Finding candidate locations for aerosol pollution monitoring at street level using a data-driven methodology

Author(s):
Moosavi, Vahid; Aschwanden, Gideon; Velasco, Erik

Publication date:
2015-09

Permanent link:
https://doi.org/10.3929/ethz-b-000106827

Originally published in:
Atmospheric Measurement Techniques 8(9), https://doi.org/10.5194/amt-8-3563-2015

Atmos. Meas. Tech., 8, 3563–3575, 2015
www.atmos-meas-tech.net/8/3563/2015/
doi:10.5194/amt-8-3563-2015

Atmospheric Measurement Techniques
Open Access

Finding candidate locations for aerosol pollution monitoring at street level using a data-driven methodology

V. Moosavi[1,2], G. Aschwanden[2,4], and E. Velasco[3]

[1]Chair for Computer Aided Architectural Design (CAAD), ETH Zurich, 8092 Zurich, Switzerland
[2]Future Cities Laboratory, ETH Zurich, 8092 Zurich, Switzerland
[3]Singapore-MIT Alliance for Research and Technology (SMART), Center for Environmental Sensing and Modeling (CENSAM), Singapore
[4]Faculty of Architecture, Building and Planning, The University of Melbourne, 3010 Melbourne, Australia

Correspondence to: V. Moosavi (svm@arch.ethz.ch)

Received: 14 October 2014 – Published in Atmos. Meas. Tech. Discuss.: 26 March 2015
Revised: 1 July 2015 – Accepted: 23 August 2015 – Published: 3 September 2015

Abstract. Finding the number and best locations of fixed air quality monitoring stations at street level is challenging because of the complexity of the urban environment and the large number of factors affecting the pollutants concentration. Data sets of such urban parameters as land use, building morphology and street geometry in high-resolution grid cells in combination with direct measurements of airborne pollutants at high frequency (1–10 s) along a reasonable number of streets can be used to interpolate concentration of pollutants in a whole gridded domain and determine the optimum number of monitoring sites and best locations for a network of fixed monitors at ground level. In this context, a data-driven modeling methodology is developed based on the application of Self-Organizing Map (SOM) to approximate the nonlinear relations between urban parameters (80 in this work) and aerosol pollution data, such as mass and number concentrations measured along streets of a commercial/residential neighborhood of Singapore. Cross-validations between measured and predicted aerosol concentrations based on the urban parameters at each individual grid cell showed satisfying results. This proof of concept study showed that the selected urban parameters proved to be an appropriate indirect measure of aerosol concentrations within the studied area. The potential locations for fixed air quality monitors are identified through clustering of areas (i.e., group of cells) with similar urban patterns. The typological center of each cluster corresponds to the most representative cell for all other cells in the cluster. In the studied neighborhood four different clusters were identified and for each cluster potential sites for air quality monitoring at ground level are identified.

1 Introduction

Air quality monitoring is needed to guide regulations for public health protection (Craig et al., 2008). At city scale it is performed through networks of monitoring stations covering large geographic areas (i.e., 2–25 km^2). The monitoring stations are placed above the urban canopy layer, where pollution measurements are not directly impacted by local emissions or obstructed wind flow (i.e., adjacent buildings). Usually rooftops provide adequate monitoring locations. They provide information about the average urban ambient pollution at district scale to which the general population is exposed, which is used for both regulatory and advisory purposes (Hidy and Pennell, 2010). However, the ambient pollutants reported by these stations do not count for the spatial variability at microscale (e.g., Moore et al., 2009; Salimi et al., 2013) and do not always represent the pollution to which people are exposed during their daily activities (Nerriere et al., 2005). Significant variations can be expected at ground level, even between sites within close proximity. The highest outdoor exposure to pollutants for many dwellers occurs while commuting or carrying out activities in proximity to emission sources (e.g., walking along busy streets). Figure 1 contrasts the difference between concentrations of particles measured along streets and at a site over the urban canopy

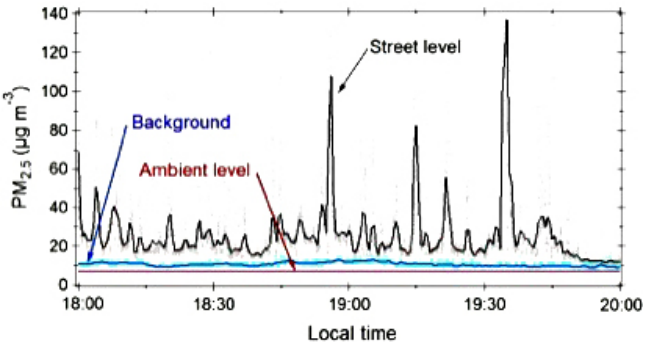

Figure 1. Time series of $PM_{2.5}$ mass concentration measured above the urban canopy (background) and along the streets of the commercial/residential neighborhood of Rochor, Singapore investigated in this work, and the hourly 24 h average concentrations reported by the local environmental agency (ambient level) on 10 July 2013 during the evening rush hour.

layer of the neighborhood of Singapore used as a case study in this work.

Despite the stark difference between ambient and ground-level pollution concentrations, monitoring networks include only a few ground level monitoring stations with the purpose of characterizing traffic emissions rather than for policy advisory. This is usually the case everywhere and not only of Singapore. The deployment of comprehensive monitoring networks at ground level is hampered by the large number of monitors and associated costs (i.e., equipment, operation and maintenance) needed to represent the urban heterogeneity in terms of land use, buildings morphology and distribution of emission. To overcome this limitation and expand existing air quality monitoring networks a new method is proposed to determine the minimum number of stations at ground level and their best potential locations.

Modeling techniques, such as Computational Fluid Dynamics (CFD) and Large-Eddy Simulation (LES) have been used to investigate the dispersion and distribution of pollutants under the urban canopy (e.g., Li et al., 2006; Tominaga and Stathopoulos, 2013). However, the complexity of the urban structure has limited their application to simplified geometries (i.e., urban morphologies), idealized atmospheric conditions and particular distributions of emission sources (e.g., Li et al., 2012; Tominaga and Stathopoulos, 2013).

With recent advancements in computational and sensing technologies, data-driven approaches, also known as inverse or empirical modeling are an alternative to solve the problem of modeling in complex systems (Kolehmainen et al., 2001; Voukantsis et al., 2010), such as those imposed by the urban heterogeneity on the distribution of air pollutants at street level. The basic idea under these models is that if there are underlying rules controlling a system, they can be found from a set of data by means of statistical and probabilistic methods. Therefore, with a statistically reasonable amount of air pollution observations and data on urban parameters, a data-driven mathematical model can be constructed to interpolate the pollutants concentration to a whole gridded domain with an acceptable level of accuracy, without a descriptive theory of the real phenomena in advance.

Considering the number of potential urban parameters controlling the pollution distribution at ground level, the modeling challenge turns into the identification of the non-linear functional relations between the urban parameters and concentration of atmospheric pollutants. This view inverts the problem of modeling from a deductive and theory-grounded approach to an inductive and data-driven approach as it is similarly described in Inverse Problem Theory (Tarantola, 2005).

The application of Self Organizing Map (SOM) as a data-driven modeling approach is used to find the association between particulate matter concentration at ground level and urban parameters in its vicinity. The model (trained SOM) is then applied to approximate the concentration of pollutants in a whole gridded domain based on the urban parameters of each particular cell. The resulting maps showing the spatial distribution of concentration of pollutants are expected to provide valuable information for epidemiological and risk assessments, as well as to identify hot spots of pollution.

