HHS Public Access

Author manuscript

J Expo Sci Environ Epidemiol. Author manuscript; available in PMC 2015 September 01.

Published in final edited form as:

J Expo Sci Environ Epidemiol. 2015 September ; 25(5): 490–498. doi:10.1038/jes.2015.1.

Effect of geocoding errors on traffic-related air pollutant exposure and concentration estimates

Rajiv Ganguly[1], **Stuart Batterman**[2], **Vlad Isakov**[3], **Michelle Snyder**[4], **Michael Breen**[3], and **Wilma Brakefield-Caldwell**[5]

[1]Department of Civil Engineering, Jaypee University of Information Technology, Solan, India

[2]Environmental Health Sciences, University of Michigan, Ann Arbor, Michigan, USA

[3]NERL, US EPA, Research Triangle Park, North Carolina, USA

[4]University of North Carolina, Chapel Hill, North Carolina, USA

[5]Community Action Against Asthma (CAAA), Detroit, Michigan, USA

Abstract

Exposure to traffic-related air pollutants is highest very near roads, and thus exposure estimates are sensitive to positional errors. This study evaluates positional and $PM_{2.5}$ concentration errors that result from the use of automated geocoding methods and from linearized approximations of roads in link-based emission inventories. Two automated geocoders (Bing Map and ArcGIS) along with handheld GPS instruments were used to geocode 160 home locations of children enrolled in an air pollution study investigating effects of traffic-related pollutants in Detroit, Michigan. The average and maximum positional errors using the automated geocoders were 35 and 196 m, respectively. Comparing road edge and road centerline, differences in house-to-highway distances averaged 23 m and reached 82 m. These differences were attributable to road curvature, road width and the presence of ramps, factors that should be considered in proximity measures used either directly as an exposure metric or as inputs to dispersion or other models. Effects of positional errors for the 160 homes on $PM_{2.5}$ concentrations resulting from traffic-related emissions were predicted using a detailed road network and the RLINE dispersion model. Concentration errors averaged only 9%, but maximum errors reached 54% for annual averages and 87% for maximum 24-h averages. Whereas most geocoding errors appear modest in magnitude, 5% to 20% of residences are expected to have positional errors exceeding 100 m. Such errors can substantially alter exposure estimates near roads because of the dramatic spatial gradients of traffic-related pollutant concentrations. To ensure the accuracy of exposure estimates for traffic-related air pollutants, especially near roads, confirmation of geocoordinates is recommended.

Correspondence: Professor Stuart Batterman, Environmental Health Sciences, University of Michigan, Ann Arbor, MI, USA; Tel: +1 734 763 2417. Fax: +1 734 763 5455. stuartb@umich.edu.

Conflict of Interest: The authors declare no conflict of interest.

Disclaimer: Mention of trade names or commercial products does not constitute endorsement or recommendation for use

Supplementary Information accompanies the paper on the Journal of Exposure Science and Environmental Epidemiology website (http://www.nature.com/jes)

Keywords

traffic; air pollution; human exposure; geocoding

Introduction

Traffic-related emissions are the major source of air pollutants in urban areas, and exposure to traffic-related air pollutants has been associated with adverse health effects including the onset and exacerbation of asthma, impaired lung function, adverse birth outcomes and cognitive decline.[1,2] Particularly susceptible groups include children with asthma[3–7] and individuals living within 200 ft to 500 ft of major roads.[8,9] An estimated 40 million people in the United States live within 100 m of major roads, railways or airports.[10] Measurements of traffic-related air pollutants demonstrate very steep concentration gradients in directions perpendicular to major roads, for example, elevated levels of $PM_{2.5}$ and ultrafine particles at the roadside levels fall to near background levels at distances of 150–200 m;[11–14] similar gradients have been shown for other pollutants, for example, volatile organic compounds and polycyclic aromatic hydrocarbons.[15–18] Given such sharp gradients, precise and highly spatially resolved information regarding pollutant concentrations and locations of individuals and emission sources is needed to estimate exposure accurately.

