Heliyon 6 (2020) e05296

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

RESILIENT: A robust statistical method for estimating multiple VOC sources from limited field measurements

Anand Kakarla [a], Asif Qureshi [b], Shashidhar Thatikonda [b], Swades De [c], Soumya Jana [a,*]

[a] Department of Electrical Engineering, Indian Institute of Technology Hyderabad, India
[b] Department of Civil Engineering, Indian Institute of Technology Hyderabad, India
[c] Department of Electrical Engineering, Indian Institute of Technology Delhi, India

ARTICLE INFO

Keywords:
Statistics
Transport process
Atmosphere modeling
Air quality
Environmental analysis
Environmental assessment
Environmental impact assessment
Environmental risk assessment
Source estimation
Pollution mapping
AERMOD
Genetic algorithm
Leave-p-out cross-validation
Limited field measurements

ABSTRACT

Air pollution due to haphazard industrialization has become a major concern in developing countries. Yet, enforcement of related norms remains problematic because violators cannot easily be pinpointed among closely situated industrial units. Accordingly, it has become imperative to equip regulatory authorities with an economical yet accurate tool that quickly locates emission sources and estimates emission rates. Against this backdrop, we propose RESILIENT, a method for Robust Estimation of Source Information from LImited field measuremENTs, which exhibits significant statistical robustness and accuracy even when the data are collected using a low-cost error-prone sensor. In our field experiment, where ground truth was unavailable, the sources estimated to be inactive based on the complete set of measurements were found inactive (up to three decimal places of accuracy) at least 72% of the time even when estimated using just 54% of random measurements. In that setting, rate estimates of active sources were also found to be statistically robust. For direct validation of RESILIENT, we considered a separate public dataset involving 10 tracer experiments, and obtained a significant correlation coefficient of 0.89 between estimated and recorded emission rates, and that of 0.99 between predicted and measured concentration levels at sensor locations.

1. Introduction

In developing countries, chemicals and pharmaceuticals along with noxious byproducts are often produced in small-scale units that are closely situated in an industrial zone adjoining residential area (Al-Wahaibi and Zeka, 2015). In such areas, emission norms for air pollutants are regularly violated, posing considerable environmental and health risks (Oyinloye, 2015, The Economic times of India, 2018). Accordingly, it has become imperative to equip the regulatory authorities with an economical yet accurate tool that quickly locates the violators (Clarke et al., 2014, Guttikunda et al., 2014). Pinpointing violators among closely situated industrial units pose considerable challenge, even with precise measurements. The task is harder, when the decision making needs to be quick and inexpensive, based on limited number of field measurements taken using low-cost error-prone sensors (Castell et al., 2017).

In case of pharmaceutical industries, a significant contributor of air pollution, a major concern arises from the unregulated emission of volatile organic compounds (VOC), especially, in a fugitive manner (European Commission - DG Environment, 2009). An expensive technique for VOC source estimation within a limited area involves differential absorption infrared laser (Robinson et al., 1995). Measurements of leaky VOC emission was also performed using expensive gas chromatography–mass spectrometry method (Suresh, 2008). In contrast, economical source estimation has been reported based on dispersion models and sparse measurements (Rao, 2007). There, the effect of source emissions on pollutant concentration profile has been described using forward atmospheric transport dispersion models, such as the Gaussian model for advection-dispersion, and its extension, AMS/EPA Regulatory Model (AERMOD) (Environmental Protection Agency, 2015, Jeong, 2011). In such framework, source information has been estimated by solving an inverse problem, i.e., by finding model parameters that minimize the mismatch between measurements and predicted concentration levels (Sanfélix et al., 2015, Thomson et al., 2007, Wang et al., 2020). An adjoint-based backward model has also been suggested (Marchuk, 1995, Pudykiewicz, 1998, Rao, 2007). Alongside aforesaid

* Corresponding author.
E-mail address: jana@iith.ac.in (S. Jana).

https://doi.org/10.1016/j.heliyon.2020.e05296
Received 27 February 2020; Received in revised form 11 September 2020; Accepted 15 October 2020

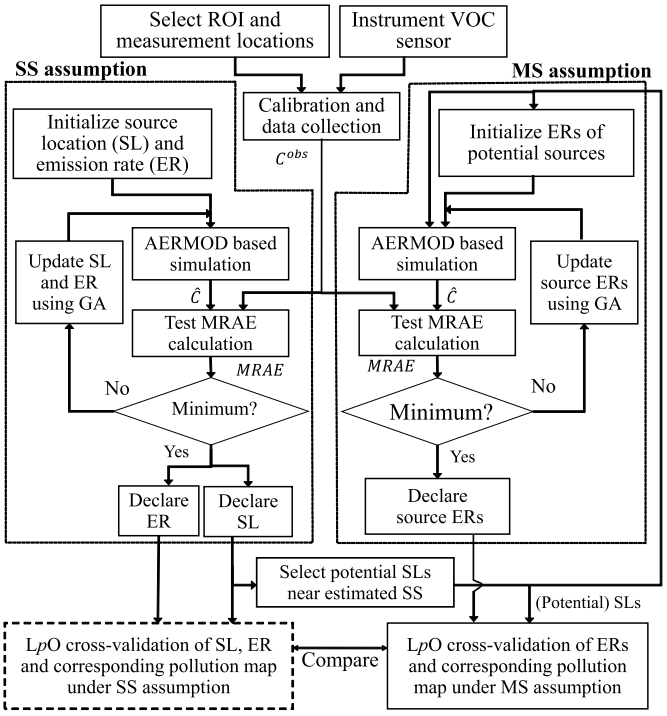


Fig. 1. Flowchart of our methodology. (SL: source location; ER: emission rate; $\hat{C}$ and C^{obs}: estimated and observed concentrations at measurement points respectively).

deterministic methods (Issartel et al., 2007, Singh and Rani, 2015), probabilistic methods have been reported as well (Bocquet, 2005, Wade and Senocak, 2013, Yee, 2012).

In most of the aforementioned works, ground truth information was collected from suitable tracer experiments. However, the ground truth on fugitive emission sources often remains unavailable (Hosseini and Stockie, 2016). In such circumstances, one operates under (or, attempts to prove) the hypothesis that it is possible to estimate sources with statistically significant levels of accuracy from a limited number of field readings taken using an error-prone sensor. In this direction, a previous attempt adopted AERMOD, and from limited number of measurements, estimated the emission source (assuming single) by minimizing the mean-squared error between the predicted and measured concentration levels (Kakarla et al., 2017). However, that attempt suffered from certain limitations such as lack of robustness and lack of thorough statistical validation.