The trained SOM is also used in combination with a clustering algorithm to determine the number of similar domains in the area, representing the optimum number of monitoring stations to cover the different urban patterns within the studied domain. The center of each cluster is the best potential location in terms of representativeness of the urban parameters.

The proposed data-driven model is tested using a data set of over 80 urban parameters and high frequency (1 or 10 s) measurements of aerosol pollution along a reasonable number of streets in a heterogeneous residential/commercial neighborhood of Singapore, selected as a case study. Fine-grained urban parameters spatially distributed in grid cells of 100×100 m include information on street networks, land-use patterns, demographics, vehicular traffic, building and street topology, etc. The aerosol pollution measurements were performed using a set of portable and battery operated sensors. The measured variables were mass concentration of particles with aerodynamic diameters ≤ 10, 2.5 and 1 µm (PM_{10}, $PM_{2.5}$ and PM_1), particle number concentration (PN), active surface area (ASA), and mass concentrations of black carbon (BC) and particle-bound polycyclic aromatic hydrocarbons (pPAHs).

It is important to point out that the study presented here is a proof of concept with the aim of testing SOM. The proposed methodology is not a receptor model. It does not determine any source apportionment. Receptor models utilize chemical measurements to calculate the relative contributions from major sources at specific locations (e.g., Viana et al., 2008).

The article first describes the main features of SOM methodology and its capabilities for multidimensional data

visualization, nonlinear function approximation, and data clustering. Then the urban parameters and aerosol pollution measurements are introduced. The application of SOM to our case is presented in three sections. The first section describes the application of SOM as a nonlinear function approximation method between urban parameters and measured aerosol concentrations. The efficiency of the approximation functions is evaluated through cross-validations between predicted and observed data. The second section explains the application of SOM to interpolate the measured pollution data from selected grid cells to the complete gridded domain. The third section describes the combination of SOM and a clustering algorithm to determine the optimum number of monitoring sites and their best locations in terms of representativeness and information gain. Maps of spatially interpolated aerosol concentrations present the results of the approach based on the SOM proposed here. The candidate locations for monitoring stations for each one of the identified types of urban patterns (i.e., clusters) are indicated in a final map showing also the representativeness of each grid cell within its respective cluster.

2 Methods

This section starts with a brief description of SOM as a data-driven modeling approach. The following section describes the selected neighborhood of Singapore as a study area and provides details of the urban parameters used for the model evaluation. Then, the aerosol pollution measurements are introduced.

2.1 Self Organizing Map

Self Organizing Map is a data-driven modeling method introduced by Kohonen (1982). From a mathematical point of view, SOM acts as a nonlinear data transformation in which data from a high-dimensional space are transformed to a low-dimensional space (usually a space of two or three dimensions), while the topology of the original high dimensional space is preserved. Topology preservation means that if two data points are similar (i.e., close) in the high-dimensional space, they are necessarily close in the new low-dimensional space. This low-dimensional space, which is normally represented by a planar grid with a fixed number of points, is called a map. Each node of this map has its own coordinates $(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$ and a high-dimensional vector $(\boldsymbol{W}_i = \{\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{in}\})$ where the original observed data are n dimensional vectors.

In comparison with other data transformation methods, SOM has the advantage of delivering two-dimensional maps visualizing smoothly changing patterns of data from the original high-dimensional space. In addition, SOM can also be used to predict values of parameters or dimensions using data of each other parameter through nonlinear approximation functions (Barreto and Souza, 2006). In the field of environmental modeling, data-driven methods, such as neural networks (e.g., Multi-Layer Perceptron Learning), Support Vector Machines (SVM) and time series forecasting methods such as the Autoregressive Integrated Moving Average (ARIMA) modeling technique, have been previously applied based on the availability of massive measured data (e.g., Kolehmainen, 2004; Kolehmainen et al., 2001). In a recent study Nguyen et al. (2014) used low-resolution satellite images in combination with SVM to estimate aerosol concentration at ground level from urban surfaces with no need for in situ measurements. However, they were not able to identify the urban parameters' influence on the aerosol concentration. Similarly, Hirtl et al. (2014) used satellite images, ground-based measurements and the support vector regression method to improve air quality forecasts at regional scale.

In summary, SOM is a generic, robust and powerful method that has been employed in several application domains (Kohonnen, 2013). It can be used for visualization of high-dimensional data and data exploration (Kolehmainen, 2004), state space modeling and clustering (Bieringer, 2013) and most importantly, as a nonlinear function approximation method without reducing the complexity of the system (Barreto and Souza, 2006).

2.2 Study area and urban parameters

The availability of parameters such as urban topology, land use, vehicular traffic, roads dimensions, etc. at fine spatial resolution makes Singapore a perfect place to investigate the influence of those parameters in the air quality at ground level. For the selected domain of 35.1 km^2, divided in cells of 100×100 m, 80 urban parameters were tested. The main categories of parameters are listed in Table 1. The complete list of parameters is provided in the Supplementary Material.

Figure 2 shows the urban area selected to test the data-driven method proposed here. This area encompasses the district of Rochor, which meets the heterogeneity requirements to investigate the nonlinear correlations between urban parameters and air pollution at ground level. The district of Rochor covers the historic neighborhood of Little India, which is characterized by two types of building typologies: shop-houses and residential towers. Shop-houses are multifunctional row houses of 3–5 stories, while the residential towers are up to 30 stories and can be built on a multi-story base with retail function. Rochor contains multiple urban land uses that range from residential to small-scale industrial workshops. The urban parameters used to train SOM were those listed in the land use section of the Singapore Master Plan 2008 within each grid cell. Land use is derived as the number of square meters for each category.

The studied area is formed by different street layouts. Some roads are eight-lane transit streets, others shopping streets or back lanes with service functionality (e.g., garbage collection). To identify the individual street typology, differ-

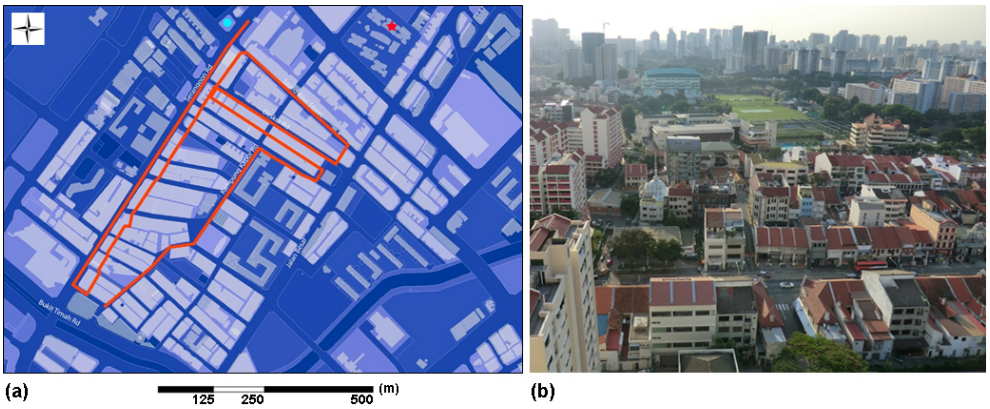

Figure 2. Location of the commercial/residential district of Rochor, Singapore, selected as an urban domain to test the data-driven method proposed here. **(a)** The streets and alleys marked in red correspond to the route followed for the aerosol measurements. The red star indicates the location of the background site and the blue dot the entrance of the subway station of Farrer Park. **(b)** Panoramic photo of Rochor showing the heterogonous landscape formed by shop houses and residential towers.

Table 1. Main categories of urban parameters with influence on the aerosols concentration at ground level of the residential/commercial neighborhood of Rochor, Singapore, used as a study case.

