Processing and Visualizing Traffic Pollution Data in Hanoi City from a Wireless Sensor Network

Dang Hai Hoang[*], Thorsten Strufe[**], Quang Duc Le, Phong Thanh Bui,
Thieu Nga Pham, Nguyet Thi Thai, Thuy Duong Le, Immanuel Schweizer[**]
[*]Ministry of Information and Communications, Hanoi, Vietnam
Email: hdhai@mic.gov.vn
[**]TU Darmstadt, Germany
Email: strufe@cs.tudarmstadt.de, schweizer@tk.informatik.tu-darmstadt.de
Hanoi University of Civil Engineering, Information Technology Faculty, Hanoi, Vietnam
Email: {quangld, phongbt, thieunga.pham, nguyettt, duonglt}@nuce.edu.vn

***Abstract*— Hanoi city is currently dealing with rapidly increasing air pollution that result from variety of sources. The main cause of pollution is exhaust gas from traffic system with a very large number of private vehicles. In order to help the city's environment authorities monitor the level of air pollution, a wireless sensor network is currently under development to collect traffic pollution data measured by a number of gas sensors. This paper focuses on how to process pollution data and visualize level of pollution relying on available datasets collected from sensor network. The volume of data collected from each area of the city can be very large and dynamic due to the number of mobile sensors deployed in the same area at the same time and their measurement frequency. First, we present a method for processing raw data using calibration and data clustering techniques. Second, we describe how measurement datasets are visually represented on the city's online map on the basis of mathematical interpolation method that corresponding to characteristics of environmental data. And then we also use computer graphic technique to improve the visualization quality. Finally, this paper show the result of those methods with sample data collected from an urban district of Hanoi City on a website by which we do not only provide to viewer the actual level of pollution by position but also by time.**

***Index Terms*— Wireless sensor networks, pollution data visualization, statistical techniques, mathematical interpolation, computer graphic technique.**

I. INTRODUCTION

In Vietnam, the process of urbanization recently causes various environmental issues in which the rapidly growing of air pollution generated by traffic is the most challenging problem that Hanoi city need to resolve. Hanoi, the capital of Vietnam, is the second largest city with highest population of approximate 8 millions. Its traffic density is really high due to continuous increment of millions of private vehicles, mostly motorcycles. Therefore, the level of air pollution is going to be seen as a new threat that directly affects life quality of citizens who have to get into traffic jam on their motorbike everyday without any protection.

Environmental pollution measurements today are carried out using static high-precision sensors. During the past few years, the field of wireless sensor network has been well-developed and widely adopted with many researches and applications [1][2]. One of the most potential applications is to monitor urban environment with a set of parameters such as noise, traffic density, temperature, air quality and exhaust gas concentration etc. The main purpose of this application is sustainable development in which the governments build up such smart cities that can detect and control environment hazard [2].

Currently, we are working on a project funded by the Alexander von Humboldt Foundation (Germany). The main goal of this project is to develop and test a smart sensor network for online monitoring of traffic-generated pollution data in Hanoi City. The main characteristics of this wireless sensor network are sensor mobility along loop paths and high frequency of measurement. However, the content of this paper does not deal with the description of this network, but concentrate on the processing and visualizing traffic pollution data which is collected from our wireless sensor network. Measurement dataset collected from wireless sensor network is an important part of the whole application. In order to accomplish data processing and data visualization, there are some challenges that we need to address, as follows.

(1) Because of multiple sensors measuring gas concentration together, those sensors need to be calibrated to provide accurate data.

(2) Large and dynamic dataset, which depends on number of mobile sensors in the same area at the same time and measurement frequency, can make measurement value more difficult to be visualized online. Therefore, we choose data clustering as a solution to evaluate typical measurement value for each location in a certain period of time.

(3) For visualizing levels of pollution on map, the measured areas should be divided into equal cells by a mesh. The level of pollution of a cell is represent by an appropriate color based on values of data points inside it. Whereas density of data points are not evenly among areas. Therefore, we use an interpolation method to calculate value of each cell and corresponding graphical technique to improve quality of visualization.

(4) Since Hanoi is a very large city, we only select an

978-1-4799-0540-9/13/$31.00 ©2013 IEEE

urban district that has high traffic density for our pilot project. In this project, we focus on how to process measurement data and how to visualize them online.

In this paper, we present data pre-processing techniques such as calibration and data clustering. And then we describe method of data representation in order to visualize gas concentration on map, using inverse distance weighted interpolation.

The paper is organized as follows. Section II will introduce measurement data pre-processing. Section III will focus on data representation and data visualization. Section IV will describe a pilot project that we carried out with measurement data collected from a district of Hanoi city. Related work is given in Section V. Section VI will wrap-up the paper.

II. Data Pre-Processing

Data pre-processing is the first and important step in any data analysis system. Before gathered-dataset is processed and represented to user, it needs to be cleaned. Raw data can be incomplete, incorrect or unreliable because of some reasons. For example, losing GPS signal makes the position information missing from measurement data. The cleaning phase is embedded in measurement process of sensor with checking GPS signal at source and eliminating irreverent values at destination. After cleaning, data is processed by two steps before being showed to user: calibration and data clustering.