This paper considers a similar setting as above, and proposes a novel method – RESILIENT (Robust Estimation of Source Information from LImited field measuremENTs) – for VOC source estimation. The proposed method is shown to achieve high accuracy and robustness, and is statistically cross-validated. The core principles of our technique are also directly validated on publicly available datasets collected from well-known tracer experiments. The rest of the paper is organized as follows. Section 2 explains the methodology, while Section 3 furnishes results. Finally, Section 4 concludes the paper.

2. Materials and methods

As alluded earlier, we aim at estimating pollution sources based on a few field measurements by solving an inverse problem. In the following, we describe in details the proposed method, RESILIENT, including experimental setup and data collection, source estimation and statistical validation. A schematic flowchart of the methodology is given in Fig. 1, and the corresponding procedure is presented in Procedure 1.

Procedure 1 The proposed RESILIENT procedure.

1. Data collection
   (a) Choose suitable set of measurement points in the ROI.
   (b) Make multiple measurements at each point and take median.
   (c) Find scale factor harmonizing CAQMS measurement with median reading of co-located sensor.
   (d) Map other median readings using same factor.
2. Source estimation under hypothetical SS assumption
   (a) Estimate SL and ER via GA-based minimization of MRAE (3).
   (b) Simulate pollution profile for estimated SL and ER via AERMOD.
3. Source estimation under practical MS assumption
   (a) Set as boundary contour of λ% of maximum value in SS profile.
   (b) Select sources within said boundary as potential ones.
   (c) Estimate ERs of selected sources via GA-based optimization (2).
4. Categorization of each source as either inactive or active:
   (a) If ER $\leq \epsilon$, mark the source inactive. Else, mark it active.

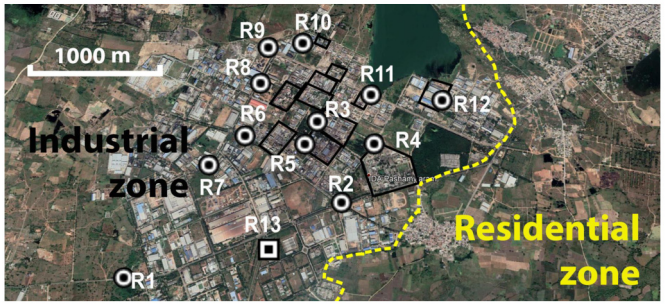


Fig. 2. Proximity of industrial and residential zones at Pashamylaram, an outskirt of Hyderabad, India. Circled dots: Locations R1, R2, ⋯, R12 of sensor-based measurement; Square within square: Location R13 of CAQMS as well as sensor-based measurement. Black borders: potential sources. [Map data sources: Google, Maxar Technologies; Map is generated using Google Earth with graphical elements and texts superimposed using Adobe Photoshop. Certain information is reproduced with permission from (Kakarla et al., 2017, Fig. 7) and (Kakarla et al., 2019, Fig. 1).]

2.1. Experimental setup and data collection

ROI and dominant VOC: Our ROI, shown in Fig. 2, and located between latitudes 17.5229 N and 17.5466 N, and longitudes 78.1610 E and 78.1952 E near Hyderabad, India, reported a frequent pungent odor, especially, at night. Noting the presence of pharmaceutical production units in the ROI, the odor was attributed to VOCs. Further, a continuous ambient monitoring station (CAQMS), located within the ROI and operated by Telangana State Pollution Control Board (TSPCB) reported ambient concentration levels of 3 VOCs, namely, toluene, xylene and benzene. During our experiment, the corresponding concentration levels were measured at the CAQMS as 45 $\mu g/m^3$, 12 $\mu g/m^3$, 3.6 $\mu g/m^3$, respectively, amounting to the respective relative proportions of 74.4%, 19.7% and 5.9%. Considering the CAQMS measurements taken at the specific time every day starting from 5 days before and ending at 5 days after, and computing relative proportions each day at hand, we found the median relative proportion for toluene to be 81.5%, with upper and lower quartile values equaling 82.5% and 79.4%, respectively. The above demonstrated that toluene was the dominant VOC (at the specific time), not only on the day of the experiment, but also over several days before and after. Accordingly, to simplify analysis, we make an assumption that the sought pollution sources emit only toluene.

Fugitive VOC emission: The ROI mainly houses pharmaceutical industries, with fugitive emission of VOC documented at various stages (European Commission - DG Environment, 2009). A prior work has also focused on quantifying such fugitive emission close to industrial units (Suresh, 2008). Against this backdrop, we consider only fugitive emission of VOC for our work, as we too collect readings reasonably close to such units (distance between any source and the nearest measurement point varied between 10 m and 70 m). Stacks for combustible

Table 1. Parameters for AERMOD simulation. U.S Environmental Protection Agency (2019).

Parameter	Value	Justification/Comments
Grid resolution	75 m	Satisfactory resolution
Grid size	48 × 36	Coverage of ROI
Surface roughness	1	Urban area
Albedo	0.35	Urban area
Bowen ratio	1.5	Urban area
Stack temperature	Ambient	Fugitive emissions
Exit velocity	0.01 m/s	Fugitive emissions
Stack height	4 m	Typical height of storage tanks
Surface frictional velocity	0.087 m/s	Estimated by AERMET
Sensible heat flux	−4.7 W/m^2	Estimated by AERMET
Monin-Obukhov length	11.8 m	Estimated by AERMET
Mixing height	61 m	Estimated by AERMET

VOC emission have heights (approximately 30 m) too tall to affect our ground-level local readings.

PID sensor: We measure VOC concentration using a photo ionization detector (PID) device named Alphasense PID-AH (Alphasense, 2012, Manes et al., 2016). It has a minimum detection limit of 5 ppb, linearity error of 3%, response time of three seconds, and sensitivity of 20 mV per 1 ppm for isobutylene. To avoid resource-intensive active calibration for toluene (or, a VOC mixture approximating the ambient one), we opted for passive calibration (Saukh et al., 2015).

Data collection and passive sensor calibration: The readings were taken (in units of mV) in March at night between 9 pm and 1 am with average temperature 27.9 degrees Celsius. As suggested in Step 1(a) of Procedure 1, $N = 13$ well separated points, denoted R1, R2, ⋯, R13, such that R13 is co-located with the CAQMS (refer Fig. 2), were chosen within the ROI. At each point, we noted the location given by Global Positioning System (GPS) and took readings approximately every 20 seconds for about six minutes from the sensor mounted at a height of about 2 m and recorded the corresponding median of those multiple readings (Step 1(b)). During the time when sensor readings were taken at R13, toluene concentration level measured by the CAQMS was reported as 45 μg/m^3 and corresponding median reading of the sensor was 8 mV, giving a scale factor of $45/8 = 5.625$ μg/m^3/mV (Step 1(c)). The scale factor maps sensor readings to concentration levels under a linearity assumption and passively calibrates the sensor (Saukh et al., 2015). It is then used to map the median readings taken at other points to the corresponding concentration levels (Step 1(d)). Numerical values of sensor readings taken at specific measurement points and mapped concentration values are given in Supplementary spreadsheet.