Category	Data source
Land use	Singapore Master Plan 2008
Street network and connectivity	Singapore Land Transport Authority
Building topology	NAVTEQ Building Footprint Singapore Master Plan 2008

ent graph measures (Hillier et al., 1976) were applied to a street graph encompassing the entire city-state of Singapore with different distance ranges to identify the major and minor roads.

2.3 Particles pollution measurements

For the evaluation of the data-driven method proposed here we measured a number of variables that characterize the aerosols pollution at ground level. Particles were chosen among the typical monitored air pollutants in cities because they are responsible for driving the worst air quality conditions in Singapore, as well as in many other cities (Velasco and Roth, 2012).

The aerosol pollution data were collected at ground level along streets, alleys and public areas of Rochor and from a site placed above the urban canopy (a balcony in a 28th floor) called thereafter background site. The purpose of this site was to measure particles concentrations at ambient level, as typical monitoring stations do. The route followed during the ground measurements and the location of the background site is shown in Fig. 2. The ground level route was designed to cover as much as possible the different land uses and urban topologies of the selected neighborhood.

Seven parameters of aerosol pollution were measured in situ using portable and battery-operated sensors. The set of sensors included two DustTrak Aerosol Monitors (TSI 8534) to measure size-segregated mass-fraction concentrations (PM_1, $PM_{2.5}$ and PM_{10}) at ground level and at the background site. Similarly, two handheld condensation particle counters (TSI 3007) were used to measure the PN concentration (only particles with a diameter $< 1\,\mu m$). Concentrations of BC and pPAHs, and the joint ASA of all particles were only measured at ground level using a Micro-Aethalometer (AE51, AethLabs), a Photoelectric Aerosol Sensor (Ecochem Analytics PAS-2000CE), and a Diffusion Charging Sensor (Ecochem Analytics DC-2000CE), respectively. All sensors were synchronized and programmed for 1 s readings, with the exception of the sensors measuring pPAHs and ASA, which were programmed for 10 s readings. For the ground level measurements, the instruments measuring mass and number concentrations were hand carried near breathing height, while the other instruments were carried in a backpack with sampling line inlets at the same height. A Global Positioning System (GPS) was used to geo-reference the aerosol pollution readings. Additional information about the instruments and data post-processing is provided in the Supplementary Material.

The measurements were limited to the evening period from 18:00 to 20:00 h on weekdays. Using commuter data from the subway station of Farrer Park located in the middle of the neighborhood of Rochor (see Fig. 2), we found that this is the period of major influx of people, and therefore of major interest from a health risk point of view. The ground level route of 3.5 km was covered 20 times along 10 days of July 2013. None of the measurement days were affected by rain or smoke-haze from wildfires in neighboring islands (e.g., Sumatra and Kalimantan). Constant meteorological conditions, as well as constant intensity of anthro-

pogenic activities (i.e., aerosol emissions) were assumed during the 2 h of measurements.

Using the location of each measurement obtained by the GPS readings properly synchronized with the particle sensors, an identification flag was assigned to each measurement point using as reference the closest grid cell and its corresponding urban parameters.

The measurements at the background site were used to verify that the ambient concentrations were constant during the 2-h measurement periods, and the ground-level measurements were used to train SOM. Two reasons explain this: (1) the differences between the concentrations measured at ground level and the background site showed a small variability and had therefore an insignificant influence on training SOM. Statistically, the combination of a random function $f(x) = \mu(x) \pm \sigma$ and a constant function $g(x) = c$ would result in $f(x) + g(x) = \mu(x) + c \pm \sigma$ that has the same variation as $f(x)$, and consequently has no effect on the function approximation problem. (2) The urban parameters are informative to explain the variations at ground level, but not at ambient level, where pollutants are usually well-mixed and chemical reactions are also important.

3 Application of SOM as a nonlinear function approximation method between urban parameters and aerosol pollution data

This section describes step by step the application of SOM as a data-driven method to approximate the nonlinear functions between the urban parameters and aerosol pollution data measured at ground level. The model is validated by cross-validations between the values predicted by SOM and the measured values. The involved steps are basically the following three:

- Step 1: Data transformation from a high to a low dimensional space;
- Step 2: Modeling the nonlinear functions between urban parameters and aerosol variables;
- Step 3: Validation and hypothesis testing.

3.1 Step 1: Data transformation from a high to a low dimensional space

A Self Organizing Map is capable of delivering two-dimensional maps in which smoothly changing patterns of the original high-dimensional space can be visualized. Figure 3 shows the interrelations between the different aerosol variables after training an SOM by simply using the averages of the measurements for each grid cell. Two patterns are observed, one linearly correlated for PM_1, $PM_{2.5}$ and PM_{10}, and one nonlinearly correlated between the other variables (PN concentration, BC, pPAHs and ASA). The first pattern starts with high values in the lower-right corner of the trained map, increasing toward the opposite upper-left corner of the map. The second pattern shows high values in the lower-left corner increasing towards the opposite upper-right corner.

The two different patterns are not surprising. In areas influenced by traffic emissions, such as the district of Rochor, ultrafine particles (UFP ≤ 100 nm in diameter) typically represent > 90 % of PN concentration (Morawska et al., 2008). The UFP emitted directly by combustion processes or formed in the air as the hot exhaust gasses are expelled from the vehicles tailpipes represent the main source of BC and pPAHs, and are strongly correlated with both PN concentration and total ASA. Because the mass concentration of PM_1, $PM_{2.5}$ and PM_{10} are several orders of magnitude larger than UFP, their concentrations do not correlate with the other measured aerosol variables. We can conclude that measurements of only two aerosol parameters would have been necessary. Considering the instruments' cost and importance of the parameters for health and risk assessments, we recommend considering only measurements of BC or PN concentration in addition to PM_1 for future studies.

In addition to the smooth pattern created by the SOM, the probabilistic distribution of the original data set (i.e., measurement vectors of each grid cell) can be also obtained from the trained map, as shown in Fig. 4. In this diagram, called a "hit-map", each hexagonal unit is a node of the SOM, where the size of the black points within each unit is relative to the number of similar observation points placed in that unit during the training phase. Hence, the data points and nodes are similar to each other in the same area of the map. This creates a smooth probabilistic pattern on top of the SOM, in which the frequency of observed patterns (proportional to the size of the black points) can be used for resampling and simulation of the observed patterns. For a detailed description of this idea one can refer to Bieringer et al. (2013).

3.2 Step 2: Modeling the nonlinear functions between urban parameters and aerosol variables

The already trained SOM in combination with algorithms such as the K-nearest neighborhood (KNN) and Radial Basis Function (RBF) represents a powerful nonlinear function approximation method. For example, using a data set $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$ in which $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $\mathbf{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$ under the assumption of $\boldsymbol{y}_j = f(X)$ for new data sets without $\boldsymbol{y}_j$, a trained SOM based on data set $\mathbf{Z}$, combined with KNN or RBF can predict with high accuracy the most likely $\boldsymbol{y}_j$ based on the observed $\mathbf{X}$ (Barreto and Souza, 2006). Hence, our assumption of nonlinear relations between $\mathbf{X}$: all urban parameters and $\mathbf{Y}$: all measured aerosol variables. To overcome the limitation imposed by not collecting aerosol data over the complete domain (only 98 out of 3510 grid cells were monitored), we considered similar aerosol concentrations for grid cells with similar urban parameters. Under these assumptions the SOM was trained only with urban parameters, producing patterns based on the grid cells' similarity in terms of

