Published in final edited form as:
Stat Med. 2010 January 15; 29(1): 116–129. doi:10.1002/sim.3732.

Integrated Exposure Modeling: A Model Using GIS and GLM

Theodore R. Holford[1,*,†], **Keita Ebisu**[2], **Lisa A. McKay**[1], **Janneane F. Gent**[1], **Elizabeth W. Triche**[3], **Michael B. Bracken**[1], and **Brian P. Leaderer**[1]

[1]Department of Epidemiology and Public Health, Yale School of Medicine New Haven, Connecticut 06520, U.S.A

[2]Yale School of Forestry and Environmental Studies, New Haven, Connecticut 06511, U.S.A

[3]Department of Community Health and Epidemiology, Brown University, Providence, Rhode Island 02912, U.S.A

Abstract

Traffic exhaust is a source of air contaminants that have adverse health effects. Quantification of traffic as an exposure variable is complicated by aerosol dispersion related to variation in layout of roads, traffic density, meteorology, and topography. A statistical model is presented which uses Geographic Information Systems (GIS) technology to incorporate variables into a generalized linear model that estimates distribution of traffic-related pollution. Exposure from a source is expressed as an integral of a function proportional to average daily traffic and a nonparametric dispersion function which takes the form of a step, polynomial or spline model. The method may be applied using standard regression techniques for fitting generalized linear models. Modifiers of pollutant dispersion such as wind direction, meteorology, and landscape features can also be included. Two examples are given to illustrate the method. The first employs data from a study in which NO_2 (a known pollutant from automobile exhaust) was monitored outside of 138 Connecticut homes, providing a model for estimating NO_2 exposure. In a second example, estimated levels of nitrogen dioxide (NO_2) from the model, as well as a separate spatial model, were used to analyze traffic-related health effects in a study of 761 infants.

Keywords

Traffic; Dispersion models; Geographic Information Systems; Generalized Linear Models; Splines

Exhaust from vehicular traffic is a source of several air contaminants, including nitrogen dioxide (NO_2) and carbon particulates from diesel engines that may exacerbate asthma symptoms [1, 2]. Traffic-related pollutant exposure at a residence depends on local traffic patterns and pollutant dispersion, which varies with distance from the source, meteorology and local landscape. Some epidemiological models specify exposure as a function of traffic density and a pollutant dispersion function that may be Gaussian [1, 3] or logarithmic [4]. Berhane *et al* [5] discuss a variety of approaches for analyzing effects of air pollution from traffic, including comprehensive models (MOBIL6 [6] and CALINE4 [7]) that incorporate detailed meteorological data (e.g., wind speed, wind direction, atmospheric stability and ambient temperature), topographical specifications (e.g., at grade, on a bridge or in a parking lot) and traffic variables (e.g., type of fuel, age of vehicle, speed of operation). While all of

[*]Correspondence to: Theodore R. Holford, Division of Biostatistics, Department of Epidemiology and Public Health, Yale School of Medicine, New Haven, CT 06520. U.S.A.
[†]theodore.holford@yale.edu

these factors may play a role in describing the dispersion of traffic-related pollutants, modeling can be challenging since data for this level of detail are generally not available.

The models mentioned above use a single measure of distance from residence to roadway, which does not capture varying traffic density due to patterns of intersecting roadways. We describe a framework that allows the use of generalized linear models (GLM) and generalized linear mixed models (GLMM) to investigate the association between traffic densities within a specified buffer surrounding a residence and a health response variable. By taking advantage of the spatial detail provided by Geographic Information Systems (GIS), our modeling framework not only allows for a more realistic pattern of highways, but enables inclusion of covariates that may modify pollutant levels, including wind direction, meteorology and topography.

While our model is similar to a distance-weighted approach, it does not assume that dispersion necessarily follows Gaussian decay function [1, 3]. Instead, it provides a mechanism for estimating the form for the relationship, including a nonparametric dispersion function. Segments of roadways are regarded as line sources, as opposed to point sources, which allows for inclusion of segment length, as well as location and orientation within the buffer. Importantly, our model also offers ways to incorporate meteorological or landscape features, factors that can modify exposures.

Point sources of exposure

The source of a pollutant can sometimes be represented as a single point, which is represented in GIS at longitude/latitude coordinates that have been projected onto a plane (e.g., Universal Transverse Mercator Projection or UTM) in order to correct for sphericity. An example of such a source is the accidental release of anthrax from a military compound in Sverdlovsk in 1979[8]. Individual exposure from such a source depends on location at a particular time, and an appropriate model must allow for dispersion. Distance from the source is an essential factor affecting dispersion, but one may also need to consider the modifying effects of prevailing winds, land use, and surrounding topography. An extensive literature discusses various approaches for refining intensity of exposure or risk surrounding a point and Lawson provides an excellent review and list of references for this work [9]. Many of these exposure models are intrinsically nonlinear, and they can include directional components that may involve the use of trigonometric functions of direction from the exposure source. In this paper we limit our discussion to intrinsically linear models in order to take advantage of the generalized linear models inferential framework.

To represent location, we use (x,y) coordinates where x represents the east-west distance from the origin and y the north-south distance. When considering a dispersion model, the origin might be considered to be either the location of the source or the location of the study subject. We introduce a prime to distinguish location when the pollutant source is the origin from the case where the residence is the origin. From a study subject located at $(0,0)$, the exposure source is given by (x,y). Similarly, taking the origin at the exposure source, $(0',0')$, the residence is at (x',y'). Notice that $x= -x'$ and $y=-y'$, so that one can readily translate from one coordinate system to the other, and the Euclidean distance between subject and source is identical from either perspective, only direction is reversed. It is convenient to think of dispersion from the perspective of the source, but when considering the cumulative effect of multiple points of exposure it is convenient to place the origin at the site of exposure recipient.

If isotropy applies (e.g., independent of wind direction), dispersion is independent of direction and can be represented as a function of distance, $\varphi(d)$. This gives rise to an exposure to an individual that is proportional to the level of pollutant released at the source and to the dispersion function, i.e., $c\ z\ \varphi(d)$, where z represents the level of release and c a constant that takes into account any modifier that could affect the amount of pollutant released into the environment. Without loss of generality, we can incorporate c into the dispersion function, so that the exposure to the individual is $z\varphi(d)$. In many instances the actual level of release is not known, so we estimate dispersion as a function of distance. The contribution of exposure from a given source on a health outcome can be estimated using a generalized linear model, in which the linear predictor is given by $\eta = \mathbf{X\beta} + z\varphi(d)$ where $\mathbf{X}$ includes a column of ones for the intercept and the remaining regressors are covariates that are to be controlled in the analysis, and $\boldsymbol{\beta}$ are unknown parameters to be estimated. It is common to use a parametric model to describe a dispersion function, $\varphi(d)$, such as the Gaussian dispersion model where $\varphi(d) \propto \exp\{-\gamma d^2\}$ [3] in which the decline in log concentration is proportion to distance squared and slope $-\gamma$. If γ is known, then calculating the exposure level for an individual is not difficult. We will describe a nonparametric model in which we estimate the dispersion function, $\varphi(d)$, as an alternative approach that does not require that the functional form be specified in advance.

If dispersion is represented as a step-function,

$$\begin{aligned}\varphi(d)=\gamma_k \quad & \text{if } \mathrm{D}_{k-1} \le d<D_k \\ =0 \quad & \text{otheriwse}\end{aligned}$$

then exposure would be identical for all points within each of two concentric circles centered at the source with diameters D_{k-1} and D_k, i.e., $z\gamma_k$. Estimates of the stepped exposure function can be obtained by including a set of covariates, $G_k = zg_k$, for observation in the generalized linear model where $g_k = I[D_{k-1} \le d < D_k]$, d is distance from the source and $I[\cdot]$ is an indicator function. The resulting estimate of the dispersion function would be $\hat{\varphi}(d) = \mathbf{g}'\hat{\boldsymbol{\gamma}}$ where $\mathbf{g}' = (g_1, g_2,\ldots g_k)$ is the vector of covariates and $\hat{\boldsymbol{\gamma}}' = (\gamma_1,\gamma_2,\ldots,\gamma_k)$ the vector of the corresponding parameter estimates obtained by fitting the model.

Within the framework of generalized linear models, one can also employ polynomial, $\varphi(d) = \gamma_0 + d\gamma_1 + d^2\gamma_2 + \ldots$, or spline functions, e.g., a cubic spline would be

$$\varphi(d)=\gamma_0+d\gamma_1+d^2\gamma_2+d^3\gamma_3+\sum\nolimits_l [d-\tau_l]_+^3\gamma_{3l}$$

where $[a]_+ = a$ for $a > 0$ and 0 otherwise, $\boldsymbol{\gamma}$ is a vector of unknown coefficients to be estimated and $\mathbf{g} = (1, d, d^2,\ldots)$ is the vector of corresponding functions of d that define $\varphi(d) = \mathbf{g}'\boldsymbol{\gamma}$. We can readily obtain estimates of the unknown parameters in the context of generalized linear models by including as covariates, $\mathbf{G} = z\mathbf{g}$. Alternatives to cubic splines may also be employed in a similar manner, including B-splines which are numerically superior because they are more nearly orthogonal [10–12]. Care is needed to avoid overfitting in the selection of knots, which can give rise to an estimated dispersion function with multiple peaks and valleys, but a penalized likelihood approach [13] and a method of knot selection such as the one suggested by Ruppert [14] offer approaches for dealing with these issues.

