Fine-Grained Air Quality Monitoring Based on Gaussian Process Regression

Yun Cheng*, Xiucheng Li*, Zhijun Li, Shouxu Jiang, and Xiaofan Jiang

Harbin Institute of Technology
{chengyun.hit,xiucheng90,fxjiang}@gmail.com,
{lizhijun_os,jsx}@hit.edu.cn

Abstract. Air quality is attracting more and more attentions in recent years due to the deteriorating environment, and $PM_{2.5}$ is the main contaminant in a lot of areas. Existing softwares that report the level of $PM_{2.5}$ can provide only the value in the city level, which may indeed varies greatly among different areas in the city. To help people know about the exact air quality around them, we deployed 51 carefully designed devices to measure the $PM_{2.5}$ at these places and present a Gaussian Process based inference model to estimate the value at any place. The proposed method is evaluated on the real data and compared to some related methods. The experimental results prove the effectiveness of our method.

Keywords: $PM_{2.5}$ concentration monitoring, non-linear regression, Gaussian Process.

1 Introduction

The air pollution is considered a major and serious problem globally, especially in some cities of developing countries, such as Beijing and New Delhi. Among the various dimensions of air quality, particulate matter (PM) with diameters less than 2.5 micron, or $PM_{2.5}$, has gained a lot of attention recently. Medical studies have shown that $PM_{2.5}$ can be easily absorbed by the lung, and high concentrations of $PM_{2.5}$ can lead to respiratory disease [1] or even blood diseases [2]. Due to its close relation to public health, it has gained a lot of attention.

Nowadays, people are looking for better ways to monitor the quality of air in their immediate environment. There are many web or smartphone applications that report publicly-available air quality data at the city or district level, however, they cannot tell the actual air quality people breath-in, which is more relevant and valuable. Actually, there is probably a significant difference between the values of $PM_{2.5}$ concentration at different locations at the district level, which has been attested by the real data as shown in Section 4. Therefore, it is necessary to develop a fine-grained air quality monitoring system.

* Corresponding authors.

C.K. Loo et al. (Eds.): ICONIP 2014, Part II, LNCS 8835, pp. 126–134, 2014.

In order to estimate the air quality of any location, two major classical ways are proposed in the past of years. One is classical dispersion models, such as Gaussian Plume models, Operational Street Canyon models, and Computational Fluid Dynamics. These models are normally a function of meteorology, street geometry, receptor locations, traffic volumes, and emission factors (e.g., g/km per single vehicle), based on a number of empirical assumptions that might not be applicable to all urban environments and parameters which are also difficult to obtain precisely [3]. The other is interpolation using reports from nearby air quality monitor stations. This method is usually employed by public websites releasing the air quality index (AQI). Recently, big data reflecting city dynamics have become widely available and a group of researchers seek to infer the air quality using machine learning and data mining techniques. In the "U-Air" paper by Yu [4], the authors infer air quality based on AQIs reported by public air quality stations and meteorological data, taxi trajectories, road networks, and POIs (Point of Interests). Since there are only a few public monitor stations in a city, their training dataset is insufficient to train a commonly used supervised learning model. Therefore they propose a co-training-based semi-supervised learning model to tackle the data sparsity problem.

To overcome the drawbacks of existing methods shown above, we present a $PM_{2.5}$ monitoring system composing of a sensor network and a inference model. The sensor network that is deployed among the area to be monitored provides the values of $PM_{2.5}$ at these places. So we are in a much different situation compared with "U-Air", since we have designed our $PM_{2.5}$ monitoring devices and deployed them at a much higher density (51 monitor stations over an 30 $km \times 30$ km urban area), which provides much more sufficient data for air quality estimation. Although our $PM_{2.5}$ devices cannot achieve the same measurement precision as the expensive public monitor stations, they are precise enough for the air quality estimation after calibration. Therefore we simply treat their readings as ground truth, and we mainly focus on the development of an effective model to estimate the value of $PM_{2.5}$ at any place using the acquired data at such an relatively higher deployment density. The paper is organized as follows: in Section 2 a description of the system is given. The inference model based on Gaussian Process is detailed in Section 3. The experiment setup and evaluation results are given in Section 4, and the conclusions are drawn in Section 5.

2 System

The system architecture of AirCloud is shown in Fig. 1. The system mainly contains two parts: 1) $PM_{2.5}$ monitoring system, which contains the AQM monitoring front-end and the backend data collection module; 2) Inference platform, which will be used to infer the unknown $PM_{2.5}$ concentrations of locations where there is not monitoring equipment. We will describe these two parts in the following subsections.