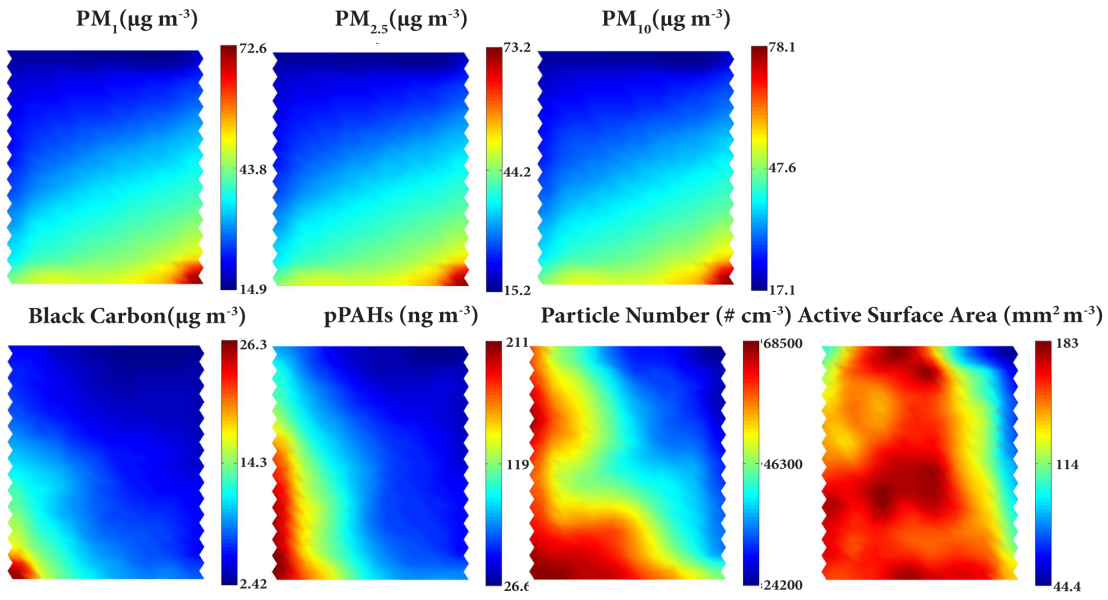

Figure 3. Visualization created by an SOM of the patterns in two-dimension maps (component planes) for the different particle variables measured at street level.

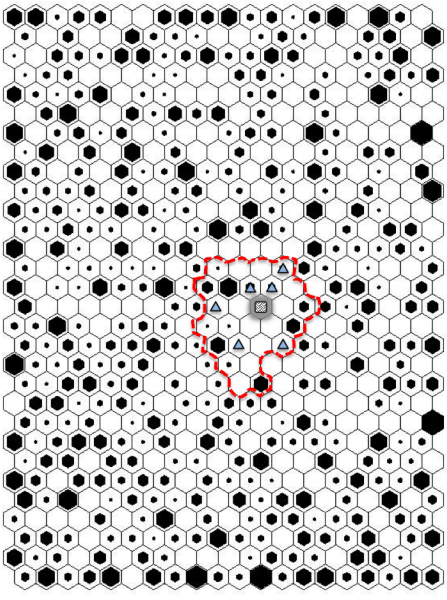

Figure 4. Distribution of training data in the trained SOM (hit-map) based on their similarity in urban parameters. Each hexagon is a node in the SOM and the size of the black points within each hexagon is proportional to the number of training data placed in that node during the training phase. The gray shaded square indicates the projection of a new grid cell on the trained SOM (only on urban parameters) within the region with K-most similar nodes after the computation of Euclidean distances between the weighted vectors in the trained SOM and the original vectors of urban parameters $\mathbf{X}_i$. The triangles indicate grid cells with direct aerosol measurements.

these parameters and not of aerosol concentrations. For a grid cell with no direct measurements, the aerosol concentrations were predicted as the weighted average of the concentrations in grid cells with measurements and similar node (urban parameters) in the trained SOM. The weighted averages were computed using normalized similarity values between cells of the same node. If a projected cell presents a null similarity with any cells with measurements, the approach cannot predict its concentrations. The following steps describe in detail the prediction process:

1. Train an SOM based only on urban parameters (with normalized values for each parameters) covering the whole domain and including grids with and without direct aerosol measurements.
2. For each grid cell i with urban parameters $\mathbf{X}_i$:
   - 2.1. Project the grid cell i into the trained SOM and find the K most similar nodes in terms of urban parameters through the computation of Euclidean distances between the weighted vectors in the trained SOM and $\mathbf{X}_i$ (see Fig. 4).
   - 2.2. Within the selected region of nodes (red contour in Fig. 4) find the grid cells with aerosol measurements ($\mathbf{X}_r$) (triangles in Fig. 4).
   - 2.3. Calculate the normalized similarity between the selected cells and those with measurements (i.e., $\mathbf{X}_i$ and $\mathbf{X}_r$). Similarity is calculated based on the Euclidean distance between each pair of high dimensional vectors.
   - 2.4. Based on the following two recommendations calculate the aerosol concentrations for cell i:

– Calculate the weighted average concentrations from the selected cells with measurements. The weight is based on the normalized similarity of the urban parameters between cell i and those cells with measurements (i.e., $\mathbf{X}_r$) for the selected node.

– If weights are close to each other with no dominant weights from a few of the selected cells, use the measurement median instead of the mean to prevent bias from extreme values when calculating the average concentrations. Geometric means are also an option.

With the assumption of existing relationships between urban parameters and aerosol concentrations, the SOM creates a smooth map of emergent urban patterns. It is worth mentioning that in the trained SOM map, the grid cells representing the spatial surface of the neighborhood in the physical space are not necessarily placed in the same region if they do not have similar urban patterns.

The number of nodes in the SOM, defined as the width and height of the trained map, is important to optimize the SOM training procedure. In our experiment we selected a map of 20×25 nodes based on the size of the training data (with 3510 grid cells and 500 nodes, on average each node will represent around 7 similar grid cell, if they follow a uniform distribution). Different grid sizes did not show to be important, but very large or very small map sizes showed direct effects on the quality of the training algorithm in terms of quantization and topographic errors (Kohonen, 2001). The number of similar nodes in neighborhood search, K, is also important to optimize the process of data-driven modeling. We tested different K values finding that values between 1 and 5 are good enough for cross-validation. Another assumption that we have in the current implementation was that all the urban parameters are equally important. However, performing the feature (i.e., urban parameter) selection and extraction in a systematic manner could also optimize the training procedure as suggested by Guyon and Elisseeff (2003). Feature selection and extraction is a computationally complex problem. The number of potential combinations of urban parameters is on the order of $2^n - 1$, where n is the number of features including all the possible transformations (e.g., $z = a + b\log(x)$). Methods such as the Genetic Algorithm can help to solve this optimization issue in a reasonable time (Niska et al., 2004).

3.3 Step 3: validation and hypothesis testing

Before applying the SOM to predict the aerosol concentrations in the entire domain, the nonlinear relationships between the different urban parameters and aerosol concentrations approximated by SOM must be tested. We performed cross-validations between the predicted values by SOM and the real values to validate the proposed data-driven modeling approach.

