sensors

MDPI

Article

RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems

Ruiyun Yu [1,*], Yu Yang [2], Leyou Yang [1], Guangjie Han [3] and Oguti Ann Move [1]

Received: 30 September 2015; Accepted: 7 January 2016; Published: 11 January 2016
Academic Editor: Leonhard M. Reindl

[1] Software College, Northeastern University, Shenyang 110819, China; yangleyou@163.com (L.Y.); annmove@swc.neu.edu.cn (O.A.M.)
[2] Department of Computer Science, Rutgers University, New Brunswick, NJ 08854, USA; yangyu.9415@rutgers.edu
[3] Department of Internet of Things Engineering, Hohai University, Changzhou 213022, China; hanguangjie@gmail.com
* Correspondence: yury@mail.neu.edu.cn; Tel.: +86-24-8368-0515; Fax: +86-24-8368-0522

Abstract: Air quality information such as the concentration of $PM_{2.5}$ is of great significance for human health and city management. It affects the way of traveling, urban planning, government policies and so on. However, in major cities there is typically only a limited number of air quality monitoring stations. In the meantime, air quality varies in the urban areas and there can be large differences, even between closely neighboring regions. In this paper, a random forest approach for predicting air quality (RAQ) is proposed for urban sensing systems. The data generated by urban sensing includes meteorology data, road information, real-time traffic status and point of interest (POI) distribution. The random forest algorithm is exploited for data training and prediction. The performance of RAQ is evaluated with real city data. Compared with three other algorithms, this approach achieves better prediction precision. Exciting results are observed from the experiments that the air quality can be inferred with amazingly high accuracy from the data which are obtained from urban sensing.

Keywords: air quality prediction; random forest; point of interest; traffic

1. Introduction

As urbanization leads to urban community growth, the transportation infrastructure dependent on fossil fuels also expands consequently [1]. The popularity in vehicle use gives rise to an increase in traffic related pollutant emissions. Urban air pollution is a major problem in both developed and developing countries, as atmospheric pollutants have a great effect on human health. Numerous illnesses such as lung cancer may be caused by various atmospheric pollutants [2]. In addition, some other serious environmental problems can also result from air pollution, such as acid rain and the greenhouse gas effect. For example, SO_2 and NO_2 are the main causes of acid rain [3], while CO_2 and N_2O are the main reasons for the greenhouse gas effect [3]. Recently, especially in China, environmental problems have become a major concern in big cities such as Beijing and Shanghai, where the primary sources of pollutants include exhaust emissions from Beijing's more than five million motor vehicles, coal burning in neighboring regions, dust storms from the north and local construction dust [4]. A particularly severe smog engulfed the Beijing for weeks in early 2013, elevating public awareness to unprecedented levels and prompting the government to roll out emergency measures [4]. Air pollution monitoring is thus becoming more and more significant. Real-time air quality information, such as the concentration of $PM_{2.5}$, PM_{10} and NO_2, is an important aspect for pollution management and protecting human beings from damages caused by air pollutants. Considering the significance of air quality, governments take measures to monitor it through establishing air quality monitoring

stations. However, because of the high expense to start up and maintain these facilities, there are not sufficient stations in cities. For example, Figure 1 shows the Google Map of Shenyang City. The red pins represent the 11 air quality monitoring stations. Among them, S1 is located in a college; S2, S3, S4, S6, S8 are located on the roofs of buildings; S5, S9 are located along roads; S10 is located in a park; S11 is located near factories. These only 11 stations that cover more than three thousand square kilometers of downtown area in Shenyang. Another example is to compare London and Beijing. The area of Beijing is 10 times bigger than London but the number of monitoring stations is less than one fourth of London's [5]. One station can only monitor an area of limited size, therefore precise air quality reports for many areas cannot be generated.

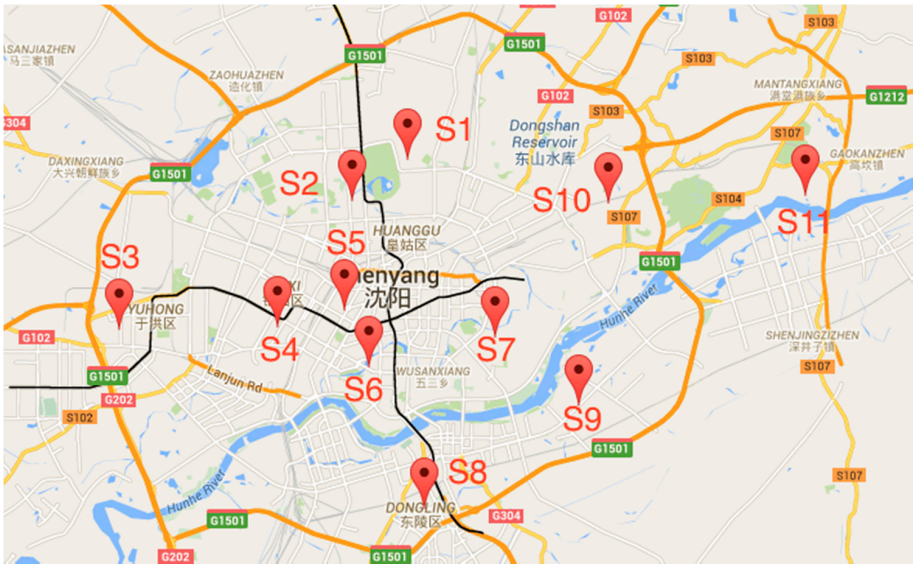

Figure 1. Monitoring station locations in Shenyang city (China).

Figure 2a shows samples of the AQI data of 10 stations in different locations. The x-axis denotes the different stations and the y-axis denotes AQI. Three bars in colors denote the AQI at different times. As demonstrated in Figure 2a, stations at different locations can differ a lot at the same time such as S7 and S8 on 6 May 2015 [6]. Air quality on continuous two days can also display big jumps such as AQI at S3 which raised from 55 to 408 in the morning between 6 May 2015 and 7 May 2015 [6]. Figure 2b shows the ways in which air quality changes follow different rules in different locations. For example, no matter whether stations are a short distance apart like S5 and S6 or a long distance like S5 and S10, they showed different changes between points in time.

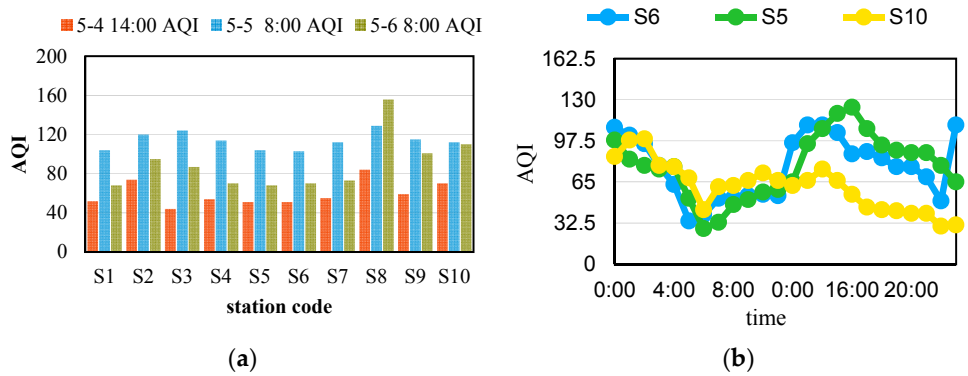

Figure 2. (**a**) AQI Samples in Shenyang; (**b**) AQI Trend on 12 May 2015 in Shenyang.