Multiple sources of a given pollutant can give rise to a cumulative exposure at a given location. If the ith source is distance d_i from away from the point of interest and z_i the concentration of pollutant emitted at this source, then the corresponding covariate vector defining the resulting dispersion from this source is $\mathbf{g}_i$. If each source has a different effect on the response, perhaps because the pollutant of interest is mixed with the other contaminant, then the combined contribution to the linear predictor would be $\sum_i z_i \mathbf{g}_i{}' \boldsymbol{\gamma}_i$. Simplification would result if $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}$ for $\forall i$, and the corresponding contribution to the linear predictor would combine corresponding covariates over the I sources, i.e., $\mathbf{G}\boldsymbol{\gamma} = \left(\sum_i z_i \mathbf{g}_i\right)' \boldsymbol{\gamma}$. However, health effects can be caused by more than one pollutant which can give rise to complexities that become difficult to analyze if chemicals react with each other or modify their effects on the health outcome.

Line sources of exposure

Traffic-related air pollution can be considered to result from a locus of points on lines representing roadways. These data are often presented as average daily traffic (ADT) estimates along highway segments, as shown by the highways displayed in Figure 1. Conceptually, the open circles on the highways in Figure 2 represent nodes dividing lines into segments with a common ADT. The ith segment ($i=1,\dots,I$) would be specified by starting and ending nodes, and when using GIS this would be presented as a line layer. To account for road curves or dispersion of a pollutant generated by a segment on exposure at a point, it may be necessary to divide segments into short subsegments. The pollution effect of a subsegment can be expressed as the product of ADT, subsegment length, and a dispersion function of distance to the residence.

If we divide the ith curve, C_i, into J_i subsegments represented by distances from the start node, these can be given by s_{i0}, s_{i1}, $s_{i2,\dots,}$ s_{iJ}. The contribution of pollution emitted by traffic at point s_{ij} on a highway to traffic-related exposure at a residence is given by the dispersion function, $\varphi(s) = \varphi(x, y)$ where (x, y) represents the coordinates for s with residence as the origin. An isotropic effect reduces the exposure from s to a function of distance, $\varphi(d)=\varphi\left(\sqrt{x^2+y^2}\right)$. Assuming level of traffic-related pollution is proportional to average daily traffic (ADT), given by Z_i for C_i, then the contribution at the residence from point s_{ij} on the segment would be given by $Z_i\varphi(s)$. To find the overall contribution of all segments within a buffer, one can cumulate line integrals over the segments,

$$\begin{aligned}\sum_i \int_{C_i} Z_i \varphi(s) ds &= \sum_i Z_i \int_{C_i} \varphi(s) ds \\ &\approx \sum_i Z_i \sum_{j=1}^{J_i} \varphi\left(s'_{ij}\right) \Delta s_{ij}\end{aligned} \quad (1)$$

where s'_{ij} is a middle point on the line between $s_{i(j-1)}$and s_{ij}, and Δs_{ij} is the length of the segment.

If dispersion is approximated by a step-function (see Figure 3), then the contribution associated with the kth step, i.e., segments where the distance (d) is $D_{k-1} \le d < D_k$, would be

$$\begin{aligned}\sum_i Z_i \int_{C_i} \varphi(s) ds &\approx \gamma_k \sum_i Z_i \sum_{\{j: D_{k-1} \le d_{ij} < D_k\}} \Delta s_{ij} \\ &= \gamma_k \left(\sum_i Z_i \{\text{length of segment } i \text{ in region}\}\right)\end{aligned} \quad (2)$$

In this case, the line integral reduces to the sum over all segments within concentric circles of the product of segment length times the corresponding ADT. The calculation is made once at the beginning of the analysis, and the resulting covariates are then entered into a generalized linear model [15] for estimation of the corresponding regression parameter that yields the estimated dispersion function.

Polynomial dispersion functions can also be considered in this setting, i.e.,

$$\begin{aligned}&\sum_i Z_i \int_{C_i} \varphi(s)ds \approx \sum_i Z_i \sum_{j=1}^{J_i} \left[\gamma_0 + d_{ij}\gamma_1 + d_{ij}^2\gamma_2 + \ldots\right] \Delta s_{ij} \\ &= \gamma_0 \left(\sum_i Z_i \sum_{j=1}^{J_i} \Delta s_{ij}\right) + \gamma_1 \left(\sum_i Z_i \sum_{j=1}^{J_i} d_{ij} \Delta s_{ij}\right) + \gamma_2 \left(\sum_i Z_i \sum_{j=1}^{J_i} d_{ij}^2 \Delta s_{ij}\right) + \ldots\end{aligned} \tag{3}$$

In this case the regressor variable associated with γ_p is

$$\sum_i Z_i \sum_{j=1}^{J_i} d_{ij}^p \Delta s_{ij} \tag{4}$$

which is again calculated once before fitting a model. Likewise, the approach can be similarly applied to a spline model of dispersion, for example, in the case of cubic splines, the line integral for the *i*th segment may be approximated by

$$\begin{aligned}\sum_i Z_i \int_{C_i} \varphi(s)ds \approx \gamma_0 &\left(\sum_i Z_i \sum_{j=1}^{J_i} \Delta s_{ij}\right) \\ &+ \gamma_1 \left(\sum_i Z_i \sum_{j=1}^{J_i} d_{ij} \Delta s_{ij}\right) \\ &+ \gamma_2 \left(\sum_i Z_i \sum_{j=1}^{J_i} d_{ij}^2 \Delta s_{ij}\right) \\ &+ \gamma_3 \left(\sum_i Z_i \sum_{j=1}^{J_i} d_{ij}^3 \Delta s_{ij}\right) \\ &+ \sum_l \gamma_{3l} \left(\sum_i Z_i \sum_{j=1}^{J_i} \left[d_{ij} - \tau_l\right]_+^3 \Delta s_{ij}\right)\end{aligned} \tag{5}$$

which reduces to individual components that are estimated at the start of the fitting process.

Within the context of GLM, covariates that modify the dispersion function can be included in our model by adding interactions with the corresponding parameters. For example, if **u** is a vector of variables that may modify dispersion, including spatially specific variables, then one such model would be $\varphi(d,\mathbf{u}) = \mathbf{g}(d,\mathbf{u})' \boldsymbol{\gamma}$. For example, to allow for directional variability or anisotropy one might introduce dummy indicators for direction to a highway point

$$\begin{array}{llll} U_1 = 1 & \text{if NW,} & =0 & \text{otherwise} \\ U_2 = 1 & \text{if NE,} & =0 & \text{otherwise} \\ U_3 = 1 & \text{if SE,} & =0 & \text{otherwise} \\ U_4 = 1 & \text{if SW,} & =0 & \text{otherwise} \end{array}$$

yielding the matrix $\mathbf{U}_{4\times 4}$. We can capture the directional variation in the effect by using a design matrix defined by the Kronecker product, $\mathbf{U}\otimes\mathbf{g}$, which results in directional variability in the dispersion function,

$$U_1\varphi_1(d)+U_2\varphi_2(d)+U_3\varphi_3(d)+U_4\varphi_4(d).$$

In this case, one would cumulate distances, or other elements needed for determining the line integral, separately for each quadrant surrounding the residence, which yields separate estimates of the dispersion function for each quadrant, $\varphi_k(d)$. If wind direction is known at a particular location, then this approach can be modified by orienting quadrants according to whether they are up- or downwind or to the side. Similarly, if a wind rose is available that provides information on duration and direction then this could be used in place of binary indicators of direction.

One can likewise introduce interaction terms for landscape features, another potential modifier of the dispersion function. By adding land-use covariates derived from satellite imagery or a topographical layer from the U.S. Geological Survey to the GIS, one can determine characteristics that affect traffic-related pollution dispersion. For example, effects of urban development, grassland, forest or water could be included using dummy variables to classify the point on the highway, the residential surroundings, or a line connecting the two. Suppose one considers the possibility that forest surrounding a residence may absorb a pollutant, and U represents the proportion of forested land within a buffer surrounding a residence, as determined by remote sensing. If reduction in exposure at a residence is directly proportional to forested land, then

$$\varphi(d,U)=(1-\delta(U))\mathbf{g}'\boldsymbol{\gamma}=\mathbf{g}'\boldsymbol{\gamma}-U\mathbf{g}'\boldsymbol{\gamma}_1$$

where $\boldsymbol{\gamma}_1$ represents the reduction in exposure resulting from forest. In reality, the modifying effect of landscape features on air pollution exposure may be much more complex than what is represented by the inclusion of interaction terms. Models used by atmospheric chemists are often mathematically complex which suggests that the data analyst must exercise considerable care when developing a statistical model that incorporates features that may change the dispersion of a pollutant.