As most individuals spend the majority of their time at home, residence location has been one of the most commonly used exposure metric in air pollution epidemiological studies.[19] The distance between a residence and emission sources, that is, proximity to highways and industry, has been used to associate human health effects with air pollution exposure.[20,21] This exposure indicator or surrogate is easy to obtain, inexpensive and potentially useful in many epidemiology studies. However, this indicator can be improved by accounting for factors that affect exposure, for example, the spatial and temporal patterns of emissions, meteorological processes governing dispersion and infiltration into buildings where individuals spend most of their time.

Given the sharp concentration gradients near roads, exposure estimates of traffic-related pollutants can be affected by positional errors of residences and other locations, as well as by errors in representing the road network. These errors depend on the geocoding method and the underlying data, for example, addresses and shape files. Typically, geocoding is performed using automated methods and commercial and proprietary software (for example, ArcGIS and SAS/GIS) or open access geocoding algorithms (for example, Bing and Google maps). Sometimes, more intensive methods are used, for example, aerial photography[22,23] and direct measurements using geographical positional system (GPS) devices. A review of geocoding methods and recommendations for epidemiology and other applications are presented elsewhere.[24–27]

Automated geocoding involves inputs, for example, addresses to be geographically referenced, processing algorithms to predict the geographic location and reference data sets to match and help confirm results.[28] Inputs must be standardized and normalized in a format compatible with the reference data set.[29,30] Matching algorithms can range from simple token parsing with lookup tables for standard abbreviations, to more complex programs

involving advanced probabilistic methods that can handle misspellings and misplacements.[29,31] Once input data have been sufficiently processed to be compatible with the reference data, matching obtains the final output. Matching efficiency can be increased by word stemming, Soundex and relaxing the need to match all attributes in the reference data set.[29,31] If a match is not obtained, then additional attributes or geocoding at lower resolution may be used.[30] If multiple matches are obtained, then users may be asked to select the best option or use additional reference data sets.

Positional errors have been characterized for several geocoding methods and settings.[23,25,26,32–35] Effects on exposure estimates, however, have received limited attention. This is a particular concern for traffic-related air pollutants, as distances can be very short and the source itself (major roads) is part of the geocoding process itself.[26] An analysis of positional errors in the CALTRANS (California Department of Transportation) and TAMN (TeleAtlas MultiNet) road networks in southern California showed differences in CALINE4[36] dispersion model predictions that reached 70% at certain receptors.[37] However, only a few analyses have examined impacts of positional errors at residences, schools, workplaces and other locations where individuals may be exposed. In comparing three spatial interpolation techniques in Grenoble, France, the choice of geocoding technique was found to influence estimates of health effects derived using a fine-scale exposure model.[38] In a study of homes in 49 US states, the accuracy of geocoding performed by four vendors differed by address characteristic and vendor, and these differences showed the potential for exposure misclassification on the basis of the concordance between address match and census tract.[27]

This study is aimed at characterizing positional errors for two automated geocoding processes, showing errors in representing road networks using a road-link-based system, and estimating concentration and exposure errors that result from these errors.

Methods

Study Overview

Positional errors were determined for residence locations of children enrolled in the Near-road EXposures and effects of Urban air pollutants Study (NEXUS), which is examining respiratory outcomes in a cohort of asthmatic children living near major roadways in Detroit, Michigan.[39] Children were recruited on the basis of the proximity of their residence to roadways in the following three exposure groups: children living within 200 m of high diesel roads, defined as having traffic that exceeds 6000 commercial vehicles/day (commercial annual average daily traffic; CAADT) and 90,000 total vehicles/day (annual average daily traffic; AADT); children living within 200 m of low diesel/high-traffic roads, defined similarly but including only roads with CAADT below 4,500; and children living in low-traffic areas, defined as those at least 300 m from any road with over 25,000 AADT. The present analysis considers 160 residences of the NEXUS participants. These homes were approximately equally distributed across the three exposure groups. The NEXUS study design oversamples homes near major roads. All study elements received approval by the University of Michigan Institutional Review Board. In addition, as a community-based

participatory research study, the study design and conduct was approved by our local community-based Steering Committee.