2.2. Source estimation

2.2.1. AERMOD-based simulation

As alluded earlier, we make use of AERMOD to predict the concentration level at any point, given source location and emission rates. Formally, we obtained concentration estimate $\hat{C}_{(x,y)}(\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^S, \{\rho_i\}_{i=1}^S)$ at location (x, y) induced by S sources located at $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^S$ with corresponding emission rates $\{\rho_i\}_{i=1}^S$. To make use of AERMOD, we needed various meteorological and other parameters. The chosen values and the rationale behind such choices are furnished in Table 1.

2.2.2. Source estimation as optimization

Using AERMOD, we next seek to estimate pollution sources which minimizes an appropriate cost function that indicates the mismatch between the simulated concentration levels and corresponding mapped median readings at points R1, R2, ⋯, R13. To give equal weightage to relative errors occurring at locations of larger as well as smaller measurements, we choose mean relative absolute error (MRAE) as the said cost function. Formally, denote by $C^{obs}_{(x,y)}$ the observed concentration at location (x, y), and by $\{(x_k, y_k)\}_{k=1}^N$, $N = 13$, the respective measurement points R1, R2, ⋯, R13. Then, we seek to minimize an average cost of the form $\frac{1}{N}\sum_{k=1}^N \phi\left(C^{obs}_{(x_k,y_k)}, \hat{C}_{(x_k,y_k)}\right)$, where

$$\phi(u, v) = \frac{|u - v|}{u}. \tag{1}$$

In general, the locations of potential sources should already have been catalogued. Hence, we were left with estimating the corresponding emission rates, i.e., (dropping the dependency of concentration estimate $\hat{C}$ on source locations, which were now specified)

$$\{\rho_i^*\}_{i=1}^S = \arg\min_{\{\rho_i\}_1^S} \frac{1}{N}\sum_{k=1}^N \phi\left(C^{obs}_{(x_k,y_k)}, \hat{C}_{(x_k,y_k)}(\{\rho_i\}_{i=1}^S)\right), \tag{2}$$

where $\phi(\cdot,\cdot)$ is given by (1), and $\{\rho_i^*\}_{i=1}^S$ indicate the desired emission rate estimates. We call this problem multiple source (MS) problem, for a reason explained next.

2.2.3. Reducing number of potential sources

In problem (2), the computation may be reduced by ignoring distant sources unlikely to influence field measurements. To this end, we first assume a hypothetical single source (SS, $S = 1$) with unknown location. Correspondingly, the MRAE problem (2) now takes the form

$$(\bar{x}^*, \bar{y}^*, \rho^*) = \arg\min_{(\rho,\bar{x},\bar{y})} \frac{1}{N}\sum_{k=1}^N \phi\left(C^{obs}_{(x_k,y_k)}, \hat{C}_{(x_k,y_k)}(\bar{x}, \bar{y}, \rho)\right) \tag{3}$$

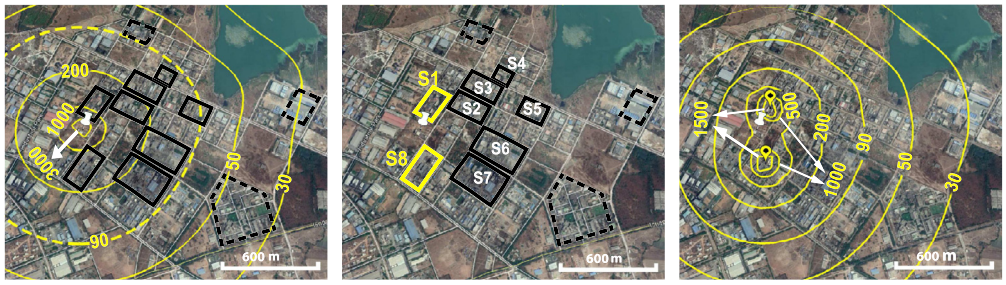
(dropping source index i), where $\phi(\cdot,\cdot)$ is given by (1), and the best-fit source location and emission rate are respectively denoted by $(\bar{x}^*, \bar{y}^*)$ and ρ^*. The estimated pollution concentration at location (x, y), given by $\hat{C}_{(x,y)}(\bar{x}^*, \bar{y}^*, \rho^*)$ under the SS assumption, was obtained via AERMOD-based simulation (Step 2(a) of Procedure 1). We considered the pollutant profile induced by the hypothetical SS (Step 2(b)), and set as the desired limiting boundary a contour of the aforementioned profile that correspond to a small percentage λ (typically, 2% or less) of the maximum concentration (Step 3(a)). Thus, we can solve problem (2) for only those potential sources (instead of for all sources in the ROI) that lie within the aforesaid boundary (Step 3(b)). We now set S to the reduced value of the number of aforesaid potential sources. The solution approach, mentioned in Step 3(c), is deferred to the next section. To distinguish from the intermediate SS problem (3), we call (2) the multiple source (MS) problem. The solution to the MS problem assigns a rate estimate to each of the potential sources under consideration. A source with negligible emission rate below an upper bound ϵ was declared inactive. Otherwise, it was marked active (Step 4). Clearly, the MS framework, using recorded source locations, should provide a better description than the hypothetical SS framework, especially when multiple sources are active.

2.2.4. Optimization using genetic algorithm

The SS optimization problem (3) can be solved even using exhaustive search, which avoids local minima (Kakarla et al., 2017). However, the MS optimization problem (2), facing more severe issues of local minima, could be too complex for exhaustive search. To solve this problem, as mentioned in Step 3(c) of Procedure 1, we employed genetic algorithm (GA) (Chambers, 2001), which is known to balance global exploration and local exploitation. Variables were encoded as real numbers with 3 decimal precision, multiple candidate solutions (chromosomes) were evaluated at once, and an improving pool of solutions was evolved according to an elitist strategy. We implemented the GA routines in FORTRAN, and employed those to solve both SS and MS problems.

2.3. Cross-validation

It is essential to ascertain whether the results of our RESILIENT procedure (described above and given in Procedure 1) are reliable when enforcement is considered. For this purpose, we cross-validate the results, as customary in absence of ground truth, per a further Procedure 2. In particular, N measurements are divided into two subsets, namely, the training subset of size $(N - p)$, and test subset of size p. For



(a) (b) (c)

Fig. 3. (a) Contour plot of estimated concentration (μg/m^3) profile along with estimated source location (SL) marked by white pin (SS assumption); Dotted contour (level 90): Contour threshold for potential source selection; solid black border: possible potential sources (b) Estimated active sources (solid yellow border), inactive potential sources (solid black border), distant sources (dashed border); (c) Contour plot of estimated concentration (μg/m^3) profile under MS assumption with locations of active sources (yellow balloons) and estimated SS location (white pin). [Map data sources: Google, Maxar Technologies; Map is generated using Google Earth with graphical elements and texts superimposed using Adobe Photoshop and AERMOD View. Source location information in (a) and (b) is reproduced with permission from (Kakarla et al., 2017, Fig. 7)].

any partition, source parameters are estimated from the training subset, and thence pollutant levels are predicted at the measurement points correspond to the test subset. The average test error serves as the basic performance index. We adopt leave-p-out cross-validation (LpO CV) (Arlot and Celisse, 2010), where averaging is performed over all partitions with fixed test subset size p. We refer to the main MS problem, unless otherwise mentioned.