It is hard to reflect these changes in a general function which can be applied to all the locations, therefore, we cannot come up with a general formula to predict the air quality in a certain time slot. Therefore, how to infer the air quality in the blank areas is a challenging and meaningful topic. In this paper, we come up with an algorithm to infer the air quality indications throughout the city. In an urban sensing system, an algorithm (RAQ) based on a random forest concept is proposed to predict the urban area air quality through the use of historical air quality data, meteorology data, historical traffic and road status as well as POI distribution information. These data are collected from all kinds of urban sensors such as weather monitoring stations. This method hides all these kinds of inaccessible factors in the traditional mathematic models. In practical applications, we cannot take all the factors such as vehicle emissions and factory emissions into count, as it is hard to get accurate data about these factors. This kind of replacement is not only good for the computation but also good for increased prediction accuracy. At the same time, all the features used in this paper are much cheaper than the accurate measured data from monitoring stations. No equipment cost is required in this approach. As for the accuracy, this algorithm performs better than some other classical ones and the overall results can provide meaningful references to citizens. Regarding the scalability and expansibility, more possible related features such as human mobility can be input into this algorithm without significant changes. The algorithm itself is also robust enough for even higher dimensions.

The remainder of this paper is organized as follow: Section 2 presents related work. The problem description and formulation are presented in Section 3. In Section 4, the system framework and the RAQ algorithm is proposed. Extensive experiments are implemented in Section 5. We conclude and outline the directions for future work in Section 6.

2. Related Work

In the past decades, many studies on air quality inference have been done using approaches such as dispersion models, satellite remote sensing and wireless sensor networks. Air pollution dispersion models are tools that use a mathematical model such as the Box model [7], Gaussian model [8], Lagrangian model [9], Eularian model [10], SLAB model [11] or some mixed models. to simulate how air pollution disperses in the atmosphere. The classical dispersion models are mainly functions of meteorology, traffic volumes, building distributions and so on. These models depend mainly on experience and the parameters above to simulate the pollution dispersion, but some other potential factors are not taken into consideration such as human mobility and concentrations. In the meantime, dispersion models depend on access to relatively accurate data, such as the strength of pollutant sources, wind speed, traffic emissions and so on, which accuracy cannot be guaranteed in certain conditions. For example, wind speed may vary a lot in different regions because of the obstructions of buildings, and their roles in determining the modified wind circulation between and over structures. Accurate traffic emissions are also hard to obtain. We can only estimate the value according to the fuel consumption and distances travelled.

Satellite remote sensing technology is another possible way to monitor air quality. Research has developed quickly using satellites to monitor air conditions in the past decades. For example, Liu *et al.* came up with an approach using satellite remote sensing technology to test the thickness of $PM_{2.5}$ on the ground [12]. Similarly, Martin *et al.* came up with a way of using satellite remote sensing technology to test some ground air pollutants, including CO, NO, SO_2 and so on [13]. Pawan *et al.* used this technology to evaluate the air conditions of every city [14]. These methods mainly use satellite remote sensing technology to directly measure the concentration of certain air pollutants by analyzing the images obtained by the satellites to estimate the concentrations of air pollutants. However, many air quality managers are not yet taking full advantage of satellite data for their applications because of the challenges associated with accessing, processing, and properly interpreting observational data. That is, a certain degree of technical skill is required on the part of the data end-user, which is often problematic for organizations with limited resources [15].

Sensor networks have also been studied extensively because of their broad applicability and enormous application potential in areas such the environmental monitoring field. A Wireless Sensor Network Air Pollution Monitor System (WAPMS) was deployed on the island of Mauritius for monitoring air quality [16]; distributed infrastructure-based wireless sensor networks and grid computing is also used for monitoring the air quality of London [17]. Rajasegarar *et al.* also used wireless sensor networks to monitor air pollutants [18]. However, sensor networks require a large number of sensor devices, and can only be deployed in a small range, such as indoors and in small areas. For a city and other large areas, if using cheap sensors with single function, we cannot get information about all kinds of air pollutants. If using sensors with complex functions such as monitoring stations, infrastructure construction and maintenance costs make it difficult to promote wireless sensor networks for a wide usage range. It is the same reason which limits the number of stations in cities of China.

Besides all the methods above, participatory sensing is also an important approach for air quality prediction. With the popularity of smart devices, participatory sensing and crowdsourcing has been a hot topic of discussion in recent years. People see unlimited possibilities in smart devices. A personalized mobile sensing system (MAQS) was proposed for indoor air quality monitoring [19]; a system based on smart phones and monitoring sensors has also been used to monitor outdoor air quality [20]; noise pollution is also monitored using mobile phones [21]. Sivaraman *et al.* used a participatory sensor system to monitor air pollutants in Sydney (Australia) [22]. However, most current smartphones does not carry air pollutant sensors, so the sensing devices required for the system need external sensing modules which leads to extra costs. Besides the high expense, user participation and the accuracy of the data are problems that remain to be solved.

Recently, urban computing has been one of the ways to solve problems in cities. Yuan Jing *et al.* proposed an algorithm to infer the functional areas of cities by using trajectories [23]; Zheng *et al.* made use of the city daily data to infer urban air quality [24,25]. However, similarly, urban computing also requires pre-installed urban sensors such as GPS devices. For instance, when inferring the air quality, Zheng made use of months of data collected from the GPS installed in taxis in Beijing. This is an important limitation that prevents the promotion of this approach because in most cities we cannot access the GPS information of taxis. Spatiotemporal data analysis is also an important aspect for air quality prediction. Chen *et al.* established a spatiotemporal data framework named BigSmog to provide China smog analysis [26]. Zhu *et al.* proposed Granger-causality-based air quality estimation with heterogeneous spatiotemporal data [27]. Some other studies [28,29] also analyzed spatiotemporal data to generate air pollutant distributions.