Geocoding Homes

Geocoordinates of study homes were measured using a handheld GPS device (GPSmap 60CS, Garmin International, Olathe, KS, USA) by our technician who stood in front of each home, generally as close as possible to the front door, and no further than 9 m away. When the indicated accuracy was 10 m or better, the location was recorded on a data entry form and as a waypoint in the device's memory. Four other GPS units (including two similar units from Garmin and two others) were used to confirm the instrument's calibration. To minimize GPS technician error, we utilized regular training, written protocols, standardized data entry forms and other measures.

Initial comparison between the on-site GPS measurements and two automated geocoders (described below) showed that 40 of the 160 residences had positional errors exceeding 50 m and sometimes much more (many kilometers). These sites were checked for possible data errors, for example, incorrect addresses and data entry mistakes. If errors remained, our technician was sent out to confirm both the address and the GPS coordinates, which lowered the number of homes with errors that exceeded 50 m to 32 homes. Again, addresses were checked; if errors remained, the technician was sent out a third time. The third set of GPS coordinates, which in all cases matched those taken the second time, was considered correct.

The first automated geocoder used publically available software, “Bing Maps” (http://www.bing.com/maps/). Each address (number, street, city and ZIP code) was manually entered into this online program, which returned its latitude and longitude based on the European Petroleum Survey Group (EPSG) code for projections, a Mercator projection and a spherical model of the earth.[40] The second geocoder used ESRI ArcMap 10.0 (Build 2414, Redlands, CA, USA), the Topologically Integrated Geographic Encoding and Referencing (TIGER) 2012 road shape files and the North American Datum for 1983. Address locations were determined using the “Geocoding Menu” and US Streets Geocode Service 10.0. This program uses a cascading sequence of geocoders, initially the Tele Atlas Address Points database, which maps 54 million residential and commercial US address records to specific physical locations, followed by Tele Atlas Street Address Range database, the 9-digit ZIP code locator and, lastly, the 5-digit ZIP code locator.[41]

Latitude and longitudes were converted to Universal Transverse Mercator coordinates for error analyses and dispersion modeling purposes, as described below.

Road Network and Proximity Analysis

An on-road emission inventory was developed for the ~ 800-km^2 study area using data obtained from the Southeast Michigan Council of Governments (SEMCOG) and the Michigan Department of Transportation (MDOT). The network used 9498 links (linear segments) to represent 3109 km of roads, which included all but small (but numerous) local roads. The network extended at least 5 km beyond the locations of the NEXUS homes and fully encompassed the city of Detroit (~355 km^2 area). Road links represented road

centerlines. Separate links were used to represent each direction of the major roads, as well as large service roads (parallel to freeways), if any, and major ramps.

Distances from study homes (using the on-site GPS measurements) to the major roads were determined using the "Near" function in ESRI ArcMap (version 10.0) and the 2012 TIGER/Line shape files for roads. A second distance measure, relevant to the emission inventory and link-based road network just described, was calculated as the perpendicular distance to the road link representing the highway segment.

Air-Quality Modeling and Emission Inventory

$PM_{2.5}$ concentrations at study homes resulting from primary exhaust emissions from on-road traffic were predicted using RLINE following the guidance for roadway sources.[42] RLINE is a steady-state dispersion model that incorporates new algorithms for predicting concentrations from line sources, including "upwind" concentrations resulting from plume meandering.[43,44] It utilizes a numerical method (or analytical approximation) that integrates multiple point sources along a line source, and dispersion parameters derived from recent field data and wind tunnel experiments. The model can predict concentrations at receptors very close to roads. Inputs to RLINE included hourly surface meteorological observations processed by AERMET for 2010 from Detroit City Airport, which was determined to be representative of the area, the road-link emission inventory described below, receptor height (1.8 m for the breathing level) and receptor locations (described below). Hourly concentrations for calendar year 2010 were predicted, from which daily (24 h) and annual average concentrations were calculated. Given our focus on local traffic-related air pollutants, concentrations due to point, area and regional emissions sources were not considered.

