Data-driven nonlinear optimisation of a simple air pollution dispersion model generating high resolution spatiotemporal exposure

Yuval*, Shlomo Bekhor, David M. Broday

Department of Civil and Environmental Engineering, Technion, Israel Institute of Technology, Haifa 32000, Israel

HIGHLIGHTS

• Fitting a simple dispersion model to high resolution air pollution monitoring data.
• Requires nonlinear regression to obtain the optimal model.
• The results are fully cross-validated.
• Providing pollution exposure estimates at a high resolution in both space and time.

ARTICLE INFO

Article history:
Received 3 March 2013
Received in revised form
30 May 2013
Accepted 3 June 2013

Keywords:
Air pollution monitoring
Exposure
Nonlinear optimisation
Traffic model output

ABSTRACT

Spatially detailed estimation of exposure to air pollutants in the urban environment is needed for many air pollution epidemiological studies. To benefit studies of acute effects of air pollution such exposure maps are required at high temporal resolution. This study introduces nonlinear optimisation framework that produces high resolution spatiotemporal exposure maps. An extensive traffic model output, serving as proxy for traffic emissions, is fitted via a nonlinear model embodying basic dispersion properties, to high temporal resolution routine observations of traffic-related air pollutant. An optimisation problem is formulated and solved at each time point to recover the unknown model parameters. These parameters are then used to produce a detailed concentration map of the pollutant for the whole area covered by the traffic model. Repeating the process for multiple time points results in the spatiotemporal concentration field. The exposure at any location and for any span of time can then be computed by temporal integration of the concentration time series at selected receptor locations for the durations of desired periods. The methodology is demonstrated for NO₂ exposure using the output of a traffic model for the greater Tel Aviv area, Israel, and the half-hourly monitoring and meteorological data from the local air quality network. A leave-one-out cross-validation resulted in simulated half-hourly concentrations that are almost unbiased compared to the observations, with a mean error (ME) of 5.2 ppb, normalised mean error (NME) of 32%, 78% of the simulated values are within a factor of two (FAC2) of the observations, and the coefficient of determination ($R^2$) is 0.6. The whole study period integrated exposure estimations are also unbiased compared with their corresponding observations, with ME of 2.5 ppb, NME of 18%, FAC2 of 100% and $R^2$ that equals 0.62.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Spatially accurate estimation of exposure to air pollutants in the urban environment is needed for most air pollution epidemiological studies (Jerret et al., 2005; Han and Naeher, 2006). Air pollution is emitted at many source locations, is dispersed and transformed by complex physical and chemical processes but is observed only at a limited sample of the locations where the population is exposed. As it is not feasible to distribute personal exposure monitors to a large number of people for long periods of time, epidemiological studies resort to various techniques to model the pollutants’ concentrations. The large spatial variability in urban traffic-related pollutant levels requires estimation methods that can produce small scale exposure variations (Jerret et al., 2005). To benefit studies of acute effects of air pollution at the personal level (e.g., Maynard et al., 2007; Van den Hooven et al., 2012), the exposure features must be provided at high temporal resolution. The ideal model for estimating the full spatiotemporal concentration field of...
an air pollutant should attempt to simulate the emission, dispersion, transformation and removal processes and, due to the inherent difficulties involved and the inevitable errors, poses a built-in data-driven mechanism to adjust the model parameters based on the available observations. To the best of our knowledge none of the current schemes used for air pollution exposure estimation employs a data-driven model that incorporates a dispersion-based simulation of air pollutants.

Gaussian dispersion models (Cyrys et al., 2005; Beelen et al., 2010; Gulliver and Briggs, 2011) and the more comprehensive meteorological–photochemical ones (Astitha et al., 2008; Borrego et al., 2010) are designed to simulate air pollution concentrations given emission and meteorological inputs. However, unlike the current meteorological models, they do not adjust the model parameters in a closed loop controlled by available observations. For example, a recent paper by Beevers et al. (2012) demonstrated an impressive performance of a model coupling the photochemical CMAQ model (Byun and Ching, 1999) with the Gaussian ADMI model (McHugh et al., 1997) in estimating hourly NOx, NO2 and O3 concentrations in London at a 20 x 20 m spatial grid resolution. Data from 80 monitoring stations were used for a comprehensive assessment of the model output. However, there was no attempt to incorporate any type of adjustment of the models using the simultaneous observations to lower the model errors and biases which were found. Various interpolation schemes (Son et al., 2010; Lee et al., 2012) take the opposite approach. The spatial air pollution exposure maps which they produce are based on observed values and statistical principles only. The sparse distribution of monitoring stations limits the ability of observed data interpolations to provide the detailed spatial exposure required for many epidemiological investigations (Yuval and Broday, 2006).