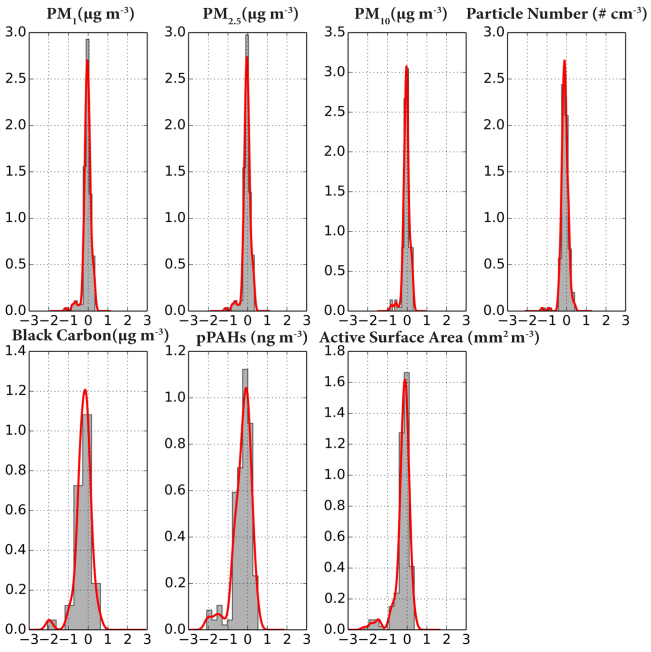

Figure 5. Relative errors distribution of the predicted aerosol concentrations based on the randomly selected validation data.

Because of the limited number of grid cells with measurements, the cross-validation was performed using 10 % of the samples in 20 iterations. This means that we removed randomly 10 % of the cells with measurements for every iteration and predicted their aerosol concentrations based on the remaining cells with measurements. The statistical metrics of the cross-validations shown in Table 2 demonstrate the ability of SOM to preserve the nonlinear relations between the urban parameters and aerosol concentrations. Figure 5 shows that the relative errors of the predicted aerosol concentrations are tightly distributed around zero with a relatively longer left tail for all of the particles, indicating a tendency to underestimate real values.

4 Application of SOM as a data-driven model to interpolate concentrations of aerosols in a gridded domain

Once the cross-validation has demonstrated satisfying results, we can proceed to interpolate the aerosol concentrations in the complete gridded domain. The interpolation methodology is essentially the same as the methodology used in the previous section for the cross-validation. The only difference is the addition of a confidence measure for the predicted concentrations. This confidence measure is based on the similarity between the urban parameters and grid cells with measurements. If no similar grid cell with direct mea-

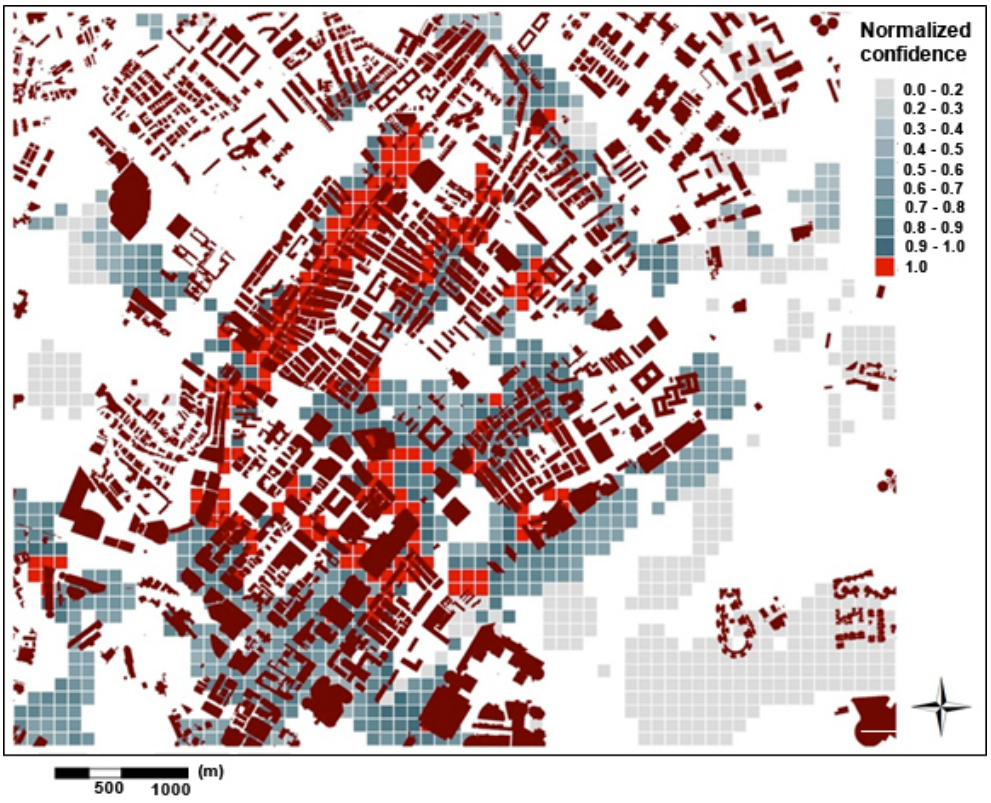

Figure 6. Distribution of the probabilistic confidence levels of the predicted aerosol concentrations using the nonlinear approximation function of SOM, overlaid on a map of the studied neighborhood of Rochor, Singapore.

surements is available for a particular set of grid cells with a similar urban pattern, a null confidence value will be obtained and no concentration will be predicted. In our study case, this situation occurred for regions with no urban similarity with the region where the measurements were performed. The confidence value for each cell was computed following the next steps:

1. The grid cells from the complete domain are divided into cells with measured data ($\mathbf{X}_{\mathrm{M}}$) and no measured data ($\mathbf{X}_{\mathrm{NM}}$).
2. The $\mathbf{X}_{\mathrm{M}}$ cells are projected into the trained SOM to calculate the Euclidean distance of each grid cell (dist_i) with its K most similar nodes of SOM.
3. The median of the calculated distances for $\mathbf{X}_{\mathrm{M}}$ cells is used as a norm for the $\mathbf{X}_{\mathrm{NM}}$ cells, (norm_dist). If the Euclidian distance for an $\mathbf{X}_{\mathrm{NM}}$ cell is smaller or equal to this norm, the confidence will be one, in contrast if it is larger (i.e., less similarity with the $\mathbf{X}_{\mathrm{M}}$ cells) the confidence will tend to zero. The confidence value for $\mathbf{X}_{\mathrm{NM}}$ cells is computed as:

$$\mathrm{Confidence}_i = \min(1, 1 - (\mathrm{dist}_i - \mathrm{norm\_dist})/\mathrm{dist}_i). \quad (1)$$

Figure 6 shows the confidence values of each grid cell overlaid on a map of the studied domain. As expected, the cells with high confidence values were those with similar patterns to cells with measured data. The distribution of cells with confidence values > 0.5 are relatively equally distributed over the built-up regions. Regions with null confidence correspond primarily to open spaces, such as public parks where measurements were not conducted. In general, we can affirm that the proposed method is capable of interpolating aerosol pollution data at ground level within the built-up areas of a heterogeneous neighborhood of 35.1 km^2 using measured data from less than 3 % of the total gridded domain.

5 Potential locations for fixed monitors based on clusters of grid cells with similar urban patterns

Similar to previous sections, after the good cross-validation results between the predicted values by the nonlinear approximations and the measured aerosol pollution data, we can apply a clustering algorithm based on the urban parameters to determine the optimum number of fixed monitoring sites and their best locations (grid cells) in terms of representativeness and maximum information gain over the whole domain.

Clustering algorithms have the task of finding the optimum number of groups from a given data set. The members (i.e., grid cells) of a group must be as much as possible similar to

Table 2. Efficiency of SOM to approximate nonlinear relationships between urban parameters and aerosol concentrations. The cross-validation was performed using 10 % of the samples in 20 interactions as explained in the text.

