Environmental Modelling and Software 128 (2020) 104701

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: http://www.elsevier.com/locate/envsoft

Model-based dense air pollution maps from sparse sensing in multi-source scenarios

Asaf Nebenzal [a], Barak Fishbain [b,*], Shai Kendler [b,c]

[a] Department of Mathematics, Technion – Israel Institute of Technology, Haifa, 320003, Israel
[b] Faculty of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa, 320003, Israel
[c] Environmental Physics Department, Israel Institute for Biological Research, 24 Lerer st, Ness Ziona, 74100, Israel

ARTICLE INFO

Keywords:
Air quality modeling
Source detection
Gaussian model
Interpolation
Spatial maps

ABSTRACT

A method for producing dense air pollution maps, based on any given air-pollution dispersion model, is presented. The scheme consists of two phases. At the first stage, sources' locations and emission rates, i.e., source term estimation, as a function of the model's parameter space are sought ("**backward computation**"). Then, the source term is used to generate the dense maps utilizing the same dispersion model ("**forward computation**"). The algorithm is model-invariant to the dispersion model, and thus is suitable for a wide range of applications according to the required accuracy and available resources. A simulation of an industrial area demonstrated that this method produced more accurate maps than current state-of-the-art techniques. The resulting dense air pollution map is thus a valuable tool for air pollution mitigation, regulation and research.

Software availability

Name of software: Multiple Sources Detection Algorithm
Developer: A. Nebenzal, Dept. of Applied Math, Technion – Israel Institute of Technology, Email: asaf.n@technion.ac.il
Year first available: 2019
Software required: MATLAB 2016 (and up)
Program language: MATLAB scripting language
Program size: 100 Kbytes
Availability: Source code available at: https://fishbain.net.technion.ac.il/home-page/projects-software/

1. Introduction

Air pollution is a major public health concern and negatively impacts the environment (WHO, 2019). Today, air pollution is considered one of the worst environmental health risks. Therefore, there is a great need to detect and monitor the various air-pollution sources and their effect on the environment. Typically, air pollution studies are based on data collected from standard air quality monitoring stations (AQMS). AQMS supply accurate and continuous measurements; are operated by professional personnel; and the data undergoes a process of quality control (Broday and Yuval, 2010). This has made AQMS the gold standard for air pollution data measurements (CDC, 2018). However, their construction and maintenance costs are high, they are bulky, and the equipment is usually immobile (Castell et al., 2017). Thus although AQMS networks are very reliable, they are spread thinly in space (Kumar et al., 2015). To illustrate, in the greater Chicago metropolitan area there are only three EPA AQMS that monitor NO_2 levels (EPA, 2019).

To evaluate the spatial variability of a concentration field, AQMS measurements do not provide sufficient coverage. Thus, any study that aims at analyzing the spatial distribution of air pollution, such as exposure assessments and epidemiological studies, needs to implement a variety of techniques to overcome measurement sparsity. The most common methods are land use regression (LUR), atmospheric dispersion models and spatial interpolation schemes. LUR techniques aim at inferring pollution levels in all non-monitored locations in the catchment region from a set of predictors such as land use, physical geography, and transportation variables (Morley and Gulliver, 2018).

Spatial interpolation is the process of finding a continuous function that best describes the whole study area. It can be classified as deterministic or geostatistical. The best-known deterministic schemes are the inverse distance weighted (IDW) and nearest neighbor (NN) algorithms. The geostatistical schemes include various types of Kriging (Li and Heap, 2014). Using these methods, Sacks et al. (2018) developed a software, which was based on the IDW algorithm for estimating how changes in

* Corresponding author.
E-mail addresses: asaf.n@technion.ac.il (A. Nebenzal), fishbain@technion.ac.il, fishbain@technion.ac.il (B. Fishbain), skendler@cv.technion.ac.il (S. Kendler).

https://doi.org/10.1016/j.envsoft.2020.104701
Received 14 April 2019; Received in revised form 16 March 2020; Accepted 16 March 2020
Available online 19 March 2020

air quality affected economic and health factors. Beauchamp et al. (2017) used a Kriging-domain estimation to investigate the representativeness of an AQMS, and the exceedance area of a pollutant, which does not always overlap.

A study analyzing exposure to air pollution in Toronto Canada implemented IDW interpolation, and used the AQMS as a reference (Buteau et al., 2017). Using both Kriging and LUR hybrid schemes, Wu et al. (2018) established a coupled Kriging and LUR model to estimate PM2.5 over Taiwan measured from 71 AQMS between 2006 and 2011. Using the hybrid method, the correlation between the actual and computed concentration was higher compared to the LUR method. The amount of improvement depended on the nature of the problem. For example, at time resolution of one month, the R^2using the conventional LUR method was 0.70 compared to 0.88 using the hybrid method.

Regardless of the interpolation method, mathematically, all interpolated values over the domain are a weighted average of the measurements. Both interpolation and LUR techniques disregard the physicochemical characteristics of pollutants, the physics governing the dispersion of the polluting materials, and meteorology. Thus, an alternative approach that uses AQMS data incorporated with a dispersion model and meteorological data may provide much more accurate, sensitive readouts than the classic interpolation techniques.

Dispersion modeling is a mathematical description of how pollutants disperse in the atmosphere using source and meteorological parameters over a defined period of time (Masey et al., 2018). These models make use of source parameters such as emission rate, source locations and stack height together with meteorological conditions such as humidity, atmospheric stability, as well as wind speed and direction. These models range from relatively simple ones such as the Gaussian Air Pollution Dispersion Model (GAPDM) (Ermak, 1977), to complex models such as AERMOD (Cimorelli et al., 2005) and CALPUFF (Scire et al., 2000).