3. Problem Description and Definition

3.1. Definition

3.1.1. Air Quality Index

An air quality index (AQI) is a number used by government agencies to communicate to the public how polluted the air is currently or how polluted it is forecasted to become [30]. As the AQI increases, an increasingly large percentage of the population is likely to be exposed, and people might experience increasingly severe health effects. Different countries have their own air quality indices, corresponding to different national air quality standards. In this paper, we use the standard of China, where the AQI is based on the levels of six atmospheric gases, namely sulfur dioxide (SO_2), nitrogen dioxide (NO_2), suspended particulates smaller than 10 μm in aerodynamic diameter (PM_{10}), suspended particulates smaller than 2.5 μm in aerodynamic diameter ($PM_{2.5}$), carbon monoxide (CO), and ozone (O_3), measured at the monitoring stations throughout each city [31]. The AQI value is calculated per hour according to a formula published by China's Ministry of Environmental Protection [31]. AQI is the maximum value of $IAQI_p$ which is a reference value of one air pollutant p:

$$AQI = max\{IAQI_1, IAQI_2, IAQI_3, \ldots, IAQI_n\} \quad (1)$$

$$IAQI_p = \frac{IAQI_{H_i} - IAQI_{L_o}}{BP_{H_i} - BP_{L_o}} \left(C_p - BP_{L_o}\right) + IAQI_{L_o} \quad (2)$$

where C_p is mass concentration value of the air pollutant p, BP_{H_i} is the high value of the concentration limit which can be checked in the reference table from the paper [31], BP_{L_o} is the low value of the concentration limit which can be checked in the reference table from [31], $IAQI_{H_i}$ is the corresponding value of BP_{H_i} in the same reference table, $IAQI_{L_o}$ is also the corresponding value of BP_{L_o} in the reference table. Table 1 shows the relationship between AQI values and air pollution levels which are marked by different colors. In this way, air quality prediction can be treated as a classification problem so that we only need to match the air quality index to different classification levels in Table 1. The six levels in Table 1 represent six AQI levels.

Table 1. AQI classification.

AQI	Air Pollution Level
0–50	Excellent
51–100	Good
101–150	Lightly Polluted
151–200	Moderately Polluted
201–300	Heavily Polluted
300+	Severely Polluted

3.1.2. Traffic Congestion Status

Traffic Congestion Status (TCS) describes the traffic conditions on a certain road. Different colors denote different levels of congestion. For example, Figure 3 shows an example of a TCS graph.

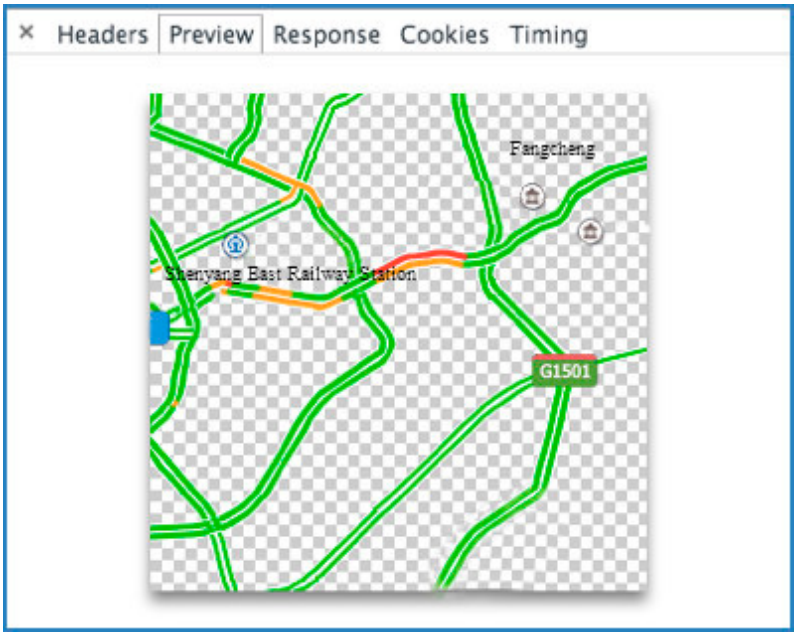

Figure 3. A TCS graph.

3.1.3. Point of Interest

A point of interest, or POI, is a specific location that someone may be interested in. For example, restaurants and shopping malls surrounding us are POI. Figure 4 presents the restaurant locations around Sanhao Street of Shenyang on Google Maps.

Figure 4. POI near the Sanhao street of Shenyang city.

3.2. Problem Formulation

This paper uses urban sensing data to solve the problem of air quality inference which means to infer the unknown air quality of areas by using all kinds of data. These data affect either the sources of air pollution such as traffic emissions and point of interest distribution or their results such as the air quality index, so establishing the relationship between these data and air quality is the key to this kind of approach. The RAQ algorithm collects several kinds of related data including air monitoring station data (AQI), meteorology data (MD), traffic (TCS), road information (RI) and POI data. All these data are fetched at intervals of one hour. We divide the city into grids (G) and each grid is regarded as one unit. Those grids (G_1) with air quality monitoring stations generate the data with the label AQI while the grids (G_2) without stations generate the data used for prediction. Data from G_1 are used for training our learning model and data from G_2 are input into the model to generate the predication value. The only difference of data from G_1 and G_2 is data from G_1 are labeled as an AQI value. The results are given as different AQI levels. If the actual value from monitoring stations belongs to this AQI level, then we know the prediction is right. Otherwise the prediction is wrong.

This problem can be formulated as follows: given a collection of grids $G = G_1 \cup G_2$ ($|G_1| \ll |G_2|$), where $g_1 \cdot AQI$ ($g_1 \in G_1$) is known and $g_2 \cdot AQI$ ($g_2 \in G_2$) is unknown, $g \cdot MD$, $g \cdot TCS$, $g \cdot RI$ and $g \cdot POI$ are known ($g \in G$), RAQ aims to predict $g_2 \cdot AQI$ at intervals of one hour.

4. RAQ Algorithm

In the RAQ algorithm, all data are collected from the urban sensing system including air monitoring station data, meteorology data, traffic data, road information and POI data and necessary features are extracted from heterogeneous data. These features are the most common data in city life. Traffic-related sources like vehicle emissions and POI like factories are the main sources for air pollutants [3]. Meteorology is the main approach for dispersion of air pollutants [3]. These data can represent well the air quality situation. The training dataset includes all the necessary features and is divided into subsets using bootstrap technology. Figure 5 shows the structure of the dataset. A decision tree is constructed on each subset, and the classification is done by aggregating the results generated from all decision trees. Figure 6 shows the procedure of the RAQ algorithm.

temperature Numeric	humility Numeric	pressure Numeric	wind Numeric	visibility Numeric	road_length Numeric	tfs Numeric	poi_number Numeric	**aqi** Nominal
5.5	89.0	758.1	2.0	14.0	2185.0	2371.0	63.0	excellent

Figure 5. Dataset structure.

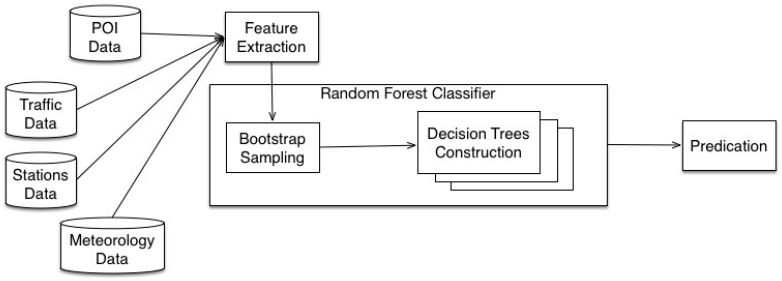

Figure 6. The procedure of RAQ.