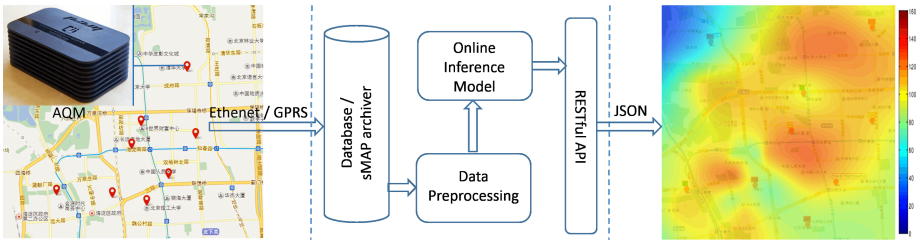

Fig. 1. The system architecture of Air-Cloud

2.1 $PM_{2.5}$ Monitoring System

The $PM_{2.5}$ monitroing system is used to collect, process and store the AQM sensor data.

$PM_{2.5}$ concentrations vary significantly over space, especially for metropolitan cities where pollution sources are multi-faceted. In addition, as we can observe from official $PM_{2.5}$ monitoring station data, $PM_{2.5}$ concentration changes at an hourly rate. As a result, direct monitoring is necessary. To solve our problem, we designed and built our own Internet-connected $PM_{2.5}$ monitors, AQM, as shown in Fig. 1. AQM station contains $PM_{2.5}$ concentration sensor, temperature and humidity sensor, the mechanical structure of the hardware is carefully designed and all the monitoring station will do the hardware calibration to remove initial hardware variations. We take an approach of using inexpensive sensors at the front-end and deployed at certain density, but rely on the reference model on the cloud to infer the $PM_{2.5}$ concentrations on the whole area.

To make the monitoring system more stable and scalable, we choose sMAP [5] as the data representation and storage system. We defined the standard specification for physical $PM_{2.5}$ sensor data, which contains the location information and the sensor readings, and use the database designed for time-series data, plus a powerful query language, provided by sMAP, as shown in Fig. 1. We use different communication approaches , Ethernet or GPRS, to connect AQM with the cloud server, and store all the data in sMAP archiver by the minute, which will be used by the inference model.

2.2 Inference Platform

The Inference platform is used to infer $PM_{2.5}$ concentrations at locations where there is not monitoring stations. We deploy the AQM monitor stations at certain density to get a general idea of the $PM_{2.5}$ concentrations around, however, to get the $PM_{2.5}$ concentration at locations where there is not monitoring stations, we have to rely on the inference platform, with the help of the Gaussian Process Inference model, we can get the accurate and fine-grained $PM_{2.5}$ concentration estimation of the whole area.

3 Air Quality Inference Model Based on Gaussian Process

In this section, we detail the Gaussian Process based inference module that estimates the value of $PM_{2.5}$ using the data from the monitoring network. First, we model the inference module as a regression after some necessary definitions. Then the problem is solved as a Gaussian Process regression and the details are presented.

3.1 Problem Definition and Regression Model

Using $\mathbf{x}_i$ to denote the coordinates of the i-th monitoring station and y_i the value of $PM_{2.5}$ at this place, the objective of the inference module is to inference the value of $PM_{2.5}$ y at any place given the data from all monitoring stations $\mathcal{D} = \{(\mathbf{x}_i, y_i),\ i = 1, \ldots, n\} \subseteq \mathcal{R}^2$ and the coordinate of the place to be estimate $\mathbf{x}$. This is a typical regression problem which is usually formulated as

$$y_i = f(\mathbf{x}_i) + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \delta^2), \tag{1}$$

where ϵ_i is the noise term and the objective is to learn a proper f from $\mathcal{D}$ which can predict a proper y for any given $\mathbf{x}$.

In our system we select Gaussian Process regression to model it, which is also known as Kirging in the spatial statistics field. Gaussian Process is a non-parametric Bayesian approach with sufficient flexibility to capture the complex and non-linear properties of the model. It has been proved to be a powerful tool in many areas and applied widely in practice. Since it is a fully probabilistic model, the objective is to learn a proper distribution of y instead of its value. We would like to detail how it is used to estimate the value of $PM_{2.5}$ at a specific place given its coordinate in following.