In the last decade the modelling approach of Land Use Regression (LUR) has been extensively used in air pollution epidemiological studies (Ryan and LeMasters, 2007; Hoek et al., 2008). An LUR fits a suit of covariates (mainly metrics of urbanisation, topography and traffic) to observed concentrations and the recovered regression parameters are then used to produce detailed spatial exposure estimation. The LUR scheme is a data-driven method but it does not attempt to mimic the actual processes that associate the model covariates with the observations to which they are fitted. Moreover, in most cases the observations used by an LUR are integrated data over a period of a couple of weeks (or a few such periods at different seasons) so the short time scale processes governing air pollution dispersion are averaged out and only the chronic exposure is estimated.

Incorporating short time scale effects has been shown to improve the accuracy of chronic exposure to air pollution. Arain et al. (2007) used a vector average of hourly wind direction data to differentiate the effects of roads down and up the wind in their LUR model. Su et al. (2008) demonstrated improvement in LUR performance while considering the hourly emission fluxes along wedge–shaped boxes whose geometry was dictated by the hourly wind direction and speed. Wilton et al. (2010) incorporated in an LUR the impact of short term meteorological variation through adding hourly dispersion model outputs as covariates. A different approach by Szpiro et al. (2010) incorporated temporal trends in long–term exposure estimation as a linear combination of empirically derived temporal basis functions. Investigating the acute effects of particulate matter on mortality, Maynard et al. (2007) included various temporal covariates like day of the week, day of year and meteorology in assessing the black carbon exposure levels in a statistical scheme akin to universal kriging. A recent study by Van den Hooven et al. (2012) accounted for temporal variations by calculating spatial PM10 and NO2 exposure patterns for eight different wind conditions and deriving hourly spatial patterns by means of interpolation between the eight characteristic spatial distributions. These patterns were subsequently adjusted for fixed temporal patterns of source activities to achieve the monthly, day of the week and time of the day concentrations at the home locations of pregnant women. None of those studies incorporated the covariates which they used in a scheme which was optimally adjusted to pollution observations at a temporal resolution close to the dominant frequency of the physical and chemical processes that govern air pollution levels.

This study introduces a new concept enabling production of detailed pollution concentration maps at very high temporal resolution. The main idea is formulating the study as an optimisation problem in which the parameters of a nonlinear scheme that simulates the spatial distribution of air pollution concentration are determined by fitting the scheme’s output to observations. This requires the incorporation of meteorological and air pollution observations at a temporal resolution equal to or lower than the dominant time scales of the emission and dispersion processes. The methodology is demonstrated using very simple means but is shown to achieve spatial variability similar to that enabled by an LUR, at the high temporal resolution of the meteorological and monitoring data. The process is packaged in a closed mathematical form and can thus be run completely automatically. This enables a quantitative assessment of its performance using a true leave-one-out cross-validation scheme. The ability to provide both the short time scale acute exposure and the long term chronic exposure is demonstrated on real data from the monitoring network along the Israeli coast.

2. Data

2.1. Study area

The proposed method is demonstrated in a strip of land about 70 km long and 25 km wide along the Israeli coast that includes the metropolitan Tel Aviv area and the cities of Netanya and Ashdod north and south of it, respectively (see Fig. 1). The total population of the area is about 3.7 million. The area includes a dense road system to which feeds much of Israel’s highway network, railway links, two airports and a seaport. The transportation fleet includes about 1.5 million private cars, 200,000 commercial vehicles, and 10,000 city and long distance buses. Industrial sources are relatively minor contributors to air pollution in the study area (Yuval and Broday, 2009). Two natural gas-powered electricity generation stations, a 450 MW station in Tel Aviv and a 1 GW station north of Ashdod, are the main industrial pollution sources.