Parameters defining the dispersion function can be estimated by introducing the terms in parentheses in equations (2), (3) or (5) along with any modifiers as regressors in a GLM. For example, if the effect of traffic is included along with other covariates, then the linear predictor would become

$$\eta=\mathbf{X}\boldsymbol{\beta}+\mathbf{G}\boldsymbol{\gamma}$$

where $\mathbf{X}$ and $\boldsymbol{\beta}$ are covariates and corresponding parameters that are not traffic related. Estimates of the traffic-related parameters, $\hat{\boldsymbol{\gamma}}$, can be used to estimate the dispersion function, $\hat{\varphi}(d)=\mathbf{g}(d)'\hat{\boldsymbol{\gamma}}$.

Area sources of exposure

Area sources of environmental exposure can arise from the spreading of a potential pollutant over a region, e.g., in agriculture the dispersion of pollutants following crop dusting.

Alternatively, area exposures may be associated with particular land use, e.g., habitats particularly conducive to a disease vector. For instance, mosquitoes responsible for malaria transmission can thrive in standing water of rice fields and in this case the likelihood of travel to a blood meal from the point of origin would be related to distance and described in a model by a dispersion function. In this case, remote sensing data from satellite imagery can be used to identify a small area or pixel predominately classified as a rice field at the appropriate stage of growth. Let $Z_i(s)$ represent a binary indicator of land use classification at a pixel centered at coordinates $s=(x,y)$ using the reference location (e.g., residence) as the origin and $\varphi(s)$ is the dispersion function for point s, then the cumulative exposure is represented integrating over the buffer, A

$$\int_A Z_i(s)\varphi(s)ds \approx \sum_j Z_i(s_j)\varphi(s_j)\Delta s_j$$

where Δs represents the area of the pixel. As with the line integrals, unknown parameters in a dispersion function represented by a linear model are cumulative sums of the product of pixel indicators, Z_i and a coefficient for the parameter in the dispersion function which can be estimated at the start of the analysis.

Generalized Linear Mixed Model with Integrated Exposures

To estimate the dispersion function, we introduce the regressors represented by row vector **G**, arising from the integration of the product of the specified dispersion function and the intensity at the source defined by row vector **Z**, into a generalized linear model that may include other covariates, **X**. Because of the spatial nature of the data, a generalized linear mixed model (GLMM) should be considered in order to account for random error not specified by the probability density of the response. A hierarchical error model may include both independent and spatially correlated elements. Let $Y = \mu(\mathbf{X}, \mathbf{Z}) + \varepsilon$ represent the response, where ε is random error with a distribution belonging to the exponential class specified by McCullagh and Nelder [15]. The link function of the mean for the i-th observation yields the linear predictor,

$$\eta_i=g(\mu_i)=\mathbf{X}_i'\boldsymbol{\beta}+\mathbf{G}_i'\boldsymbol{\gamma}+\upsilon_i$$

where $\mathbf{X}_\mathrm{i}$ are covariates that are not related to dispersion and $\mathbf{G}_\mathrm{i}$ are regressor variables described in the previous section which were derived to characterize the integrated dispersion function, with corresponding unknown parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ that must be estimated. The additional hierarchical error term for the i-th individual, υ_i, may include random elements that are independent, $\boldsymbol{\upsilon} \sim \mathrm{MVN}(\mathbf{0},\sigma^2\mathbf{I})$ and/or are spatially correlated, $\boldsymbol{\upsilon} \sim \mathrm{MVN}(\mathbf{0},\sigma^2\boldsymbol{\Sigma})$, as described by Breslow and Clayton [16].

Development of statistical methods for fitting GLMM continues to be an active area of research which was motivated in part by Breslow and Clayton's seminal paper. The text by Lawson [9] provides an excellent overview of techniques for handling spatially correlated responses which may be caused in part by point sources of pollution, and these can be readily applied in the context of exposure derived from integration of line or area sources. Generally, the strategies include specification of the covariance structure which is then included in a variation of likelihood based inference, such as the penalized quasi-likelihood approximation [16]. For example, Braimoh and Onishi [17] make use of a restricted maximum likelihood technique to obtain estimates of both fixed effects in the model, and parameters that specify the structure of the covariance matrix. Bayesian inference, in which Markov Chain Monte Carlo techniques are used to obtain samples from the posterior distribution of model parameters, offer the possibility of even more flexible GLMMs. Boyd *et al* [18] compared several approaches to the analysis of geographically referenced observation, including a Bayesian hierarchical approach that allowed for correlated errors in a GLMM.

In our first example, we apply the integrated exposure method to predict a pollutant level that is known to result primarily from traffic. Our objective is to demonstrate that the method does provide reasonable estimates and that we can validate the estimate. The model used is a standard regression model which does not make use of other aspects of the more general GLM. This example also illustrates how our approach can be used to estimate a particular traffic-related exposure where a direct measure of the pollutant was not obtained. Our second example demonstrates the approach using an ordered logistic model for a health response where the exposure estimate generated is used to examine the association between of this pollutant and respiratory symptoms.

Example 1—Estimating Distribution of Traffic-Related Pollution

A Connecticut study investigated the effects of acid aerosol exposure on respiratory symptoms in infants for which NO_2 levels were measured outside each subject's residence [19]. The primary source of outdoor NO_2 is thought to be vehicular engine exhaust, and thus related to ADT[20]. A full description of the health study methods is provided elsewhere [19, 21]. Families were recruited from mothers delivering babies at seven Connecticut hospitals between 1993 and 1996. Of 138 families enrolled, 129 had outdoor NO_2 measurements and 126 (97.7%) were successfully geocoded. Point locations for the residences were obtained by geocoding each address against ESRI's® Streets USA database [22]. A Universal Transverse Mercator (UTM) projection was used to determine appropriate distance for Connecticut latitudes.

Traffic volume data available for Connecticut consists of estimated ADT volumes for segments of the highway where significant traffic volume changes occur [23]. ADT associated with a section of highway estimates the number of vehicles passing through that section on an average day for both directions combined (except for on- and off-ramps, or roadways designated one-way). Data are collected only on highways maintained by the State Department of Transportation (DOT), and we used data for the year 2000. Lengths of the 5197 highway segments range from 0.01–7.64 mi. The median is 0.46 miles and the quartile range is 0.20–1.00 miles. Visits to the DOT revealed that estimates for each segment are based on surveys at selected points on highways conducted every three years on a rotating basis. By studying a map that includes the survey points, major arteries were identified as locations where change in volume were likely to have occurred, thus defining the highway segment for which the estimate at the traffic survey point applies. ADT is reported in this form for 17 of the lower 48 states, and one can use the approach described to construct similar files for those states that only report data at monitoring stations.

Data are reported for highway segments and include segment length (in miles) making these data particularly well-suited for use in a line source model. Prior to merging with traffic density, highway network data were converted from a simple polyline system to a measured route system, so that traffic events could be associated with a highway section according to where those events occurred. Total road length was established using beginning and ending mile values calibrated to values determined by the highway network. Additional adjustment points (such as the mile at which a road intersects another feature) were used to further calibrate each route. For this project, measurements at every town boundary and major road intersection between beginning and end points were used in ESRI®'s GIS software, Arcview®. Traffic data were added using mile-to-mile measurements associated with each highway segment, thus providing the ADT value for each segment.

The line integral calculation used an approximation in which highway segments were divided into 50 m subsegments. While the integral approximation would generally improve as the length of the subsegments go to 0, the amount of data extracted from GIS for each subject would quickly grow, thus increasing the calculation effort. Part of the rationale for selecting 50 m subsegments was heuristic and based on limitations resulting from the GIS representation of highways by polylines. Interstate highways have huge traffic volume compared to other roads, and the width of one lane is 12 ft. A highway in an urban area with at least three lanes in each direction would be at least 72 ft or 22 m of driving surface and the line feature would also include the median so the overall width of the road would usually be 25 m or more. In rural areas, these highways usually have two lanes in each direction, but the median is much wider. Given that GIS uses a line to represent a highway, we felt that there was not sufficient precision to be gained by using subsegments smaller than twice the width of the major sources of traffic exposure. As a second calculation to confirm our choice of 50 m subsegments, we performed numerical integration assuming Gaussian dispersion over a line passing through the center of a 2 km buffer, where 95% of pollutant at the source would have been dispersed by 250 m, 500 m and 1000 m, which cover the range found in this analysis. The definite integral calculated using a numerical approximation with 50 m subsegments was within .005% of the value found using 1 cm subsegments.

The method used to determine NO_2 levels outside of participants' homes was passive sampling using Palmes tubes [24]. At the enrollment home visit, the NO_2 monitoring tube was placed in an inverted funnel-shaped metal weather protector, then hung from a tree branch or outdoor clothes line at least 5 feet off the ground and as close to the home as possible. The monitor was left in place for 10 to 14 days.