Procedure 2 LpO Cross-validation procedure (under MS assumption).

1. Validation of predicted concentration:
   (a) Obtain Med-RAE and IQR-RAE at each p between $p_{\min}$ and $p_{\max}$.
   (b) Record minimum Med-RAE, corresponding p^* and IQR-RAE.
   (c) If Med-RAE $< \tau$ and IQR-RAE $< \theta$, proceed to step 2. Else, declare all estimates unreliable. STOP.
2. Validation of each source identified as inactive at $p = p_{\max}$:
   (a) Compute Q_3 of estimated ER values over all training subsets.
   (b) If $Q_3 \leq \epsilon'$, validate the source as inactive. Else, declare identification of inactivity as unreliable.
3. Validation of each source identified as active:
   (a) Compute Q_2 at $p = p_{\max}$. Also, compute QCD of estimated ER values over all training subsets for each p between $p_{\min}$ and $p_{\max}$.
   (b) If $|Q_2 - \rho^*|/\rho^* < \gamma$ at $p = p_{\max}$, QCD $< \delta$ at $p = p_{\max}$, and QCD(p)−QCD($p-1$) $< \Delta$ for each $p \geq p^*$, declare ER estimate ρ^* as reliable. Else, declare it unreliable.

2.3.1. Cross-validation of predicted concentration

In the present context, the number of aforesaid partitions equals ${}^N C_p$ ($N = 13$) for test subset size p, i.e., 13, 78, 286, 715, 1287, 1716, respectively, corresponding to $p = 1, 2, 3, 4, 5, 6$. In particular, the average test and the training errors were calculated as follows. For each partition corresponding to a given p, the relative absolute error (RAE) between predicted and observed concentration levels was noted at each of the $(N - p)$ training as well as the p test points. We considered one measurement point out of R1,...,R13 at a time, and only those partitions, in which the said point belongs to the training subset, collected corresponding RAE values, and calculated point-specific median of RAE (Med-RAE) and inter-quartile range (IQR) of RAE (IQR-RAE) for training. We adopted the median (instead of mean) and the IQR (instead of standard deviation) as respective measures of central tendency and dispersion in view of their insensitivity to outliers (Weisberg, 1992). For the same point, next considering the rest of the partitions, where the said point belongs to the test subset, we similarly computed test Med-RAE and test IQR-RAE. Average test error (Med-RAE) was computed over measurement points R1,...,R13, and plotted as a function of p between suitable bounds $p_{\min}$ and $p_{\max}$ (Step 1(a) in Procedure 2). The value of p^* minimizing average test Med-RAE was noted, alongside the corresponding average test IQR-RAE (Step 1(b)). Here, a low average test Med-RAE indicates a desirably accurate method. For a method to be robust, the average test IQR-RAE should be low. We compare each to a suitable upper bound of acceptability (τ and θ, respectively), and do not proceed unless both conditions are met (Step 1(c)). We repeat Steps 1(a) and 1(b) for the SS problem too, to verify our expectation that the MS framework provides a superior model compared to the SS framework.

2.3.2. Cross-validation of emission rate estimates

We turn to the cross-validation of inactive sources and emission rate estimates of active sources (both obtained using the complete data for training, i.e., $p = 0$). As shown in Step 2 of Procedure 2, we individually validated inactive sources in the worst case scenario of $p = p_{\max}$. For each such source, we obtained one emission rate estimate for each of the ${}^N C_p$ training sets. Thence, we computed the third quartile value, Q_3, of those rate estimates (Step 2(a)). If Q_3 is less than ϵ', ϵ' being a suitably small threshold, then the source is validated as inactive (Step 2(b)). One may or may not set ϵ' and ϵ (in Step 4(a) of Procedure 1) to the same value. For each active source, we validated the rate estimate in Step 3 of Procedure 2. If the relative change in the median $|Q2 - \rho^*|/\rho^*$ at $p = p_{\max}$ with respect to rate estimate ρ^* at $p = 0$ is less than some upper bound γ, the quartile coefficient of dispersion QCD $= (Q_3 - Q_1)/(Q_3 + Q_1)$ (computed in Step 3(a), and Q_1 and Q_3 respectively denoting the first and the third quartiles) is less than δ at $p = p_{\max}$, and the rate of increase in QCD with respect to p is less than an upper bound Δ for each $p \geq p^*$ (Step 3(b)), emission rates are declared as reliable. Here, γ, δ and Δ are taken as suitably small thresholds.

3. Results

We first consider the field experiment described in Section 2, present the results of the Procedure 1, and cross-validate per the Procedure 2. We then directly validate our method based on a public dataset arising from certain tracer experiments.

3.1. Results from field experiment

3.1.1. Source estimation under SS and MS assumptions

Solving the intermediate SS problem (3), the latitude and longitude of the hypothetical source was estimated as (17.5382 N, 78.1774 E) with emission rate 3.013 g/s (Step 2(a) in Procedure 1). Hence, simulating the pollution map (Step 2(b)), presented in Fig. 3(a), we observed the maximum predicted concentration level of 4516 μg/m^3, and chose as the desired boundary (Step 3(a)) the contour corresponding to $\lambda = 2\%$ of 4516 μg/m^3, i.e., 90 μg/m^3. $S = 8$ of the catalogued sources, denoted S1, S2, ⋯, S8, were found to lie either entirely or partially within such boundary (Step 3(b), see Fig. 3(b)), and were considered potential ones, for which problem (2) was solved. As shown in Table 2, up to three decimal places of accuracy, rate estimates of six sources S2, S3, ⋯, S7 were each 0.000 g/s, and those of the remaining two sources S1 and S8

Table 2. Rate estimates of candidate sources for $p = 0$ (all measurements) and their second and third quartiles for $p = 6$.