2.2. Traffic model output

The traffic data is produced by the Metropolitan Tel Aviv Traffic Model developed by Cambridge Systematics (Cambridge Systematics, 2008) during 2005–2008 using the EMME/2 software platform (INRO, 2012). The model was applied and is maintained by an expert team employed by the Netavey Ayalon Company, a joint venture of the Israeli Government and the Tel Aviv municipality. The road network which the model uses includes 11,553 road segments (Fig. 1). Private car traffic is simulated by an activity based model. Bus tours follow the true bus network and the commercial fleet (trucks and pick-ups) is modelled by a separate module. The model output includes 23 traffic attributes produced for five periods during the day: morning, morning rush hour, middle day, afternoon rush hour and evening. For this work only the morning rush, middle day and afternoon rush outputs were available.
2.3. Monitoring and meteorological data

Air pollution monitoring data in the study area are observed by an air quality network of 38 stations and are available from the Israeli Ministry for Environmental Protection’s website. We used for the model development the 2009 NO2 data from the 25 stations that observed the pollutant and that are designed to meet the requirement of the EU Council Directive 1999/30/EC for protection of human health. NO2 data from additional six road-side monitoring stations were used only for additional model testing. The wind data used by the model were created from the set of wind observations in the monitoring stations following the representative wind method of Yuval and Broday (2009). The representative wind data have full temporal coverage. They were compared and found very similar at most time points to observations at 60 m elevation above ground from a meteorological mast located close to the centre of the study area (184.3 East; 670.8 North). The air quality and meteorological data are all available as half-hourly means of the continuous measurements. A series of quality control steps was carried out prior to using the data to ensure highest fidelity.

3. Methods

3.1. Description of the proposed model

The first step in transforming the traffic model outputs, given for road sections, to spatially distributed air pollution concentrations is assigning these outputs to a regular grid. We used a regular grid that divided the area covered by the traffic model into square cells. The contribution $E$ of a road segment to the emission rate from a cell through which it passes can be assumed to be proportional to the multiplication of its traffic volume $V$ (vehicles per unit time) by the mean time $\delta t$ that it takes the traffic to cross the cell via the road segment,

$$E = V\delta t = VL/S,$$

where $L$ is the length of the road segment through the cell and $S$ is the mean speed along it. The length $L$ for each segment in each cell was calculated given the grid and the road network geometries. For the purpose of air pollution study, the traffic volume $V$ should be considered as summation of the numbers of the different vehicle types passing through each segment, weighted by their emission rates. The traffic model output which we used calculates private vehicles, trucks, commercial vehicles and bus volumes for each of the road segments in the network. The trucks and commercial vehicles volume data are considered of lower accuracy and we thus used in our traffic volume estimation only the private vehicle and bus numbers i.e.,

$$V = V_a + R_c V_b,$$

where $V_a$ and $V_b$ are the traffic model’s private car and bus numbers per hour in the segment, respectively, and $R_c$ is the ratio between the mean bus to the mean private car emission of the air pollutant of interest. Following Romilly (1999), $R_c$ in our case was taken as 20. The volumes of the trucks and commercial vehicles are assumed to be proportional to $V$. The proxy to the total emission rate from a grid cell is the summation of the contributions of the road segments that go through it,

$$T_j = \sum_{k=1}^{n} E_{jk},$$

where $T_j$ is the traffic emission proxy at the $j$th cell in the grid and $E_{jk}$ is the contribution of the $k$th road segment that passes through it.

Each grid cell can be considered an air pollutant source of intensity proportional to $T_j$. The impact of the contribution of $T_j$ on the concentration in any cell of the grid depends on the direction between the source and receptor cells relative to the regional wind direction, and the distance between them. The traffic-related concentration $C_i$ in the $i$th cell at a given time point is a superposition of the contributions from the $M$ cells of the grid. A reasonable, albeit admittedly simplistic, way to specify that concentration is by

$$C_i = p_1 + p_2 \sum_{j=1}^{M} \frac{T_j f(\theta_{ij})}{(D_{ij} + p_3)^{p_4}},$$

where $D_{ij}$ is the distance between the $i$th and $j$th cells, $\theta_{ij}$ is the angle between the regional wind direction and the direction between the $i$th and $j$th cells, $f(\theta_{ij}) = \cos(\theta_{ij})$ for $\theta_{ij} \geq 90^\circ$ and zero otherwise, and $p_1$, $p_2$, $p_3$, $p_4$ are unknown parameters. A schematic diagram of the setting of Equation (4) is given in Fig. 2. In a matrix
the vector of concentrations in all the grid cells, $C$, is given by

$$C = \mathcal{W}(p)|T,$$

where $T_i = 1; T_j = 2,3,\cdots,M + 1$ are the emission proxies for the $M$ grid cells; $p = [p_1, p_2, p_3, p_4, p_5]^T$ and $\mathcal{W}$ is a weighting matrix such that $\mathcal{W}_{ij} = p_1$ and $\mathcal{W}_{ij}, j = 2,3,\cdots,M + 1$ is given by