The response used in the model, Y, is the level of NO_2 measured outside a residence and we assume that the errors about the mean are independent and $N(0,\sigma^2)$. The effect of traffic density within a 2000 m buffer surrounding the residence was used in three models for the dispersion function (step, polynomial, and regression spline) by creating the corresponding regressor variables and using the resulting parameters to estimate the dispersion function, $\hat{\varphi}(d)$. *Step-function*— Initially, we considered a single step (constant) model in which all highway segments within the 2000 m buffer contribute equally (Figure 4) and the effect for distances farther than 2000 m are constrained to 0. A significant association with outdoor level of NO_2 was found ($F_{1,112}$=93.09, P<.0001) with R^2 =.454 . To determine whether the effect varied by distance to highway, a model with the steps occurring at 400, 800, 1200, and 1600 m was fitted by creating regressor variables, **G**, defined in equation (2). A plot of residuals against the fitted value indicated the adequacy of the constant variance assumption, and a normal probability plot indicated that this distribution function was appropriate. In addition, a variogram did not indicate spatial correlation among the errors for the distances between residences observed in these data, which indicates that the independence assumption is reasonable in this example.

As expected, the highest step for the estimated dispersion function, (least dispersion of NO_2) occurs in the first 400 m (Figure 4). An overall significance test for this model indicates there is an association between traffic density and NO_2 ($F_{5,108}$=22.47, P<.0001, R^2 =.510). Compared to the constant dispersion model, using additional steps significantly improves the fit ($F_{4,108}$=3.08, P=.019), which suggests effect variability within the 2000 m buffer, but the increases with distance are well within the precision of the estimates as indicated by the confidence limits. Predicted estimates of NO_2 can be obtained by integrating the highway lines as in equation (1). Figure 1 shows a map in which the background was color coded according to the results from this integration for a region that includes New Haven County. The dots locating residences have also been color coded by observed level of NO_2, and it is clear that sites with high (or low) observed levels predominate in the areas predicted to be high (or low) by our model.

To assess isotropy for NO_2 dispersion, the buffer surrounding each residence was divided into quadrants along the north/south and east/west axes, and creating regressor variables defined in equation (2) for each one. Estimates of the dispersion function by quadrant are shown in Figure 5. An overall test of variability of effect among the four quadrants was of borderline significance ($F_{15,93}$=1.64, P=.078), which may be due in part to multiple parameters diluting the focus of the test. The largest positive effect is in the first 400 m of the SW quadrant (Figure 5), which represents highways that are generally upwind (t_{93}=1.95, P=.054). The estimated NE (downwind and away from the residence) effect for 0–400 m is negative, although not significantly different from 0 (t_{93}= −1.59, P=.115). The NW and SE quadrants appear to have effects smaller than SW, although the small effects for 1600–2000 m were statistically significant (t_{93}=5.31, P<.001 and t_{93}=3.48, P<.001 respectively). This could result from the increasing width of the plume as one goes farther away, so that slight variation in wind direction would result in these highway segments exposing a residence.

Polynomial

A polynomial dispersion function was fitted using regressor variables created using equations (3) and (4). Order three appears to capture well the shape of the dispersion function implied by the step function (Figure 4), offering significant improvement over the constant model ($F_{3,109}$=3.34, P=.022, R^2=.500). A series of hierarchical tests revealed a significant quadratic effect only (t_{110}=3.10, P=.003) which suggests that a single monotone dispersion function may not provide the best description of these data.

Spline

For a spline model, we first considered linear splines or a broken line for the dispersion function, with knots set at 400 and 800 m using regressor variables defined in equation (5) (see Figure 4). This model included three more parameters than the constant model and provided marginally significant improvement in fit ($F_{3,109}$=3.73, P=.013, R^2=.505). We also fitted similar linear spline models with separate contributions for each quadrant (R^2=.566) (Figure 5). The strongest positive effects appear to be in the SW and SE quadrants, but the dispersion function in other quadrants take negative or implausible values, especially for the NE, although these are not statistically significant. We fitted a model with a linear spline for the SW quadrant only, with constants for the others (see Figure 5). The functions are now positive with a spike occurring within the first 400 m for the SW quadrant (R^2=.511).

In order to examine possible overfitting in our model we split our data into learning and validation sets. We randomly selected three quarters of the data (N=114) as the learning set in which we fitted the step function. The remaining one third (N=38) served as the validation set in which observed and estimated values were compared. For the learning set we obtained R^2=.52. For the validation set the correlation between observed and estimated values was .666 (r^2=.44), which is somewhat less than when the data are used in the parameter estimation, but still considerably better than results achieved using other approaches (Table 1).

Various approaches have been proposed for using GIS and ADT to determine associations with health risk, including distance to the closest highway [25, 26], ADT on the closest highway, ADT on the busiest highway within a buffer, distance-weighted measures assuming Gaussian dispersion, and the distance-weighted sum for all road segments within a buffer [1, 27]. Table 1 provides R^2 values that were obtained using some of these alternative measures in our data. The best of these models used the sum or the highest ADT, but R^2 was only about .1, and other models were considerably lower. None of these were close to .5, the values of R^2 from the model proposed here. Differences in performance measures may reflect differences in the methods of collecting traffic and site-specific exposure data and demonstrates the inherent danger of using a GIS exposure model without validation at the study site.

Example 2—Effects of Traffic on Respiratory Symptoms

The objectives of our second example are first to illustrate how our integrated exposure model can be directly applied by GLM to a health outcome. This can be useful if one does not wish to specify a particular pollutant as the putative agent affecting health. Secondly, we demonstrate how our technique can be used to analyze the effect of geographically varying environmental exposures. This is accomplished by employing the model developed in the first example (above) then using it in a different study where individual pollution levels were not determined.

The second study, conducted between 1997 through 1999, recruited 1002 infants from families residing in CT and southwestern MA who also had at least one sibling with physician-diagnosed asthma [28–30]. Of infants enrolled, 833 residing in CT were selected for the current analysis since traffic data on road segments were not available for MA. Addresses and respiratory symptom severity scores could not be determined for 20 subjects, leaving 811. Of these, 47 had less than seven months of symptom information and three had missing confounding variables leaving 761 for analysis. The study was approved by the Human Investigations Committee of Yale University, and the mothers of study subjects assented prior to participation.

A summary of the demographic characteristics of the subjects is provided in Table 2. Individuals who identified themselves as black or Hispanic make up 35.7% of this population. The majority of mothers, 87.8%, had at least a high school education, and 58.8% of the children had more than one sibling.

Mothers recorded their infants' daily respiratory symptoms (wheeze and persistent cough) using a calendar provided [28]. Quarterly reports of the number of days for each symptom were collected by a phone call to the mother. We report here the results for days of wheeze in the first year of life categorized as 0 if none, 1 if 1–30 days and 2 if more than 30 days were reported for the year. This ordinal response was fitted to an ordered logistic model using PROC LOGISTIC in SAS® which assumes that the responses are independent, which is consistent with spatial independence that we observed in the first example. We directly estimated the dispersion function from traffic exposure associated with wheeze severity using 500 m steps within a 2000 m buffer and defined regressor variables using equation (2). In Table 3, odds ratios are shown for wheeze severity category at a given level or higher compared to lower severity per 100,000 vehicle kilometers within a specified buffer using areas farther than 2000 m from a residence at the reference. The estimated unadjusted OR for wheeze for a highway within 500 m of a residence is 0.79 for 100,000 vehicle km (95% CT=0.49–1.25). Likewise, the estimate for the 501–1000 step yields an estimated OR of 1.20 (95% CI=1.01–1.43) per 100,000 vehicle km. For the 1000–1500 m and 1501–2000 m steps, the OR's are close to 1, which suggests a return to the background level, i.e., distances more than 2000 m away. Adjusting for the covariates had little effect on the estimates.

As an alternative analysis, we also considered the association of wheeze with NO_2, a traffic-related pollutant. In Example 1, we described the use of the integrated exposure model in a separate study to estimate NO_2 levels across Connecticut. This model was used to estimate the outdoor level of NO_2 for each subject's residence, $\hat{X}_i$, as well as the corresponding error variance associated with predicting each individual exposure, s_i^2. Carroll et al [31] provide an excellent discussion of the various techniques that have been developed for dealing with covariates such as exposure estimates that have been determined with measurement error. The techniques include regression calibration [32–34], simulation extraction (SIMEX) [35] and Bayesian inference [36]. In order to avoid biased estimates of association with disease outcome that would arise from using each fitted value, we use $\hat{X}_i + e_i$ where $e_i \sim N\left(0, s_i^2\right)$ was randomly generated. We repeated this process 1000 times to provide multiple imputations, i.e., regression calibration estimates of exposure [31, 33, 37]. The 1000 data sets with randomly imputed NO_2 exposures values were fitted with an ordered logistic model with wheeze severity as the outcome using PROC LOGISTIC in SAS® [38]. The reported model parameters are the mean of the 1000 estimates, and the standard errors are $Var(\bar{\beta}) = 1000^{-1}\Sigma_i[Var(\hat{\beta}_i)] + 999^{-1}\Sigma_i(\hat{\beta}_i - \bar{\beta})^2$.