Cambridge Environmental Research Consultants' (CERC) Atmospheric Dispersion Modeling System Ver. 5 (ADMS-5) is a popular variant of the GAPDM. The ADMS-5 can handle several kinds of terrains including urban, coastal or mountain areas. ADMS has been used for environmental studies including densely populated areas.(Carruthers et al., 1994). Choosing the dispersion model requires a balance between several factors: 1) the nature of the problem, i.e., climatic condition, terrain size, and topography; 2) the required precision; and 3) available resources to perform the computations (Leelőssy et al., 2014). While these models facilitate, in many cases, a reliable and accurate representation of the dense pollution filed, they require a comprehensive knowledge on the sources and on the problem's physicochemical attributes. These data are often not available. Thus, the need for mechanisms that provide accurate pollution estimations with limited datasets.

Here we present an interpolation scheme that generates dense spatial pollution maps by integrating dispersion models into the process. This allows for more accurate dense pollution maps than the state-of-the-art interpolation schemes, as physicochemical model is integrated in the process, while alleviating the requirement for exact knowledge about the problem's characteristics. This is achieved by utilizing the model for source term estimation (backward computation). Then, the source term is used for calculating the pollution dense maps (forward computation). This concept was presented for the theoretical case of a single source by Nebenzal and Fishbain (2017). Here, this methodology is extended to a source term with multiple sources having different attributes. There are no constraints on the type of the dispersion model, such that any dispersion model regardless of complexity can be used, i.e., this methodology is *model invariant*.

2. Methodology

2.1. Notation

The following notation is used for formal description of the problem and is briefly defined here. Let Ωbe the research area. Let $\{S\}$be a set of sources in Ω, where each $s \in \{S\}$, is located in $\omega_s \in \Omega$, and its emission rate is q_s. The number of sources is $|S|$.Let $\{R\}$ be the set of receptors (sensors) located in Ω, where each $r \in \{R\}$is located at $\omega_r \in \Omega$and the pollution level measured by $r$is denoted by c_r. The number of sensors is $|R|$. Let m_{rs}be the pollution transfer function of the dispersion model, which associates sensor r's readings with the emissions of source s:

$$c_r = m_{rs} \cdot q_s \tag{1}$$

For a multiple sources scenario, each sensor's reading consists of the contribution of all sources, i.e.:

$$c_r = \sum_{s\in\{S\}} m_{rs} \cdot q_s \tag{2}$$

For the set $\{R\}$, the sources' contribution can be formulated as a matrix formulation:

$$\vec{c} = M\vec{q}^{\,t} \tag{3}$$

where $\vec{c}$is the row measurement vector, M is the transfer matrix consisting of m_{rs}, and $\vec{q}^{\,t}$is the emission column vector. The values of $m_{rs}$for each source-sensor combination are determined by the dispersion model. Note that in order to obtain accurate estimation, one must include in Ωall sources that might affect c_r.

Let Ebe the inverse operation of M; i.e., the inverse dispersion transfer function (**backward computation**). For each c_r, Eprovides the corresponding emission rate of $q_s$for source s. For a single source case:

$$q_s = e_{sr} \cdot c_r \tag{4}$$

And for multiple sources case, the matrix formulation of Eq. (4) is:

$$\vec{q} = E\vec{c}^{\,t} \tag{5}$$

Finding E is an ill-posed problem (Kabanikhin, 2008), since the number of variables is significantly larger than the available measurements.

Regardless, if $\vec{q}$is obtained, then one can apply a **forward computation** of the dispersion model and determine the ambient pollution level, $c_\omega$for all $\omega \in \Omega$:

$$c_\omega = \sum_{s\in\{S\}} m_{\omega s} \cdot q_s \tag{6}$$

Using the above notation, the interpolation scheme is detailed below. Section 2.2 presents the theoretical case of a single source detection as in Nebenzal and Fishbain (2017). Section 2.3 describes the extension of this algorithm to a more realistic scenario with multiple sources.

2.2. Single source interpolation

The algorithm consists of two phases. In the first stage, using the sensor measurements, $\vec{c}$and the inverse dispersion transfer function, E, the source's location and emission rate are obtained. For a simple case with a single source, this is done by computing the estimated emission rate, $q_\omega$as described in Eq. (4) for all possible source locations $\omega_s \in \Omega$. This procedure is carried out separately for each sensor $r \in \{R\}$, resulting in $|R|$estimated emission values for all locations $\omega \in \Omega$.

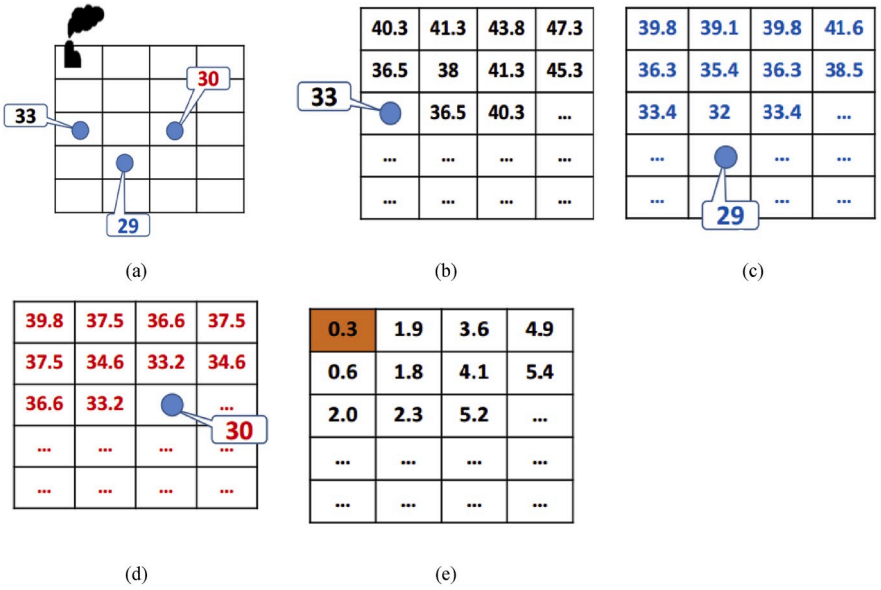

Fig. 1. Qualitative (not in scale) illustration of single source detection, grid origin is at the bottom left corner. (a) The domain **Ω**, with one source, located at (1,5) and 3 sensors. (b–d) estimated emission rate of $s_1$over **Ω**as reflected from sensors 1–3 respectively. (e) The STD of each cell, according to the estimations from $\{\boldsymbol{R}\}$, the point (1,5) has the smallest STD.