Sources	Rate estimate (g/s)			
	$p = 0$	$p = 6$		
		Q1	Q2	Q3
S1	1.361	0.923	1.282	1.510
S2	0.000	0.000	0.000	0.000
S3	0.000	0.000	0.000	0.000
S4	0.000	0.000	0.000	0.000
S5	0.000	0.000	0.000	0.000
S6	0.000	0.000	0.000	0.004
S7	0.000	0.000	0.000	0.000
S8	1.323	1.198	1.324	1.817

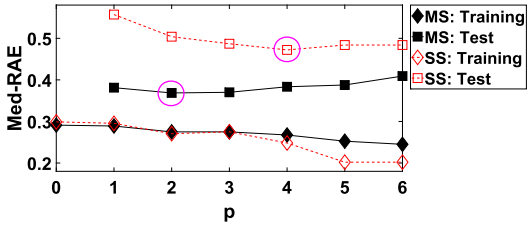


Fig. 4. LpO CV under SS and MS assumptions: Plot of point-wise average of Med-RAE against p.

were respectively 1.361 g/s and 1.323 g/s (Step 3(c)). Setting a small threshold $\epsilon = 0.004$ g/s, we identified the former six sources as inactive, and the latter two as active (Step 4(a)). The two active sources S1 and S8 happen to be the ones closest to the hypothetical SS location (see Fig. 3(b)). Pollution profiles under SS and MS assumptions were obtained, and respective contour plots are also presented in Figs. 3(a) and 3(c), which are different, especially at high concentration levels.[1] We now turn to cross-validation of estimates.

3.1.2. Cross-validation

In absence of ground truth, we now turn to LpO CV of the aforementioned results per the Procedure 2.

Cross-validation of predicted concentration levels: In particular, We considered the RAE criterion, and values of p between the limits $p_{\min} = 1$ and $p_{\max} = 6$. We obtained and plotted test Med-RAE against p and recorded minimum value at $p^* = 2$ for the MS problem (Steps 1(a) and 1(b) in Procedure 2). For the sake of comparison, repeating those steps for the SS problem, the analogous minimum was observed at $p^* = 4$ (Fig. 4). In general, the test error was smaller for the MS problem, as expected. We also obtained analogous plots of training (instead of test) Med-RAE. The training Med-RAE decreased with p in each case, as learning a model is easier with more data. Further, the training error is larger in the SS case for smaller p (more training data), but smaller for larger p (less training data). This is consistent with the fact that the simpler SS model fits less (resp. more) training data better (resp. worse) than the complex MS model.

For the MS (resp. SS) problem, test Med-RAE and test IQR-RAE were respectively obtained as 0.37 (resp. 0.50) and 0.06 (resp. 0.43) at $p^* = 2$ (resp. $p^* = 4$) (Steps 1(a) and 1(b) in Procedure 2). This shows overall (coarse-grain) superiority of the MS framework in terms of both accuracy and robustness. Here, a test error (Med-RAE) of 37% for the MS problem should be considered satisfactory in the face of error-prone sensors, imperfect modeling and imperfect knowledge of parameters, and other inaccuracies. A test IQR-RAE of 0.06 indicates significant robustness. One may choose $\tau = 0.4$ (40%) and $\theta = 0.1$ (10%) Step 1(c), and proceed further.

[1] However, difference in Med-RAE taken over all measurement points (0.299 versus 0.291, respectively) between SS and MS scenarios appears insignificant.

Table 3. Observed and estimated values along with error statistics at locations R1 through R13 with MS ($p = 2$) and SS ($p = 4$) assumptions during LpO CV.

Loc	C_{obs}	Med-$\hat{C}_{\text{test}}$	Med-RAE	IQR-RAE
		MS (SS)	MS (SS)	MS (SS)
R1	22.5	12.4 (17.5)	0.45 (**0.42**)	**0.05** (0.74)
R2	33.8	54.0 (54.2)	**0.60** (0.61)	**0.13** (0.38)
R3	506.3	229.6 (121.0)	**0.55** (0.76)	**0.03** (0.25)
R4	47.8	51.1 (40.2)	**0.07** (0.28)	0.21 (**0.20**)
R5	405.0	551.7 (125.2)	**0.36** (0.69)	**0.11** (0.29)
R6	196.9	159.2 (307.0)	**0.19** (0.58)	**0.03** (1.17)
R7	78.8	55.3 (74.5)	0.30 (**0.25**)	**0.04** (0.79)
R8	253.1	81.7 (114.1)	0.68 (**0.56**)	**0.02** (0.13)
R9	84.4	72.3 (107.2)	**0.14** (0.36)	**0.01** (0.77)
R10	90.0	104.8 (90.0)	**0.16** (0.19)	**0.02** (0.20)
R11	56.3	74.8 (80.4)	**0.33** (0.48)	**0.03** (0.46)
R12	84.4	25.4 (25.9)	0.70 (**0.69**)	**0.03** (0.09)
R13	45.0	33.0 (33.4)	0.27 (**0.27**)	**0.03** (0.19)
Avg			**0.37** (0.50)	**0.06** (0.43)
Std			0.21 (**0.19**)	**0.06** (0.33)

To obtain a fine-grain insight, we furnish point-wise information in Table 3. Under each assumption, we furnish for each measurement point median value Med-$\hat{C}_{\text{test}}$ of simulated concentration levels taken across all partitions where the said point belonged to the test subset, along side associated Med-RAE and IQR-RAE values. At 8 (resp. 12) of 13 measurement points, Med-RAE (resp. IQR-RAE) is lower under MS assumption than that under SS assumption, showing even fine-grain superiority of the former framework. The standard deviation in Med-RAE over measurement points under MS assumption remained close to that under SS assumption, perhaps reflecting the inherent imprecision of the underlying data. However, the standard deviation in IQR-RAE under MS assumption remained significantly lower than that under SS assumption, indicating even fine-grain robustness of the MS framework.

Cross-validation of inactive and active sources: We now cross-validate the inactive and active sources identified earlier. Consider $p = p_{\max} = 6$ (Step 2). For each of the ${}^{N}C_{p} = 1716$ training subsets of size $N - p = 7$, we estimated emission rates of sources S1, S2, …, S8. We plot cumulative relative frequency function of those rate estimates for each source in Fig. 5(a) and compute Q_1, Q_2 and Q_3 (Table 2). Choosing $\epsilon' = \epsilon = 0.004$ g/s for the sake of simplicity, we had $Q_3 \le \epsilon'$ for each of the six inactive sources S2, S3,..., S7, which were thus validated as inactive (Steps 2(a) and 2(b)). For each such source, the 72 percentile mark equals essentially zero (up to three decimal places of accuracy) emission as seen in Fig. 5(b). For each of the active sources S1 and S8, Q_1, Q_2, Q_3 values for rate estimates were plotted against p in Fig. 5(c). Observe in the figure that the change in these quantities with increasing p is gradual, rather than drastic, as desired. For each of those sources, the relative change in the median $|Q_2 - \rho^*|/\rho^*$ is low for any p (Fig. 5(d)), and is less than $\gamma = 5\%$ even at $p = p_{\max} = 6$. Further, the QCD is also plotted against p in the same figure. At $p = p_{\max} = 6$, QCD values of S1 and S8 were 20.5% and 24.1%, respectively, each less than a sensible choice of $\delta = 25\%$. Moreover, the rate of change in QCD for $p \ge p^* = 2$ remained less than $\Delta = 8\%$ for either source. Thus, for the aforementioned choice of parameters, both the active sources S1 and S8 are validated as active (Steps 3(a) and 3(b) in Procedure 2).