Aerosol variable	Based on median values of similar grid cells		Based on arithmetic mean values of similar grid cells	
	Median of accuracy (%)	Mean of accuracy (%)	Median of accuracy (%)	Mean of accuracy (%)
PM_1	92.18	85.81	93.03	87.13
$PM_{2.5}$	92.16	85.87	92.98	87.21
PM_{10}	92.64	87.27	93.21	87.20
Particle number	89.34	85.97	90.24	86.48
Black carbon	78.08	69.37	79.10	67.34
pPAHs	78.72	70.67	82.42	70.98
Active surface area	88.36	75.06	88.67	76.81
Average accuracy (%)	87.35	80.00	88.52	80.45
Min accuracy (%)	78.08	69.37	79.10	67.34
Max accuracy (%)	92.64	87.27	93.21	87.21

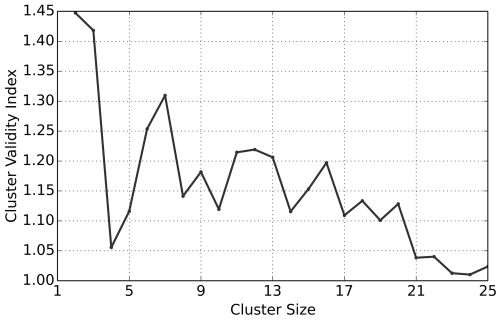

Figure 7. Application of the heuristic elbow method to identify the optimal number of clusters in which the grid cells of the studied domain of Rochor can be grouped. The drastic decrease of the clustering index with four clusters suggests that additional clusters will not significantly improve the clustering quality.

each other and dissimilar to members of other groups. Each cluster must represent an individual group with a specific set of parameters. The clustering algorithm must also be capable of identifying the most informative (representative) members of that cluster within each cluster.

The K-means clustering algorithm is frequently used in combination with a SOM. The SOM acts like a first step filtering and smoothing of the data points and then K-means is applied to nodes of the trained SOM knowing the number of clusters in advance.

In practice, this number is determined by heuristic methods, such as the elbow method (Tibshirani et al., 2001). For our case study, the elbow method suggests that four clusters are enough for the whole gridded domain. It means that four main types of urban settings define the neighborhood of Rochor. As shown in Fig. 7, the clustering index (metric to evaluate the clusters compactness and separation) decreases drastically with four clusters. More clusters do not represent any major improvement.

Figure 8a shows the distribution of grid cells within the trained map by SOM. The size of the black spot is proportional to the number of training data in each node. After applying the K-means clustering algorithm the cells grouped in four clusters can be observed in Fig. 8b. The size of the internal spots indicates the representativeness degree of the cells to their clusters. The centroid point (mean value across all the dimensions) of each cluster can be considered as its most representative point. In this context, the grid cells with the highest representativeness degree within each cluster should be considered as candidate locations for monitoring air quality stations.

6 Results

The methodology based on the data-driven model of SOM developed in this work offers two outcomes of potential relevance for the air quality management in cities. Maps showing the spatial distribution of aerosol concentrations at ground level within the whole gridded domain represent the first outcome. The second outcome and main goal of this work is the finding of the optimum number of fixed monitoring stations and their potential best locations to cover the different types of urban settings (i.e., clusters) of the studied neighborhood.

Figure 9 shows the spatial distribution of $PM_{2.5}$ within the gridded domain including measured and estimated data as an example of the first outcome. Similar to Fig. 6, the grid cells with no data correspond to cells with null interpolation confidence, where no prediction was possible due to the lack of similarity in terms of urban parameters with the grid cells with direct measurements. Constraining the analysis to grid cells with measured and interpolated data with confidence values ≥ 0.5, the average $PM_{2.5}$ concentration at

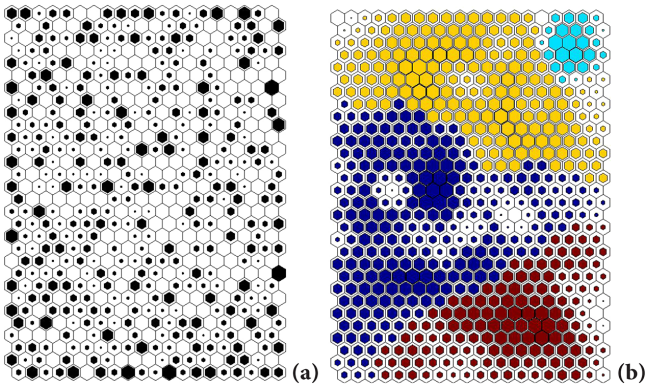

Figure 8. (a) Distribution of grid cells after applying the K-means clustering algorithm. Each cluster is represented by a different color. The size of the internal spots indicates the representativeness degree of each cell to its cluster. Cells with larger spots are more representative. **(b)** Distribution of grid cells within the trained map by SOM. The size of the black points is proportional to the number of training data, which are placed in the same area of SOM based on their similarities.

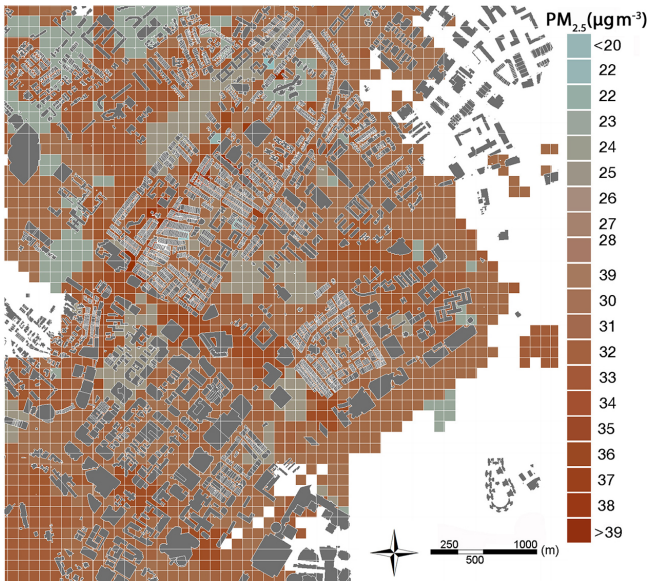

Figure 9. Spatial distribution of $PM_{2.5}$ at the studied neighborhood of Rochor during the evening rush hour (18–20 h) on weekdays, including measured and interpolated data. The data interpolation was based on the aerosol pollution measurements conducted along the ground level route of 3.5 km marked in Fig. 2 during 10 days of June 2013.

ground level in the neighborhood of Rochor is 31.3 µg m^{-3} during the evening period from 18:00 to 20:00 h. This concentration is 1.25 times higher than the 24 h average concentration of 25 µg m^{-3} recommended as guideline by the World Health Organization (WHO). Similarly, only 10 % of the grid cells report concentrations below the WHO guideline and 75 % are higher than 31.0 µg m^{-3}. In comparison with the average concentration of 18.8 µg m^{-3} measured at the background site, only two cells present smaller concentrations. Table 3 shows detailed statistics of the aerosol concentrations predicted for the whole gridded domain.