3.2 Gaussian Process Regression

Gaussian Process for Regression. A Gaussian process is a collection of random variables, any finite number of which have consistent joint Gaussian distributions. In Gaussian process regression problems, latent function f behaves following a Gaussian distribution (Normal distribution) when conditioning on $\mathbf{x}$

$$\mathbf{P}(f_1, f_2, \ldots, f_n | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \mathcal{N}(0, K)$$

where $f_i = f(\mathbf{x}_i)$ is latent function and K is a covariance matrix with entries given by the covariance function, $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. $k(\mathbf{x}_1, \mathbf{x}_2)$ can be any valid kernel function satisfying Mercer's condition [6].

In inference, the training and test latent values is denoted as $\mathbf{f} = [f_1, f_2, \ldots, f_n]$, $\mathbf{f}_* = [f_{*1}, f_{*2}, \ldots, f_{*n}]$ separately, we combine the prior with the likelihood function via Bayes rule obtaining the posterior distribution:

$$\mathbf{P}(\mathbf{f}, \mathbf{f}_* | \mathbf{y}) = \frac{\mathbf{P}(\mathbf{f}, \mathbf{f}_*)\mathbf{P}(\mathbf{y}|\mathbf{f})}{\mathbf{P}(\mathbf{y})} \tag{2}$$

The desired posterior predictive distribution can be produced by marginalizing out the training set latent variables $\mathbf{f}$ in equation (2):

$$\mathbf{P}(\mathbf{f}_*|\mathbf{y}) = \frac{1}{\mathbf{P}(\mathbf{y})} \int \mathbf{P}(\mathbf{f}, \mathbf{f}_*)\mathbf{P}(\mathbf{y}|\mathbf{f})\, \mathrm{d}\mathbf{f} \tag{3}$$

since the prior and the likelihood function are mutually independent and both follow Gaussian distribution as

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K_{\mathbf{f},\mathbf{f}} & K_{*,\mathbf{f}} \\ K_{\mathbf{f},*} & K_{*,*} \end{bmatrix}\right), \quad \mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \delta^2 I) \tag{4}$$

where δ^2 is the noise variance and I is the identity matrix. Then the integral in equation (3) can be computed in close form and the result is also a Gaussian distribution [7] with $\mathbf{f}_*|\mathbf{y} \sim \mathcal{N}(\mu_*, \mathbf{\Sigma}_*)$.

$$\mu_* = K_{*,\mathbf{f}}(K_{\mathbf{f},\mathbf{f}} + \delta^2 I)^{-1}\mathbf{y} \tag{5}$$

$$\Sigma_* = K_{*,*} - K_{*,\mathbf{f}}(K_{\mathbf{f},\mathbf{f}} + \delta^2 I)^{-1}K_{\mathbf{f},*} \tag{6}$$

where μ_* is the predictive mean and Σ_* is the corresponding covariance which indicate us the uncertain of the predictive value in the locations (we use $*$ as shorthand for f_*). In our scenario, μ_{*i} will be used as the predictive value y_i.

4 Experiment and Results

In experiment the real deployment dataset of more than one month was used to evaluate the performances of Gaussian Process Inference. There are totally 51 monitor stations deployed in an area with the size of 30 $km \times 30$ km and each station reports its measurements every 30 minutes, the deployment map is shown in Figure 2-(**A**). We deliberately remove one station as ground truth and infer its

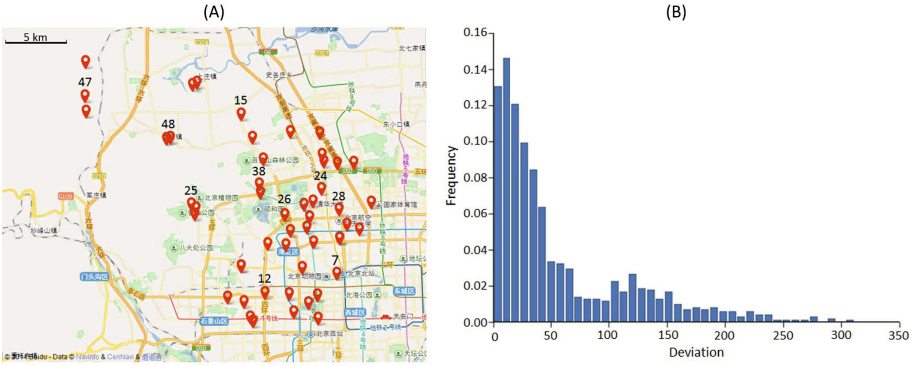

Fig. 2. (**A**) The deployment map of the monitor stations; (**B**) The distribution of the deviation between station S_{26} and S_{27} over one month

value using the remaining stations' reading at each timestamp. U-Air provides a public online visualization of its inference result in [8]. However, the AQI inferred by U-Air is just five standard levels specified by United States Environmental Protection Agency, at most cases the inferred results all stay in the same level in our deployment region, therefore it is unmeaningful to compare with it, and the linear and cubic spline interpolation are selected as the baseline methods. The following parts are organized as following: 1) we first discussed the setting of the covariance function as well as its parameter; 2) then the comparison between GP and the baseline methods was presented.

