Identification of Air Pollution Sources using Predictive Models and Vehicular Sensor Networks

Aleksandar Gavrić*, Aleksandar Stamirović*, Leonid Stoimenov*
* Faculty of Electronic Engineering, University of Niš, 18000 Niš, Serbia
agavric@elfak.rs, {aleksandar.stanimirovic, leonid.stoimenov}@elfak.ni.ac.rs

***Abstract*— Observation of air pollution levels at certain points in space and time is done by using mobile and static sensor networks. The values of air pollution levels at points where no measurements were made are mostly assumed by numerous types of interpolation between known values at measured points. The authors of this paper propose techniques for predicting air pollution levels in points in space where there are no measurements. The proposed techniques are based on the analysis of measurements from the sensor network that are affected by the same sources of pollution. Three approaches for identifying unknown air pollution sources by collecting measures from sensors mounted on public service vehicles are defined, implemented, and evaluated. The first approach can be treated as the optimization problem, the second approach is based on clustering in a multidimensional space and the third one is a fast and light method for a specific simplified case of the problem. The system is also implemented for a distributed computer cluster that applies machine learning algorithms over data streams for efficient estimation of dominant pollution sources in real-time.**

I. INTRODUCTION

According to Global Alliance on Health and Pollution [1], 175 people per 100,000 inhabitants have suspected death every year as a result of polluted air in Serbia, which puts Serbia in 1st place in Europe and 9th in the world in terms of mortality from air pollution. With different sensors for measuring air pollution, we can determine how much pollution has been detected in the air that has flowed through a point in space where the sensor is planted during a short interval of time. However, the intensity of air pollution at points of space where the measurements were not performed should be assumed by methods of prediction and approximation as shown in Figure 1. Identification and localization of pollution sources in an urban area is not a trivial problem if we consider wide and inaccessible areas with obstacles. Therefore, interpolation algorithms are often used to approximate pollution in unmeasured spaces from which the locations and intensities of pollution sources can be inferred.

By applying machine learning predictive models, new methods can be addressed for more precise identification of pollution sources. This may be of particular interest in large urban areas, during periods with intense air pollution, when it is necessary to take urgent measures in order to permanently or temporarily eliminate sources of pollution (usually large combustion plants without filters or industrial plants).

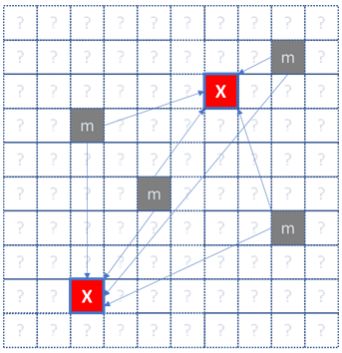

Figure 1 – Illustration of matrix that represent geographical area that is observed. The "m" elements represent geographical regions where air pollution measurements are performed and the "X" elements represent locations of air pollution sources.

Detection of sources of pollution could enable the competent public services to react immediately and reduce environmental disruption by certain measures.

Additionally, it is necessary to discuss and provide the system for efficient data collection and its maintenance and robustness, the possibility of evaluating the results of system and to take into consideration the cost of system. In order to select the most appropriate method of data collection, different sensor networks should be considered especially those that are attached to mobile subjects in a large metropolitan area. For the purpose of simulating data collection, the authors created a simulation of bus traffic in the city of Nis through buses that carry sensors on defined fixed routes as shown in Figure 2. Sensors on mobile vehicles measure air pollution in certain locations in space and time at the moment when the bus was passing through that location. The authors propose methods to analyze the dependence between the obtained measurements in order to determine how many sources of air pollution there are and what their characteristics are.

