited to significantly improve the energy-efficiency in WSN [10]. Further, stochastic interpolation methods are used to predict the values at unmeasured locations based on the data spatial autocorrelation and to estimate the prediction accuracy. In particular, kriging has been used for the spatial analysis of soil bulk density [11], temperature mapping [12], estimation of rainfall [13], as well as air pollution [14]. Tyagi et al. [15] use ordinary kriging to estimate the pollution in areas without measurements in Agra (Dayalbagh) region, similar as Shad et al. [16] who use fuzzy spatial prediction techniques to determine pollution concentration areas in practical situations where observations are imprecise and vague.

6 Conclusion

This paper deals with the statistical approaches which can be used to estimate the pollution at arbitrary points by exploiting the spatial and temporal coverage of mobile sensors in combination with the accuracy of fixed stations. In particular, we discuss the model-based interpolation with a focus on the kriging approach. We have proposed an

algorithm to fuse fixed and mobile air quality sensors and get harmonized sensor readings. However, due to the variance in the mobile sensor measurements together with the limited availability of fixed station data (note that this data is most often only available as aggregates over several minutes: Swiss -10 minutes, Austria and Germany – 30 minutes, Croatia – 1 hour time scale), the initial experiments did not provide usable results.

We also experimented with relating mobile sensors to each other. Interestingly this experiment showed that the often asserted inaccuracy of mobile sensors might be less than usually assumed. At least our finding was that we had no need for extensive compensation of the sensors.

We plan to repeat the experiment in a more controlled manner (like having some mobile sensors near the fixed station). Furthermore, we plan to simulate sensor networks combined from real stations and mobile sensors. Moreover, we want to integrate street canyon models into the system.

7 Acknowledgments

This work is supported in part by the H2020 symbIoTe project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 688156. This work has been supported in part by Croatian Science Foundation under the project 8065 (Human-centric Communications in Smart Networks).

8 References

1. Directive 2008/50/EC of the European Parliament and the Council of 21 May 2008 on ambient air quality and cleaner air for Europe.
2. HEI Panel on the Health Effects of Traffic-Related Air Pollution. Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects; HEI Special Report, 17.
3. Krige, D. G., "A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand", Journal of the Chemical, Metallurgical and Mining Society of South Africa, vol. 52(6), pp. 119-139, 1951.
4. Antonić, Aleksandar, Bilas, Vedran, Marjanović, Martina, Matijašević, Maja, Oletić, Dinko, Pavelić, Marko, Podnar Žarko, Ivana, Pripužić, Krešimir and Skorin-Kapov, Lea, Urban crowd sensing demonstrator: Sense the Zagreb Air, in Proceedings of the 22nd International Conference on Software, Telecommunications and Computer Networks, SoftCOM2014, 2014.
5. Gummadi, Jayaram, A comparison of various interpolation techniques for modelling and estimation of radon concentrations in Ohio, Theses and Dissertations, Paper 86, 2013.
6. Kravchenko, Alexandra and Bullock, Donald G., A comparative study of interpolation methods for mapping soil properties, Agronomy Journal, vol. 91, no. 3, pp. 393-400., 1999.

7. Li, Jin and Heap, Andrew D., A Review of Spatial Interpolation Methods for Environmental Scientists, Canberra: Geoscience Australia, vol. 137, 2008.
8. Robinson, T. P. and Metternicht, G., Testing the Performance of Spatial Interpolation Techniques for Mapping Soil Properties, Comput. Electron. Agric., Elsevier Science Publishers B. V., vol. 50, no. 2, pp. 97-108, 2006.
9. Akkala, Arjun and Devabhaktuni, Vijay and Kumar, Ashok, Interpolation techniques and associated software for environmental data, Environmental Progress & Sustainable Energy, vol. 29, no. 2, pp. 134-141, 2010.
10. Vuran, Mehmet C., Akan, zgr B., Akyildiz, Ian F., Spatio-temporal correlation: theory and applications for wireless sensor networks, Computer Networks, vol. 45, issue 3, pp. 245-259, 2004.
11. Sajid, A.H., R.P. Rudra and G. Parkin, Systematic Evaluation of Kriging and Inverse Distance Weighting Methods for Spatial Analysis of Soil Bulk Density, Canadian Biosystems Engineering, vol. 55, pp. 1.1-1.13, 2013.
12. Hudson, Gordon and Wackernagel, Hans, Mapping temperature using kriging with external drift: theory and an example from Scotland, International journal of Climatology, vol. 14(1), pp. 77-91, 1994.
13. Goovaerts, P., Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall, Journal of hydrology, vol. 228(1), pp. 113-129., 2000.
14. Janssen, Stijn, Dumont, Gerwin, Fierens, Frans and Mensink, Clemens, Spatial interpolation of air pollution measurements using CORINE land cover data, Atmospheric Environment, vol. 42, no. 20, pp. 4884-4903, 2008.
15. Tyagi, Aman and Singh, Preetvanti, Applying Kriging Approach on Pollution Data Using GIS Software, International Journal of Environmental Engineering and Management, vol. 4, no. 3, pp. 185-190, 2013.
16. Shad, Rouzbeh, Mesgari, Mohammad Saadi, Abkar, Aliakbar, and Shad, Arefeh, Predicting air pollution using fuzzy genetic linear membership kriging in GIS, Computers, Environment and Urban Systems, vol. 33, no. 6, pp. 472-481, 2009.