A. Calibration

In our system, measurement data in an area are sent from sensors to data processing server via sensor network. Those sensors have some advantages such as small size, mobility, and low-cost that makes them suitable to be mounted on public transport vehicle. The main drawbacks of low-cost sensors are their limited accuracy, low stability and poor selectivity [11]. In our experiment with a certain gas concentration, different sensors, with the same type, do not give the same measurement values. The deviation of measurement values between two sensors could be large enough to make dataset become unreliable. Thus, measurement values from different sensors need to be calibrated to make sure they have approximate values if measuring in the same environment. Calibration is performed by modifying value of appropriate parameters for each sensor. Each type of sensor has its own parameters that can be used for calibration. For example, Carbon Monoxide sensors TGS2442 produced by Figaro [12] has Load Resistance parameter.

Fig. 1 shows the result after calibrating measurement values from three sensors. We collect measurement data in raw format (Voltage output) and use parameter of each sensor to calculate gas concentration values (ppm unit). In Fig. 1a, we use the same Load Resistance parameter for three sensors. In Fig. 1b, data values from sensor 2 and sensor 3 are adjusted by modifying Load Resistance parameter to values of sensor 1, approximately. In order to ensure accuracy of measurement data, sensor calibration should be based on reliable measurement from a high quality sensor. This kind of sensor is usually very expensive and bulky, however. Therefore, in our scenario, the high quality sensor is a calibrated sensor to ensure it can give measurement data closest to average values of many sensors if they all measure a same gas concentration. Before deployment, all sensors measures with the same interval in the same conditions to determine the average measured value. Then we choose parameter for each sensor to make sure the deviation between measured values of the sensor and the average values is less than a given threshold. Consider sensor z takes measurement $p(t_i)$ at time t_i and the average measurement of all sensors is $\tilde{q}(t_i)$. That deviation can be evaluated by the mean absolute error M [11].

$$M=\frac{1}{|B(T_S,T_E)|}\cdot\sum_{t_i\in B(T_S,T_E)}|p(t_i)-\tilde{q}(t_i)| \qquad (1)$$

Where $B(T_S,T_E)=\{t_i, T_S<t_i\leq T_E\}$ is the set contains all measurement time between T_S and T_E. Thus, we choose a sensor as a standard for keeping in storage and bring the other sensors for deployment. To reduce the influence of sensor aging, the deployment sensors are calibrated with standard sensor with a periodic cycle of 3-6 months.

B. Data Clustering

Data clustering is basically a method that deals with the problem how to divide data objects into groups such that similar objects should belong to the same group.

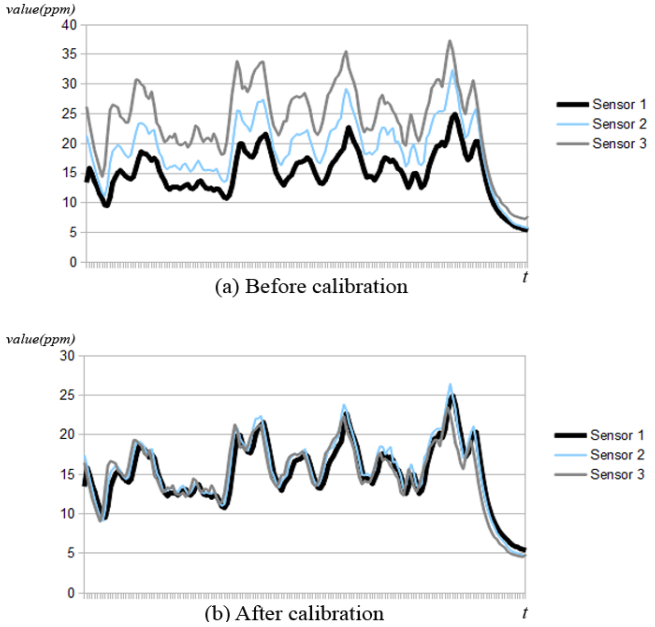


Fig. 1 Calibration sensors TGS2442 by choosing appropriate load resistance

In measurement dataset, each data point is characterized by spatial attributes (latitude, longitude), timestamp and gas concentration value. Therefore, clustering measurement dataset should be carried out not only on the basis of data point density but also on spatial limit of each group.

There are two basic types of clustering algorithms [24], partitioning and hierarchical algorithms. Hierarchical algorithms creates a hierarchical tree, so called *dendrogram,* in which each node is a subset of dataset and every leaf only has one instance. In this tree, each node is a data cluster. This algorithm requires a stopping criterion that determines how

deep the dendrogram should be created and how data clusters can represent natural attributes of dataset. In our application, we can choose stopping criterion such that the maximum distance of each group does not exceed given Dmax. However, in term of complexity, hierarchical algorithms have the minimum complexity of $O(n^2)$ so that it is not suitable with large dataset.

Partitioning algorithms create a partition with *k* clusters from *n-instance* dataset. One of the most popular partitioning algorithms is K-means [25]. Computational complexity of K-means is O(n.k.t) in which n is size of dataset, k is number of clusters and t is the number of iteration. If k and t are both relatively small, then K-means complexity can be considered as linear. As a result, with regard to complexity, K-means is seemingly a better choice for large dataset in comparison with hierarchical algorithms.

We also try to cluster data using DBScan [25], which is a density-based algorithm. In the case of our application, data points distribute with high density along a street. The result shows that clusters have large size and locate in a large space. Therefore, it needs a method to partition clusters such that each cluster locates in a region that has smaller radius.