Quartiles for the estimated level of NO_2 at a residence obtained in Example 1 are 11.18, 15.57 and 20.20 ppb. Table 3 provides odds ratios for wheeze severity at a particular level or higher compared to a lower score using the ordered logistic regression model. Both unadjusted and adjusted results suggest a dose response relationship with NO_2, although adjusting for ethnicity, gender, mother's education and number of siblings reduced the estimated OR for fourth compared to first quartiles from 1.95 to 1.62. When NO_2 was included in the model as a continuous variable, the unadjusted analysis for the effect on wheeze resulted in an estimated OR for 10 ppb change to be 1.35 (95% CI=1.12–1.63), and the adjusted analysis yielded an OR for 10 ppb change of 1.22 (95% CI=0.98–1.50), putting the lower bound below the null value. Adjusting for error in the modeled exposure measure using the multiple imputation approach, we obtained results shown in Table 3, and it appears that this had little effect on either the estimates of association or the corresponding confidence intervals.

Discussion

The proposed methodology provides a way to estimate effects of exposure to traffic-related pollution using publicly available data. We describe and demonstrate models that can be used to analyze the effect of traffic density on any outcome, including disease risk. Our approach is related to the use of a distance weight that allows for dispersion, but does not require that dispersion weights be known. Estimating the dispersion function does not impose a particular model, such as Gaussian. For example, the distribution of particulate matter may differ from the dispersion of a gas like NO_2, so the same distance weights may not apply. In addition, there may be variability due to regional factors such as prevailing winds and/or landscape features which are not captured in some of the exposure dispersion models in current use. Our method yields direct estimates of the dispersion function using a step, a polynomial or a spline model. To fit these models, it is necessary to first calculate specified line integrals, which account for roadway patterns, distance and direction to residence, and other landscape features. This is accomplished prior to model fitting, then relevant covariates can be introduced into a generalized linear model, permitting both continuous and categorical representations of disease status. These calculations are readily performed using available GIS software, such as ArcGIS®.

Using GIS to estimate exposure is not always ideal and in some instances personal monitors may be better suited for the needs of a study [39]. An advantage of using GIS to estimate exposure in an epidemiological study can be to reduce cost. For example, an investigator might use a two-stage design in which more expensive personal monitors are used on a subset along with the necessary geographic coordinates. These data would then be used to estimate exposures for the subjects who were not monitored, but do provide spatial information. In addition, this approach would provide a means of generating estimates of exposure in which relevant address information is available but the adverse health event has already occurred.

We have limited our discussion to intrinsically linear dispersion models, but studies of the dispersion of air pollution from point sources suggest that nonlinear models may be more appropriate [9]. Further work is needed to develop statistical tools for fitting these models because integrating a function that involves additive distance-dependent contributions would no longer be relevant. Instead, one would need to integrate the dispersion function at each step. A Bayesian approach that employs Markov Chain Monte Carlo techniques offers considerable flexibility for GLMMs, but in the present context computational needs required if the numerical integration must be performed at each step requires further study.

English et al. [1] geocoded residential addresses in a case-control study of the effect of traffic flow on clinic visits for asthma symptoms. Citing results from dispersion models[40, 41], they assumed that a 80–90% decay in traffic-related pollutant concentration occurs between 150 to 200 m of a roadway, and thus developed exposure measures based on average daily traffic (ADT) within a 167.6 m. (550 ft) buffer surrounding a residence. Various estimates of total traffic exposure were considered: segments of roadway within the buffer with highest traffic volume, traffic volume on the road nearest the residence and the sum of ADT on all roads within the buffer. However, none of these measures captures the complexity of the distance and density relationship near a residence. Our estimated dispersion function resembles English's method[1] in that points close to a residence tend to have greatest effect as indicated by higher values for the dispersion function, but our data also suggest effects may persist well beyond a 167.6 m. buffer. In addition, we found evidence for anisotropic effects, which can easily be incorporated into our model. Models like MOBILE6 [6] and CALINE4 [7] can provide for effects of wind direction, but they also require detail that is often not available, such as presence of land fill, vehicle age and fuel type, thus requiring use of a default value. The studies used to develop MOBILE6 [6] and CALINE4 [7] were conducted in California during the 1970s, so may not be directly applicable to the current vehicle fleet in Connecticut.

Simply adding all estimates of ADT within a buffer ignores effects of highway length. While a sum weighted by segment length within a buffer offers an improved measure, it still fails to capture roadway patterns within a buffer, as well as the dispersion function of distance. The step and the spline models we propose for the dispersion function are nonparametric and flexible. A step function has an advantage in that it can capture even a complex function by making steps arbitrarily small, although precision can deteriorate quickly with small numbers of data points in each step. A spline function has an advantage in this regard because it allows for relatively smooth transitions, but splines can also be over-fitted if too many knots are used, giving the estimated dispersion function unrealistic peaks and valleys. Penalized likelihood methods can help to avoid this problem [14]. For example, one could apply such an approach by creating a relatively fine grid for the knots, then selecting among them using a stepwise selection procedure. A variable associated with a knot would be added if it resulted in the greatest increase in a penalized likelihood, where the penalty is proportional to number of dispersion parameters in the model. At each step, an included variable might also be dropped and perhaps replaced by a variable associated with a different knot. The process would be continued until the penalized likelihood no longer increased.

In these examples, we did not constrain the dispersion function to be positive in order to allow the form for the function to emerge without *a priori* restriction. One can extend our approach to models that only allow non-negative values, e.g., $\varphi(d) = \exp\{h(d)\}$ where $h(d)$ is either a step, polynomial or spline function that involves a log link function. However, the computational advantage of one-step data preparation prior to model fitting is no longer available. We have modeled data on measured residential (outdoor) NO_2, a traffic-related pollutant in which motor vehicles account for 55% of nitrogen oxides in the air [20], but the approach can also be used to examine associations between exposure to traffic itself and health outcomes.

Further work is needed to refine the dispersion function. A carefully designed study would be useful, not only to improve the estimates, but to determine the contribution of effect modifiers, including meteorology, season, topography, and land use. Improvement in prediction is limited by availability of existing traffic density data that for most roads is an annual average. Using data from an actual epidemiological study provides observations from points that are most relevant for health effects analyses. In addition, data from actual epidemiological studies cover a wider spatial range than typically available from atmospheric chemistry studies. On the other hand, an atmospheric chemistry study will typically provide far more temporal and meteorological precision (e.g., hourly measurements) than is practical in any epidemiological study described here. Further work is needed to develop statistical tools that make optimal use of the strengths of both approaches to data collection.

There are several limitations in the data used in these examples. Lack of more frequent measurements of traffic variability, resulting in additional exposure measurement error, can generally be expected to attenuate estimated associations with outcome. In addition, details on automobile, truck and bus traffic, i.e., distribution of diesel engines, which would be useful when studying health effects, was unavailable.

Data used in our example include traffic density from the year 2000 and NO_2 measurements from 1994. While overall level of traffic in Connecticut continues to increase, the level on different highway segments relative to the overall level tends to be less, although development in a particular geographic area can have a noticeable effect on traffic density. One would expect temporal concordance between traffic density and air pollution measurement dates to improve the model fit. Further work is needed to explore the properties of this approach, as well as to understand the additional effect of temporal variation. In spite of these limitations, our model explained about 50% of the variability in $NO_{2,}$ a pollutant by-product of automobile exhaust, a fit considerably better than alternative methods. Our model also conformed to expectations by yielding stronger associations with upwind highways. This model provides a practical tool for analyzing epidemiological studies by fitting generalized linear models.

We found evidence for an association between exposure to traffic and risk of wheeze during the first year of life. An exposure-response relationship was observed using NO_2 exposure estimates derived from a spatially-integrated model, although the effect was difficult to disentangle from ethnicity due to collinearity. This suggests that exposure to traffic could account for some of the socioeconomic differences in symptom severity that have been observed [42–51]. The US standard for ambient levels of NO_2 is 53 ppb [52] and the WHO standard is 20 ppb [53] which are higher than the levels for which we found significant effects on wheeze.

We also directly analyzed the effect of traffic exposure on wheezing by estimating a dispersion function associated with traffic density within specified buffers surrounding a residence. We found consistent risks to respiratory health associated with exposure to traffic 500–1000 m from residences. Estimates were unaffected by covariate adjustment. One might expect the pattern to be monotone, i.e. highest near the origin and decreasing to zero when dispersion is essentially complete, but a trend that is low near a highway then reaches a peak before declining could arise if a compound was the result of a reaction occurring after emission from a tailpipe. Approximately 95% of NO_x at the tailpipe is NO, which is subsequently converted to NO_2 in the presence of ozone, one example of such a delayed chemical reaction. In addition, many hydrocarbons at the tailpipe are highly reactive, which could delay production of a particular traffic-related compound with a putative health effect. A negative association for the closest distance category seems implausible and could be due to chance (the confidence interval includes the null value, but just barely), or selection bias either as a consequence of families of asthmatic children choosing to live away from a highway, or of families living nearest to highways being more likely to keep asthmatic children indoors. Only 360 subjects lived within 500 m of a highway, which resulted in estimates that had considerably lower precision than the other distance categories.