Assuming the dense pollution maps are a collection of isolines, the estimated emission rate values based on all sensor's readings should agree in one grid location (Ballard, 1991). For a single source scenario, the location $\omega^* \in \Omega$with the highest agreement among all the sensors is said to be the source's location. Agreement here means that all sensors assess the emission rate, q_ω, of the source at location ω^*as having roughly the same value. Using the standard deviation (STD) of the estimates as an agreement measure, the location with the lowest standard deviation is the approximate location of the source:

$$\boldsymbol{\omega}^* = \min_{\omega \in \Omega} \boldsymbol{std}\left\{\left[\boldsymbol{q}_\omega^{r=1}, \boldsymbol{q}_\omega^2, \ldots, \boldsymbol{q}_\omega^{|R|}\right]\right\} \tag{7}$$

Once $\omega^*$is found, the emission rate is evaluated by the average estimates:

$$\widehat{\boldsymbol{q}}_{\omega^*} = \frac{1}{|\boldsymbol{R}|}\sum_{r=1}^{|R|} \boldsymbol{q}_{\omega^*}^r \;\Big|\; \forall \boldsymbol{r} \in \{\boldsymbol{R}\} \tag{8}$$

Fig. 1 illustrates this process, where, for simplicity's sake, Ωis divided into a 2-D regular grid, under the assumption that each grid cell is small enough so the pollution level, all over the cell, is uniform. For the illustrations in Figs. 1 and 2, the grid size is 20 × 20m. An important step in future implementation of this method, is to assess the grid size according to the problem specific conditions and requirements. In this example Ωcontains a single source, s_1, and three sensors, $R = \{r_1, r_2, r_3\}$ as depicted in Fig. 1a. r_1, indicates a pollution level of 33 $\mu g/m^3$(i.e. $c_1 =$ 33) and is located at grid cell (1,3); r_2, is located at (2,2), measures 29 $\mu g/m^3$; and r_3, at (3,3) measures a level of 30 $\mu g/m^3$.

Next, we plug in Eq.(4) to estimate the source's emission rate over Ωgiven only r_1(**backward computation**). Since each grid cell presents a uniform pollution level, the estimated emission value, q_ω, is the same for the entire cell. Fig. 1b illustrates the process for a simple exponential decay dispersion model. A source in cell (2,4), would have yielded an estimated emission rate, $q_{2,4}$, based on Sensor r_1's measurement, of 38. If the source had been located at (4,5), then $q_{4,5}$, based on Sensor 1, would have been 47.3 $\mu g/m^3$. Fig. 1c and 1d are the estimation grids generated in the same way as Fig. 1b, for Sensor $r_2$and Sensor $r_3$respectively.

To evaluate the agreement between sensors, for each grid cell, we compute the standard deviation of the estimates of the three sensors. The lower the STD, the higher the agreement, as described by Eq. (7). This is illustrated in Fig. 1e. The smallest STD is for location (1, 5), where, in this example, the source is located.

Once $\omega^*$is obtained, we can evaluate $\widehat{q}_{\omega^*}$by plugging in Eq. (8), which is 39.9. Now, having the estimated source's location, $\omega^*$and its corresponding emission rate, q_{ω^*}, we can estimate the dense pollution map over all Ω(**forward computation**) using the dispersion transfer function in Eq. (6).

2.3. Multiple source interpolation

Given the single source formulation, we next present the extension of the methodology to a scenario with multiple sources, where the number of sources, $|S|$, as well as sources' emission rates, $q_s$for all $s \in \{S\}$, are unknown. In this formulation, the problem is ill-posed (Kabanikhin, 2008). One approach to cope with this challenge is to solve it by the enumeration of the solution space; i.e., by placing the sources in all possible locations with all possible emission rates. Then, for each of the configurations, Eq. (3) is computed to get the model's estimation of the sensors' readings; i.e., $c_r$for all $r \in R$at the sensors' locations, $\omega_r \in \Omega$. These are then compared against the actual readings. The configuration with the minimum discrepancy is the one selected. This process, however, results in $\frac{|\Omega|!}{(\Omega-|S|)!} \cdot |Q|^{|S|}$different source configurations. For a small $|\Omega|$and small $|S|$, the entire solution space can be enumerated, but as $|\Omega|$, and $|S|$increase, the enumeration of all possible configurations becomes infeasible.