Base on the fact we have mentioned above, we choose K-Means for clustering data. However, it is difficult to determine *k,* particularly very large spatial dataset. If inappropriate choice may result in a poor partition. Moreover, if we set a limit for cluster's radius, then finding an appropriate k is more difficult..

We developed an algorithm called Limited-Distance-K-means based on K-means, in order to:

(1) determine parameter k and initial clusters that have radius less than a given distance value.
(2) have acceptable computational complexity.

This Limited-Distance-K-means algorithm includes two stages. First, we create set *S* of *k* initial clusters. Second, we make adjustment to set *S* to remove unnecessary cluster with a modified K-means algorithm. After the first stage of algorithm, number of cluster can be large because of arbitrarily choosing instance from dataset. Subsequent instance tends to be selected into new cluster. Therefore, we can make clustering more effectively by modifying K-means algorithm to remove a cluster if all of its instance can be assigned to the other clusters.

Distance function is used in the algorithm is Euclidean distance function.

Algorithm:
Input:
D: Dataset
radius: the radius restricted for each cluster
t: the interval time
Output:
clusters S= $\{S_j\}$ with S_j is a cluster.
Variables:
m_j: the mean to be the center of cluster.
$\boldsymbol{x_i}$: The instance of dataset
belong: the determination whether the instance belong to any cluster or not.
D_t: The dataset including data in the given interval time

Initiate **S**={empty};
Select $\mathbf{D_t}$ from **D**;
for each instance $\boldsymbol{x_i}$ of $\mathbf{D_t}$
    set belong = false;
    for each $\mathbf{S_j}$ of **S**
    call **center;**
    Calculate distance from i to $\boldsymbol{m_j}$ of $\boldsymbol{Sj}$;
    if (belong = false && distance < radius)
        add i to S_j;
        belong =true;
    end if
    end for
    if (not belong)
// mean that the instance is not assigned to any cluster.
        Initiate new $\mathbf{S_{(j+1)}}$**;**
        add $\boldsymbol{x_i}$ to $\mathbf{S_{(j+1)}}$;
    end if
end for
repeat
    call **center;**
    for each instance $\boldsymbol{x_i}$ of D_t
        assign $\boldsymbol{x_i}$ to the cluster whose center $\boldsymbol{m_i}$ is closest to it.
    end for
    for each $\mathbf{S_i}$ of **S**
        if($\mathbf{S_i}$ is empty) $\mathbf{S}\backslash\{\mathbf{S_i}\}$
    end for
until there is no change on the clusters.
return S;
function center:
    Calculate the new means $\boldsymbol{m_j}$ for instances in the new clusters

$$m_j = \frac{1}{|S_j|}\sum_{x_i \in S_j} x_i$$

    return $\boldsymbol{m_j}$

Computational complexity of Limited-Distance-K-means is O(k.n.t) in which k is number of clusters, n is size of dataset and t is number of iteration.

III. VISUALIZATION

After pre-processing, measurement data are organized as a set of data points in which one point represent a pollution level in a certain location. Each point has properties of latitude, longitude, measured data and measurement time. In order to visualize pollution level on map, each data point can be assigned to a colored cell [15]. But dataset has variety of densities according to location where the measurement takes place. It can be dense on traffic-lines and sparse on other areas. If each point is represented by a colored cell, the cells can be overlapped or too sparse to give effective visualization. Therefore, in order to represent level of pollution independently of density of data points, we use a regular grid to divide entire measurement area into equal cells. The value of a

cell is calculated from data points inside it. There are three regular tessellations including rectangular grid, hexagonal grid and triangular grid. In our approach, hexagonal grid is used because of its advantages. In order to determine value for each cell, we apply Inverse Distance Weighted interpolation method [7]. The inverse distance weighted procedure is versatile, easy to program and understand, and is fairly accurate under a wide range of conditions [10], include visualizing pollution data [9].

A. Hexagonal Grid

The rectangular grid is more widely used than other grids because of its symmetrical, orthogonal co-ordinate system. It is particularly suitable for use of Geographic Information Systems in many systems [4]. However, the hexagon grid maps have some visual advantages compared to rectangular grid maps [3]. Maps in a hexagonal grid is less ambiguous than maps in a rectangular grid. The grid lines in rectangular grid maps are vertical and horizontal lines crossing the surface. The human vision is especially sensitive to vertical and horizontal lines. Therefore, the grid lines in rectangular grid maps are distracting [4][5]. Cells in a hexagonal grid are aligned along three axes rather than just two in a rectangular grid. Thus, the outlines of groups of cells in a hexagonal grid can form more varied, less rectilinear shapes than groups of cells in a rectangular grid [4][6]. Moreover, hexagon shape is closer in shape to circle than square. This characteristic is useful when determining how to get value point for applying interpolation. Instead of to determine which data points inside hexagonal cell, we just need to find data points inside the circumcircle of hexagonal cell.

B. Inverse Distance Weighted (IDW)

Inverse distance weighting models use the notion in which points further away should have a lower contribution than those near the point of interest. Value of target point is can be interpolate by observations in the following equation:

$$p(x_0) = \frac{\sum_{i=1}^{n} \frac{1}{d_i^{\alpha}} p(x_i)}{\sum_{i=1}^{n} \frac{1}{d_i^{\alpha}}} \quad (2)$$

The choice of power parameter α is arbitrary [13]. Increasing α, the data points closer to x_0 have more contribution to value of x_0. Decreasing α to 0, interpolation becomes calculating average value of n surrounding points [14]. We choose the most popular choice $\alpha = 2$ because it not only gives seemingly satisfactory empirical results for purpose of general surface mapping and description, but also presents the easiest calculator [7].