Estimating exposure to traffic using either a model of traffic density or a model of NO_2 suggests an effect of traffic on severity of respiratory symptoms in the first year of life. The NO_2 exposure estimate, which is only one of many traffic-related pollutants, showed an effect after 1000 m and was still positive 1600–2000 m from the highway. The elemental components of traffic-related particulates, for example, could disperse quite differently from gases. Further study using more detailed particle-pollution data is required to obtain better estimates of traffic-related factors that can affect asthma symptoms. It appears likely that a Gaussian dispersion model may not adequately characterize dispersion of factors related to traffic in the Connecticut area.

Health effects from air pollution can be caused by different contaminants, and exposure for individual contaminants can vary in both space and time. We have discussed an approach for analyzing geographic variation in air pollution exposure that may affect respiratory health, which provide exposure estimates averaged over time. However, an extensive literature also exists for temporal variation in air pollutants. Detailed data on temporal variation in pollution is usually only available at relatively few monitoring stations representative of a specific geographic space. Both of these approaches have similar challenges and further work is needed in order to enhance the study of pollutants that vary both temporally and spatially. Challenges documented in temporal models that are likely to also apply to spatial models are the effects of model selection and the uncertainty that necessarily exists in the choice of model [54, 55]. In the consideration of traffic exposure, there is also a parallel to the temporal lag in the effect [56] perhaps due to chemical reactions that occur after gases exit the tailpipe. Further work is needed to better understand the effects of these factors on studies that seek to estimate health effects resulting from exposure to environmental air pollutants.

Acknowledgments

This work was supported in part by grants ES05410, ES07456 and ES11013 from the National Institute of Environmental Health Sciences (NIEHS). We are grateful to the Connecticut Department of Transportation for making available the data on traffic density for highway segments.

References

1. English P, Neutra R, Scalf R, Sullivan M, Waller L, Zhu L. Examining associations between childhood asthma and traffic flow using a geographic information system. Environmental Health Perspectives. 1999; 107:761–767.10.2307/3434663 [PubMed: 10464078]

2. Friedman M, Powell K, Hutwagner L, Graham L, Teague WG. Impact of changes in transportation and commuting behaviors during the 1996 Summer Olympic Games in Atlanta on air quality and childhood asthma. Journal of the American Medical Association. 2001; 285:897–905.10.1001/jama. 285.7.897 [PubMed: 11180733]

3. Pearson RL, Wachtel H, Ebi KL. Distance-weighted traffic density in proximity to a home is a risk factor for leukemia and other childhood cancers. Journal of the Air and Waste Management Association. 2000; 50:175–180. [PubMed: 10680346]

4. Pleijel H, Karlsson GP, Gerdin EB. On the logarithmic relationship between NO_2 concentration and the distance from a highroad. Science of the Total Environment. 2004; 332:261–264.10.1016/ j.scitotenv.2004.03.020 [PubMed: 15336908]

5. Berhane K, Gauderman WJ, Stram DO, Thomas DC. Statistical issues in studies of the long-term effects of air pollution: The Southern California Children's Health Study. Statistical Science. 2004; 19:414–449.10.1214/088342304000000413

6. US Environmental Protection Agency. User's Guide to MOBILE6.1 and MOBILE6.2: Mobile Source of Emission Factor Model. US EPH National Vehicle and Fuel Emissions Laboratory; Ann Arbor, MI: 2003.

7. Benson, P. CALINE4 - A dispersion model for predicting air pollution concentration near roadways. Office of Transportation Laboratory, California Department of Transportation; Sacramento, CA: 1989.

8. Guillemin, J. Antrax: The Investigation of a Deadly Outbreak. University of California Press; Berkeley: 1999.

9. Lawson, AB. Statistical Methods in Spatial Epidemiology. 2. John Wiley & Sons, Ltd; Chichester, England: 2006.

10. de Boor, C. A Practical Guide to Splines. Springer-Verlag; New York: 1978.

11. Hastie T, Tibshirani R. Generalized additive models. Statistical Science. 1986; 1:297–318.10.1214/ ss/1177013604

12. Hastie T, Tibshirani R. Generalized additive models for medical research. Statistical Methods in Medical Research. 1995; 4:187–196.10.1177/096228029500400302 [PubMed: 8548102]

13. O'Sullivan F, Yandell BS, Raynor WJ. Automatic smoothing of regression functions in generalized linear models. Journal of the Americal Statistical Association. 1986; 81:96–103.10.2307/2287973

14. Ruppert D. Selecting the number of knots for penalized splines. Journal of Computational & Graphical Statistics. 2002; 11:735–757.10.1198/106186002853

15. McCullagh, P.; Nelder, JA. Generalized Linear Models. 2. Chapman and Hall; London: 1989.

16. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993; 88:9–25.10.2307/2290687

17. Braimoh AK, Onishi T. Geostatistical techniques for incorporating spatial correlation into land use change models. International Journal of Applied Earth Observation. 2007; 9:438–446.10.1016/ j.jag.2007.02.005

18. Boyd HA, Flanders WD, Addiss DG, Waller LA. Residual spatial correlation between geographically referenced observations: A Bayesian hierarchical modeling approach. Epidemiology. 2005; 16:532–541.10.1097/01.ede.0000164558.73773.9c [PubMed: 15951672]

19. Triche E, Belanger K, Beckett W, Bracken M, Holford T, Gent J, Jankun T, McSharry J, Leaderer B. Infant respiratory symptoms associated with indoor heating sources. American Journal of Respiratory and Critical Care Medicine. 2002; 166:1105–1111.10.1164/rccm.2202014 [PubMed: 12379555]

20. US Environmental Protection Agency. Evaluating Ozone Control Programs in the Eastern United States: Focus on the NO_x Budget Trading Program, 2004. Environmental Protection Agency; Washington, DC: 2005.

21. Pettigrew MM, Gent JF, Triche EW, Belanger KD, Bracken MB, Leaderer BP. Infant otitis media and the use of secondary heating sources. Epidemiology. 2004; 15:13–20.10.1097/01.ede. 0000101292.41006.2e [PubMed: 14712142]

22. ESRI. Street Map USA. Environmental Systems Research Institute; Redlands, CA: 2003.

23. Connecticut Department of Transportation. 2000 Traffic Volumes, State Maintained Highway Network. Bureau of Policy and Planning; 2001.

24. Palmes ED, Gunnison AF, DiMattio J, Tomczyk C. Personal sampler for nitrogen dioxide. American Industrial Hygene Association Journal. 1976; 37:570–577.10.1080/0002889768507522

25. Hoek G, Fischer P, Van Den Brandt P, Goldbohm S, Brunekreef B. Estimation of long-term average exposure to outdoor air pollution for a cohort study on mortality. Journal of Exposure Analysis and Environmental Epidemiology. 2001; 11:459–469.10.1038/sj.jea.7500189 [PubMed: 11791163]

26. Hoek G, Brunekreef B, Goldbohm S, Fischer P, van den Brandt PA. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. The Lancet. 2002; 360:1203–1209.10.1016/S0140-6736(02)11280-3

27. Wilhelm M, Ritz B. Residential proximity to traffic and adverse birth outcomes in Los Angeles County, California, 1994–1996. Environmental Health Perspectives. 2003; 111:207–216. [PubMed: 12573907]

28. Belanger K, Beckett W, Triche E, Bracken M, Holford T, Ren P, McSharry J-E, Gold D, Platts-Mills T, Leaderer B. Symptoms of wheeze and persistent cough in the first year of life: associations with indoor allergens, air contaminants and maternal history of asthma. American Journal of Epidemiology. 2003; 158:195–202.10.1093/aje/kwg148 [PubMed: 12882940]

29. van Strien RT, Gent JF, Belanger K, Triche E, Bracken MB, Leaderer BP. Exposure to NO_2 and nitrous acid and respiratory symptoms in the first year of life. Epidemiology. 2004; 15:471–478.10.1097/01.ede.0000129511.61698.d8 [PubMed: 15232409]

30. Gent JF, Ren P, Belanger K, Triche E, Bracken MB, Holford TR, Leaderer BP. Levels of household mold associated with respiratory symptoms in the first year of life in a cohort at risk for asthma. Environmental Health Perspectives. 2002; 110:A781–A786. [PubMed: 12460818]

31. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, C. Measurement Error in Nonlinear Models. 2. Chapman & Hall/CRC; Boca Raton: 2006.

32. Carrol RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. Journal of the American Statistical Association. 1990; 85:652–663.10.2307/2290000

33. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. American Journal of Epidemiology. 1990; 132:734–745. [PubMed: 2403114]

34. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Statistics in Medicine. 1989; 8:1051–1069.10.1002/sim.4780080905 [PubMed: 2799131]

35. Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. Journal of the Americal Statistical Association. 1994; 89:1314–1328.10.2307/2290994

36. Richardson, S. Measurement error. In: Gilks, WR.; Richardson, S.; Spiegelhalter, DJ., editors. Measurement error. Chapman & Hall; London: 1996.

37. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. 2. John Wiley & Sons, Inc; Hoboken, NJ: 2002.