An hourly link-based $PM_{2.5}$ emission inventory was compiled for the road network described earlier and the year 2010. This involved merging traffic data (from MDOT) with traffic demand model outputs (from SEMCOG) to estimate AADT and speed for each road link. AADT was adjusted using temporal allocation factors (TAFs) for month, day-of-week and hour-of-day. On the basis of the SMOKE model,[45] a bi-modal TAF pattern portrayed morning and evening rush hours on weekdays; a unimodal pattern represented a broad afternoon peak on Saturdays and Sundays. TAFs depended on the road type, which were designed as National Functional Class (NFCs) 11, 12, 14, 16, 17, 19 or 0, respectively, representing interstate, other freeway, other principal arterial, major collector, minor collector and other road types. Fleet mix (modeled using eight vehicle classes) also depended on the NFC link, as well as hour-of-day and day-of-week. Emission factors were generated using MOVES2010,[46] the monthly average temperature, link speed, vehicle class, and the local vehicle age and fleet distributions. Finally, hourly emissions $E_{i,t}$ (g/m-s) for each link i and hour t were calculated as:[47]

$$E_{i,t} = AADT_i \, TAFM_{i,t} \, TAFD_{i,t} \, TAFH_{i,t} \sum_j FM_{i,j} \, EF_{i,j,t} \quad (1)$$

where $AADT_i$ = annual average daily traffic (vehicles/day); $TAFM_{i,t}$, $TAFD_{i,t}$ and $TAFH_{i,t}$ = monthly, daily and hourly TAFs, which depend on NFC and time *t* (month, day and hour;

dimensionless); $FM_{i,j}$ = fleet mix for link *i* and vehicle class *j* (dimensionless); and $EF_{i,j,t}$ = MOVES2010 emission factor for link *i*, which depends on vehicle speed, vehicle class *j* and time *t* (using season and average monthly temperature, dimensionless).

Concentration Errors

We first demonstrate concentration errors due to positional errors that result from miscoded home locations and approximations in the configuration of the road network using a sensitivity analysis and simplified test case. This modeled a single road link (north–south orientation, 1 km length, NFC = 11, AADT = 200,000, hourly $PM_{2.5}$ emissions for 2010) and a set of 22 receptors in a transect perpendicular and centered across the road (east–west orientation, 50 m intervals to 500 m from the road). Annual average and 24-h concentrations were predicted using RLINE and 2010 meteorology.

Effects of geocoding errors for the 160 home locations were evaluated by comparing dispersion model predictions at two sets of receptor locations: the on-site GPS measurements, considered the “gold standard” and coordinates given by the Bing-automated geocoder for the same homes. Because the ARC-GIS geocoder gave results comparable to the Bing decoder, dispersion modeling was not conducted a third time.

Error Analysis

Positional errors for the 160 study homes were defined as the distance between the GPS measurement and the automated geocoding estimates. Geocoding errors associated with the road-link network were estimated as the difference in home-to-road distances determined using the GPS measurements and the TIGER shape files for the roads, and home-to-road distances determined using the same home coordinates and the road-link network. Only the major highways used to classify homes were considered.

Concentration errors because of geocoding errors for the home locations were estimated as differences between RLINE predictions using on-site GPS measurements and the automated geocoding coordinates. Both annual average and maximum 24-h concentrations for 2010 were considered. Descriptive statistics, including relative absolute differences (RADs), were calculated for positional errors and concentration differences. To examine effects of highway proximity, concentration errors were stratified by distance to major roads. Homes with RADs exceeding 25% were mapped and received further analysis. The relationship between positional and concentration errors was explored using several regression models.