4.1. Data Collection and Feature Extraction

4.1.1. Meteorology Data

Meteorology data such as temperature and humidity are very important factors that severely affect the concentration and spread of air pollutants. Understanding the behavior of meteorological parameters in the planetary boundary layer is important because the atmosphere is the medium in which air pollutants are transported away from the source, which is governed by the meteorological parameters such as atmospheric wind speed, wind direction, and temperature [32]. In this paper, we use weather monitoring stations as one part of the urban sensing system. Considering the accessibility of the data, we use following meteorology data features: temperature (F_{mt}, °C), humidity (F_{mh}, %), barometric pressure (F_{mp}, mmHg), wind speed (F_{mw}, m/s) and visibility (F_{mv}, m).

4.1.2. Traffic and Road Data

Traffic is one of the most important factors that affect the air quality. Figure 3 is a sample of the original data that is available from map service providers. In this paper, we rely on two important characteristics of traffic, which are road length (F_{rl}) and traffic congestion status (F_{tcs}). If the road is very long and traffic congestion is relatively light, exhaust gas emissions can be at a high level because of the total number of vehicles on this road. Similarly, if a road is short and traffic congestion is heavy. However, we do not have a method or accurate data to quantify these two characteristics directly. Most map service providers offer online maps and real-time traffic status. They do not publish public application interfaces (APIs) for third party developers to access these data, but we can still get some useful hints through analyzing the web http requests of the map. Essentially, these data are collected from GPS equipment installed in cars or speed measurement sensors. These data denote another important part of the urban sensing systems. Figure 7 shows the http request records of a typical Baidu map when we invoke the traffic widget.

As we know, a picture is composed of many pixels, so a picture can be digitized into a matrix. We use the colored pixel distribution to represent the information of road length and congestion status. For each tile grid, we count the quantity of pixels to represent the road. The larger the quantity of pixels, the greater the length of the road is in one tile grid.

As shown in Figure 3, traffic congestion status is denoted by different colors (green, orange and red) in the pixels which represent roads. According to the traffic volume of different congestion levels, different weights are assigned to the numbers of pixels in different colors (1, 2 and 5). In Figure 8, the weighted tcs value is calculated by formula $a + 2b + 5c$, where a is the number of pixels in green, b is the number of pixels in orange and c is the number of pixels in red.

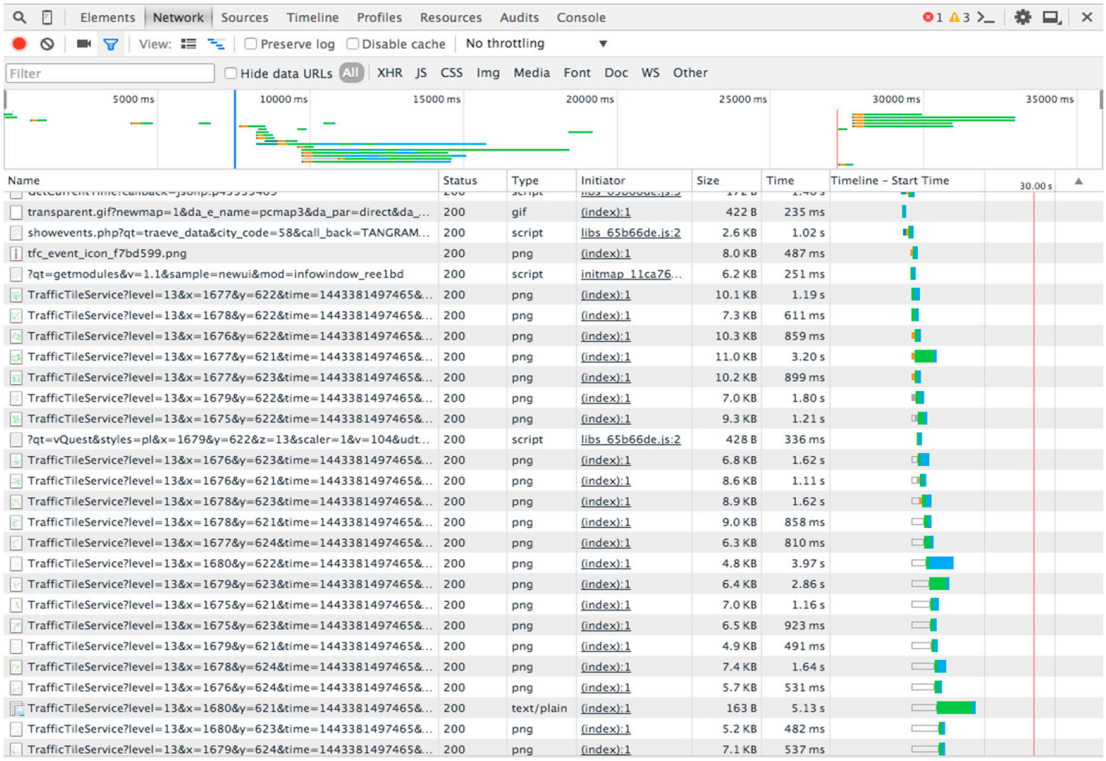

Figure 7. HTTP request analysis by Chrome developer tool.

id	city_id	station_id	tcs	time
193	1	746	1	2015-05-09 20:00:00
194	1	747	1	2015-05-09 20:00:00
195	1	750	4	2015-05-09 20:00:00
196	1	751	2	2015-05-09 20:00:00
197	1	748	2	2015-05-09 21:00:00
198	1	749	3	2015-05-09 21:00:00
199	1	742	1	2015-05-09 21:00:00
200	1	743	1	2015-05-09 21:00:00
201	1	746	1	2015-05-09 21:00:00
202	1	747	1	2015-05-09 21:00:00
203	1	750	4	2015-05-09 21:00:00
204	1	751	2	2015-05-09 21:00:00
205	1	748	2	2015-05-09 22:00:00
206	1	749	3	2015-05-09 22:00:00
207	1	742	1	2015-05-09 22:00:00
208	1	743	1	2015-05-09 22:00:00
209	1	746	1	2015-05-09 22:00:00

Figure 8. Traffic congestion status.

4.1.3. POI Data

The category of POIs and their density in a region indicate the land use and the function of the region as well as the traffic patterns in the region, therefore contributing to the air quality inference of the region [24]. For example, shopping streets are more likely to gather more people than parks so there will be more human-related air pollution sources like vehicles. Schools always have more green areas than factories so there are more plants to absorb the air pollutants. Therefore, POI distribution has a strong effect on air quality. These data also imply the significance of human activities in urban sensing systems. In this paper, the number of POI is counted in each tile grid. According to the searching results of Baidu maps and Google Map, the majority of POI are divided into ten categories. Table 2 shows the categories and Figure 9 presents the number of POI (F_{pn}) in each category.

Table 2. POI categories.

Code	POI Category
P1	Transportation
P2	Entertainment
P3	Restaurant
P4	Education
P5	Residential District
P6	Park
P7	Company
P8	Factory
P9	Shopping mall
P10	Gas station

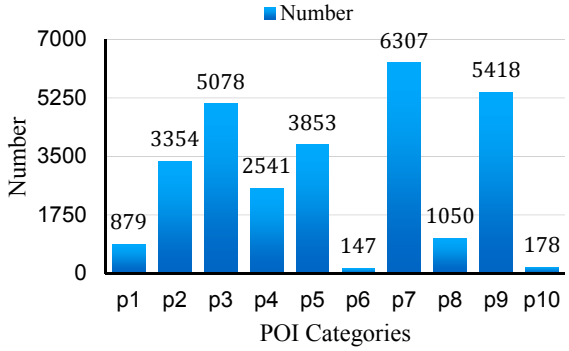

Figure 9. Numbers of POI in Shenyang city.