With each hexagon cell, we assume the value of cell is value of center point of cell. If connect center point of each cell to center points of all neighbor cells, we can construct a triangular grid from hexagon grid (Fig. 2). Hence, calculating values of hexagon cells turns into interpolating grid points on triangular grid. In order to interpolate value for grid point x_0, we need to determine set of data points surrounding x_0 involved the interpolation equation (Eq. 2). Denote r is the max distance from a data point x_i to x_0 determine if x_i belong to set of observations of x_0. In other word, if data point x_i is inside circle $I(x_0, r)$, x_i is used to interpolate value of x_0. In our approach, we determine value of radius r by the distance d between two adjacent grid points and a coefficient l.

$$r = d \times l \quad (3)$$

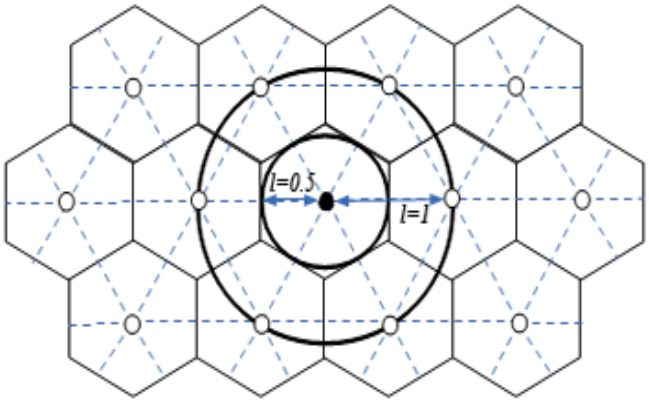


Fig. 2 Determine radius r by parameter l

Because of characteristics of equilateral triangular grid, each grid point is equidistant to its entire neighbor so that parameter $\boldsymbol{d}$ also controls granularity of the grid. If decreasing $\boldsymbol{d}$, cells are smaller and the visualization is smoother. Fig. 3a and Fig. 3b show the difference between two values of $\boldsymbol{d}$ effect to granularity level of display. However, if $\boldsymbol{d}$ is too small, the number of grid points is very large that make interpolating time rapidly increasing. In our experiment, we choose $\boldsymbol{d}$ small enough to satisfy both visual requirement and computational complexity. Parameter $\boldsymbol{l}$ is a coefficient to control the range of data points for interpolation. Relation between $\boldsymbol{l}$ and radius $\boldsymbol{r}$ is presented in Fig. 2. If $\boldsymbol{l}$ is equal to $\boldsymbol{1/2}$, circle $\boldsymbol{I}$ is the incircle of hexagonal cell. If $\boldsymbol{l}$ is equal to $\boldsymbol{1/\sqrt{3}}$, $\boldsymbol{I}$ become the circumcircle. Increasing $\boldsymbol{l}$, the area for collected data point is larger so that there are more data points collected to calculate. Hence, the visualization becomes smoother with fewer steep gradients between two points. Fig. 3c and Fig. 3d show the visual effect if keeping $\boldsymbol{d}$ fixed and changing $\boldsymbol{l}$.

The choosing l depends on data density and requirement of user. If l is small, there may be only a few grid points that are close enough to data points have interpolated value while the others do not. In our dataset, most of data points distribute around traffic-line. Thus, the cells that have interpolated value also should be converged along the streets. In this case, parameter l controls the width of cluster of cells that cover data points on the streets.

When increasing l, the region with data view is expanded in both side of road so we call l is spread level. Fig. 4a and Fig.4b show that the representation of data points is spread broader if l is set to a higher value.

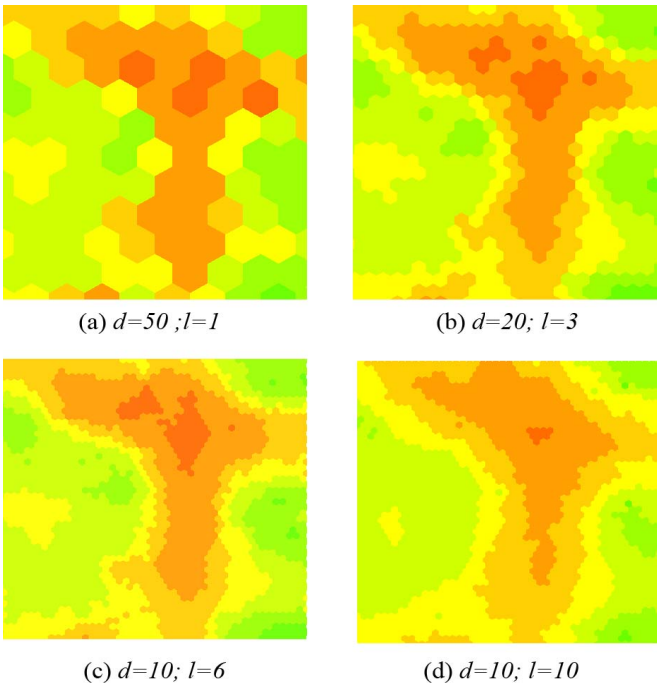
C. Showing Pollution Data on Map

In our system, pollution data is represented on website using Google Map service. One of the projects using Google Map service is NoiseTube[15]. Users can access NoiseTube website and download KMZ files containing measurement data

and open them by Google Earth.