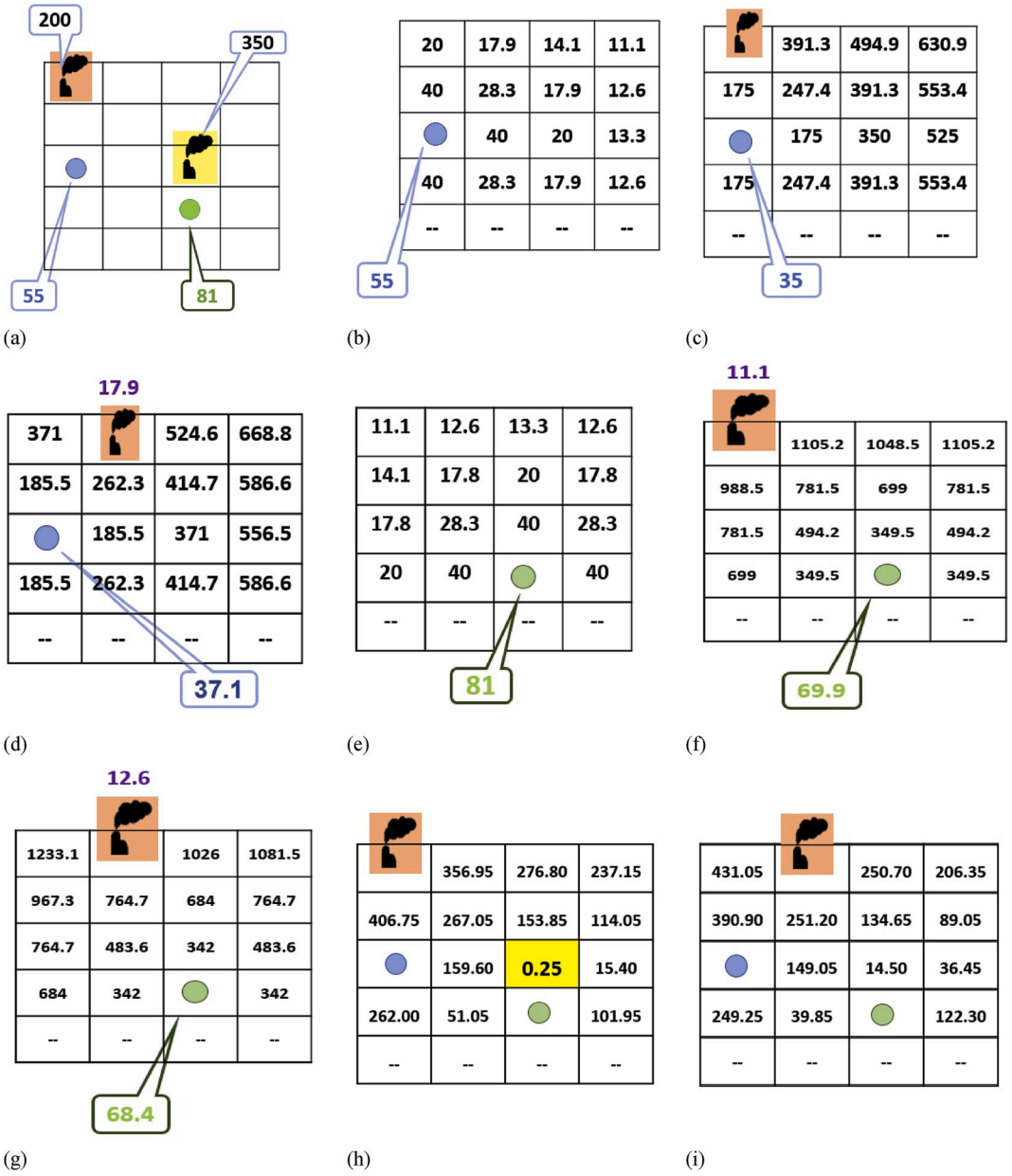

Fig. 2. Illustration of multiple source detection. (a) The domain **Ω**, with two sources, and 2 sensors. (b) Estimated relative contribution of s_1 to the pollution data measured by sensor 1. (c–d) Estimated emission rate of the second (last) source s_2, when s_1 is located at (1,5) and (2,5) respectively. (e) Estimated relative contribution of s_1 to the pollution data measured by sensor 2. (f–g) Estimated emission rate of the second (last) source s_2, when s_1 is located at (1,5) and (2,5) respectively.(h–i) The STD of estimated emission rate of s_2, given s_1 located at (1,5) and (2,5) respectively. Overall, the lowest STD, 0.25 is when s_1 is located at (1,5) and s_2 located at (3,3), (highlighted at (h)).

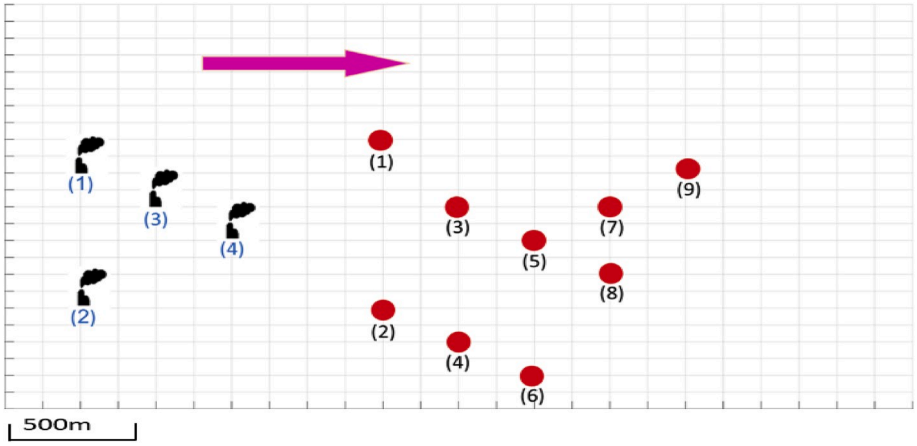

Fig. 3. The 12 km^2 study area. The four sources are marked with a black factory icon, and numbered in blue. The 9 sensors are marked by red circles and numbered in black. The wind direction is 270° and is represented by the pink arrow. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Ambient data measured by the sensors (units are in μg/m^3) for the GAPDM model.

Sensor #	Value	Sensor #	Value
(1)	8.01	(6)	0.39
(2)	165.03	(7)	85.13
(3)	119.02	(8)	34.36
(4)	16.07	(9)	60.23
(5)	134.56		

Using the single source interpolation scheme above, and by limiting the upper and lower bounds of the possible emission rate for each source, the solution space can be significantly reduced. Further decrease in computation time can be made if some of the sources parameters are known in advance and only few of them are unknown. Such an approach is very effective for regulators who may possess some knowledge regarding the interrogated area. To do so, the estimated pollution rate of $|S|-1$sources is derived from a Pollutant Release and Transfer Registers (PRTRs) regulatory reporting system (Kerret and Gray, 2007; Sullivan and Gouldson, 2007; Ayalon et al., 2015). If no PRTR records are available in the target region, one can still derive the upper and lower bounds on the emission rates of each source based on the sensors' measurements and the dispersion model in use. This can be done through nested enumeration; in other words, by making an initial run to estimate the source set, and then performing a fine-tuning phase to obtain more accurate sets. Regardless of the emission rate estimates, the source locations, $\omega_s \in \Omega$, are unknown.