4.2. Random Forest Classification

The Random Forest is a general term for ensemble methods using tree-type classifiers $\{h(x, \theta_k), k = 1, ...,\}$ where the $\{\theta_k\}$ are independent identically distributed random vectors and x is an input pattern, $h(x, \theta_k)$ is a generated classifier [33]. It uses recursive partitioning to generate many trees and then aggregate the results. Each tree is independently constructed using a bootstrap sample of the training data, which subdivides the parameter set first into several parts depending on one of the parameters, and subsequently repeats the process for each part.

4.2.1. Bootstrap Aggregating (Bagging)

There is usually a single data sample in each class for training. A simple method is to divide the dataset into non-overlapping subsets and construct the trees independently. However, this requires a huge amount of data and it cannot always be guaranteed in different situations. A better way is sampling the original dataset with replacement for a certain times to produce a bootstrap sample. This method ensures that the samples' distributions are statistically identical with the original data sample [34]. There are n records in the original dataset and so the probability of each record is constantly $1/n$. The probability of not selecting a certain record is $(1 - 1/n)$, which results in $(1 - 1/n)^n$ when repeated n times.

Assuming the sample size tends to be infinite, the probability can be expressed as $\lim_{n\to\infty}(1 - 1/n)^n$ which is equal to e^{-1}. Therefore, the probability of selecting one record is $(1 - e^{-1}) \approx {}^2/_3$. Thus, in each bootstrap sample there are about ${}^2/_3$ original samples for training.

4.2.2. Tree Growing and Splitting

As we know, a decision tree starts with one root node. In the following process, the samples are split into different spaces using one of the features including monitoring station data (AQI), meteorology data(MD), traffic(TCS), road information(RI) and POI data.. Therefore, how to select the

feature in each split is of great significance for the performance of a decision tree. Information gain [35] is usually used as the criterion for classifiers.

The features selection for each bootstrap sample is randomized. According to bagging theory, random forest is strong classifier based on multiple weak classifiers. Therefore, both the number of data and the number of features of the subset are smaller than original dataset's. We need T subsets with m features. According to Brieman's suggestions [33], m is much less than the number of all the features. Brieman suggests three possible values for m: $\frac{1}{2}\sqrt{m}$, $\sqrt{m}$, $2\sqrt{m}$. In the evaluation section, we would show four features and 400 subsets are best for our model and dataset.

When splitting the dataset, for each feature candidate, entropy is calculated as in Equation (3):

$$Entropy(c) = -\sum_{i=1}^{k} p(c_i)\log_2 p(c_i) \tag{3}$$

$$p(c_i) = \frac{N_i}{\sum_{i=1}^{k} N_i} \tag{4}$$

where c_i is the AQI level i which is specified in Table 1, the probability $p(c_i)$ is calculated through Equation (2) where Ni is the quantity of records in different AQI level and k is the number of AQI levels. Therefore, the information gain is defined as shown in Equation (5):

$$Gain(f_i) = Entropy(c) - \sum_{j=1}^{w} \frac{\left|f_i^j\right|}{|f_i|} Entropy(f^j{}_i) \tag{5}$$

where f_i represents records of the i_{th} level of tree, $f_i{}^j$ are records in jth node of the ith level of tree, and w is the number of nodes in this level.

The process of splitting stops when: (a) the records in one node fall below the threshold value defined by users; (b) the node is pure which means all the records fall into one class. For the terminated node has unordered records, the percentage of different classes are calculated and so the predicted class is defined as in Equation (6):

$$C(i) = Max(p(c_i)) \tag{6}$$

4.3. Prediction

After all the trees are constructed, the unlabeled data are input into all decision trees. For each tree, $p(c_i)$ is the estimated probability of the AQI level i. The final probability of the AQI level i $p'(c_i)$ in the random forest is defined in Equation (7), where T is the number of decision trees as mentioned before:

$$p'(c_i) = \frac{1}{T}\sum_{k=1}^{T} p(c_i) \tag{7}$$

The final result is determined by Equation (8):

$$C'(i) = Max(p'(c_i)) \tag{8}$$

The pseudocode of RAQ algorithm is described in Algorithm 1.

Algorithm 1. RAQ

Input:	A dataset S with features: F_{mt}, F_{mh}, F_{mp}, F_{mw}, F_{mv}, F_{ri}, F_{tcs}, F_{pn} and labeled AQI level; unlabeled dataset U; trees quantity T; features quantity m;
Output:	AQI level
1	for T trees
2	randomly select m features from S;
3	for m features in each node
4	calculate information gain by Equation (3);
5	choose maximum gain to split the dataset in the node;
6	remove used feature from feature candidates;
7	input unlabeled data into trees;
5	get predicted AQI level according to Equations (5) and (6);

5. Evaluation

5.1. Dataset

In the experiments, one-month data from 4 May 2015 to 5 June 2015 is collected and the following four datasets of Shenyang are used which are all available to the public. In our testing period, we use a total of 2701 data to test this algorithm and Shenyang is divided into 1258 grids corresponding to 34 rows and 37 columns. Because all the grids belong to the main city area, all data including meteorology data, traffic data, road information and POI data in these grids are accessible from our data sources. Air quality data is accessible in the areas covered by air monitoring stations.

5.1.1. Monitoring Station Data

The air quality information from the Shenyang monitoring stations includes AQI, the concentrations of CO, NO_2, SO_2, O_3, PM_{10} and PM_{25} and timestamp. Table 3 shows the format of the monitoring station data. Table 4 shows the locations of all the monitoring stations. All the data are collected from the public website [36] whose data are produced by National Department of Environmental Protection. We use the Java programming language to access the API interface hourly and store all the data into a MySQL database.

Table 3. Data samples of monitoring stations.

Station_id	Aqi	CO (μg/m^3)	NO_2 (μg/m^3)	SO_2 (μg/m^3)	O_3 (μg/m^3)	PM_{10} (μg/m^3)	PM_{25} (μg/m^3)	Time
747	77	1.802	70	69	63	104	52	2015-05-24 03:00
750	139	2.233	62	70	57	125	106	2015-05-24 03:00
751	82	1.706	73	58	69	100	60	2015-05-24 03:00
741	85	1.942	80	64	43	94	63	2015-05-24 03:00
748	63	1.024	61	62	68	76	37	2015-05-24 04:00
749	67	1.358	60	29	62	81	48	2015-05-24 04:00
742	88	1.646	97	82	12	125	14	2015-05-24 04:00
743	84	0.808	68	167	45	117	52	2015-05-24 04:00
744	98	1.718	66	56	43	92	73	2015-05-24 04:00
745	86	1.333	78	72	9	121	37	2015-05-24 04:00
746	66	1.229	66	24	48	82	45	2015-05-24 04:00
747	63	1.175	58	48	70	75	36	2015-05-24 04:00

Table 4. Locations of monitoring stations.