Another project also using online map services for visualizing noise pollution is NoiseMap[16]. Unlike NoiseTube, website of NoiseMap project allows user can interact with the online map on website. The basic idea of this method is that visualization is restricted by the map view on user's device. Whenever user changes the map region, such as pan or zoom, the image showing pollution level for current region is generated and display for user as an overlay on map.

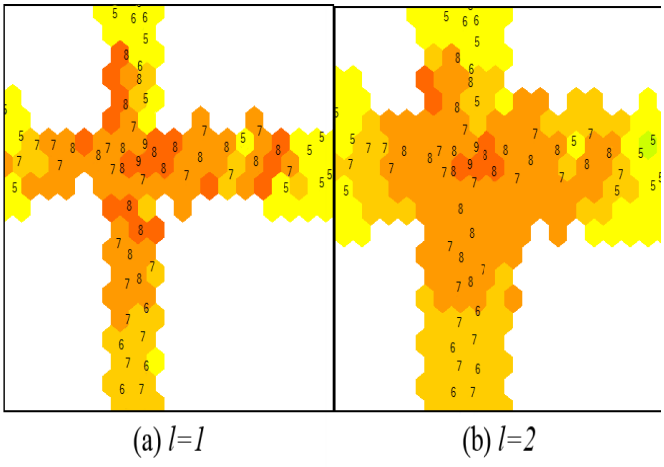
In our approach, the overlay image is a hexagonal grid constructed from all data points available in the visible region by interpolation method we mentioned above. Fig. 5 is showing pollution level on a map region with d in (Eq. 3) is 1/40 width of current viewport and *spread level* l is 2.



(a) *d=50 ;l=1* (b) *d=20; l=3*

(c) *d=10; l=6* (d) *d=10; l=10*

Fig. 3 Visual effects when changing d and l



(a) *l=1* (b) *l=2*

Fig. 4 The drawing region expanded when increasing spread level l

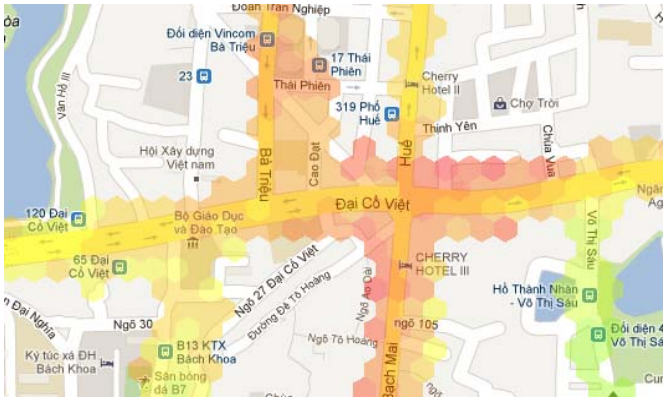


Fig. 5 Hexagonal grid showing pollution data

IV. The Pilot Project

In this section, we present result of a pilot project with data measured from an urban district in Hanoi city.

Hai Ba Trung is one of the two districts of Hanoi city that have highest traffic density. There are more than seventy-five thousands of resident families in which each own at least two motorbikes. Besides, many personal vehicles come from other regions since Hai Ba Trung district has a complex of a garment industry park, two river ports and three largest universities including Hanoi University of Science and Technology, University of Civil Engineering and National University of Economics.

Recently, high traffic density in this district is said to cause severe air pollution that makes a serious impact on local resident. Studies show that 70% air pollution in this district in particular and Hanoi in general is caused by traffic. At intersections that frequently have traffic congestion, exhaust emissions are usually very high. Therefore, we choose streets with large traffic density to set up a pilot project.

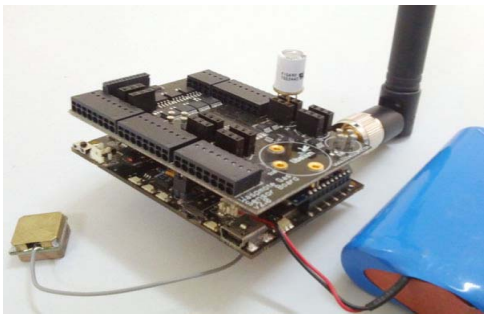


Fig. 6 Waswpmote with CO sensor

In our pilot project, we use 10 wireless sensors that cover 23 main streets on 8 routes of the district to measure CO gas concentration. The routes were design to give Waspmotes the ability to send data to each other and forward data to the base station. Thus, there should be one common street distance between any two routes. Therefore, the number of collected data may be very large. In consequence, processing data becomes more necessary.

The reason why we choose CO is that CO is a typical

exhaust gas generated from traffic. CO is also classified as toxic gas that can seriously damage people health. During measurement time, each sensor is attached on a volunteer motorcycle running at the speed of 10 km per hour. Measurement value is captured from CO sensor for every 30 seconds, and then stored on local storage.

The Carbon Monoxide TGS 2442 sensors is mounted on Waspmote, which is a compact, versatile and high mobility device, developed by Libelium [26] (Fig. 6). Parameters of each sensor is chosen to ensure the mean absolute error (Eq. 1) from the standard sensor must not exceed 1.5 ppm threshold. Fig. 7 shows the change in the mean absolute error when changing the parameters of the sensor. Based on the graph, in order to keep the error below 1.5 ppm, we can choose the value of the LR parameter is 20 kΩ.