Let the set $\{S^{\dagger}\}$be a set of $\{S-1\}$sources, out of the $|S|$ sources in Ω. As mentioned above, the emission rate of these sources is estimated using PRTR or by an enumeration process. For $\{S^{\dagger}\}$, the contribution of the last source $s \in S \backslash S^{\dagger}$to sensor r is given by:

$$c_r^{residual} = c_r - \sum_{s \in S^{\dagger}} m_{rs} \cdot q_s \tag{9}$$

The residual is computed for all possible source locations in Ω. After having found the residual, the location and emission of $s \in S$can be computed as described above in section 2.2. This process is illustrated by the simple configuration in Fig. 2a with two sources and two sensors. The source $s_1$is located at (1,5) and $s_2$at (3,3). The emission rates of $q_1$and $q_2$are 720 kg/h and 1260 kg/h respectively. Assuming a simple decay dispersion model; r_1, (located at (1,3)) measures 55 $\mu g/m^3$and r_2, (at (3,2)) records 81 $\mu g/m^3$. Now, because both $\omega_s$and $q_s$are unknown for all $\{S\}$, PRTR or the enumeration process must be used to evaluate the contribution of $\{S^{\dagger}\}$to $c_r,\ r \in \{R\}$. In this illustration we assume two sources (by the PRTR/enumeration process): $s_1$with a known emission rate of 720 kg/h and $s_2$with an unknown emission data, both sources locations are unknown in this example. Fig. 2b depicts the contribution of $s_1$to r_1, if it had been located in each of the grid cells. For example, if $s_1$had been located in (1,5), its contribution to $r_1$would have been 20 $\mu g/m^3$. Had it been located at (2,5), its contribution would have been 17.9, and so on.

For the case, where $s_1$is located at (1,5) and its contribution to $r_1$is 20$\mu g/m^3$, the residual, $c_{r=1}^{residual}$is 35. Once 35 $\mu g/m^3$has been deduced, the single source methodology; namely, computing the estimated source's emission rate for all $\omega_s \in \Omega$applying Eq. (4) (backward computation) is executed. This is shown in Fig. 2c and d for $s_1$located in (1,5) and (2,5) respectively. The same process is repeated for all $r \in R$. In this example, the process is repeated for $r_2$and is described in Fig. 2e through g, which corresponds to Fig. 2b through d.

Like the single source estimation procedure, the standard deviation of the source estimations, considering all sensor and source locations, are computed. This is presented in Fig. 2h and i, where the standard deviation for the source estimation was found for $s_1$located at (1,5) and at (2,5) respectively. The lowest standard deviation was found for location (3,3) in Fig. 2h, which corresponds to the case, of $s_1$situated at (1,5). Thus, given only the estimated emission rate of s_1, the algorithm can successfully identify the locations of both $s_1$and s_2.

Once the source locations have been determined, the estimated emission rate of $s_2$can be deduced from Eq. (8) (forward computation). In this example, $q_2 \cong 349.7$, where the true value is 350 $\mu g/m^3$. This process is also described in the pseudo-code provided in Alg. 1.

Alg. 1. A pseudo code depicting the major process involved in this search procedure. Main process are in bold font.

Stage 1: calculate partial contribution of each source $s_i \in S^{\dagger}$ to each sensor at R:

Input:

- Set of sources, $s_{1\ldots n-1}$ with unknown location and known emission rate notated as $S^{\dagger}$;
- Set of sensors $r_{1\ldots m} \in R$

Output:

- set of $c_r^{residual}$, the contribution of s_n to each $r_i \in R$.

1. FOR s_i in $S^{\dagger}$:
2. FOR each cell in the grid:
3. FOR r_i in R:
4. $c_{r_i}^{residual} \leftarrow$ The partial contribution of s_i to r_i

// e.g. if s_1 is located at point (1,5) its partial contribution to the measurement of r_1 is 20ppm,while if its location is (2,5) the contribution is 17.9ppm (Fig. 2c).

Stage 2: calculate the estimated emission rate of s_n:

Input:

- The output of the previous stage
- The last source s_n with unknown emission rate and location;

Output:

- q_s, the estimated emission rate of s_n at each cell in the grid

5. FOR each point in the grid:
6. FOR r_i in R:
7. $q_s \leftarrow$ The estimated emission rate of s_n based on r_j

// e.g. based only on the measurement of r_1, the estimated emission rate of s_2, as s_1 is located at (1,5) is 391.3 and 494.9 as s_1 location is at (2,5), (Fig. 2c), while the same estimation based on the measurements of r_2 yields to 1105.2 and 1048.5 respectively (Fig. 2f).

Stage 3: calculate the standard deviation the estimated emission rate of s_n

Input:

- The output of the previous stage

Output:

- The standard deviation of s_n at each cell in the grid.

8. FOR each cell in the grid
9. STD $s_n \leftarrow$ $std(\, q_s \; from \; R)$

// e.g. As s_1 is located at (1,5) and s_2 at (2,5), the s_2 estimated emission rate based on $r_1$is 391.3 (Fig. 2c), and r_2 is 1105.2 (Fig. 2f), so its std is 356.95 (Fig. 2h).

2.4. Simulation study

The simulated region of interest (ROI), Ω, was a 12 km^2 area, with 4 sources and nine sensors, as depicted in Fig. 3.