38. SAS Institute Inc. SAS OnlineDoc®*9.1.3.*. 4. SAS Institute Inc; Cary, NC: 2004.

39. Briggs, DJ. Exposure assessment. In: Elliot, P.; Wakefield, JC.; Best, NG.; Briggs, DJ., editors. Exposure assessment. Oxford University Press; Oxford: 2000.

40. Versluis AH. Methodology for predicting vehicle emissions on motorways and their impact on air quality in the Netherlands. Science of the Total Environment. 1994; 146/147:359–364.10.1016/0048-9697(94)90257-7

41. Fraigneau YC, Gonzalez M, Coppalle A. Dispersion and chemical reaction of a pollutant near a motorway. Science of the Total Environment. 1995; 169:83–91.10.1016/0048-9697(95)04636-F

42. Mannino DM, Homa DM, Pertowski CA, Ashizaxa A, Nixon LL, Johnson CA, Ball LB, Jack E, Kang DS. Surveillance for asthma prevalence-United States, 1960–1995. Morbidity and Mortality Weekly Report, CDC Surveillance Summaries, 1998. 1998; 47:1–27.

43. Carr W, Zeitel L, Weiss K. Variations in asthma hospitalizations and deaths in New York City. American Journal of Public Health. 1992; 82:59–65.10.2105/AJPH.82.1.59 [PubMed: 1536336]

44. Gergen P, Mullally D, Evans RI. Changing patterns of asthma hospitalization among children: 1979–1987. Journal of the American Medical Association. 1990; 264:1688–1692.10.1001/jama.264.13.1688 [PubMed: 2398608]

45. Weiss KB, Wagener DK. Changing patterns of asthma mortality: Identifying target populations at high risk. Journal of the American Medical Association. 1990; 264:1683–1687.10.1001/jama.264.13.1683 [PubMed: 2398607]

46. Weiss KB, Wagener DK. Asthma surveillance in the United States. Chest. 1990; 95:179S–184S. [PubMed: 2226006]

47. Weitzman M, Gortmaker SL, Sobol AM, Perrin JM. Recent trends in the prevalence and severity of childhood asthma. Journal of the American Medical Association. 1992; 268:2673–2677.10.1001/jama.268.19.2673 [PubMed: 1304735]

48. Schwartz J, Gold D, Dockery DW, Weiss ST, Spiezer FE. Predictors of asthma and persistent wheeze in a national sample of children in the United States: Association with social class, perinatal events, and race. American Reviews of Respirtory Disease. 1990; 142:555–562.

49. Gold DR, Rotinitzky A, Damokosh AI, Ware JH, Speizer FE, Ferris BG Jr, Dockery DW. Race and gender differences in respiratory illness prevalence and their relationship to environmental exposures in children 7–14 years of age. American Review of Respiratory Disease. 1993; 148:10–18. [PubMed: 8317784]

50. Evans R III. Asthma among minority children: A growing problem. Chest. 1992; 101:368S–371S. [PubMed: 1591933]

51. Wissow LS, Gittelsohn AM, Szklo M, Starfield B, Mussman M. Poverty, race and hospitalization for childhood asthma. American Journal of Public Health. 1988; 78:777–782.10.2105/AJPH.78.7.777 [PubMed: 3381951]

52. US Environmental Protection Agency. AIRTrends 1995 Summary: Nitrogen Dioxide (NO_2). 2007

53. World Health Organization. WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide. WHO Press; Geneva: Switzerland: 2006. p. 27

54. Dominici F, Wang C, Crainiceanu C, Parmigiani G. Model selection and health effect estimatioin in environmental epidemiology. Epidemiology. 2008; 19:558–560.10.1097/EDE.0b013e31817307dc [PubMed: 18552590]

55. Peng RD, Dominici F, Louis TA. Model choice in time series studies of air pollutioni and mortality. Journal of the Royal Statistical Society Series A. 2006; 169:179–203.10.1111/j.1467-985X.2006.00410.x

56. Welty LJ, Peng RD, Zeger SL, Dominici F. Bayesian distributed lag models: Estimating effects of particulate matter air pollution on daily mortality. Biometrics. 2008; 65:282–291.10.1111/j.1541-0420.2007.01039.x [PubMed: 18422792]

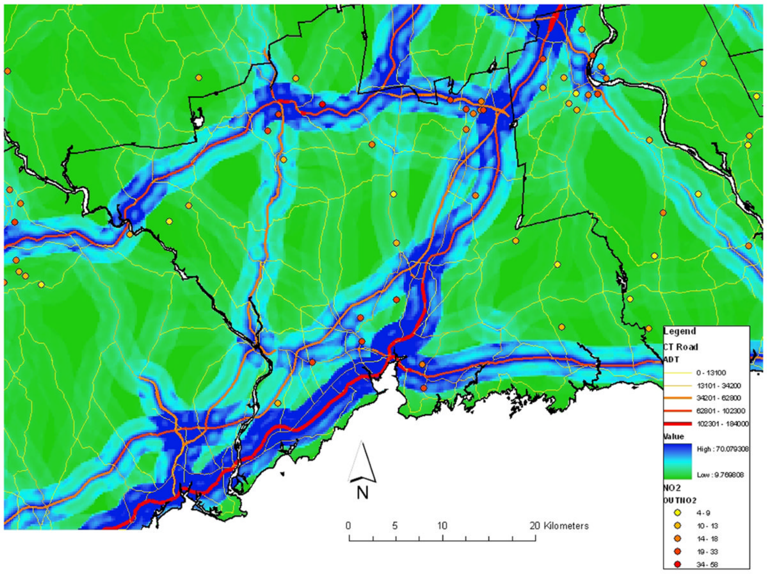

Figure 1.
Location of residences (dots), highways (color and weight coded by ADT) and the estimated NO_2 exposure surface in the greater New Haven area, 2000.

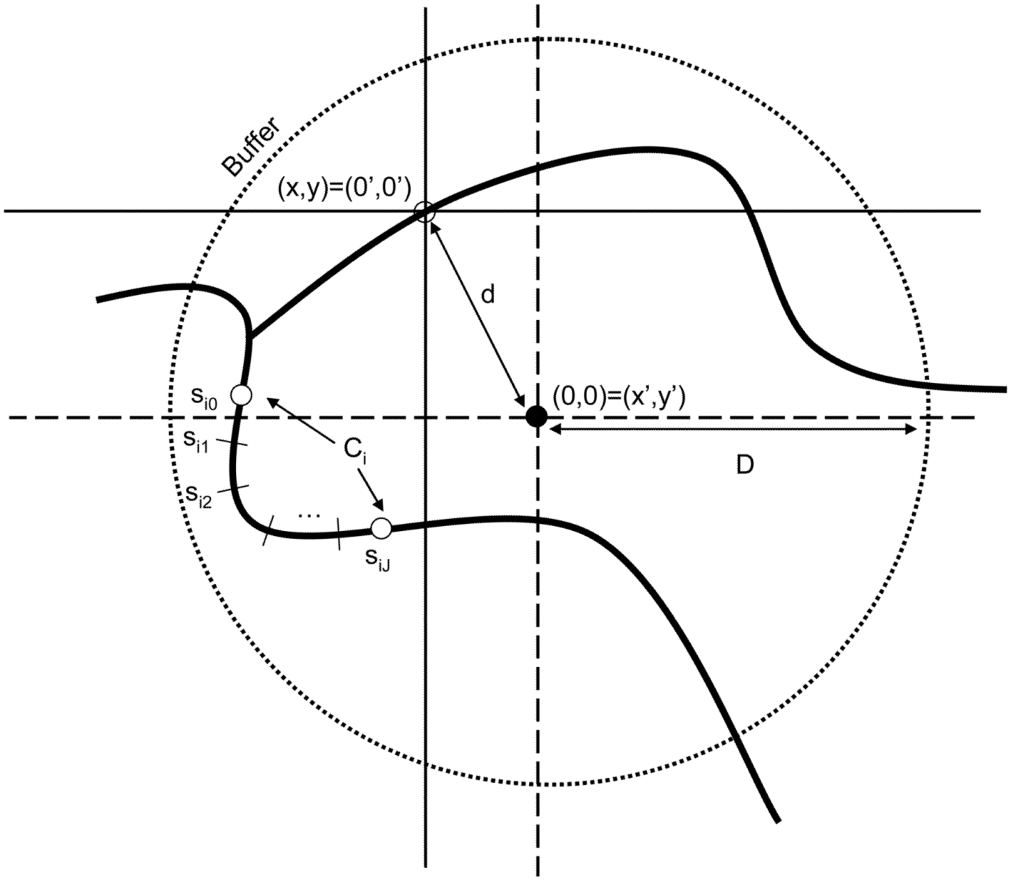

Figure 2.

Illustration of the relationship between exposure source points (open circles) on a highway (heavy line) and the location of a residence (large solid circle) along with the corresponding axes. Solid circles on highway lines demark segments, and short lines intersecting highways demark subsegments.