Figure 2 – The map of City of Niš and the simulated routes of public city buses that are carrying sensors

II. Related work

The most popular methods for identification of the pollution sources are mostly based on the analysis of the spatial distribution of pollution using interpolation algorithms. However, there are some approaches for air pollution source identification based on neural networks with Bayesian optimization [2]. The authors of this approach first create an air-pollution sensing network in a relatively smaller grid area to collect environmental data, such as air pollution concentration, wind speed and wind direction. The model built using diffusion models of pollutants is tuned when necessary, in order to prevent impacts from outliers and other unstable factors like wind direction. After that, the built model is applied to predict the sources of air pollution in a wide area. A novel, machine learning based, dense air pollution estimation system is introduced in [3]. This system utilizes historical data both from (sparse) government monitoring sites and (dense) wireless sensor network. The authors choose seven regression models and compare their estimation performances. As a result of these comparisons, the authors selected the Support Vector Regression (SVR) as the best candidate for further consideration. In the air pollution surface estimation part, estimates using SVR correspond well with the sensing interpolation map and can more clearly highlight the most polluted area compared with other regression models. These results indicate that the proposed system can generate accurate air pollution estimations, and highly increase the air pollution map resolution. Realtime high-resolution urban air pollution maps using mobile sensor nodes and drones are discussed in [4], and [5].

The authors of this paper propose an approach for measuring air pollution using mobile sensors installed on public service vehicles such as buses and taxis. Also, the authors propose different techniques for making predictive estimations. Unlike methods that use a predictive model to generalize behavior using data collected from the small geographic area, our proposed techniques are based on combining measured data with the pollution simulation model to describe pollution in a specific geographic area and discover knowledge.

The main objective of this paper is to present a prototype implementation and evaluate proof of concept results of the proposed techniques for identifying unknown air pollution sources using data collected from mobile sensors installed on public service vehicles. Additionally, the authors will discuss the architecture and robustness of the system for efficient data collection from mobile sensors.

III. Methodology

The authors present three methods for solving the problem of estimating air pollution sources and increasing air pollution map resolution. The first method comes down to solving the optimization problem, which is to find arguments for which the nonlinear cost function described in Figure 1 has a minimum value. As a first step, the authors initialize the simulation model parameters to random values to simulate the pollution map using the Gaussian Plume Model library [6]. In order to estimate air pollution map and air pollution sources, the proposed system needs to tune the model parameters and minimize the error between the real sensor measurements map and the simulated map. Figure 3 represents the first method where X and Y are vectors of x and y coordinates of sources, Q is the vector of masses emitted per sources (ug/m3) and H is the vector of heights of sources. Filter matrix is a matrix whose element is 1 if the air pollution is measured by sensors on the drones on the corresponding segment of the space or 0 otherwise. The measurement matrix is a matrix whose element represents the mean value of air pollution in the corresponding segment of space if the measurements with sensors were performed over that segment of space and if not, then the element of this matrix is undefined. Generated matrix is created using the Gaussian plume model function that takes the given values of X, Y, Q, H vectors as arguments. The upper part of the Figure 3 shows the multiplication of the filter matrix and the simulated map obtained by a function that initially takes random arguments. The lower part of the Figure 3 illustrates the search for arguments of the simulation function with which there is a minimal difference between the measured values and the simulated ones.

In the second method as illustrated in Figure 4, the authors divide the space into k segments and calculate the measured pollution mean value over those segments. Therefore, every possible value of pollution map is represented as a point in k-dimensional space. In the training phase, system states generated by the simulation are clustered as illustrated in Figure 5. When the real state of the system is observed using data collected from sensors, the values of the modeling parameters that can simulate the closest state are selected. The disadvantage of this method is that a large number of cases that can occur must be generated.