Results and Discussion

Errors in Geocoding Homes

Of the 160 home locations (mapped in Figure 1), 36 (23%) were located within 100 m of the major roads, 46 (29%) within 100–200 m, 5 (3%) within 200–300 m, 2 (0.6%) within 300–400 m and 70 (43.8%) were at distances exceeding 500 m. The few homes in the 200- to 500-m bins (n = 8) were pooled for subsequent analysis.

Table 1 summarizes positional errors for the two automated geocoding techniques, and Figure 2 compares the two cumulative distributions. For the Bing geocoder, positional errors

averaged 32 ± 32 m (SD); errors exceeded 50 and 100 m at 32 (20%) and 7 (4%) homes, respectively, and the maximum error was 157 m. The ArcGIS geocoder gave slightly larger but generally very similar errors. Positional errors were not affected by proximity to a specific named highway (that is, I-75, I94, M-10 and M-39), the distance to the highway, city region or the initial classification into the three exposure groups. Scatterplots showed small biases in the east–west direction (median of 0.89 and −1.39 m for Bing and ArcGIS, respectively), and larger biases in the north–south direction (−2.72 and 8.03 m; Supplementary Figure S1 in the supplementary material). These biases represent systematic errors that may result from the legacy of the Public Land Survey System that used a rectangular network of surveys, or other displacement and projection errors.[26] Errors made by the two geocoders also were weakly correlated ($r = 0.25$), suggesting that some errors had a common source, for example, incorrect street numbering, or possibly issues with the GPS measurements, despite being triple-checked.

Errors and differences among automated geocoders have been documented in several studies. For 1000 residences in upstate New York, USA, the mean, median and 95th percentile differences between MapMarker plus 6.0 and Geographic Data Technology software were 58, 38 and 152 m, respectively.[32] In 100 homes in urban areas of New York, the median error was 32 m and 89% of homes had errors below 100 m.[48] For 234 urban residences in south central Iowa, US, the median error was 56 m and 81% of homes had errors below 100 m as determined using ArcView 3.2 and TIGER files; a commercial firm using proprietary software obtained slightly smaller errors (median of 50 m, 84% below 100 m).[23] For 21,890 residences in Sydney, Australia, differences between MapInfo 5.5 and StreetWorks 5.0 averaged 47 m, and 31 m after removing outliers (5% of the data).[49] For 104,865 residences in suburban Orange County, FL, USA using ArcGIS 9.1, the median and 95th percentile errors were 41 and 137 m, respectively.[33] For 135 rural addresses in West Virginia, USA, several automated geocoding methods showed comparable results (for example, the median errors from 39 to 62 m, and 84% within 100 m of GPS measurements with E911 conversion).[35] For 354 homes in France, the median errors ranged from 26 m to 36 m, depending on interpolation technique.[38] Overall, these studies are quite consistent and show average positional errors from 47 m to 58 m, and median errors from 26 m to 62 m, 95th percentile errors from 137 m to 152 m and that errors exceed 100 m at roughly 10% to 20% of sites. Recent review papers considering a large number of geocoding studies and applications, including those with greater geographic and environmental diversity, show larger errors.[25,26]

Positional errors can be affected by address type, projection method, parcel size, road orientation, technician errors and other factors.[24–26] Errors are larger in rural areas where geocoding uses PO boxes[32,50] or other substitutes for addresses that cannot be geocoded.[33] Projections in online maps may not align with the more sophisticated projections in the GIS software, for example, the EPSG code used by Bing has changed several times and may be out of date.[40] In addition, GPS measurements can be affected by satellite geometry, atmospheric conditions, loss of signal and multipath errors (because of reflections from nearby tall buildings, water bodies and dense clusters of trees).[51] Home coordinates should lie within the property parcel boundary, but mapped positions may not correspond to the

geocoded site (near the front door in the present study); such errors may increase for larger buildings or parcel sizes.