Station_id	Latitude	Longitude
741	41.841445	123.65436
742	41.758166	123.533761
743	41.71694	123.451378
744	41.788094	123.288852
745	41.838551	123.549754
746	41.855605	123.442396
747	41.773208	123.421573
748	41.785295	123.489395
749	41.79609169	123.4084114
750	41.789429	123.373275
751	41.83933982	123.4126515

5.1.2. Meteorological Data

We collect meteorological data including temperature, humidity, barometric pressure, wind speed and visibility from the public website [37]. As Table 5 illustrates, the data format is presented as temperature (F_{mt}), humidity (F_{mh}), barometric pressure (F_{mp}), wind speed (F_{mw}) and visibility (F_{mv}).

Table 5. Meteorological samples.

Temperature (F_{mt}, °C)	Barometric Pressure (F_{mp}, mmHg)	Humidity (F_{mh}, %)	Wind Speed (F_{mw}, m/s)	Visibility (F_{mv}, m)	Time
18.8	748.6	56	2	16.0	2015-05-14 11:00:00
18.3	746.4	50	7	26.0	2015-05-14 08:00:00
17.0	744.6	63	3	12.0	2015-05-14 05:00:00
18.4	743.0	58	1	16.0	2015-05-14 02:00:00
19.7	743.9	63	1	18.0	2015-05-13 23:00:00
18.0	742.6	72	0	7.0	2015-05-13 21:00:00

5.1.3. Road and Traffic Data

There are no public websites that offer statistical road and traffic data. Therefore, we cannot directly get available formatted data. However, most of the map service providers offer online maps and real-time traffic status. They do not publish public API interfaces for third party developers to access these data, but we can still get some useful tips through analyzing the map web http requests. From map services providers [38,39], we collect the traffic map tiles every hour.

5.1.4. POI

Thank to Baidu map and Google map service, we can easily get these data from a public interface. Each POI record contains name, latitude, longitude, tag and located tile grids. Figure 10 shows about 28,000 records in the MySQL database.

id	name	lat	lng	x	y	region	tag
11	jiaxingxiaochibu	41.925921	123.302196	5001	13403	0_3	Restaurant
12	shenyangxingyouzhuzaochang	41.926659	123.304037	5001	13403	0_3	Company
13	bosishengjianzhugongchenggongsi	41.92639	123.303329	5001	13403	0_3	Company
14	shenyangshixinxianghejianzhucailiaochang	41.92639	123.303202	5001	13403	0_3	Company
15	shenyangxingyouzhuzaochang	41.926659	123.304037	5001	13403	0_3	Company
16	shenyangshixinxianghejianzhucailiaochang	41.92639	123.303202	5001	13403	0_3	Company
17	shenyangzhengyishiyeyouxiangongsi	41.926397	123.344853	5001	13407	0_7	Company
18	shenbeijiancaidashichangmucaiqu	41.926974	123.349024	5001	13408	0_8	Shopping mall
19	shenbeijiancaidashichang	41.927625	123.34898	5001	13408	0_8	Shopping mall
20	shenyangqitizhizaoyouxiangongsi	41.92584	123.352822	5001	13408	0_8	Company
21	shenyangletaijixieyouxiangongsi	41.927679	123.354225	5001	13408	0_8	Company
22	fengyuanjixiejiagongchang	41.927679	123.354225	5001	13408	0_8	Company
23	lvxingdabaicaiyanjiusuogongsi	41.925666	123.35087	5001	13408	0_8	Company
24	bangdezhinengjinshuzhizaogongsi	41.925968	123.354066	5001	13408	0_8	Company
25	shenyangshigujinjiajuchang	41.928021	123.348288	5001	13408	0_8	Company
26	fengyuanjixiejiagongchang	41.927679	123.354225	5001	13408	0_8	Company
27	shenyangshigujinjiajuchang	41.928021	123.348288	5001	13408	0_8	Factory
28	wangshiwangluo	41.926142	123.359175	5001	13409	0_9	Entertainment

Figure 10. POI Samples in Shenyang.

5.2. Evaluation Method

The most accurate criterion for air quality measure is the air quality information from monitoring stations. In this experiment, we use the AQI data from monitoring stations as the reference standard. To construct a random forest, we need to determine two parameters which are the numbers of trees and the number of features used to construct each tree. To choose the best parameters, we use OOB (Out-of-Bag) [33] error to compare RAQ accuracy based on different parameters pairs <#features, #trees> which means the number of features used to construct each tree and the number of trees that are constructed in the random forest. In random forests, the error is estimated internally during the construction of trees. Each tree is constructed using a different bootstrap sample from original data, which about one-third are left out of the bootstrap sample. The one-third sample is used as test cases to be input into the tree and get the classification of each test case. At the end of the run, take the class *j* that got most of the votes every time case *n* was oob [40]. The proportion of times that j is not equal to the true class of *n* averaged over all cases is the oob error estimate [40]. The smaller number of oob, the high accuracy of the model. For the number of features, we increase by one each time from 2 to 8 (total number of features is 8 specified in algorithm). For the quantity of trees, we increase by 100 from 100 to 1000. Because of the time consumption with more number of trees, we ignore the trees number greater than 1000 and 100 gap is suitable to balance performance and accuracy. To compare this algorithm with others, we use cross-validation method to judge the performance.

5.3. Results

5.3.1. Effects of Parameters on Prediction Error Rate

There are two important factors that affect the performance of a random forest, which are the number of trees and features. Figure 11 shows how the OOB error changes along with the number of features and trees. X-axis is the number of features and Y-axis is the number of trees.

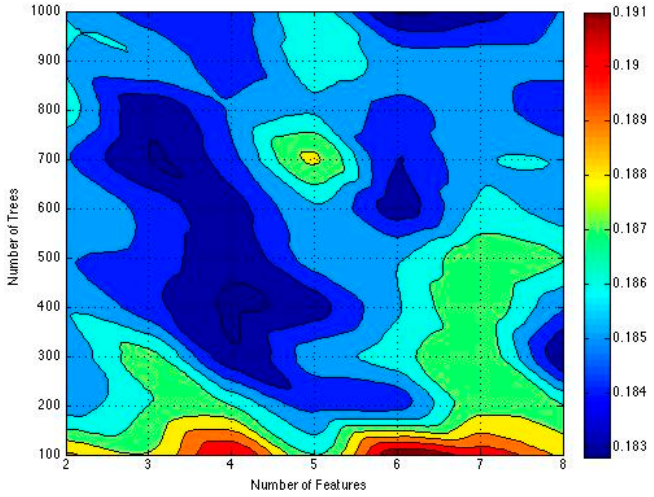

Figure 11. OOB error result distribution.

Empirically, for our experiment, we choose integer as the number of features and 100 interval integer as the number of trees, so only the discrete coordinate values such as (2,100), (3,200) are meaningful in this graph. Different colors mean different OOB error values. The deeper the color is, the smaller the oob is. As the graph shows, the OOB errors reach the best when the parameters pairs are <4, 400> and <6, 1000>. Considering less time consumption, we choose <4, 400> as the best parameters pair.