3.2. Direct validation based on tracer experiments

We have considered a field experiment (introduced in Section 2.1), where ground truth remains unavailable, and presented results along with cross validation. Now, we turn to direct validation of our method based on publicly available data from a well-known tracer experiment.

3.2.1. Prairies grass tracer experiments

In each of the 68 Prairies Grass Tracer (PGT) experiments, performed during July and August, 1956 (Barad, 1958), a single SO_2 source

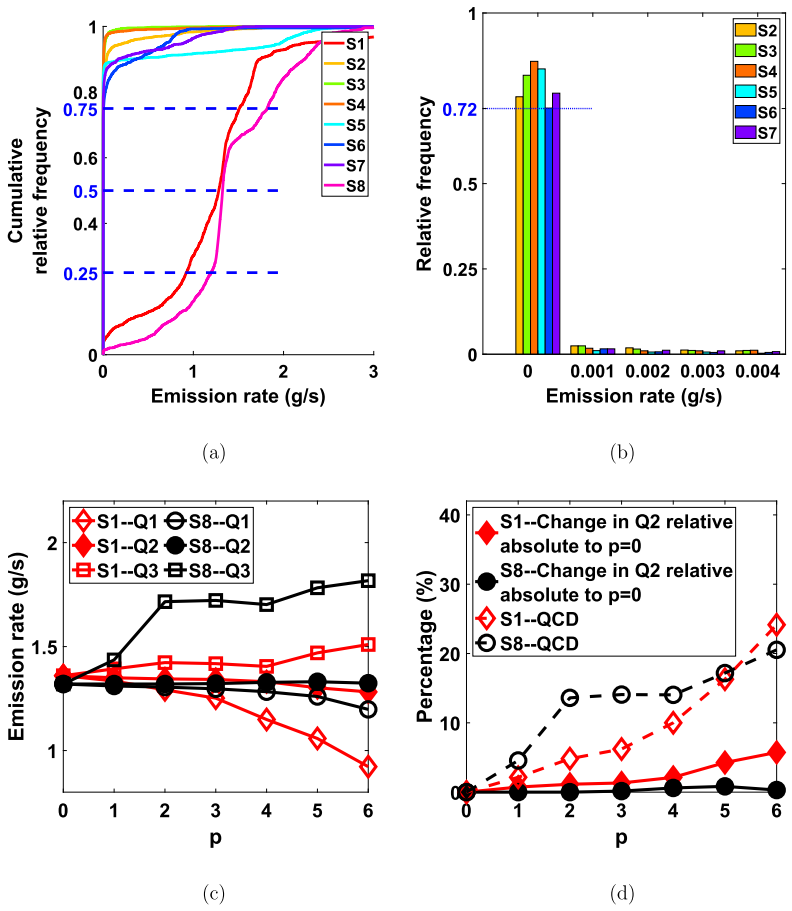


Fig. 5. (a) Cumulative relative frequencies of rate estimates of candidate sources and (b) relative frequencies of rate estimates of inactive sources of LpO CV at $p = 6$; Plot of (c) quartile values (Q_1, Q_2 and Q_3) and (d) dispersion statistics of emission rates of active sources S1 and S8 against p.

was placed at a height of 46 cm, and SO_2 concentration was measured at various points on arcs at distances of 50 m, 100 m, 200 m, 400 m and 800 m from the origin. The source emission rate was varied among those experiments. Environmental conditions varied, too. The wind velocity, averaged over 10 min interval, was measured at heights of 0.25 m, 0.5 m, 1 m, 2 m, 4 m, 8 m and 16 m above the ground. While sensors were placed on the entirety of aforesaid arcs, readings from only the downwind ones were recorded alongside wind velocity as well as other meteorological quantities. The settings of the PGT experiments and our experiment described in Section 2.1, while bearing certain similarities, had important differences. While our Pashyamylaram experiment only considered significant measurements, many sensors (not in the downward region) in the PGT experiments recorded zero measurement. So, the earlier relative error measures did no longer apply, and accordingly, we now consider mean absolute error (MAE) in place of MRAE. Accordingly, in both the SS problem (3) and the MS problem (2), we now use $\phi(u, v) = |u - v|$ (instead of the earlier $\phi(u, v) = |u - v|/u$).

3.2.2. Results and direct validation

Single experiment: We first validate the proposed method — specifically, Steps 2, 3 and 4 in Procedure 1 — for PGT experiment 13, chosen *ad hoc*. Solving SS problem (3) (now with $\phi(u, v) = |u - v|$), the source location and emission rate were estimated as $(1, -5)$ and 62 g/s, respectively (step 2(a)). We simulated the corresponding concentration profile (dashed contours in Fig. 6(a), Step 2(b)). We chose a boundary contour (black dashed and dotted in Fig. 6) corresponding to $\lambda = 0.5\%$ of the maximum concentration value of 9520 mg/m^3 (Step 3(a)), and selected multiple ($S = 8$, to maintain parity with our field experiment) source locations within said boundary including origin (0,0) and seven more equispaced locations along the major axis (Fig. 6(a), Step 3(b)). Solving (2), emission rates of the said $S = 8$ sources were estimated (Step 3(c)), with rate estimate for the source at origin equaling 50.5 g/s. Indeed, setting a low threshold $\epsilon = 2.5$ g/s, we declared the source at origin to be active, and the remaining candidate sources to be inactive (Step 4(a)). Indeed, the emission rate estimate of the source at origin turned out to account for about 90% of the sum total of all emission rate estimates. Turning to ground truth, it is known that the only source was located at origin with an emission rate of 56.5 g/s, which closely matched the estimate of 50.5 g/s. The concentration profile corresponding to the estimated source is shown by contours with solid lines in Fig. 6(a). Of course, solid MS contours nearly matched dotted SS contours, because in reality there was only one active source, and the estimate of the hypothetical SS location was quite close to the actual location, i.e., the origin.