Parcel sizes in Detroit averaged 431 ± 249 m^2 (± SD), including typically small front and backyards; the floor area of houses in our cohort averaged 132 ±45 m^2, and the house footprint averaged 77 ± 29 m^2 (Data Driven Detroit, personal communication). Errors because of parcel size itself are estimated as 11 m, on the basis of one-half of the average parcel dimension (assuming one-fourth of the sum of the parcel's length and width, the mean parcel size area and a typical 2:1 parcel dimension). The components of positional errors were estimated using a simple Gaussian quadrature analysis that assumed independence between error components. The total error (as a SD) was apportioned to parcel size variation (11 m), GPS uncertainty (5 m using the typical indicated GPS precision), projection errors (up to 8 m on the basis of the systematic differences) and address errors (5–10 m, calculated as the balance of the error). Although approximate, this analysis suggests that parcel size and address errors are the most significant error sources.

Overall, the positional errors resulting from automated geocoding in Detroit were somewhat smaller than those reported previously for residences. This may result from our double and triple checking, the regular and stable street numbering and the small parcel size. Still, a fraction of homes had positional errors exceeding 100 m or more, a large amount considering the steep concentration gradients near major roads, as demonstrated in the following section. Automated geocoding produces larger positional errors in rural areas and in urban areas with large parcel sizes, irregular address numbering or rapid growth. As shown later, errors in the direction perpendicular to roads are the most significant; these will occur for addresses on roads that are perpendicular to highways. To avoid large errors, we recommend confirming geocoded coordinates whenever feasible.

Distances to Roads

Distances to the major roads for the high-traffic homes determined using the TIGER road shape files and the road-link network were highly correlated (r = 0.80, n = 82), and the average difference between the two methods was 23 ± 17 m. Distances using the TIGER shape files were shorter in most (82%) cases, as expected, as these represent distances to the road edge, whereas calculations using the link-based network reference the road centerline (on each side of the highway). In addition, the link- based network used a simplified road geometry, for example, curves were represented using a minimal number of linear segments. These two factors accounted for most differences, including cases where the Tiger file estimates were larger. Automated calculations also may calculate distances to access ramps, typically labeled with the main road's name. Substantial differences occurred for a subset of homes, for example, on the basis of the fractional bias (deviation from the mean), the two sets of distances differed by at least 25% for 44% of the homes, and by at least 50% for 16% of the homes. As discussed below, this could cause large differences in exposure estimates, given the sharp concentration gradients near roads.

This analysis highlights differences in proximity estimates using road edges *versus* road centerlines, and the accuracy of the road-link network. A difference of 10 –20 m is reasonable, given that most highways in Detroit have three or four lanes in each direction, a

right-hand shoulder and sometimes a left-hand shoulder. In addition, many highways have a two-lane service drive on each side separated from the highway by a low-profile barrier and buffer. However, most service roads have only modest traffic, and thus should not be counted as part of the highway. The level and significance of traffic on ramps (for example, vehicle acceleration causing higher emissions) must be evaluated on a case-by-case basis. The largest highways in Detroit, which contain 12 and 14 lanes (excluding shoulders, service lanes and ramps), were represented by parallel groups of 3 or 4 lanes in the road-link network, and did not result in larger errors.

Exposure metrics using distance from the road edge may be advantageous compared with metrics using the road centerline, as the former is the distance from the emission source, particularly if vehicle-induced turbulence fully mixes exhaust emissions across the road width. However, this assumption may be less valid for wide roads, low-traffic speeds, low-traffic conditions and if the edge is defined by a shoulder or buffer. Notably, the distance to the road edge may approach zero for buildings that are immediately adjacent to the road, which causes inverse distance measures to “blow-up” mathematically. (Several Detroit homes had very small distances.) Overall, this analysis highlights the need to carefully formulate and evaluate proximity measures used either directly as an exposure surrogate, or indirectly as an input in a dispersion, land use regression or other types of exposure model.