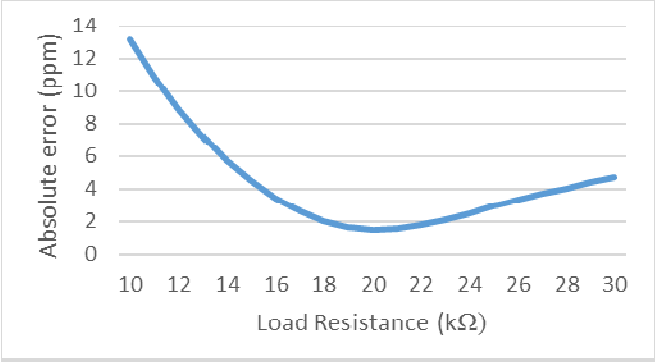


Fig. 7. Mean absolute error when changing the parameter

Every hour, our server receives thousands of measurement records of data. We applied a Limited-Distance-K-means algorithm for clustering data and we could find out typical values for each location every hour using this algorithm. The locations are separated from each other by a distance of radius about 50m.

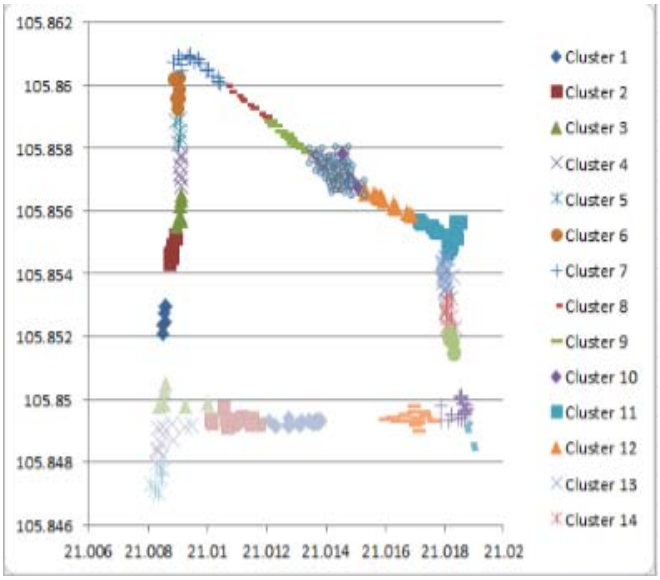


Fig. 8. Distribution of data under clustering

Figure 8 presents an example of data clustering for a route along four streets. Dataset for clustering in one hour has 897 records, each record includes the attributes: waspId, latitude, longtitude, time, measuring value. For this example we get following result: Dataset is divide to 62 cluster with the radius of 50 m.

After clustering, each cluster will be represented by the average value of the measured values of the points in the cluster. Graph in Fig. 9 shows the difference of CO concentration levels at different times, particularly in comparison with peak hour. Graph in Fig. 10 indicates the difference of concentration measured at different positions along a street. Levels of CO gas at intersections, which are represented by point 1,2, 3 in Fig. 10, are significantly higher than those at other position.

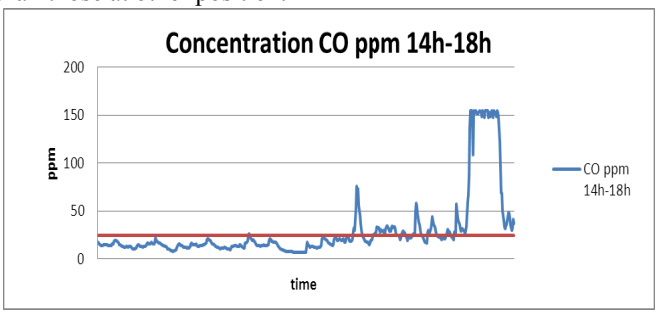


Fig. 9. Concentration of CO at different times

Measurement data collected from all 23 streets on 8 routes have been processed and visualized on an online map available at http://han-sense.vn (Fig. 11). The level of pollution, explicitly based on CO concentration, is represented by a range of colors on map in which each color matches with a certain value of concentration. User will be able to choose options that display measurement data by a certain time period, as well as by type of gas. Currently, we only measure CO concentration but we plan to measure more exhaust gases in the future. Providing measurement data on a public website does help user have more useful information about air pollution in traffic.

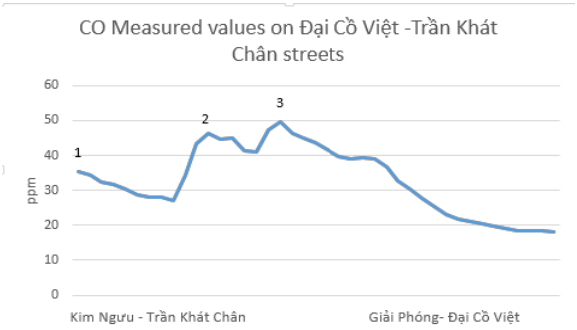


Fig. 10. Concentration of CO at different positions along Đại Cồ Việt street

Moreover, user can compare the level of pollution between different areas (Fig. 11).

The QCVN05 Standard of Vietnam defines the limit of CO concentration as follows:

TABLE I. LIMIT OF CO CONCENTRATION

Unit	Avg. an hour	Avg. 3 hours	Avg. a day
$\mu g/m^3$	30000	10000	5000
ppm	24.3	8.11	4.05

In comparison to the defined limit in QCVN05 Standard, the CO concentration level we measure at peak hour is almost 7 times higher. That high concentration of CO as we can see on data graph shows that traffic air pollution in Hai Ba Trung district is possibly over the threshold that can seriously affect health and quality of life.