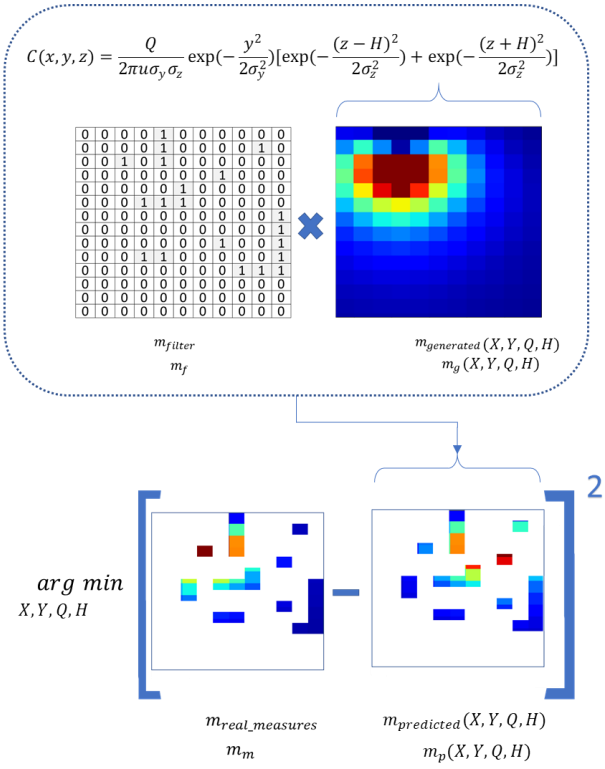

Figure 3 – Visual illustration of the optimization problem that is used to determine parameters that describe the full air pollution map of the observed area.

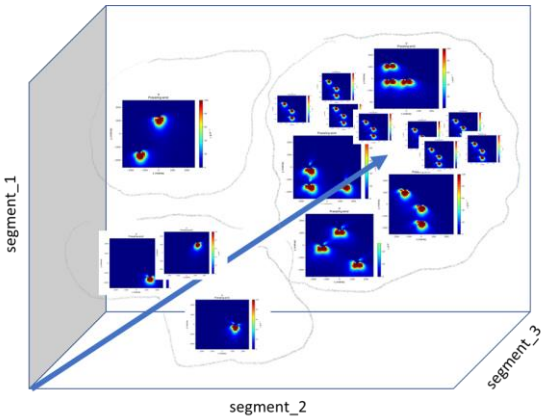

Figure 4 – Illustration of the clustering of generated air pollution maps in a three-dimensional coordinate system. The implemented platform uses 22-dimensional coordinate system to represent air pollution map.

a) 3x3	*b) 4x4*	*c) 5x5*	*d) 6x6*
$\binom{9}{3} = 84$	$\binom{16}{3} = 560$	$\binom{25}{3} = 2300$	$\binom{36}{3} = 7140$
$\binom{9}{2} = 36$	$\binom{16}{2} = 120$	$\binom{25}{2} = 200$	$\binom{36}{2} = 630$
$\binom{9}{1} = 9$	$\binom{16}{1} = 16$	$\binom{25}{1} = 25$	$\binom{36}{1} = 36$
$\Sigma = 129$	$\Sigma = 696$	$\Sigma = 2625$	$\Sigma = 7806$

$$C_3^n = \frac{n(n-1)(n-2)}{3!} \leq L$$

Figure 5 – Samples that illustrate the number of generated maps if we use grid of 3x3 up to 6x6 possible locations where the location where the air pollution source may be located. With integer L we set the upper limit of samples that we want to generate for training.

Generating examples for network training grows with exponential complexity but can be calculated once and new examples can be added incrementally and then predictions are generated in real time in almost constant complexity. On the other hand, the first method proposed does not have a training phase, but predictions are expected to find global minima of complex functions in real time, which has no guarantee that it will give satisfactory results in a sufficiently short expected time.

Attributes of K-Medoids medoids represent the generalization of attributes for members of corresponding clusters.

The third method that authors propose is a fast and light simplified system that is trained with simple classification and regression models. The third method is applicable if it is assumed that certain conditions are valid which can be expected in the problem of estimation of a small number of dominant pollution sources and in a system where measurements are collected from previously known locations. The authors also provide a comparison of their methods with air pollution maps and sources estimation using linear and cubic interpolation methods.

For the problem of collecting measurements from sensors and data representation, the authors propose two approaches as illustrated in Figure 7. One applies Bresenham's straight-line scanning algorithm and is applicable to a system in which the sensors are mounted on buses that are supposed to drive on predefined known routes and the other is based on the K-Means algorithm and is more general but more computer-intensive.