Similarly to the analysis of Nebenzal and Fishbain (2017), the ambient pollution level, c_r, in a Cartesian coordinate system, $r = [x, y]$, generated by a source, $s \in S$, over Ω, was simulated by the GAPDM (Ermak, 1977):

$$c_r\left(x,y,z\right)=\frac{q_s}{2\pi\sigma_y\sigma_z\bar{u}}\exp\left(-\frac{y^2}{2\sigma_y^2}\right)\cdot\left[\exp\left(-\frac{(z-H)^2}{2\sigma_z^2}\right)+\exp\left(-\frac{(z+H)^2}{2\sigma_z^2}\right)\right] \tag{10}$$

In this formula, x is coaligned with the wind direction and y is the crosswind direction. z is the vertical distance from the source; $\bar{u}$is the time-averaged wind speed at the height of release H; and $\sigma_y$and $\sigma_z$represent the standard deviations of the crosswind and the vertical Gaussian distribution of the pollutant concentration, respectively. The model also assumes full reflection from the ground.

The ambient data measured by the sensors were captured by a single measurement acquired at a given point in time. The values are derived from the GAPDM model and are summarized in Table 1. In this example, emission rate of sources number 1–3 are known – 25.2, 28.8 and 18 kg/h, but their locations are unknown. The fourth source location and emission rate (21.6 kg/h) are unknown. The results show exact estimation of all four sources locations. The computed emission rate of the fourth source is underestimated by less than 5% from the set value.

The suggested method can handle, in principle, any number of sources by setting a source in every grid cell with a lower bound for possible emission of 0 kg/h. Since run time scales as $|S|^{2\cdot|S|}$($|S|$ being the number of sources), the problem becomes intractable quickly as the number of sources increases. This can be handled by parallel computing or the use of heuristics, which are the common practice is such situations (Lee and El-Sharkawi, 2007; Altinel et al., 2008; Burke et al., 2013; Fishbain et al., 2013; Shanmugasundaram et al., 2014). Such modification is beyond the scope of this work which focusses on the proof-of-concept.

3. Results and discussion

After setting up the problem and the search method, the accuracy of the pollution dense map computed in this method is compared to the classical methods.

Simply by using the data collected from the sensors, a concentration field over Ωwas generated using four methods: IDW, Ordinary Kriging, NN and the proposed source detection methodology. The results are displayed in Fig. 4.

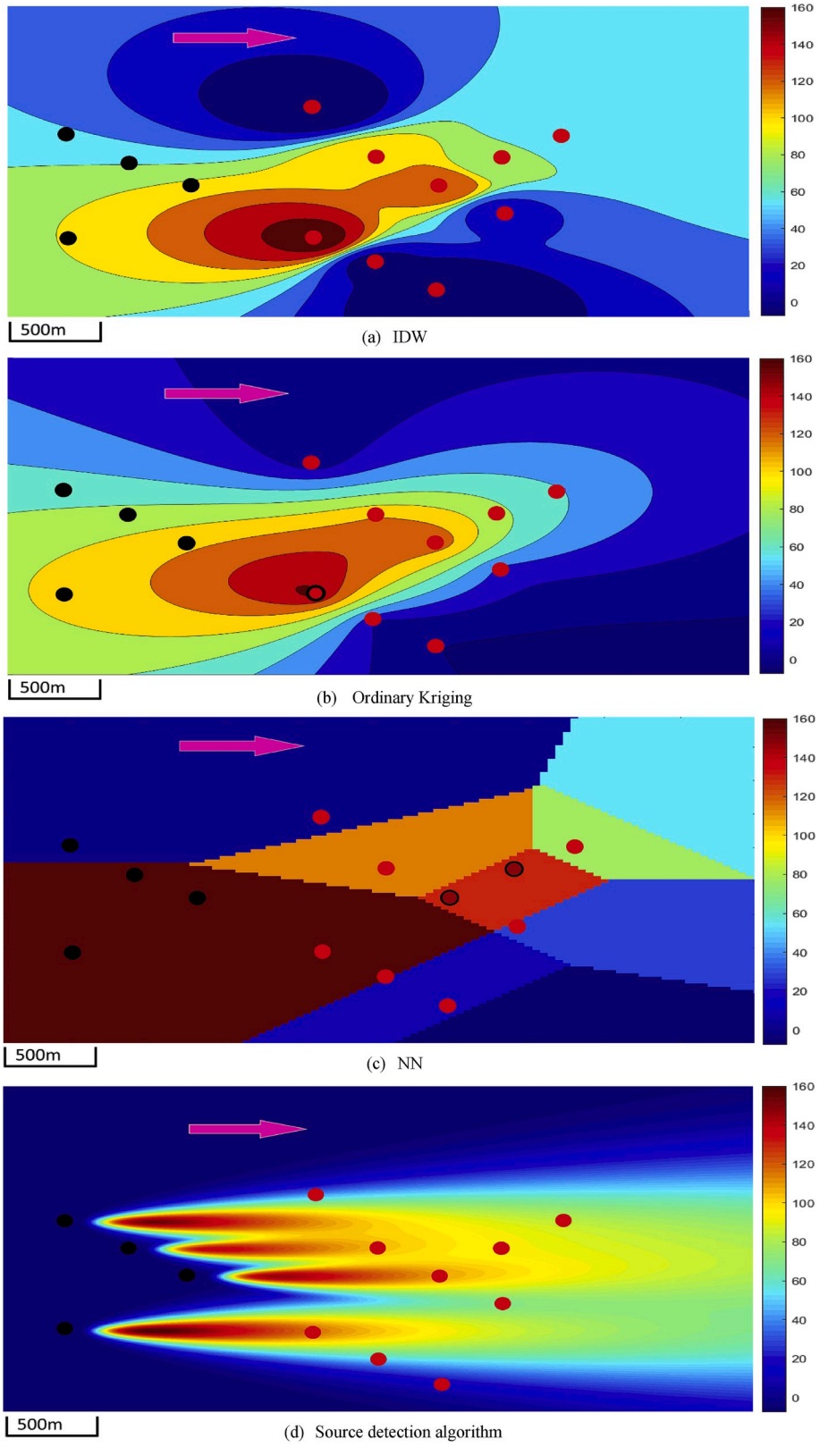

Fig. 4. Spatial maps based on the GAPDM model sensor measurements. (a) IDW (b) Ordinary Kriging (c) NN (d) Source detection algorithm. Black dots indicate sources and red dots are sensors, the pink arrow represents the wind direction. The pollution level is represented on a blue (low)- to red (high) color scale. The computed sensors' reading are presented in Table 1 above are in excellent agreement (errors less than 5%) with the expected value. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