The covariance function plays a significantly import role in Gaussian Process, the training points that are close to a test point should be informative about the prediction at that point. From the Gaussian process view it is the covariance function that defines nearness or similarity [9]. In experiment we mainly investigated the following squared exponential covariance functions

$$k(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{1}{2\ell^2}\|\mathbf{x}_1 - \mathbf{x}_2\|_2\right)$$

Here ℓ is the horizontal scale over which the function changes. When the horizontal scale ℓ becomes large, the corresponding feature dimension is deemed irrelevant and the contrary is also true. If a relatively larger ℓ provides us a better inference result, it would imply that the distribution of the $PM_{2.5}$ concentrations in space is much smoother otherwise it would mean that the change of $PM_{2.5}$ concentrations among space is rapid. Therefore a suitable value of ℓ could reflect the variation degree of the $PM_{2.5}$ concentrations among the urban.

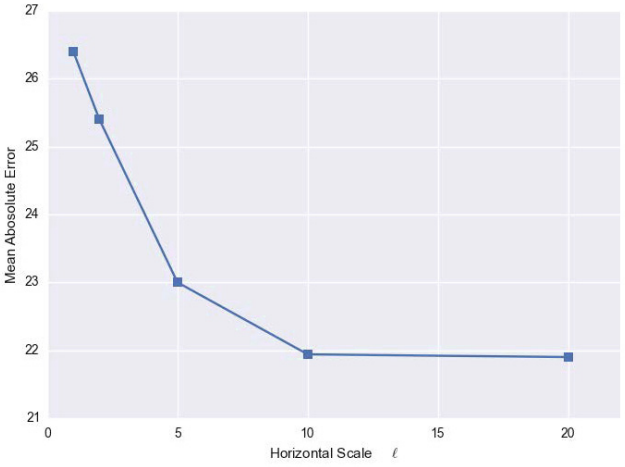

Fig. 3. The relationship between horizontal scale and mean absolute error

In Figure 3 we present the relationship between the covariance function parameter ℓ and the mean absolute error (using the data of all monitor stations).

With the increment of ℓ the error goes from 27.6 down to 21.9. This implies that the distribution of $PM_{2.5}$ concentrations is not smooth and the concentration in one location would highly differ from the one which departing in a long distance away from it. The Figure 2-(**B**) also shows the distribution of deviation between our two monitor stations, S_{26} and S_{28}, from May. 1, 2014 to Jun. 1, 2014. The geospatial distance of the two stations is about 6 km shown in Figure 2-(**A**), over 21% cases have a deviation greater than 100.

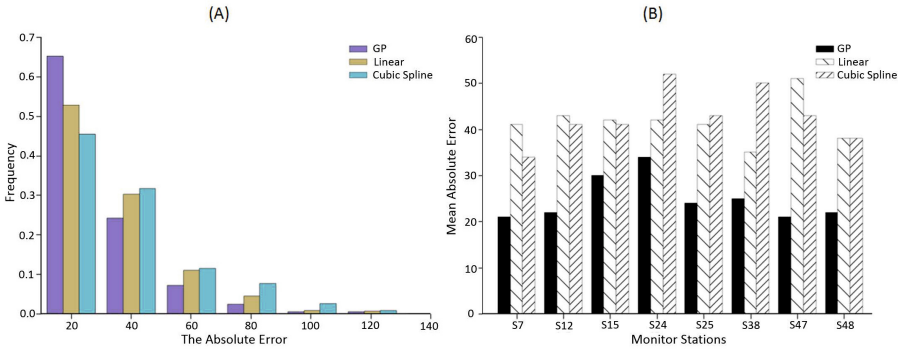

Fig. 4. (**A**)-The distribution of absolute error (using all dataset); (**B**)-The mean absolute error of eight monitor stations over one month