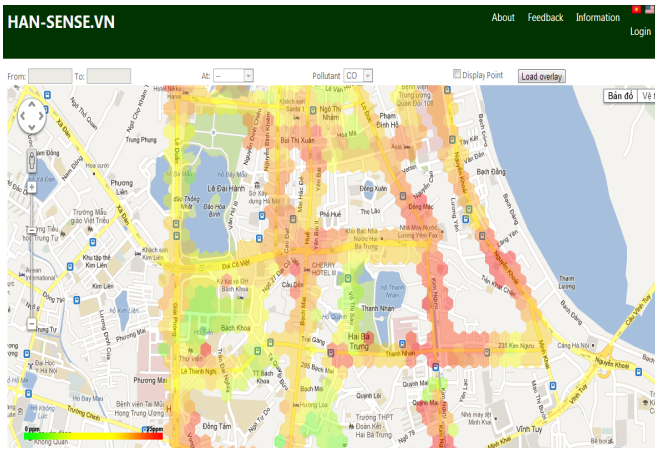


Fig. 11 Pollution data on map Han-sense.vn

We designed the routes for Waspmotes, so that the total road each packet, starting at the location of collecting and ending at the base station, should be always less than 10km. Since Waspmotes move with a speed of 10km/h, we can ensure that data from the farthest Waspmotes can be forwarded to the center. Thus, the base station is ensured to receive updated data after every hour.

Project's monitoring results provide city government a trustworthy view about current traffic pollution so that leaders could make appropriate policies such as improving public transportation, reducing personal vehicle, developing city infrastructure and moving industry parks and universities to outside of the city. On the other hand, they can also apply new technologies and environment standards for transport vehicles. The more people know about the project's website, the more awareness and responsibility they have about their living environment and how to protect themselves from pollution.

With achieved results, the pilot project demonstrated the feasibility of the project in terms of scientific, technical and practical significance in monitoring the traffic pollution. In the future, instead of having the data that are collected by volunteers on motorcycles in some designed trajectories, we might mount sensors onto the buses moving around some fixed routes. However, it leads to some problems as following:

- We must have the support of the city leaders and the permission of the transport sector to implement the project.

- The project scale would be much larger, because the scope is not only in one district, but also in many districts of Hanoi, especially in the inner districts, since the buses cross several districts.

- The installation of the sensors on the bus is complicated. It's not easy to protect the sensor boards against tropical weather of Vietnam that is hot and humid with much rain. Moreover, we must prevent the breakdown and the loss of sensors from the variety of the traffic participants.

V. RELATED WORKS

Several research projects on the field of wireless sensor networks have been initiated, for instance [16-23,27], which aim to investigate various aspects of WSN technologies and techniques for an integrating communication infrastructure that enables diverse applications in the civil life.

NoiseMap [16] is a real-time participatory sensing application developed by a research group at TU Darmstadt and implemented in the DA-SENSE project. NoiseMap is built upon a sensor network connected with a data platform. It also provided a Web-based map that displays real-time noise density all over Darmstadt city.

RESCATAME [17] is a smart city project developed in Salamanca, Spain. Its objective is to improve environment monitoring and sustainable management of traffic. The most important milestone of the project is wireless sensor network deployment. Each sensor captures data of temperature, CO and NO2 concentration level and then transmits them to data server either via base station or via 3G/GPRS connection. However, the project is still under development and lack of data providing service like NoiseMap so that users are not able to access any information.

The PermaSense project [18] aims to gather environmental data on high-mountain permafrost in the Swiss Alps. In this wireless sensor network, the base station collects data via GPRS/EDGE system.The backend consists of a server for storing collected data in a data base and for management.

The OpenSense Zurich project [27] is a big project developed at ETH Zurich for monitoring air pollution aiming to increase public awareness of urban air pollution and to involve general public into environmental monitoring. OpenSense nodes are mounted on trams and public buses to record air pollution. Personal data collectors (e.g. smart phones) may participate in environmental monitoring to upload data to the system and have access to real time air pollution data from system. Several challenging issues have been addressed including on-the-fly-calibration of sensor nodes, calibration accuracy, measurement accuracy,detection and filtering of systematic and transient sensor errors, area coverage problem, routing selection. The most drawbacks of low-cost sensors are their limited accuracy and resolution. Thus, the sensors requires frequently recalibration in order to provide precise pollution recordings and to increase sensing accuracy. Using improved temporal and spatial resolution the platform can provide fine-grained air pollution maps.

Although there is significant number of related works on wireless sensor networks, most of them usually use sensor nodes installed on public transport networks or sensor nodes may be part of mobile sensor networks.

VI. Conclusion

In order to continuously measure and monitor air quality of city traffic, tradition method is not adequate, particularly in a big crowded city with high traffic density like Hanoi. Therefore, using wireless sensor networks comes up as a new approach. Because of some characteristics of wireless sensor networks such as sensor mobility, sensor density, and measurement frequency, dataset can be so large and dynamic. In this paper, we have chosen calibration and data clustering techniques for pre-processing of measurement data and suitable methods for visualizing data on a map, and successfully applied to a pilot project for monitoring pollution data of an urban district in Hanoi City.