One can easily observe that IDW (Fig. 4(a)), Ordinary Kriging (b) generate spatial maps whose maximum is obtained near sensors #2 and #5, the two highest measuring points. Nearest Neighbor, NN (c), while is commonly used (Li and Heap, 2014; Wiemann et al., 2016), is completely off. Furthermore, all three methods represent an isotropic dispersion, radial-like, formation near each sensor, although this is not likely to have been the case given the wind and the decay of pollution. According to these maps, pollution is assumed to be located upwind from the sources, which is clearly not possible. This is expected, since they do not consider the atmospheric conditions or physicochemical

features of the area(Nebenzal and Fishbain, 2017). Fig. 4(d) depicts the pollution field derived from the sensor readings using the new method presented here. Clearly it detects the source locations, emission rates, and reconstructs the pollution field over Ωaccurately.

The results presented in Fig. 4(d), show that the suggested method is accurate. Computed sources' locations match the actual locations and the computed emissions rates are accurate (under-estimation less than 5%). This results in a highly accurate prediction of the sensors' readings and the pollution dense map. Such accuracy is certainly superior compared to LUR and interpolation method, as also presented in Fig. 4. However, we note that real-life scenarios are considerably more complicated. For example, sources characteristic, dispersion models, and sensors' attributes will be a significant factor in the capability to provide such an accurate estimation of the sources' locations and emission rates. We anticipate that uncertainties regarding the dispersion phenomena, including sources properties, such as deposition and chemical reactions, the effect of the terrain topography, will reduce the estimation accuracy. Additionally, sensors' attributes, such as the minimum detectable level, will have to be considered. This real-life complexity is a part of our ongoing study in which the suggested method is adapted to account for these important issues.

4. Conclusion

This paper introduces a methodology for estimating a complex source term with different attributes to generate accurate dense pollution maps from sparse sensing. Unlike popular interpolation schemes such as LUR, IDW or Ordinary Kriging that do not consider dispersion phenomena explicitly, the method presented here incorporates a dispersion model into the process which results in a more accurate and exploitable dense pollution field from sparse sensor networks.

The method, described in this manuscript, needs an input regarding pollution concentrations in various locations at all times. This is achieved by using a dispersion model that calculates this input using climatic conditions and emission rates. The dispersion model, GAPDM, applied here serves as a proof of concept. The GAPDM is simple and requires minimal computational resources, which makes it attractive for real-time risk assessment. Having said that, it is important to note that the suggested methodology allows for the use of any dispersion model.

Future work will incorporate more advanced models, (e.g. (Agirre--Basurko et al., 2006; Sousa et al., 2007; Hill et al., 2011; Li and Heap, 2014; Reis et al., 2015; Lauret et al., 2016), into the scheme. Such sophisticated dispersion models take fine details regarding the source term into account such as exhaust velocity, temperature, deposition rate, and terrain effect on the flow field. The incorporation of these considerations into our computational scheme should generate accurate results even for a complex terrain. This method is also applicable for leak detection. In this case the known sources are all accounted for, and the marginal contributions are then used for allocating unknown sources, i.e., the leaks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the Israeli Ministry of Science and Technology Research Program.

References

Agirre-Basurko, E., Ibarra-Berastegi, G., Madariaga, I., 2006. Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area. In: Environmental Modelling and Software. Elsevier, pp. 430–446. https://doi.org/10.1016/j.envsoft.2004.07.008.

Altinel, I.K., et al., 2008. Binary integer programming formulation and heuristics for differentiated coverage in heterogeneous sensor networks. Comput. Network. 52 (12), 2419–2431. https://doi.org/10.1016/j.comnet.2008.05.002.

Ayalon, O., Lev-On, M., Lev-On, P.P., 2015. Greenhouse gas emission mitigation plan for the State of Israel: strategies, incentives and reporting. Clim. Pol. 15 (6), 784–800. https://doi.org/10.1080/14693062.2014.968763. Taylor & Francis.

Ballard, D.H., 1991. Generalizing the Hough transform to detect arbitrary shapes. Pattern Recogn. 13 (2), 183–194.

Beauchamp, M., de Fouquet, Malherbe, L., 2017. Dealing with non-stationarity through explanatory variables in kriging-based air quality maps. Spatial Statistics 22, 18–46. https://doi.org/10.1016/j.spasta.2017.08.003.

Broday, D.M., Yuval, 2010. 'Studying the time scale dependence of environmental variables predictability using fractal analysis', *environmental science & Technology*. American Chemical Society 44 (12), 4629–4634. https://doi.org/10.1021/es903495q.

Burke, E.K., et al., 2013. Hyper-heuristics: a survey of the state of the art. J. Oper. Res. Soc. 64 (12), 1695–1724. https://doi.org/10.1057/jors.2013.71.

Buteau, S., et al., 2017. Comparison of spatiotemporal prediction models of daily exposure of individuals to ambient nitrogen dioxide and ozone in Montreal, Canada. Environ. Res. 156 (March), 201–230. https://doi.org/10.1016/j.envres.2017.03.017. Elsevier Inc.

Carruthers, D.J., et al., 1994. UK-ADMS: a new approach to modelling dispersion in the earth's atmospheric boundary layer. J. Wind Eng. Ind. Aerod. 139–153. https://doi.org/10.1016/0167-6105(94)90044-2.