The Figure 4-(**A**) shows the absolute error distribution of the three methods (generated using all monitor stations over one month). It can be seen quite clearly that the Gaussian Process (GP) outperforms the baseline methods with small errors over 65% while the linear interpolation and cubic spline reaches 52% and 46% respectively. It is also worthwhile to note that the linear interpolation achieves a much better result than the cubic spline method. But this does not imply the $PM_{2.5}$ concentrations vary linearly among the urban, and Figure 4-B depicts the mean absolute error of 8 monitor stations which showing relatively large mean error. In station S_{24}, S_{25}, S_{38}, the linear interpolation performs much better than cubic spline method, however in station S_7, S_{12}, S_{47}, the cubic spline shows a more desirable performances than the linear interpolation. This might indicate us that in certain small local areas the distribution of $PM_{2.5}$ concentrations is more likely to be linear, but in the other areas it tends to be non-linear. Since the Gaussian Process always outperforms the baseline methods, it also

Table 1. Inference Errors

Measure / Method	$\|x\|_1$	$\frac{1}{n}\|x\|_1$	$\|x\|_2$	RMSE	$\|x\|_\infty$
Linear	929135.20	33.54	5991.73	36.00	266.92
Cubic Spline	975562.63	35.21	6379.94	38.33	266.92
Gaussian Process	585229.04	21.12	4101.57	24.64	154.14

proves the flexibility of the Gaussian Process in spatial inference of $PM_{2.5}$ concentrations.

Table 1 lists the inference errors of the three methods measured via different rules(assume that x is the absolute error vector). Gaussian Process beats all baseline methods, especially the Chebyshev norm $\|x\|_\infty$ achieved by both the linear and cubic spline interpolation could be as large as 266.92 while the Gaussian Process obtains a much smaller value 154.14, which proves that the Gaussian Process is much more stable in the inference of $PM_{2.5}$ concentrations.

5 Conclusion

In this paper, we present an ambient $PM_{2.5}$ concentrations monitoring and estimation system using Gaussian Process Inference model. We deployed 51 our designed air quality measuring devices among the area to be monitored and the $PM_{2.5}$ at these places are continuously sent back. We use the Gaussian Process Inference model to estimate the $PM_{2.5}$ concentrations at locations where monitor stations are unavailable and the proposed method is compared with two baseline models: linear and cubicle spline interpolation. The result shows that GP Inference model performs much better in different situations than the other two baseline models, which proves the flexibility of Gaussian Process in spatial inference and that it is indeed suitable for estimation of $PM_{2.5}$ concentration among the urban area. Since an exact inference in Gaussian Process involves computing K^{-1}, the computation cost is $O(n^3)$ (n is the number of the training cases), so when the deployment scale growing out of 1000 stations, we will resort to the approximation schemes, such as sparse approximations [7][9]. Additionally, we also analyzed the effect on inference of altering the parameter in covariance function, the results indicated us that the distribution of $PM_{2.5}$ concentrations is not that smooth and it is necessary to resort to dense deployment in order to monitor the fine-granularity $PM_{2.5}$ pollution.

Acknowledgement. We thank the anonymous reviewers for their valuable comments. This work was jointly supported by the National Natural Science Foundation of China under Grant No. 61300210 and No. 61370214.

References

1. Boldo, E., Medina, S., Le Tertre, A., Hurley, F., Mücke, H.G., Ballester, F., Aguilera, I.: Apheis: Health impact assessment of long-term exposure to pm2. 5 in 23 european cities. European Journal of Epidemiology 21(6), 449–458 (2006)
2. Sørensen, M., Daneshvar, B., Hansen, M., Dragsted, L.O., Hertel, O., Knudsen, L., Loft, S.: Personal pm2. 5 exposure and markers of oxidative stress in blood. Environmental Health Perspectives 111(2), 161 (2003)
3. Vardoulakis, S., Fisher, B.E., Pericleous, K., Gonzalez-Flesca, N.: Modelling air quality in street canyons: a review. Atmospheric Environment 37(2), 155–182 (2003)

4. Zheng, Y., Liu, F., Hsieh, H.P.: 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. ACM (2013)
5. Dawson-Haggerty, S., Jiang, X., Tolle, G., Ortiz, J., Culler, D.: smap: a simple measurement and actuation profile for physical information. In: Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems, pp. 197–210. ACM (2010)
6. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT Press (2012)
7. Quiñonero-Candela, J., Rasmussen, C.E.: A unifying view of sparse approximate gaussian process regression. The Journal of Machine Learning Research 6, 1939–1959 (2005)
8. Zheng, Y.: Urban air, `http://urbanair.msra.cn/`
9. Rasmussen, C.E.: Gaussian processes for machine learning (2006)