5.3.2. Comparison

For the contrast tests, Naïve Bayes, Logistic Regression, Single Decision Tree and ANN are chosen. Here we use Weka [41] as the tool to conduct all the comparison tests. For Naïve Bayes, there are eight features which are F_{mt}, F_{mh}, F_{mp}, F_{mw}, F_{mv}, F_{ri}, F_{tcs}, F_{pn} and six classification categories (C) which are specified in Table 1. In Weka, this algorithm is denoted as weka.classifiers.bayes.NaiveBayesMultinomial. For Logistic Regression, we choose Multinomial Logistic Regression because of the multi AQI levels. In Weka, this algorithm is denoted as weka.classifiers.functions.Logistic. For Single Decision Tree, we choose all the features to construct one single tree for classification. In Weka, this algorithm is denoted as weka.classifiers.trees.REPTree. For ANN, we choose back-propagation neural network with one hidden layer for its simplicity and generality. In Weka, this algorithm is denoted as weka.classifiers.functions.MultilayerPerceptron.

After realizing different algorithms, tests are carried out. Table 6 shows the results of the test cases in which Y means correct predictions and N means incorrect predictions. The precision is calculated by the formula Y/(Y + N) where Y is the number of correct predictions and N is the number of incorrect predictions. Figure 12 illustrates how the prediction precision changes as the data size changes. This figure shows RAQ performs steadily, even when the data size is relatively small. Other algorithms are less accurate at all time.

Table 6. Precision table of different algorithms.

Algorithm	Precision	Y	N
NaïveBayes	52.1%	1408	1293
Logistic	66.2%	1790	911
Decision Tree	77.4%	2092	609
ANN	71.8%	1940	761
RAQ	81.5%	2203	498

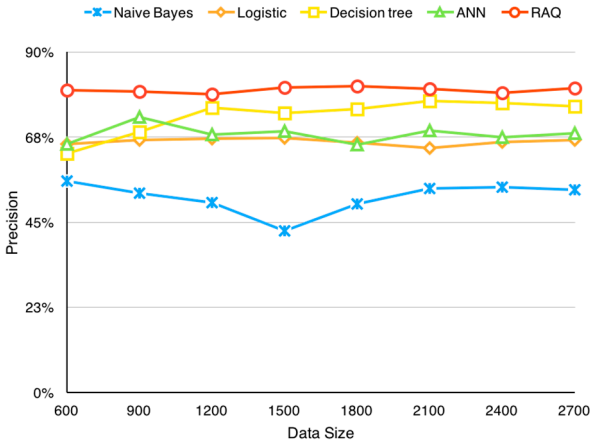

Figure 12. Precision changes according to data size.

Besides the precision measurement, we also refer to other measurements including Recall, F-score, Relative Absolute Error (RAE) and Receiver Operating Characteristic (ROC). Recall is the proportion of instances classified as a given class divided by the actual total in that class. F-score is a combined measure for precision and recall calculated as $2{*}Precision{*}Recall/(Precision + Recall)$ where Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Relative absolute error is calculated by the following formula:

$$RAE = \frac{\sum_{i=1}^{N} |\theta_i - r_i|}{\sum_{i=1}^{N} \left|\overline{\theta} - r_i\right|}$$

where θ_i is the estimated value, r_i is the real value, $\overline{\theta}$ is the average value, N is the number of test cases. ROC shows how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples [42].

Table 7. Indexes of different algorithms.

Algorithm	Recall	F-Score	ROC	RAE
Naive Bayes	0.521	0.529	0.7	84.9%
Logistic	0.663	0.649	0.785	75.8%
Decision Tree	0.775	0.769	0.888	47.4%
ANN	0.718	0.707	0.829	60.9%
RAQ	0.816	0.814	0.928	36.9%

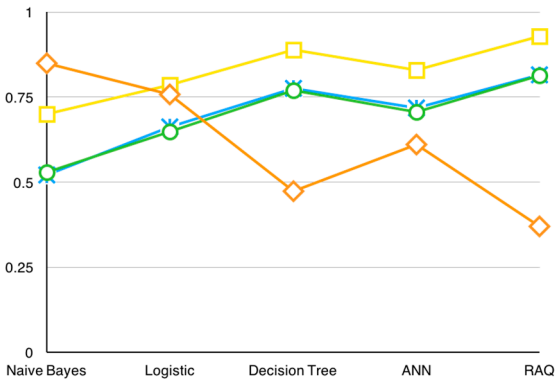

Figure 13. Indexes chart of different algorithms.

Based on our dataset, these measurements also show that RAQ performs better than others in this specific problem. Table 7 shows the original data of the experiments result and Figure 13 illustrates these data in chart form.

6. Conclusions

In this paper, with the public data in the urban sensing system, our model predicts the AQI of all the regions in Shenyang based on the AQI published by 11 air quality monitoring stations, meteorology data reported by weather stations, road information and real-time traffic status collected from Baidu Map and Google Maps and the POI distributions provided by Baidu Map and Google Maps. We use a random forest algorithm to predict all the uncovered regions in the downtown area. In Shenyang, this algorithm finally results in an overall precision of 81% for AQI prediction. This experimental result outperforms that of Naïve Bayes, Logistic Regression, single decision tree and ANN. All of these data are directly or indirectly available on the Internet. This shows that the algorithm could be easily applied for other cities. RAQ makes use of historical data for model training but ignores the real-time data. Our work will be extended to support online learning so daily data can be used to improve the performance of the air prediction algorithm.

Acknowledgments: This work is in part supported by the National Natural Science Foundation of China under Grant No. 61272529; Ministry of Education-China Mobile Research Fund under Grant No. MCM20130391; the Fundamental Research Funds for the Central Universities under Grant No. N130817003.

Author Contributions : Ruiyun Yu proposed and developed the idea, designed the algorithm together with the other authors, conducted the coordination of the research activities and coordinated the revision activities. Yu Yang and Leyou Yang co-created the research design, conducted the simulations and contributed to the manuscript writing and revisions. Guangjie Han co-supervised the research activities and contributed to the manuscript revisions. Oguti Ann Move contributed to the manuscript revisions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Devarakonda, S.; Sevusu, P.; Liu, H.; Liu, R.; Iftode, L.; Nath, B. Real-time air quality monitoring through mobile sensing in metropolitan areas. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, New York, NY, USA, 11 August 2013.
2. Mage, D.; Ozolins, G.; Peterson, P.; Webster, A.; Orthofer, R.; Vandeweerd, V.; Gwynne, M. Urban air pollution in megacities of the world. *Atmos. Environ.* **1996**, *30*, 681–686. [CrossRef]
3. NASA. Available online: http://climate.nasa.gov/causes/ (accessed on 23 December 2015).
4. South China Morning Post. Available online: http://www.scmp.com/topics/beijing-air-pollution (accessed on 5 May 2015).
5. People. Available online: http://politics.people.com.cn/GB/1026/17220033.html (accessed on 25 November 2015). (In Chinese)
6. PM_{25}.in. Available online: http://www.pm25.in/shenyang (accessed on 5 May 2015). (In Chinese)
7. Ragland, K.W. Multiple Box Model for Dispersion of Air Pollutants from Area Sources. *Atmos. Environ.* **1973**, *7*, 1017–1032. [CrossRef]
8. Bosanquet, C.H.; Pearson, J.L. The spread of smoke and gases from chimneys. *Trans. Faraday Soc.* **1936**, *32*, 1249–1263. [CrossRef]
9. Zannetti, D.P. Lagrangian Dispersion Models. In *Air Pollution Modeling*; Springer U.S.: Boston, MA, USA, 1990; pp. 185–222.
10. Pai, P.; Karamchandani, P.; Seifneur, C. Simulation of the regional atmospheric transport and fate of mercury using a comprehensive eulerian model. *Atmos. Environ.* **1997**, *31*, 2717–2732. [CrossRef]
11. Ermak, D.L. *User's Manual for SLAB: An Atmospheric Dispersion Model for Denser-Than-Air-Releases*; Lawrence Livermore National Laboratory: Livermore, CA, USA, 1990.
12. Liu, Y.; Sarnat, J.A.; Kilaru, V.; Jacob, D.J.; Koutrakis, P. Estimating Ground-Level $PM_{2.5}$ in the Eastern United States Using Satellite Remote Sensing. *Environ. Sci. Technol.* **2005**, *39*, 3269–3278. [CrossRef] [PubMed]
13. Martin, R.V. Satellite remote sensing of surface air quality. *Atmos. Environ.* **2008**, *42*, 7823–7843. [CrossRef]