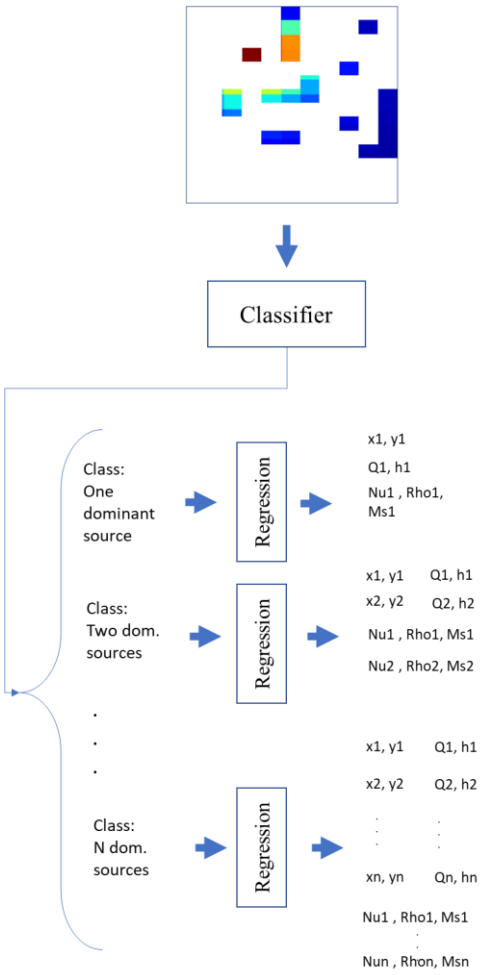

Figure 6 – Illustration of a light approach for predicting air pollution sources parameters if it is assumed that there are no more than N dominant sources in observed area.

The goal is to group measurements from geographic space to finally many groups. The data were selected to be aggregated by mean to 22 groups.

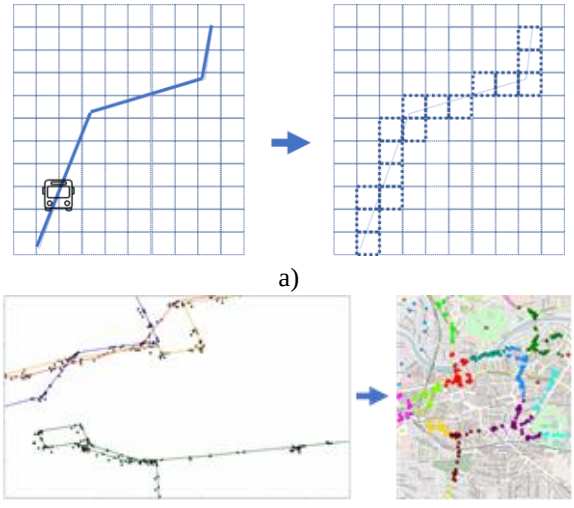

Figure 7 – The methods used to assign sensor data to one of the finite many regions used to represent space. a) Method one: Brasenham's straight line scanning and b) Method two: K-means Clustering.

IV. CONCLUSION

The authors have shown by simulations that it is possible to predict with high accuracy up to three dominant sources of air pollution using the proposed system. A system for collecting data from sensors mounted for buses in a mobile sensor network is presented. The problem of power supply of sensors and regulations regarding permits and maintenance of sensors has been solved through the selected type of vehicle carrying the sensor, and those are buses that are the property of the companies that realizes public city transport. In addition to sensors for measuring PM2.5 air pollution, buses also carry GPS sensors to locate in the area where pollution measurements were carried out. Figure 8 shows the result of air pollution estimation for a test situation that was randomly selected. It can be noticed on diagram that shows the route of the bus and Measured locations in Figure 8 that pollution is not measured at every point of space through which the bus passes, which depicts the real application because the frequency with which the sensor can measure time must be taken into account and different spatial distances occur for different bus speeds between two consecutive measurements of the same sensor. Since the fixed time that elapses between two consecutive measurements of one sensor (respecting the measurement frequency of that sensor) and the bus carrying the sensor does not move at a constant speed, the absolute spatial distances between two consecutive measurements of the same sensor differ.