$$\mathcal{W}_{ij} = \frac{p_2}{p_4 + p_4} f(\theta) p_1.$$

To find the unknown model parameters we can write

$$\hat{\mathcal{W}}(p)T - \hat{C} = \epsilon,$$

where $\hat{C}$ is the vector of $N$ observed concentrations at the grid cells where monitoring reside and $\hat{\mathcal{W}}(p)$ are the corresponding rows of $\hat{\mathcal{W}}(p)$. The unknown parameters are found by minimising the cost function

$$\Phi = \|\epsilon\|^2 = \|\hat{\mathcal{W}}(p)T - \hat{C}\|^2,$$

subject to the constraints $p_1, p_2, p_3, p_5 \geq 0$ and $p_4 > 0$. The constraints ensure a physically plausible model (increasing impact of a source cell with decreasing $\theta$ and $D$), and non-negative and bound concentrations. We used for the constrained minimisation the Matlab\textsuperscript{a} fmincon function (The MathWorks Inc., 2010). The solution $p$ of Equation (8) is plugged in Equation (5) to calculate the full grid spatial distribution of the pollutant concentration. Note that for simplicity of writing, the time dependence was not explicitly expressed in Equations (4)–(7) but in fact $C = C(t)$, the spatial distribution at time point $t$. Moreover, $p$ is also time dependent and a different parameters set $p(t)$ is recovered at each time point.

### 3.2. Assessment of the results

In addition to a qualitative inspection of the concentration maps produced by the proposed method, we quantitatively assessed the model performance using leave-one-out cross-validation. This process involves the removal of one observation from the set $C$, solving Equation (8) to find the model parameters and using the relevant row of Equation (7) to estimate the concentration value at the grid cell of the observation. Repeating this process by turn for all the observations at a time point results in a set of independently produced concentration estimations which can be compared to the corresponding observations. The process is then carried out consecutively for all time points. A second assessment of the model was by a comparison of the values observed in road-side stations (completely independent and not used in the model production) to those assigned by the model to the grid cells in which they reside. The assessment of the model performance against the observations is by a set of statistical measures recommended by the dispersion models statistics review of Simon et al. (2012). The measures are the mean bias (MB), the mean error (ME), the normalised mean error (NME), the ratio of modelled values within a factor of two of the observations (FAC2), and the coefficient of determination ($R^2$). Definitions of these measures and discussion of their properties are given by Simon et al. (2012). The proposed method’s performance was also compared to the performance of three benchmark methods (a) Assigning at each time point to all the grid locations the mean of the available observed concentrations. (b) A simple LUR that uses the traffic emission proxy $T$ and the wind speed as the explanatory variables and is calibrated at each time point by the same observations used by the proposed method. (c) A spatial interpolation scheme calculated for each time point. The three comparison methods are described in detail in an appendix given in the supplementary electronic Supporting Material.

### 4. Results

The proposed model is designed to run for consecutive time points of a study period. For this study we demonstrate only the results for 07:30, 12:00 and 17:00 on all the working days in 2009. Testing against data from road-side stations, half of which started operations only after the beginning of 2010, are for the same hours of the day in 2011. The model was run on a $500 \times 500$ m grid covering the traffic model road network. The reported concentrations are for the centres of the rectangular grid cells. The $500$ m spatial resolution was selected so that it provides exposure at the postal code scale but yet not too prohibitive in terms of computation time and computer memory requirements.

Fig. 3 presents three examples of half-hourly spatial NO\textsubscript{2} distribution which the model produced. The time points were selected for their different times of the day and very different ambient conditions. Fig. 3a shows the concentrations in an autumn morning, with a typically light wind from the south (180°, 2.2 m s\textsuperscript{-1}), Fig. 3b is for a spring noon with vigorous wind from the west (270°, 4.6 m s\textsuperscript{-1}) and Fig. 3c is for a winter evening with mild wind from the northwest (315°, 2.5 m s\textsuperscript{-1}). The small scale concentration variability is clearly seen in all three plots, with locally higher values along the roads. The spatial patterns in Fig. 3 reflect the ambient conditions at the time points for which the plots were produced and demonstrate how the simple modelling scheme captures the expected dispersion pattern. The main feature in the plots is of high concentrations in the Tel Aviv area and downwind from it. In Fig. 3a (wind from the south) relatively high concentrations can be seen anywhere along the coastline north of Tel Aviv. In Fig. 3b and c (onshore wind from the west and northwest, respectively) the concentrations along the coastline are very low.
and the plume of high NO2 values expands very clearly downwind, east and southeast from Tel Aviv. However, the concentrations in Fig. 3c, for 27 December at 17:00, are much higher then in Fig. 3b (17 April at noon), reflecting the fact that around sunset and close to the winter solstice the photochemical rates are low and the evening Tel Aviv rush hour emissions result in high concentrations down the wind.