Although the measurements were conducted during the period of major influx of people, and therefore of major interest from a health risk point of view, the nonlinear relations between urban parameters and aerosol pollution data obtained by SOM cannot be representative for the whole diurnal course. They are only representative of the rush-hour period monitored. The nonlinear relationships will vary throughout the day as a consequence of the variability in the emissions' intensity within the studied neighborhood. However, the aerosol measurements during the evening rush-hour had the unique purpose of testing SOM. The proposed methodology is expected to be used in the design of future long-term studies. Maps showing the spatial distribution of pollutants concentration at ground level in fine-grained domains will provide valuable information for epidemiological and risk assessments. We already discussed the poor ability of typical ambient monitoring stations to represent the pollution levels at the height where urban dwellers carry out the majority of their activities. This is of particular concern in the ubiquitous environments affected by vehicular emissions. Many epidemiological studies have found significant health effects due to exposure to vehicular traffic (e.g., Lipfert and Wyzga, 2008). Although these studies have investigated various exposure criteria, including traffic intensity and proximity, control strategies have generally not yet been proposed on a widespread basis, in part due to the lack of long term air pollution monitoring at street level, as well as of a methodology to understand the relationships between pollutant concentrations and urban parameters. In a following article we will discuss the features and roles of those parameters in the air quality of the studied neighborhood of Rochor. The understanding of these relationships might be also useful for urban planning, in particular when designing strategies to improve urban mobility promoting walking and cycling as a means to cover the so called first and last miles (distances that commuters must cover in getting to and from public transportation).

Figure 10 summarizes the application of SOM as a data-driven method to find the optimum number of monitoring stations and their potential locations in terms of representativeness. The top candidate grid cell(s) for each one of the four different urban settings that form the residential/commercial neighborhood of Rochor are marked over the individual maps of each urban setting. The grid cells were brought back to the real two-dimensional space using as reference their latitude and longitude data. The next step in the selection of sites is to visit the candidate locations and verify that enough space is available for a fixed monitoring station, the security conditions and continuous access to power. The location must fulfill the criteria to assure the proper performance of the air

Table 3. Statistics of the aerosol concentrations predicted by SOM for the complete gridded domain of the neighborhood of Rochor on weekdays during the evening rush hour (18–20 h). The analysis considers only grid cells with measured or extrapolated data with confidence values ≥ 0.5.

Aerosol	Mean	SD	Min	25 %	50 %	75 %	95 %	Max
PM_1 ($\mu g\,m^{-3}$)	30.96	3.32	14.60	30.66	31.02	32.88	35.23	66.37
$PM_{2.5}$ ($\mu g\,m^{-3}$)	31.26	3.34	14.83	30.95	31.33	33.18	35.57	67.58
PM_{10} ($\mu g\,m^{-3}$)	33.98	3.35	15.86	33.72	34.29	35.70	38.46	71.73
Particle number (# cm^{-3})	46585	7154	20427	42659	44044	53311	57932	92240
Black carbon ($\mu g\,m^{-3}$)	6.82	2.11	1.94	4.93	6.40	8.65	10.42	25.85
pPAHs ($ng\,m^{-3}$)	78.26	27.14	27.72	57.14	73.16	102.57	119.26	345.00
Active surface area ($mm^2\,m^{-3)}$	149.09	17.23	43.55	133.57	153.99	162.20	178.95	225.50
Confidence	0.85	0.16	0.50	0.73	0.90	1.00	1.00	1.00

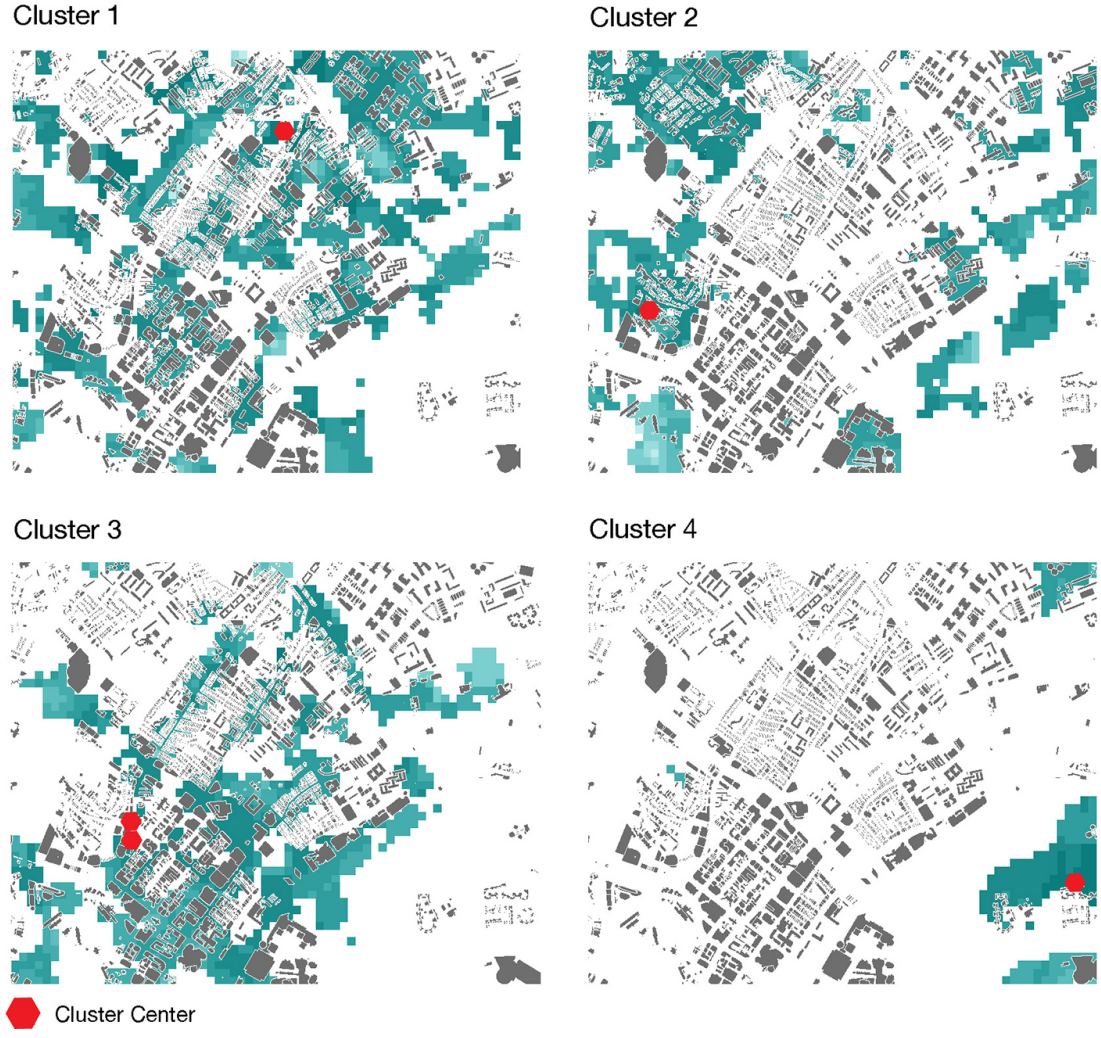

Figure 10. Distribution of the grid cells in the real two-dimensional space grouped in the four different clusters (i.e., urban settings) that form the neighborhood of Rochor, Singapore. The color intensity indicates the representativeness degree. Grid cells colored more intensely represent better the urban parameters of their corresponding clusters. The most representative grid cells of each cluster (highest value(s)) are marked in red as the candidate locations for fixed air quality stations. The gray areas correspond to buildings.

quality monitors. For further guidance the reader may refer to handbooks for air quality monitoring (e.g., US-EPA, 2013). If we select more than one candidate location for each cluster (say three top candidates, each selected independently), another optimization step would be necessary to find the best four monitoring stations out of 12 candidate points (three locations for four clusters) in a way to minimize the overlap between stations of different clusters and maximize the total physical coverage of the monitoring stations.