Concentration Gradients for Test Road

Concentration transects for the test case are shown in Figure 3. For this large highway (AADT = 200,000 vehicles/day), the roadside annual and maximum 24-h average $PM_{2.5}$ predictions were 17.6 and 129 $\mu g/m^3$, respectively. Concentrations rapidly dropped with distance: the annual average fell by 11 and 25 times at distances of 100 and 500 m, respectively; the 24-h average fell by 8 and 16 times. Note that transects may not be symmetrical because of the influence of meteorology (for example, prevailing winds) and other emission sources. In realistic cases, gradients at larger distances will be lower than the test case (which modeled a single road link), as concentrations will include contributions from other roads, as well as regional and urban background levels, which are usually substantial for $PM_{2.5}$. In addition, positional changes that are parallel to the road will not affect concentrations unless another road is encountered.

The sharp concentration gradients in Figure 3 suggest the influence of even small positional errors. We quantified the sensitivity to concentration errors by expressing concentration changes in terms of a 10-m change in the receptor-to-road distance (using the derivative of the concentration transect, that is, sensitivity = $\Delta\mu g/m^3$ per Δ10 m). Both absolute and relative concentration changes were calculated after averaging the eastern and western halves of the transect. For the annual average, for example, increasing the receptor-road distance from 50 to 60 m decreased concentrations by ~ 1 $\mu g/m^3$ or 19%. Relative concentration changes for 10 m distance changes from 50 to 200 m (the range appropriate for NEXUS high-traffic homes) were 6% to 19% for annual averages, and slightly lower for 24-h averages, 5% to 16%. Sensitivity was highest at shorter distances. Considering the overall range of sensitivity (5% to 19% concentration change per 10 m), concentration errors from 13% to 49% might be expected for homes near highways because of the median

geocoding error (26 m, Table 1), and from 10% to 40% considering the differences between road edge and centerline (the median difference of 21 m). However, these estimates are overstated, as positional errors occur in all directions (not just perpendicular to roads), other roads contribute $PM_{2.5}$ (especially considering the density of Detroit's road network), and some homes are at greater distances where sensitivity is lower. Thus, positional errors must be evaluated using a large-scale simulation and the full road network, described next.

$PM_{2.5}$ Concentrations at Study Homes

Predicted annual average $PM_{2.5}$ concentrations due to traffic exhaust emissions ranged from 0.45 to 5.2 $\mu g/m^3$ and averaged 1.27 $\mu g/m^3$ across the 160 homes (using the on-site GPS coordinates). Receptors near the major roads had the highest concentrations. The highest annual (5.2 $\mu g/m^3$) and 24-h average (26.3 $\mu g/m^3$) concentrations occurred at the same home, which was located 42 m from I-75 (AADT = 132,800). Annual average concentrations at homes more than 300 m from the major roads ranged from 0.4 to 1.7 $\mu g/m^3$ and averaged 0.83 $\mu g/m^3$.

Concentration predictions for 2010 were lower than monitored levels and source apportionments for Detroit performed for earlier years. In the past 10 years, annual average $PM_{2.5}$ concentrations monitored in Detroit have ranged from 10 to 17 $\mu g/m^3$, depending on site and year. The major $PM_{2.5}$ sources have been identified as coal combustion, road traffic emissions, municipal waste incinerators, oil refineries, oil sewage sludge incinerators and iron/steel manufacturing.[52] On the basis of chemical mass balance (CMB) apportionments and samples collected in summer and winter seasons in 2004–2006, the estimated average traffic contribution at a population-oriented monitor located in northeast Detroit (Allen Park) was 5.3 $\mu g/m^3$ or 30% of $PM_{2.5}$.[53] In another CMB study monitoring ambient air near six residences in Detroit in summer and winter 2004–2006, the estimated traffic contribution was slightly higher, 5.9 $\mu g/m^3$ or 33% of $PM_{2.5}$.[53] Both studies were conducted during periods when $PM_{2.5}$ levels at some sites exceeded the annual National Ambient Air-Quality Standards, then set at 15 $\mu g/m^3$. By 2010, however, annual average $PM_{2.5}$ concentration across Wayne County (11 monitoring sites) encompassing Detroit had fallen to 10.1 $\mu g/m^3$,[54] compared with 14.5 $\mu g/m^3$ during the 2004–2006 period used for CMB apportionments.[53]