Fig. 4 shows scatter plots of all the cross-validated half-hourly time points. The NO2 concentrations, simulated for grid points where ambient monitoring stations reside, are plotted against the corresponding observations, split by their time of the day. In all three plots the points are concentrated along the 1:1 ratio line, pointing to generally minimal bias in the model simulations. Most of the points are within a factor of two of the observations. The ones that are not are mainly unduly high modelled values of very low observed concentrations. The plot for 07:30 (Fig. 4a) shows also many too high modelled values at the higher concentration range. Table 1 provides a summary of the model performance measures calculated for the data points in Fig. 4. As could be figured out from Fig. 4, the MB is indeed very small, less than 10% of a ppb. However, the scatter around the 1:1 ratio line results in ME values of 6.4, 4.1, and 4.1 ppb for the 07:30, 12:00 and 17:00 time points, respectively. The higher ME in the morning may partially be due to poorer model performance in typical morning conditions, as depicted by the Fig. 5. The higher ME in the morning may partially be due to poorer model performance in typical morning conditions, as depicted by the Fig. 5. Moreover, the patterns of the performance measures are also clearly different at the two different times of the day. For example, the ME in the twin stations around location (188 East, 655 North) is very large in the morning (Fig. 5b), when the wind is usually from the south or southeast, but is average in the evening hour (Fig. 6b) when the wind usually blows from the west or northwest and brings to these stations emissions from the Tel Aviv area. In this case the unduly low morning modelled concentrations are probably a result of underestimation of the traffic volumes in the major highway along which the stations are situated (see Fig. 1).

To provide a visual impression of the temporal modelling which gave rise to the performance measures in Table 1 and Figs. 5 and 6, the observed concentrations and their modelled simulations in one of the northern Tel Aviv stations (179.6 East, 665.7 North) is given in Fig. 7. The seasonal differences are clearly captured by the modelled values, with low concentrations and low variability during the summer months and higher concentrations with very large variability in the winter. The general level and many of the true NO2 peaks are very well modelled in the noon and evening plots (Fig. 7b and c, respectively). The morning modelled values (Fig. 7a), although reasonably well correlated with their corresponding observations (see Fig. 5d), are generally overestimating the observations as depicted by the high positive bias in Fig. 5a.

Table 2 provides the summary of model performance measures calculated for observations in road-side stations whose data were not used in the modelling. The ME and NME are much larger compared to the corresponding values in Table 1 which shows the cross-validated measures for the ambient monitoring stations.
inadequate to capture the high NO$_2$ concentrations in a very close proximity to roads. It must be noted though that similar poor performance in simulating road-side concentrations was noted also by Beevers et al. (2012) who used a 20 $\times$ 20 m grid.

The half-hourly modelled concentrations discussed thus far can be integrated to provide the accumulated exposure during any desired time periods. Fig. 8 shows long term mean concentrations maps produced by integrations of the proposed model’s half-hourly output. The plots are for the 07:30, 12:00 and 17:00 time points during 2009. They were produced by taking for each grid cell centre the mean of its modelled concentration time series during the specified hour of the day. The plots capture the high variability in the chronic NO$_2$ exposure, which is expected in an urban environment. They also demonstrate the significant differences in NO$_2$ exposure during different hours of the day. The morning concentrations are generally higher and the spatial pattern is different from the noon and evening patterns. The differences in the concentration range and patterns are due to the different emission rates, photochemical rates, and the breeze cycle which determines most of the daily wind direction variability in the area. Fig. 9 shows scatter plots of the mean values of the cross-validated modelled values and their corresponding means of the observations during the 07:30, 12:00 and 17:00 h of 2009. Table 3 provides model performance measures for simulating the long-term mean of observations measured at the roadside stations (which were not used in the modelling). Figs. 8 and 9, and the information in Tables 3 and 4 enable assessment of the ability of the model to estimate the long term chronic exposure to NO$_2$. The maps in Fig. 8 are detailed enough for specifying the exposure at the postal code scale and there is very little bias in the cross-validated modelled values. The $R^2$ of the model is 0.62, in the middle of the range reported for NO$_2$ LUR models (Ryan and LeMasters, 2007; Hoek et al., 2008). It is clear though that using a relatively coarse grid and not incorporating road-side station data in the modelling resulted in limitations on the model’s ability to estimate the exposure at the very close proximity to roads.