7 Conclusions

The capability of SOM as a data-driven modeling method to approximate nonlinear relationships between multiple urban parameters and air pollution data at ground level was demonstrated using a database of urban parameters spatially distributed in high-resolution grid cells created with purposes different to air quality monitoring (e.g., urban planning) and aerosol pollution data collected during a short field study. The good agreement between measured and predicted aerosol concentrations showed that the group of urban parameters used in this work provides a good indirect measure of aerosol pollution at ground level within the studied neighborhood. The same methodology can also be used for any gaseous pollutant. Every pollutant, depending on its origin and physical and chemical characteristics will present different nonlinear relationships.

The satisfying results of SOM to approximate nonlinear relationships from multidimensional data gave the opportunity to apply SOM as a method to interpolate aerosol pollution data in a complete gridded domain, including grid cells with no direct measurements. In the same context, SOM in combination with a clustering algorithm was used to determine the optimum number of locations for monitoring sites to cover the different urban settings or clusters forming the studied neighborhood, as well as to find their best location in terms of representativeness of urban patterns within their clusters.

The data-driving modeling methodology developed in this work as a proof of concept must be relatively easy to implement to other urban domains if such urban parameters as street networks, land-use patterns, demographics, transportation data, and building and street topology are available in databases of high spatial resolution. The aerosol pollution measurements should not represent a major cost if portable and battery-operated sensors are used, as in this work. We evaluated seven different aerosol parameters, but measurements only of black carbon or particle number concentration in addition to PM_1 would have only been necessary according to the nonlinear correlations between aerosol parameters, identified visually by SOM.

The Supplement related to this article is available online at doi:10.5194/amt-8-3563-2015-supplement.

Acknowledgements. This research was supported by the National Research Foundation Singapore (NRFS), through the Singapore MIT Alliance for Research and Technology's CENSAM research program and the Singapore-ETH Centre for Global Environmental Sustainability (SEC) co-funded by NRFS and ETH Zurich.

Edited by: P. Xie

References

Barreto, G. A. and Souza, L. G. M.: Adaptive filtering with the self-organizing map: a performance comparison, Neural Networks, 19, 785–798, 2006.

Bieringer, P. E., Longmore, S., Bieberbach, G., Rodriguez, L. M., Copeland, J., and Hannan, J.: A method for targeting air samplers for facility monitoring in an urban environment, Atmos. Environ., 80, 1–12, 2013.

Craig, L., Brook, J. R., Chiotti, Q., Croes, B., Gower, S., Hedley, A., Krewsky, D., Krupnik, A., Kryzanowski, M., Moran, M. D., Pennell, W., Samet, J. M., Schneider, J., Shortreed, J., and Williams, M.: Air pollution and public health: a guidance document for risk managers, J. Toxicol. Env. Heal. A, 71, 588–698, 2008.

Guyon, I. and Elisseeff, A.: An introduction to variable and feature selection, J. Mach. Learn. Res., 3, 1157–1182, 2003.

Hidy, G. M. and Pennell, W. T.: Multipollutant air quality management, J. Air Waste Manage., 60, 645–674, 2010.

Hillier, B., Leaman, A., Stansall, P., and Bedford, M.: Space syntax, Environ. Plann. B, 3, 147–185, 1976.

Hirtl, M., Mantovani, S., Krüger, B. C., Triebnig, G., Flandorfer, C., Bottoni, M., and Cavicchi, M.: Improvement of air quality forecasts with satellite and ground based particulate matter observations, Atmos. Environ., 84, 20–27, 2014.

Kohonen, T.: Self-organized formation of topologically correct feature maps, Biol. Cybern., 43, 59–69, 1982.

Kohonen, T.: Self-Organizing Maps, Vol. 30, Springer, 105–176, 2001.

Kohonen, T.: Essentials of the self-organizing map, Neural Networks, 37, 52–65, 2013.

Kolehmainen, M., Martikainen, H., Hiltunen, T., and Ruuskanen, J.: Neural networks and periodic components used in air quality forecasting, Atmos. Environ., 35, 815–825, 2001.

Kolehmainen, M. T.: Data Exploration with Self-Organizing Maps in Environmental Informatics and Bioinformatics, Helsinki University of Technology, Helsinki, 54–64, 2004.

Li, X. X., Liu, C. H., Leung, D. Y. C., and Lam, K. M.: Recent progress in CFD modeling of wind field and pollutant transport in street canyons, Atmos. Environ., 40, 5640–5658, 2006.

Li, X. X., Britter, R. E., Norford, L. K., Koh, T.-Y., and Entekhabi, D.: Flow and pollutant transport in urban street canyons of different aspect ratios with ground heating: large Eddy Simulation, Bound.-Lay. Meteorol., 142, 289–304, 2012.

Lipfert, F. and Wyzga, R. E.: On exposure and response relationships for health effects associated with exposure to vehicular traffic, J. Expo. Sci. Env. Epid., 18, 588–599, 2008.

Moore, K., Krudysz, M., Pakbin, P., Hudda, N., and Sioutas, C.: Intra community variability in total particle number concentrations

in the San Pedro Harbor area (Los Angeles, California), Aerosol Sci. Tech., 43, 587–603, 2009.

Morawska, L., Ristovski, Z., Jayaratne, E., R., Keogh, D. U., and Ling, X.: Ambient nano and ultrafine particles from motor vehicle emissions: characteristics, ambient processing and implications on human exposure, Atmos. Environ., 42, 8113–8138, 2008.

Nerriere, É., Zmirou-Navier, D., Blanchard, O., Momas, I., Ladner, J., Le Moullec, Y., Personnaz, M. B., Lameloise, P., Delmas, V., Target, A., and Desqueyroux, H.: Can we use fixed ambient air monitors to estimate population long-term exposure to air pollutants? The case of spatial variability in the Genotox ER study, Environ. Res., 97, 32–42, 2005.

Nguyen, T. N. T., Ta, V. C., Le, T. H., and Mantovani, S.: Particulate matter concentration estimation from satellite aerosol and meteorological parameters: data-driven approaches, in: Knowledge and Systems Engineering, Springer International Publishing, 351–362, 2014.

Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., and Kolehmainen, M.: Evolving the neural network model for forecasting air pollution time series, Eng. Appl. Artif. Intel., 17, 159–167, 2004.

Salimi, F., Mazaheri, M., Clifford, S., Crilley, L. R., Laiman, R., and Morawska, L.: Spatial variation of particle number concentration in school microscale environments and its impact on exposure assessment, Environ. Sci. Technol., 47, 5251–5258, 2013.

Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation, SIAM, 1–37, 2005.

Tibshirani, R., Walther, G., and Hastie, T.: Estimating the number of clusters in a data set via the gap statistic, J. Roy. Stat. Soc. B, 63, 411–423, 2001.

Tominaga, Y. and Stathopoulos, T.: CFD simulation of near-field pollutant dispersion in the urban environment: a review of current modeling techniques, Atmos. Environ., 79, 716–730, 2013.

Velasco, E. and Roth, M.: Review of Singapore's air quality and greenhouse gas emissions: current situation and opportunities, J. Air Waste Manage., 62, 625–641, 2012.

Viana, M., Kuhlbusch, T. A. J., Querol, X., Alastuey, A., Harrison, R. M., Hopke, P. K., Winiwarter, W., Vallius, M., Szidat, S., Prévôt, A. S. H., Hueglin, C., Bloemen, H., Wåhlin, P., Vecci, R., Miranda, A. I., Kasper-Giebl, A., Maenhaut, W., and Hitzenberger, R.: Source apportionment of particulate matter in Europe: a review of methods and results, J. Aerosol Sci., 39, 827–849, 2008.

Voukantsis, D., Niska, H., Karatzas, K., Riga, M., Damialis, A., and Vokou, D.: Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece, Atmos. Environ., 44, 5101–5111, 2010.