Statistics for multiple experiments: We repeat the proposed method on 9 more PGT experiments, numbered 8, 14, 15, 23, 24, 25, 26, 36 and 37, and study the statistics of all 10. In each, the only active source was located at origin. However, as stated earlier, other aspects such as the emission rate and meteorological quantities such as wind velocity differed among experiments. For each experiment, we estimated the emission rate estimate of the source at origin, took the maximum of the rate estimates among the rest of the sources, and normalized those quantities by the sum of rate estimates. Considering all 10 PGT experiments at hand, we calculated the cumulative relative frequencies of such normalized rate estimates, and plot in Fig. 6(b). While the first quartile of normalized emission rate for the source at origin is recorded as 91.6%, the third quartile of the aforesaid maximum is only 4.8%, establishing the strong predominance of the former, consistent with known facts. Plotting estimated versus recorded (ground truth) emission rates for the active source over the 10 experiments in Fig. 6(c), we ob-

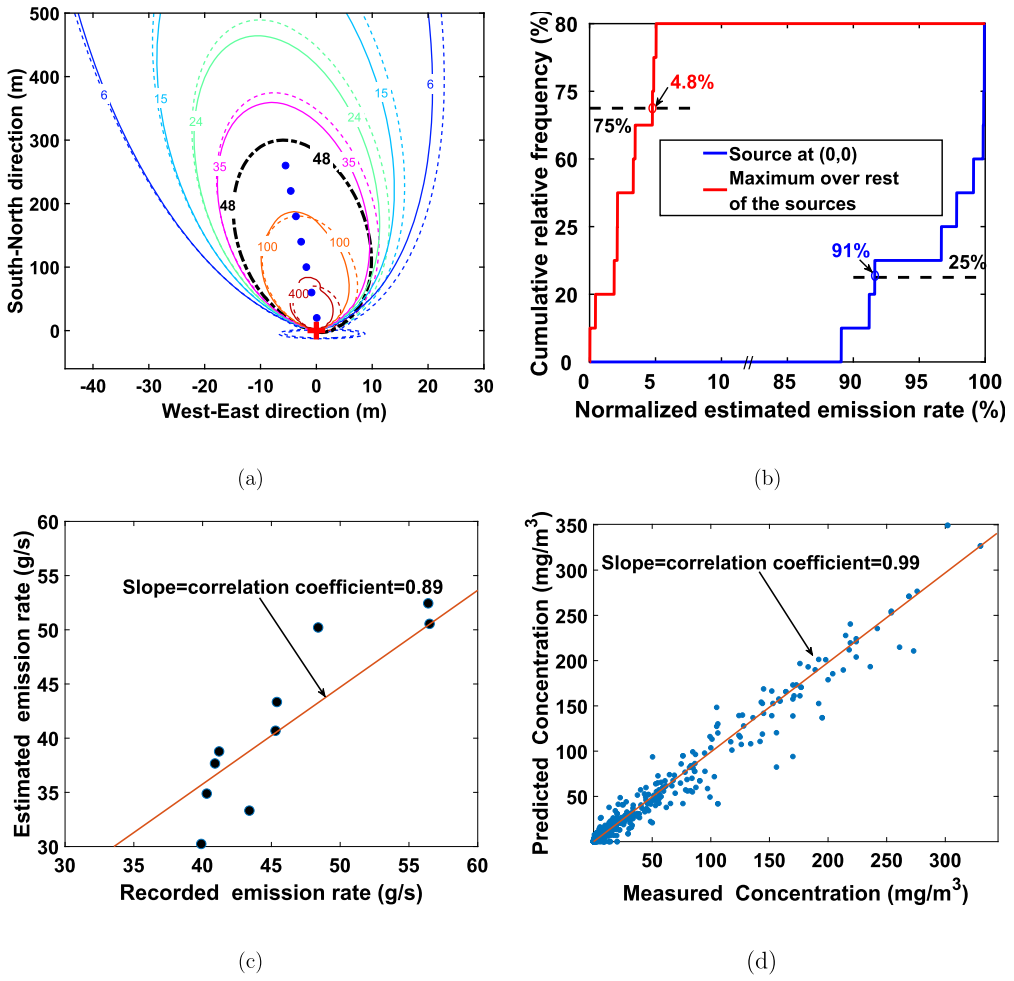


Fig. 6. PGT experiments and validation statistics: (a) Boundary contour selection and estimated SO_2 profiles for experiment 13 (conducted on 23rd July, 1956, between the interval 19 : 55 hrs and 20 : 15 hrs): Dotted contour – Estimated SO_2 profile under SS assumption; Contour in dash and dot – Selected boundary contour at value of 48 (corresponding to $\lambda = 0.5\%$) under SS assumption; Red plus – Known source location, i.e., origin (0,0); Blue filled circle – Additional candidate sources (chosen to be uniformly distributed within the boundary contour for the purpose of illustration); Contour with solid lines – Estimated SO_2 profile under MS assumption; (b) Plots of cumulative relative frequency of relative rate estimates over 10 selected PGT experiments: Blue corresponds to the (active) source at origin (0,0); Red corresponds to the maximum taken over rest of the (inactive) sources.; (c) Correlation plot of recorded and estimated emission rates; (d) Correlation plot of recorded and estimated concentrations at measurement (sensor) points.

tained a significant correlation coefficient of 0.89. Depicted in Fig. 6(d), a similar analysis of predicted versus measured (ground truth) concentration levels at various sensors over the 10 experiments yields a high correlation coefficient of 0.99. These high correlation values provide statistical validation of the accuracy of the proposed method.

4. Conclusions

In this paper, we proposed RESILIENT, a robust statistical method for estimating pollution sources based on limited number of field measurements made by a low-cost error-prone sensor. Assuming that all industrial units had already been catalogued by the authorities, our task consisted in identifying the active sources among those, and estimating the corresponding emission rates. Sources with negligible emission rates were declared inactive, and those with high emission rates were identified as active. Finally, the match between the measurements and predicted concentration values, and source inactivity/activity were validated using L*p*O CV. The core principles of RESILIENT were also validated using publicly available datasets collected from well-known tracer experiments.

To translate the proposed source estimation method into a regulatory tool, one will require additional steps. Extensive field trials (beyond the scope of the present work) must be conducted to tune and test the proposed procedures under varying conditions. The various thresholds chosen *ad hoc* in this paper, need to be ultimately decided based on statistics collected from numerous field trials. Further, the accuracy of the proposed method can conceivably be improved as follows. Recall that our method presently makes use of AERMOD, which ignores complex fine-scale chemical and other interactions during pollutant advection and diffusion, assumes uniform wind field and other simplifications, and thus models the steady state behavior. However, VOC pollutants at hand are expected to interact with ambient atmosphere, realistic modeling of which potentially improves source estimation performance. The wind field in complex urban regions is significantly influenced by buildings and other structures, and cannot truly be considered constant throughout a domain. Use of more sophisticated models, possibly based on computational fluid dynamics, should improve performance. Moreover, considering larger ROIs, and including combustible VOC emissions would make the proposed tool more comprehensive.

Declarations

Author contribution statement

Anand Kakarla, Soumya Jana: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data;

Contributed reagents, materials, analysis tools or data; Wrote the paper. Asif Qureshi, Swades De: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper. Shashidhar Thatikonda: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Funding statement

This work was partly supported by the Department of Electronics and Information Technology, Ministry of Communications and Information Technology, India (13(2)/2012-CC&BT).