5. Discussion

This study proposes the formulation of air pollution exposure estimation as an optimisation problem in which the parameters of a model, simulating the spatial air pollution distribution, are determined by fitting its output to observations. Carrying out the proposed methodology requires (a) data regarding the pollution sources, (b) a model to transform emissions (or their proxies) to pollutant concentrations, (c) a spatially representative sample of observations at the temporal resolution of the dominant dispersion processes and (d) a regression algorithm to fit the modelling output to the observations. The nonlinear nature of air pollution physics and chemistry most probably necessitates the use of a nonlinear

<table>
<thead>
<tr>
<th>Time of day</th>
<th>$n$</th>
<th>Mean (ppb)</th>
<th>MB (ppb)</th>
<th>ME (ppb)</th>
<th>NME (%)</th>
<th>FAC2</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>07:30</td>
<td>12,950</td>
<td>19.5</td>
<td>0.00</td>
<td>6.2</td>
<td>32</td>
<td>0.87</td>
<td>0.49</td>
</tr>
<tr>
<td>12:00</td>
<td>6831</td>
<td>11.3</td>
<td>0.01</td>
<td>4.7</td>
<td>41</td>
<td>0.74</td>
<td>0.55</td>
</tr>
<tr>
<td>17:00</td>
<td>6895</td>
<td>12.3</td>
<td>0.00</td>
<td>4.9</td>
<td>39</td>
<td>0.74</td>
<td>0.60</td>
</tr>
<tr>
<td>All</td>
<td>20,580</td>
<td>14.4</td>
<td>0.00</td>
<td>5.2</td>
<td>36</td>
<td>0.78</td>
<td>0.59</td>
</tr>
</tbody>
</table>
regression. The concept is demonstrated using very simple means for each of the components of the scheme but it is able to produce acute NO$_2$ exposure favourably compared with that obtained by comprehensive dispersion models (e.g., Beevers et al., 2012), and chronic exposure at a level commensurate with the LUR models discussed by the review papers of Ryan and LeMasters (2007) and Hoek et al. (2008). Application of the method to pollutants other than NO$_2$, especially non-primary ones, is more complicated and may require incorporation of chemical reaction calculations and a good pollutant background estimation. The advantages of high resolution spatiotemporal exposure are many and to our opinion worth the efforts. Exposure at a given location can be integrated for any desired period e.g., produce daily means for acute health outcome studies, weekly or monthly means for pregnancy studies, and annual or longer term means for chronic exposure. In addition, refining personal exposure by tracking a person’s mobility is made possible thanks to the option to calculate a person’s exposure as a time-weighted mean of the exposure at home, work, etc. The high temporal resolution may also enable investigations of the possible existence of threshold effects in the impact of air pollutants on health.

The proposed method’s performance was compared to that of three benchmark methods. The benchmarks’ performance measures are given in the appendix in the electronic Supporting Material. The exposure estimated by our data-driven model outperforms by a very large margin the exposure based on assigning to each location at each time point the mean of the simultaneous observations. This shows that the additional variability provided by the proposed method is indeed contributing a beneficial spatial variability and not additional spatial noise. Similar advantage was achieved in the comparison with the exposure produced by linear LUR models computed for each time point and using as explanatory

![Fig. 5. The statistical measures of model performance at the 07:30 h, calculated separately for each station. The measures’ values are presented as colour-coded diamonds located at the stations’ locations, with the colour coding key given in the colourbars. (a) Mean Bias (ppb), (b) Mean Error (ppb), (c) Normalised Mean Error, (d) Coefficient of Determination ($R^2$), (e) Ratio of modelled values within factor of two from the observations (FAC2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)](image1)

![Fig. 6. Like Fig. 5 but for the 17:00 h.](image2)
variables the same traffic emission proxy used by our model and the wind speed. This justifies our approach of embedding a dispersion model in a regression which is nonlinear and thus more demanding in terms of computing resources. The comparison of the proposed model’s cross-validated performance measures (Tables 1 and 3) to the corresponding ones obtained by a spatial interpolation (Tables 9 and 11 in the Supporting Material appendix) is quite balanced. However, the proposed model clearly outperforms the interpolation in estimating the exposure at the road-side stations whose data were not used by either scheme (compare Tables 2 and 3).