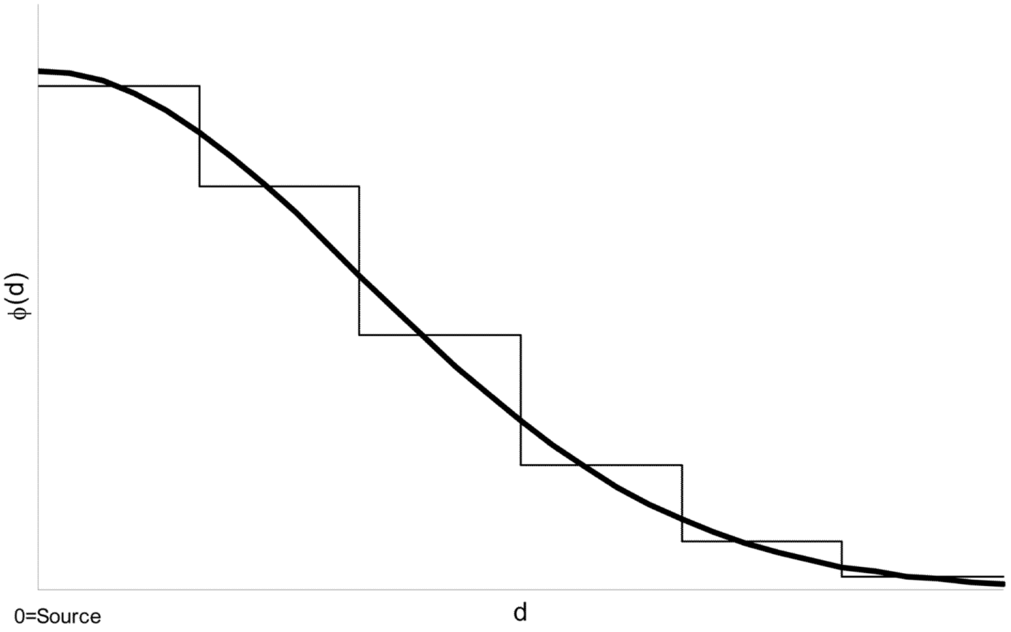

Figure 3.
Gaussian dispersion (heavy line) and step function (light line) models for relative concentration from a point source.

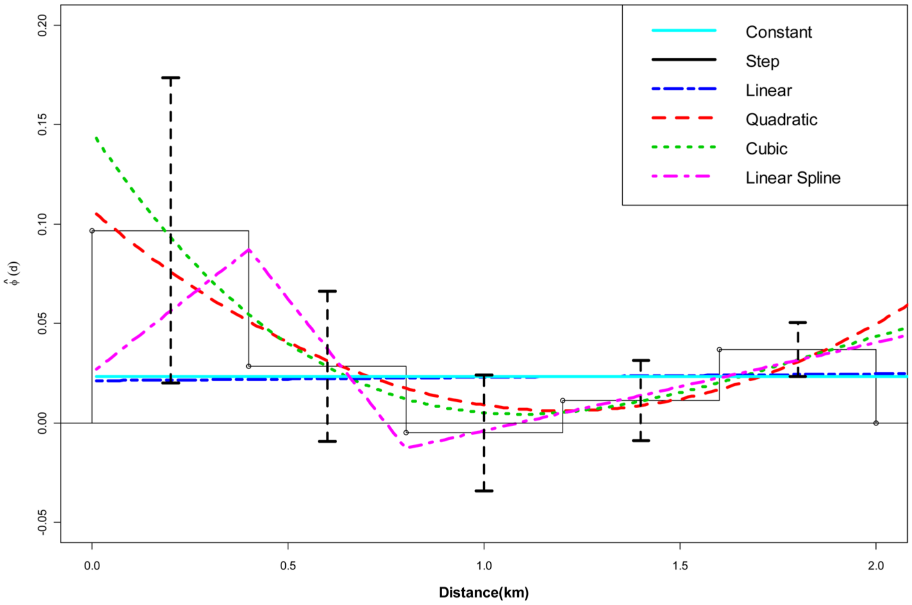

Figure 4.
Constant, step-function (95% CI), polynomial and linear spline estimates of the dispersion function, $\hat{\varphi}(d)$, which provides the weights for the effect of ADT on measured NO_2 levels as a function of distance, *d*.

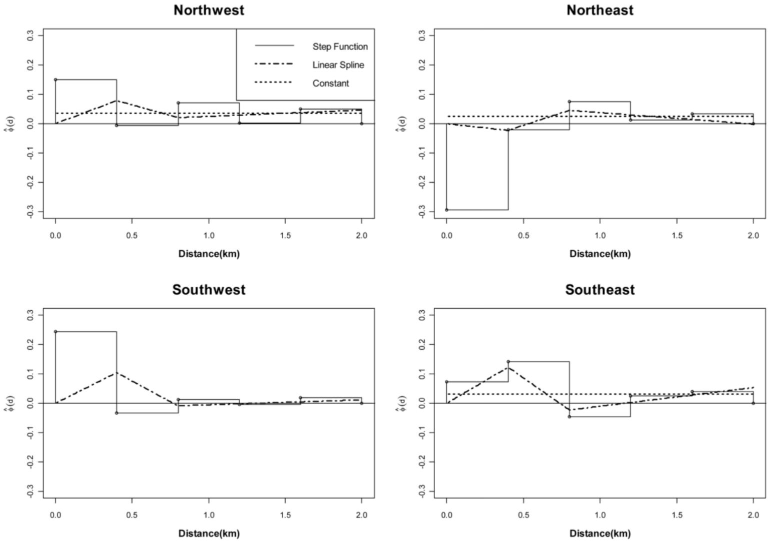

Figure 5.

Dispersion functions, $\hat{\varphi}(d)$, which provides the weights for the effect of ADT on measured NO_2 levels as a function of distance, *d* by directional quadrants about a residence.

Table 1

Summary (R^2 and P-value) of association between NO_2 and alternative measures of traffic exposures using GIS and ADT for roadways within a 750 ft buffer of residence.

	Unweighted[1]		**Distance weighted**[2]	
Traffic exposure measure	**R^2**	**P-value**	**R^2**	**P-value**
ADT on nearest road[3]	.017	.1527	.018	.1449
Highest ADT in buffer	.105	.0003	.044	.0220
ADT sum for all segments	.115	.0002	.069	.0040
Integrated exposure model (isotropic) using step function dispersion	-	-	.510	<.0001
Integrated exposure model (anisotropic): SW Cubic spline and constants for others	-	-	.513	<.0001

[1] Measure is unadjusted for distance.

[2] Measure is weighted using distance in meters to residence and the corresponding Gaussian dispersion weights used in other studies, $W_i = \exp\left\{-d^2/7432.24\right\}/(0.4\sqrt{2\pi})$. [1,27]

[3] Alternative transformations of distance were considered, include the log [4], but R^2 was not substantially improved.

Table 2

Demographic characteristics of study subjects.

Variable	Total
Number	761 (100.0)
Sex	
Male	372 (48.9)
Female	389 (51.1)
Ethnicity	
White	457 (60.1)
Black	103 (13.5)
Hispanic	169 (22.2)
Other	32 (4.2)
Mother’s education	
<HS	93 (12.2)
≥HS	668 (87.8)
#Siblings	
1	314 (41.3)
2	295 (38.8)
3 or more	152 (20.0)

Table 3

Estimated odds ratio for a one unit change in wheeze score and two measures of traffic exposure: traffic-related NO_2 or amount of traffic around a residence.

	OR-- uncorrected		**OR—corrected for exposure error**	
	Unadjusted (95% CI)	**Adjusted (95% CI)**	**Unadjusted (95% CI)**	**Adjusted (95% CI)**
Dispersion function steps (OR for a change of 100,000 vehicle km):				
0–500 m (n=360)	0.79 (0.49–1.25)	0.76 (0.48–1.22)		
501–1000 m (n=564)	1.20 (1.01–1.43)	1.22 (1.02–1.46)		
1001–1599 m (n=691)	0.96 (0.83–1.12)	0.93 (0.80–1.09)		
1501–2000 m (n=731)	1.14 (1.03–1.26)	1.10 (0.98–1.22)		
> 2000 m (n=761)	1 Ref.	1 Ref.		
Estimated NO_2:				
Q1[†]: ≤11.18 ppb (n=222)	1 Ref.	1 Ref.	1 Ref.	1 Ref.
Q2: >11.18, ≤ 15.57 ppb (n=188)	1.19 (0.81–1.75)	1.13 (0.76–1.68)	1.13 (0.69–1.86)	1.10 (0.67–1.82)
Q3: >15.57, ≤ 20.20 ppb (n=185)	1.43 (0.97–2.10)	1.24 (0.82–1.87)	1.38 (0.87–2.19)	1.22 (0.75–1.98)
Q4: >20.20 ppb (n=166)	1.95 (1.32–2.89)	1.62 (1.03–2.53)	1.92 (1.22–3.03)	1.58 (0.95–2.63)
Trend (OR for 10 ppb change)	1.35 (1.12–1.63)	1.22 (0.98–1.50)	1.34 (1.10–1.62)	1.20 (0.97–1.49)

[*] Adjusted for ethnicity, gender, mother’s education and number of siblings.

[†] Quartile of estimated NO_2 exposure.