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2020.e05296.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.heliyon.2020.e05296.

References

Al-Wahaibi, A., Zeka, A., 2015. Health impacts from living near a major industrial park in Oman. BMC Public Health 15, 524.

Alphasense, 2012. Photo ionization detector. Available at http://www.alphasense.com/WEB1213/wp-content/uploads/2016/09/PID-AH2.pdf. (Accessed 14 September 2020).

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79.

Barad, M.L., 1958. Project Prairie Grass, a Field Program in Diffusion. Volume 1. Technical Report Air Force Cambridge Research Labs Hanscom Afb, MA.

Bocquet, M., 2005. Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: theory. Q. J. R. Meteorol. Soc. 131, 2191–2208.

Castell, N., Dauge, F.R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., Bartonova, A., 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environ. Int. 99, 293–302.

Chambers, L.D., 2001. The Practical Handbook of Genetic Algorithms: Applications. CRC Press.

Clarke, K., Kwon, H.-O., Choi, S.-D., 2014. Fast and reliable source identification of criteria air pollutants in an industrial city. Atmos. Environ. 95, 239–248.

Environmental Protection Agency, 2015. AERMOD modeling system. Available at https://www.epa.gov/scram/air-quality-dispersion-modeling-preferred-and-recommended-models. (Accessed 14 September 2020).

European Commission - DG Environment, 2009. Guidance on VOC substitution and reduction for activities covered by the VOC solvents emissions directive. Available at https://www.varam.gov.lv/sites/varam/files/content/files/voc_guidance_210509.pdf. (Accessed 14 September 2020).

Guttikunda, S.K., Goel, R., Pant, P., 2014. Nature of air pollution, emission sources, and management in the Indian cities. Atmos. Environ. 95, 501–510.

Hosseini, B., Stockie, J.M., 2016. Bayesian estimation of airborne fugitive emissions using a Gaussian plume model. Atmos. Environ. 141, 122–138.

Issartel, J.-P., Sharan, M., Modani, M., 2007. An inversion technique to retrieve the source of a tracer with an application to synthetic satellite measurements. Proc. R. Soc. A, Math. Phys. Eng. Sci. 463, 2863–2886.

Jeong, S.-J., 2011. CALPUFF and AERMOD dispersion models for estimating odor emissions from industrial complex area sources. Asian J. Atmos. Environ. 5, 1–7.

Kakarla, A., Munagala, V.S.K.R., Qureshi, A., Thatikonda, S., De, S., Ishizaka, T., Fukuda, A., Jana, S., 2019. Comprehensive air quality management system for rapidly growing cities in developing countries. In: 2019 IEEE Glob. Humanit. Technol. Conf., pp. 1–7.

Kakarla, A., Qureshi, A., Shashidhar, T., De, S., Singh, S.G., Jana, S., 2017. Source localization via AERMOD-based simulation under mean squared error criterion: demonstration using field data. In: IEEE Int. Geosci. and Remote Sensing Symp., pp. 6201–6204.

Manes, G., Collodi, G., Fusco, R., Gelpi, L., Manes, A., Di Palma, D., 2016. Realtime gas emission monitoring at hazardous sites using a distributed point-source sensing infrastructure. Sensors 16, 121.

Marchuk, G.I., 1995. Main and adjoint equations. Perturbation theory. In: Adjoint Equations and Analysis of Complex Systems. Springer Netherlands, Dordrecht, pp. 9–94.

Oyinloye, M.A., 2015. Environmental pollution and health risks of residents living near Ewekoro cement factory, Ewekoro, Nigeria. Int. J. Environ. Ecol. Geol. Geophys. Eng. 9, 108–114.

Pudykiewicz, J.A., 1998. Application of adjoint tracer transport equations for evaluating source parameters. Atmos. Environ. 32, 3039–3050.

Rao, K.S., 2007. Source estimation methods for atmospheric dispersion. Atmos. Environ. 41, 6964–6973.

Robinson, R.A., Woods, P.T., Milton, M.J., 1995. Dial measurements for air pollution and fugitive-loss monitoring. In: Air Pollution and Visibility Measurements. In: Int. Soc. for Opt. and Photonics, vol. 2506, pp. 140–150.

Sanfélix, V., Escrig, A., López-Lilao, A., Celades, I., Monfort, E., 2015. On the source inversion of fugitive surface layer releases. Part I. Model formulation and application to simple sources. Atmos. Environ. 109, 171–177.

Saukh, O., Hasenfratz, D., Thiele, L., 2015. Reducing multi-hop calibration errors in large-scale mobile sensor networks. In: Proc. of the 14th Int. Conf. on Info. Proc. in Sensor Netw., pp. 274–285.

Singh, S.K., Rani, R., 2015. Assimilation of concentration measurements for retrieving multiple point releases in atmosphere: a least-squares approach to inverse modelling. Atmos. Environ. 119, 402–414.

Suresh, S., 2008. Environmental sampling and analysis of volatile organic compounds at Tarapur, Navi Mumbai, Dombivali, Chandrapur and Aurangabad industrial areas. Available at https://www.mpcb.gov.in/sites/default/files/focus-area-reports-documents/MPCB_VOC_Report_ver5.pdf. (Accessed 14 September 2020).

The Economic Times of India Over Rs 83 lakh penalty imposed for violating pollution norms: CPCB. Available at https://economictimes.indiatimes.com/politics/over-rs-83-lakh-penalty-imposed-for-violating-pollution-norms-cpcb/articleshow/66500454.cms?from=mdr. (Accessed 14 September 2020).

Thomson, L.C., Hirst, B., Gibson, G., Gillespie, S., Jonathan, P., Skeldon, K.D., Padgett, M.J., 2007. An improved algorithm for locating a gas source using inverse methods. Atmos. Environ. 41, 1128–1134.

U.S Environmental Protection Agency, 2019. User's guide for the AERMOD meteorological preprocessor (AERMET). Available at https://www3.epa.gov/ttn/scram/7thconf/aermod/aermet_userguide.pdf. (Accessed 14 September 2020).

Wade, D., Senocak, I., 2013. Stochastic reconstruction of multiple source atmospheric contaminant dispersion events. Atmos. Environ. 74, 45–51.

Wang, J., Zhang, R., Li, J., Xin, Z., 2020. Locating unknown number of multi-point hazardous gas leaks using principal component analysis and a modified genetic algorithm. Atmos. Environ. 230, 117515.

Weisberg, H.F., 1992. Central Tendency and Variability, Vol. 83. Sage.

Yee, E., 2012. Probability theory as logic: data assimilation for multiple source reconstruction. Pure Appl. Geophys. 169, 499–517.