14. Gupta, P.; Christopher, S.A.; Wang, J.; Gehrig, R.; Lee, Y.; Kumar, N. Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmos. Environ.* **2006**, *40*, 5880–5892. [CrossRef]
15. Duncan, B.N.; Prados, A.I.; Lamsl, L.N.; Liu, Y.; Streets, D.G.; Gupta, P.; Hilsenrath, E.; Kahn, R.A.; Nielsen, J.E.; Beyersdorf, A.J.; *et al*. Satellite data of atmospheric pollution for US air quality applications: Examples of applications, summary of data end-user resources, answers to FAQs, and common mistakes to avoid. *Atmos. Environ.* **2014**, *94*, 647–662. [CrossRef]
16. Khedo, K.K.; Perseedoss, R.; Mungur, A. A Wireless Sensor Network Air Pollution Monitoring System. *Int. J. Wirel. Mob. Netw.* **2010**, *2*, 31–45. [CrossRef]
17. Ma, Y.; Richards, M.; Ghanem, M.; Guo, Y.; Hassard, J. Air Pollution Monitoring and Mining Based on Sensor Grid in London. *Sensors* **2008**, *8*, 3601–3623. [CrossRef]
18. Rajasegarar, S.; Zhang, P.; Zhou, Y.; Karunasekera, S.; Leckie, C.; Palaniswami, M. High resolution spatio-temporal monitoring of air pollutants using wireless sensor networks. In Proceedings of the 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Singapore, 21–24 April 2014; pp. 1–6.
19. Jiang, Y.; Li, K.; Tian, L.; Piedrahita, R.; Yun, X.; Mansata, O.; Lv, Q.; Dick, R.P.; Hannigan, M.; Shang, L. MAQS: A personalized mobile sensing system for indoor air quality monitoring. In Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing, China, 17–21 September 2011; pp. 271–280.
20. Hasenfratz, D.; Saukh, O.; Sturzenegger, S.; Thiele, L. Participatory air pollution monitoring using smartphones. In Proceedings of the 1st International Workshop on Mobile Sensing: From Smartphones and Wearables to Big Data, Beijing, China, 16 April 2012.
21. Maisonneuve, N.; Stevens, M.; Ochab, B. Participatory noise pollution monitoring using mobile phones. *Inf. Polit.* **2010**, *15*, 51–71.
22. Sivaraman, V.; Carrapetta, J.; Hu, K.; Luxan, B.G. HazeWatch: A participatory sensor system for monitoring air pollution in Sydney. In Proceedings of the 2013 IEEE 38th Conference on Local Computer Networks Workshops (LCN Workshops), Sydney, NSW, Australia, 21–24 October 2013; pp. 56–64.
23. Yuan, J.; Zheng, Y.; Xie, X. Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 186–194.
24. 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, Chicago, IL, USA, 11–14 August 2013; pp. 1436–1444.
25. Hsieh, H.-P.; Lin, S.-D.; Zheng, Y. Inferring Air Quality for Station Location Recommendation Based on Urban Big Data. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 437–446.
26. Chen, J.; Chen, H.; Pan, J.Z.; Wu, M.; Zhang, N.; Zheng, G. When big data meets big smog: A big spatio-temporal data framework for China severe smog analysis. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, Orlando, FL, USA, 5–8 November 2013; pp. 13–22.
27. Zhu, J.Y.; Sun, C.; Li, V.O.K. Granger-Causality-Based Air Quality Estimation with Spatio-Temporal (S-T) Heterogeneous Big Data. In Proceedings of the 2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Hong Kong, China, 2015; pp. 612–617.
28. Song, L.; Pang, S.; Longley, I.; Olivares, G.; Sarrafzadeh, A. Spatio-temporal $PM_{2.5}$ prediction by spatial data aided incremental support vector regression. In Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China, 6–11 July 2014; pp. 623–630.
29. Hasenfratz, D.; Saukh, O.; Walser, C.; Hueglin, C.; Fierz, M.; Thiele, L. Pushing the spatio-temporal resolution limit of urban air pollution maps. In Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom), Budapest, Hungary, 24–28 March 2014; pp. 69–77.
30. United States Environmental Protection Agency. Available online: http://www3.epa.gov/airnow/aqi_brochure_02_14.pdf (accessed on 25 December 2015).
31. China's Ministry of Environmental Protection. Available online: http://kjs.mep.gov.cn/hjbhbz/bzwb/dqhjbh/jcgfffbz/201203/W020120410332725219541.pdf (accessed on 27 November 2015). (In Chinese)
32. Wexler, H. *The Role of Meteorology in Air Pollution, Monograph Series*; World Health Organization: Geneva, Switzerland, 1961; pp. 46–49.

33. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
34. Efron, B. Bootstrap Methods: Another Look at the Jackknife. In *Breakthroughs in Statistics*; Springer New York: New York, NY, USA, 1992; pp. 569–593.
35. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660–674. [CrossRef]
36. PM_{25}.in: Air Quality Data Provider. Available online: http://pm25.in (accessed on 8 January 2016). (In Chinese)
37. RP5.ru: Weather for 243 Countries of the World. Available online: http://rp5.ru (accessed on 8 January 2016).
38. Baidu Map. Available online: http://map.baidu.com (accessed on 8 January 2016).
39. Google Map. Available online: http://map.google.com (accessed on 8 January 2016).
40. Breiman, L.; Cutler, A. Random Forests. Available online: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr (accessed on 25 November 2015).
41. Weka 3: Data Mining Software in Java. Available online: http://www.cs.waikato.ac.nz/ml/weka/ (accessed on 25 November 2015).
42. Davis, J.; Goadrich, M. The Relationship between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning, Corvallis, OR, USA, 25–29 June 2006; pp. 233–240.