Acknowledgment

This research project has been supported by the Alexander von Humboldt Foundation (Germany). This work has been co-funded by the DFG as part of the CRC 1053 MAKI. We would like to thank the sponsors for offering the opportunity, by which German and Vietnamese researchers can cooperate in this joint project.

References

[1] Kavi K. Khedo1, Rajiv Perseedoss, Avinash Mungur. Wireless sensor Network Air pollution Monitoring system. International journal of wireless and mobile networks, vol2, no2, May 2010.

[2] G. Wener-Allen, K. Lorincz, M. Ruiz, O. Marcillo, J. Johnson, J. Lees,M. Walsh, Deploying a wireless sensor network on an active volcano, Data-Driven Applications in Sensor Networks (Special Issue), IEEE Internet Computing, March/April 2006.

[3] Carr, D.B., Olsen, A.R., White, D., 1992. Hexagon mosaic maps for display of univariate and bivariate geographical data. Cartogr. Geograph. Inform. Syst. 19, 228–236.

[4] Colin P.D. Birch, Sander P. Oom, Jonathan A. Beecham. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. Ecological modelling 2 0 6 (2007)347–359.

[5] Coppola, D.M., Purves, H.R., McCoy, A.N., Purves, D., 1998. The distribution of oriented contours in the real word. Proc. Natl.Acad. Sci. U.S.A. 95, 4002–4006.

[6] Overton, W.S., White, D., Stevens Jr., D.L., 1990. Environmental monitoring and assessment program: design report. Report No. EPA/600/3-91/053, U.S. Environmental Protection Agency, Environmental Monitoring and Assessment Program, Washington, DC, 52 pp.

[7] Shepard, Donald. A two-dimensional interpolation function for irregularly-spaced data. 1968. Proceedings of the 1968 ACM National Conference. pp. 517–524.

[8] Willmott, C. J. and Matsuura, K. (1995): Smart interpolation of annually averaged air temperature in the United States. J. Appl. Meteorol.,34: 2577-2586.

[9] Dilip Kumar Jha, M.Sabesan, Anup Das, N.V.Vinithkumar R. Kirubagaran. Evaluation of Interpolation Technique for Air Quality Parameters in Port Blair, India. Universal Journal of Environmental Research and Technology.Volume 1, Issue 3: 301-310.

[10] Lam, N. S. 1983. Spatial interpolation methods review. The American.Cartographer10: 129-149.

[11] David Hasenfratz, Olga Saukh, and Lothar Thiele.On-the-fly Calibration of Low-cost Gas Sensors.Lecture Notes in Computer ScienceVolume 7158,2012,pp 228-244.

[12] Figaro Group.Technical Information for Carbon Monoxide Sensors. Revised 07-2007.

[13] Webster, R. and Oliver, M., 2001. Geostatistics for Environmental Scientists. John Wiley & Sons, Ltd, Chichester, 271 pp.

[14] Brus, D.J. et al., 1996. The performance of spatial interpolation methods and choropleth maps to estimate properties at points: a soil survey case study.Environmetrics, 7: 1-16.

[15] N. Maisonneuve, M. Stevens, and B. Ochab. Participa-tory noise pollution monitoring using mobile phones. Information Polity, 15(1):51–71, 2010.

[16] Immanuel Schweizer, Roman Bärtl, Axel Schulz, Florian Probst, and Max Mühlhäuser. NoiseMap - Real-time participatory noise maps. In Second International Workshop on Sensing Applications on Mobile Phones, 2011.

[17] Marta Gómez. RESCATAME ProjectPervasive Air-Quality Sensors Network for an Environmental Sustainable Urban Traffic Management. Intergroup meeting: AIR POLICY. 26/09/2012.

[18] J. Beutel, S. Gruber, A. Hasler, R. Lim, A. Meier, C. Plessl, I. Talzi, L. Thiele, C. Tschudin, M. Woehrle, and M. Yuecel. PermaDAQ: A scientific instrument for precision sensing and data recovery in environmental extremes. In Processing of the International Conference on Sensor Networks (IPSN), April 2009. IEEE Computer Society, pp.265-276.

[19] Project “da-sense” (Darmstadt Sensor Network), http://www.da-sense.de/.

[20] Project “Cocoon” (Cooperative sensor communication – a project at TU Darmstadt), http://www.cocoon.tu-darmstadt.de/cocoon/index.en.jsp.

[21] Project “ESNA–European Sensor Network Architecture” with 21 partners from 7 European countries, http://www.sics.se/esna/.

[22] Project “Wireless Sensor Networks for Real-Time Continuous Monitoring and Assessment of Water and Wastewater Pipes”, National Research Council Canada, http://www.nrc-cnrc.gc.ca/eng/projects/irc/waste-water.html.

[23] Project “KleeNet - Symbolic Execution of Sensornets”, RWTH Aachen, Germany, http://www.comsys.rwth-aachen.de/ research /projects/kleenet.

[24] Leonard Kaufman, Peter J. Rousseeuw, Finding Groups in Data-An Introduction to Cluster Analysis, 1990, John Willey & Sons, p.37-38.

[25] Harvey J. Miller (Editor), Jiawei Han (Editor),Geographic Data Mining and Knowledge Discovery, Second Edition, 2009, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series), p.156-157, p.170-173.

[26] Libelium, Waspmote datasheet version v.4.2, Technical Document, 4/2013.

[27] http://data.opensense.ethz.ch