Table 2
Statistical measures of model performance in producing the half-hourly NO2 observations in 2011 at grid cells in which transportation stations reside. The table provides the number of observation and modelled value pairs (n), the mean observed concentration value (Mean), the mean bias (MB), mean error (ME), normalised mean error (NME), the ratio of modelled values within factor of two from their corresponding observations (FAC2) and the coefficient of determination ($R^2$).

<table>
<thead>
<tr>
<th>Time of day</th>
<th>n</th>
<th>Mean (ppb)</th>
<th>MB (ppb)</th>
<th>ME (ppb)</th>
<th>NME (%)</th>
<th>FAC2</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>07:30</td>
<td>1723</td>
<td>34.2</td>
<td>-7.8</td>
<td>10.1</td>
<td>30</td>
<td>0.88</td>
<td>0.20</td>
</tr>
<tr>
<td>12:00</td>
<td>1681</td>
<td>19.8</td>
<td>-6.6</td>
<td>7.9</td>
<td>40</td>
<td>0.71</td>
<td>0.56</td>
</tr>
<tr>
<td>17:00</td>
<td>1705</td>
<td>20.0</td>
<td>-6.1</td>
<td>7.9</td>
<td>40</td>
<td>0.71</td>
<td>0.60</td>
</tr>
<tr>
<td>All</td>
<td>5109</td>
<td>24.7</td>
<td>-7.0</td>
<td>8.7</td>
<td>35</td>
<td>0.77</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Fig. 7. The time series of observed and cross-validated modelled NO2 concentrations (green circles and black crosses, respectively) in station Antokolski (179.6 East, 665.7 North) in 2009. (a) The 07:30 time points. (b) The 12:00 time points. (c) The 17:00 time points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 8. The long term mean NO2 concentration in the study area. The concentrations are colour-coded, with the colour key given by the colourbars in ppb units. (a) Mean values of all the time points in 2009 at 07:30. (b) Mean values for of all the time points in 2009 at 12:00. (c) Mean Values for of all the time points in 2009 at 17:00. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Table 3
Statistical measures of model performance in producing the mean 2009 observation values using the corresponding means of the cross-validated half-hourly modelled concentrations. The table provides the number of observation-modelled value pairs (n), the mean observed concentration value (Mean), the mean bias (MB), mean error (ME), normalised mean error (NME), the ratio of modelled values within factor of two from their corresponding observations (FAC2) and the coefficient of determination ($R^2$).

<table>
<thead>
<tr>
<th>Time of day</th>
<th>n</th>
<th>Mean (ppb)</th>
<th>MB (ppb)</th>
<th>ME (ppb)</th>
<th>NME (%)</th>
<th>FAC2</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>07:30</td>
<td>25</td>
<td>19.5</td>
<td>-0.27</td>
<td>4.2</td>
<td>21</td>
<td>1.00</td>
<td>0.54</td>
</tr>
<tr>
<td>12:00</td>
<td>25</td>
<td>11.3</td>
<td>-0.09</td>
<td>2.2</td>
<td>20</td>
<td>0.96</td>
<td>0.49</td>
</tr>
<tr>
<td>17:00</td>
<td>25</td>
<td>12.3</td>
<td>-0.13</td>
<td>2.1</td>
<td>17</td>
<td>0.96</td>
<td>0.56</td>
</tr>
<tr>
<td>All</td>
<td>25</td>
<td>14.4</td>
<td>-0.16</td>
<td>2.5</td>
<td>18</td>
<td>1.00</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Fig. 9. Scatter plots of the annual mean observed vs. cross-validated modelled NO2 concentrations in the ambient monitoring stations during 2009. (a) Rush hour, 07:30. (b) Noon, 12:00. (c) Evening rush hour, 17:00.

Table 4
Statistical measures of model performance in producing the mean 2011 observation values at grid cells in which transportation stations reside using the corresponding means of the half-hourly modelled concentrations. The table provides the number of observation-modelled value pairs (n), the mean observed concentration value (Mean), the mean bias (MB), mean error (ME), normalised mean error (NME), the ratio of modelled values within factor of two from their corresponding observations (FAC2) and the coefficient of determination ($R^2$).