Castell, N., et al., 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environ. Int. 99, 293–302. https://doi.org/10.1016/j.envint.2016.12.007. Pergamon.

CDC, 2018. Air Quality Measures on the National Environmental Health Tracking Network.

Cimorelli, Alan J., Perry, Steven G., Venkatram, Akula, Weil, Jeffrey C., Paine, Robert J., Wilson, Robert B., Lee, Russell F., Peters, Warren D., Brode, Roger W., 2005. AERMOD: A dispersion model for industrial source applications. Part I: General model formulation and boundary layer characterization. Journal of applied meteorology 44 (5), 682–693.

EPA, 2019. Air Data: Air Quality Data Collected at Outdoor Monitors across the US. EPA.

Ermak, D.L., 1977. An analytical model for air pollutant transport and deposition from a point source, 1967 Atmos. Environ. 11 (3), 231–237. https://doi.org/10.1016/0004-6981(77)90140-8. Elsevier.

Fishbain, B., Hochbaum, D.S., Yang, Y.T., 2013. Real-time robust target tracking in videos via graph-cuts. In: Proceedings of SPIE - the International Society for Optical Engineering, pp. 1–19. https://doi.org/10.1117/12.2002947.

Hill, D.J., et al., 2011. A virtual sensor system for user-generated, real-time environmental data products. Environ. Model. Software 26 (12), 1710–1724. https://doi.org/10.1016/j.envsoft.2011.09.001. Elsevier.

Kabanikhin, S.I., 2008. Definitions and examples of inverse and ill-posed problems Survey paper. J. Inv. Ill-Posed Problems 16, 317–357. https://doi.org/10.1515/JIIP.2008.069.

Kerret, D., Gray, G.M., 2007. What do we learn from emissions reporting? Analytical considerations and comparison of pollutant release and transfer registers in the United States, Canada, england, and Australia. Risk Anal. 27 (1), 203–223. https://doi.org/10.1111/j.1539-6924.2006.00870.x. John Wiley & Sons, Ltd (10.1111).

Kumar, P., et al., 2015. The rise of low-cost sensing for managing air pollution in cities. Environ. Int. 75, 199–205. https://doi.org/10.1016/J.ENVINT.2014.11.019. Pergamon.

Lauret, P., et al., 2016. Atmospheric dispersion modeling using Artificial Neural Network based cellular automata. Environ. Model. Software 85, 56–69. https://doi.org/10.1016/J.ENVSOFT.2016.08.001. Elsevier.

Lee, K.Y., El-Sharkawi, M.A., 2007. Modern Heuristic Optimization Techniques: Theory and Applications to Power Systems, Modern Heuristic Optimization Techniques: Theory and Applications to Power Systems. https://doi.org/10.1002/9780470225868.

Leelőssy, Á., et al., 2014. Dispersion modeling of air pollutants in the atmosphere: a review. Open Geosci. 6 (3) https://doi.org/10.2478/s13533-012-0188-6.

Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: a review. Environ. Model. Software 53, 173–189. https://doi.org/10.1016/j.envsoft.2013.12.008. Elsevier Ltd.

Masey, N., Hamilton, S., Beverland, I.J., 2018. Development and evaluation of the RapidAir® dispersion model, including the use of geospatial surrogates to represent street canyon effects. Environ. Model. Software 108, 253–263. https://doi.org/10.1016/J.ENVSOFT.2018.05.014. Elsevier.

Morley, D.W., Gulliver, J., 2018. A land use regression variable generation, modelling and prediction tool for air pollution exposure assessment. Environ. Model. Software 105, 17–23. https://doi.org/10.1016/J.ENVSOFT.2018.03.030. Elsevier.

Nebenzal, A., Fishbain, B., 2017. Hough-Transform-Based Interpolation Scheme for Generating Accurate Dense Spatial Maps of Air Pollutants from Sparse Sensing, pp. 51–60. https://doi.org/10.1007/978-3-319-89935-0_5.

Reis, S., et al., 2015. Integrating modelling and smart sensors for environmental and human health. Environ. Model. Software 74, 238–246. https://doi.org/10.1016/j.envsoft.2015.06.003. Elsevier.

Sacks, J.D., et al., 2018. The Environmental Benefits Mapping and Analysis Program – community Edition (BenMAP–CE): a tool to estimate the health and economic benefits of reducing air pollution. Environ. Model. Software 104 (2), 118–129. https://doi.org/10.1016/j.envsoft.2018.02.009.

Scire, J.S., et al., 2000. A User's Guide for the CALPUFF Dispersion Model. Earth Tech, Inc. Concord, MA.

Shanmugasundaram, S.K., Prasad, R., Fear, J., 2014. Optimization of complex water supply network. Procedia Engineering 70, 1524–1530. https://doi.org/10.1016/J.PROENG.2014.02.168. Elsevier.

Sousa, S., et al., 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ. Model. Software 22 (1), 97–103. https://doi.org/10.1016/j.envsoft.2005.12.002. Elsevier.

Sullivan, R., Gouldson, A., 2007. Pollutant release and transfer registers: examining the value of government-led reporting on corporate environmental performance. Corp. Soc. Responsib. Environ. Manag. 14 (5), 263–273. https://doi.org/10.1002/csr.148. John Wiley & Sons, Ltd.

WHO | Ambient (outdoor) air quality and health, 2019. WHO | Ambient (Outdoor) Air Quality and Health. Who.

Wiemann, S., et al., 2016. Environmental Modelling & Software Design and prototype of an interoperable online air quality information system. Environ. Model. Software 79, 354–366. https://doi.org/10.1016/j.envsoft.2015.10.028. Elsevier Ltd.

Wu, C. Da, Zeng, Y.T., Lung, S.C.C., 2018. A hybrid kriging/land-use regression model to assess PM2.5 spatial-temporal variability. Sci. Total Environ. 645, 1456–1464. https://doi.org/10.1016/j.scitotenv.2018.07.073. Elsevier B.V.