The lower part of Figure 8 shows the results of success of different methods of estimating the characteristics of pollution sources and real values from test examples, which can be concluded by visual analysis since subjective assessment of similarity is expected since quantitative evaluation of interpolated methods is beyond the scope of this paper.

The lower part of Figure 8 shows the results of success of different methods of estimating the characteristics of pollution sources and real values from test example, which can be concluded by visual analysis since quantitative evaluation of interpolated methods is beyond the scope of this paper. The success of locating the source of pollution is visually shown on the map by marking the real location of the source of pollution and the predicted location of the source of pollution. The evaluation of the success of predicting the location and number of pollution sources analyzed on a retrospective example of 10,000 generated random maps is only 58% for classification by number of pollution sources and 68% for regression of pollution source location when the number of sources is known. It is considered that the set of simulated pollution maps used for network training is inadequate from the point of view that pollution is very locally concentrated and that pollution values measured in most locations are not influenced by generated sources since the sources are not dominant enough. Therefore, further research is planned related to pollution simulation models and the creation of more adequate data sets for models training. When the error is expressed with physical quantities, locating the pollution is unsuccessful for the order of hundreds of meters, which for many practical applications is not a problem of precision and can be used to take measures against sources of pollution. The code of the implemented system can be found on the repository: https://github.com/alex-gavric/icist21.

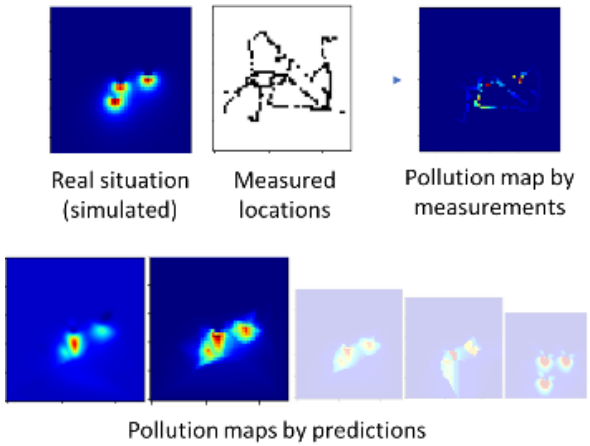

Figure 8 – Predicted maps for an instance of a simulated test data.

The system is based on a Gaussian plume model simulation but other models of air pollution can be further examined. The system assumes that the dominant sources of air pollution are stationary. That is always true for a sufficiently short moment of time thus air pollution sources can be considered immobile.

Further improvements to the system may predict the mobile sources of pollution and simulate obstacles that prevent the expected spatial distribution of pollution.

REFERENCES

[1] Richard Fuller, Karti Sandilya, David Hanrahan, “The 2019 Pollution and Health Metrics: Global, Regional and Country Analysis report”, Global Alliance on Health and Pollution (GAHP), January 2021, Retrieved from https://gahp.net/wp-content/uploads/2019/12/PollutionandHealthMetrics-final-12_18_2019.pdf

[2] L. Fang-Yie, "Keynote: Air pollution source identification by using Neural Network with Bayesian Optimization," 2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Tangerang, Indonesia, 2019, pp. 82-82,

[3] Hu, K., Rahman, A., Bhrugubanda, H., & Sivaraman, V. (2017). “HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation from Fixed and Mobile Sensors,” IEEE Sensors Journal, 17(11), 3517–3525.

[4] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele, “Deriving high-resolution urban air pollution maps using mobile sensor nodes,” Pervasive and Mobile Computing, vol. 16, Part B, pp. 268 – 285, 2015.

[5] B. Predic, Z. Yan, J. Eberle, D. Stojanovic, and K. Aberer, “Exposuresense: Integrating daily activities with air quality using mobile participatory sensing,” in Pervasive Computing and Communications Workshops (PERCOM Workshops), IEEE International Conference on, March 2013, pp. 303–305

[6] Paul Connolly, "Computer Practical: Gaussian Plume Model", October 2017, Retrieved from https://personalpages.manchester.ac.uk/staff/paul.connolly/teaching/practicals/material/gaussian_plume_modelling/gp_notes.pdf