<table>
<thead>
<tr>
<th>Time of day</th>
<th>n</th>
<th>Mean (ppb)</th>
<th>MB (ppb)</th>
<th>ME (ppb)</th>
<th>NME (%)</th>
<th>FAC2</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>07:30</td>
<td>6</td>
<td>34.2</td>
<td>-7.8</td>
<td>7.8</td>
<td>23</td>
<td>1.0</td>
<td>0.14</td>
</tr>
<tr>
<td>12:00</td>
<td>6</td>
<td>19.8</td>
<td>-6.7</td>
<td>6.7</td>
<td>34</td>
<td>1.0</td>
<td>0.44</td>
</tr>
<tr>
<td>17:00</td>
<td>6</td>
<td>20.0</td>
<td>-6.2</td>
<td>6.3</td>
<td>31</td>
<td>1.0</td>
<td>0.46</td>
</tr>
<tr>
<td>All</td>
<td>6</td>
<td>24.7</td>
<td>-7.0</td>
<td>7.0</td>
<td>28</td>
<td>1.0</td>
<td>0.22</td>
</tr>
</tbody>
</table>

4 for the proposed model with Tables 10 and 12 in the Supporting Material for the interpolation). This advantage is a quantitative manifestation of the more physically-sensible approach of our modelling and the higher spatial variability in the maps which it produces (e.g., compare Figs. 3 to 5 in the Supporting Data appendix).

There are potentially many options to enhance our model’s performance, which can be relatively easy to incorporate. Refining the spatial grid resolution may yield a significantly better estimation in the areas where the traffic network is dense. Point source emissions from grid cells where large industrial plants exist can be integrated in the dispersion scheme and account for what is usually the second largest source of air pollution in modern urban cities. The traffic emission proxies could be refined using better traffic volume estimations and incorporating also vehicle speeds information. In the near future comprehensive specification of traffic intensity in a road network may become possible thanks to the ubiquity of GPS mounted smartphone. A more challenging and promising possible improvement, one that is actively sought by the authors, is incorporating a recursive scheme to minimise the modelling errors using information from past time points.

The advantage of the proposed scheme comes from the integration of various data sources via a regression scheme. The possible difficulty in obtaining all these components is a disadvantage. Traffic models, providing proxies for emissions, are available for many regions. For example, Henderson et al. (2007) used output of traffic model similar to the one we used (the EMME/2) as a traffic density input in their LUR. Data from cellular telephone networks is a very promising traffic information source which will soon be available almost anywhere. Lack of monitoring data at a sufficiently high spatial and temporal resolution is a more severe limitation. Our methodology relies on detailed source specification to provide the high spatial variability. However, to properly transform this input into pollutant concentrations a sample of observations that represents well the concentration distribution in the area must be available. Many urban areas, especially in developing countries, do not have adequate air quality monitoring network. Many examples of such networks do exist however (e.g., Henderson et al., 2007; Beelen et al., 2010; Beevers et al., 2012) and the emergence of cheap mobile monitoring devices may result in dense air quality network becoming more common. The lesser accuracy of such devices compared to established monitoring is less of a concern while incorporated in our method because their data are not used directly as in an interpolation scheme.

Additional issue to bear in mind is that running a nonlinear regression may be costly in terms of computer resources. For example, in our very simple scheme evaluating the full grid concentration field at each time point took approximately 7 s on a desktop machine with an Intel® Core2 processor and required at least 6 GB of random access memory (RAM). Carrying out careful cross-validation is of prime importance for a nonlinear regression
scheme with potentially many model parameters. In our case, with only five model parameters and no more than 25 observations, producing the cross-validation results for each time point took about 55 s. Adding complexity, and model parameters, or refining the grid may increase significantly the computation time and the RAM requirements. Moreover, a hidden danger with complex models is that they might overfit the observations. An inner cross-validation scheme, separate from the final cross-validation to test the model performance, may be needed in such cases in order to arrive at the optimal model parameters. The general cross validation function (Golub et al., 1979) has been successfully used in the past for that purpose to control the fit of nonlinear schemes to observations (Yuval, 2000). We believe that the technical obstacles in incorporating optimisation schemes in the field of air pollution exposure can be overcome and enhance our capabilities in carrying out air pollution epidemiological studies.

Acknowledgements

This study has been supported by the Technion Centre of Excellence in Exposure Sciences and Environmental Health. The authors would like to thank two anonymous reviewers for their constructive comments.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.atmosenv.2013.06.005.

References


Ryan, P.H., LeMasters, G.K., 2007. A review of land-use regression models for characterizing intraurban air pollution exposure. Inhalation Toxicology 19 (Suppl. 1), 127–133.