Even accounting for the decline in $PM_{2.5}$ concentrations, dispersion model predictions were considerably lower than the CMB apportionments. This likely results from the substantial reductions in local $PM_{2.5}$ emissions and concentrations achieved in past years, especially in the diesel fleet; the exclusion of secondary pollutants, pavement, brake and tire wear, and entrained dust from RLINE predictions; differences between CMB-monitoring sites and the dispersion model receptors; model input errors, especially vehicle emission rates;[55] and meteorological influences. Despite these limitations and uncertainties in dispersion modeling results, our analysis of the influence of positional errors on pollutant contributions from traffic sources should not be greatly affected, as the models are used in a comparative manner, for example, concentration errors are estimated by comparing predictions at two sets of receptors, and as we expressed results using relative differences.

Concentration Errors at Homes

Table 2 shows absolute and relative concentration errors at the 160 study homes resulting from geocoding errors, that is, differences between receptors determined using the automated (Bing) geocoder and the on-site GPS measurements. The maximum errors were 0.88 and 6.79 $\mu g/m^3$ for annual average and 24-h maximum $PM_{2.5}$ concentrations, respectively. Concentrations using automated geocoordinates were within a factor of 1.5 of those using GPS measurements for annual averages, and within a factor of 2 for daily averages (Figure 4).

Considering the RADs, positional errors caused errors of 8 ± 10% (maximum of 54%) for annual averages and errors of 8 ± 12% (maximum of 87%) for 24-h maxima. The distribution of errors, shown in Figure 5, includes several homes with large errors, which are mapped in Figure 6. Of the 13 homes with the largest differences (RAD 425%), three were located near M-39, four were near M10, five were near I-94 and one was near the junction of other secondary roads, in addition to the major roads. Although spread across the study area, three homes were within 100 m of major roads, and the remainder were between 100 and 200 m.

Both absolute and relative errors tended to decrease with distance from the road. Concentration errors depended on positional errors, and a power law relationship using the absolute positional error explained a portion of the concentration error ($r = 0.68$; Figure 7a). Model fit improved ($r = 0.79$) when only distance errors in the direction perpendicular to the road and homes within 500 m were considered (Figure 7b). Errors expressed as RADs tended to decrease with distance from the road, for example, errors in annual average concentrations fell from 10 ± 10% for homes within 100 m of highways, to 5 ± 8% for homes over 500 m from highways. However, the highest errors did not follow this trend, and RAD exceeded 20% for ∼ 5% of the homes in each distance bin (Table 2).

Large concentration errors resulted from two reasons. First, for homes near the major roads, concentrations were very sensitive to small changes in the distance to the road. The steepness of concentration gradients near roads, particularly within the first 200 m of the road, was shown earlier in Figure 3. Second, positional errors may locate the home near to a secondary road (not necessarily the high-traffic road used to classify the residence), or move the home away from such sources. However, these events were uncommon.

Strengths and Limitations

This paper benefits from its use of a modern dispersion model specifically designed for road sources, the availability and quality of data collected for NEXUS, and the effort made to correctly geocode locations and confirm results. Further, Detroit is likely to be representative of many older cities in the United States of America. Several limitations are recognized. First, the sample consisted of 160 home locations in one, Midwest, suburban-to-urban US locale. Other settings may have greater geographic or environmental diversity that might affect and likely increase both positional and concentration errors. In consequence, positional errors for Detroit were somewhat smaller than those reported elsewhere, and thus our estimates of concentration differences are likely to be

underestimated. Errors in TIGER files that represented the road network[37] were not evaluated. The GPS measurements were assumed to provide accurate locations, and errors are unlikely to exceed 5 or 10 m. However, on-site measurements may not correspond to the parcel centroids or locations mapped by the automated geocoders. While dispersion model predictions were compared to monitored and apportioned concentrations, predicted concentrations were assumed to be accurate on both an absolute and a relative base. However, the use of relative errors and the model-to-model comparisons of the analyses diminish these limitations. Finally, the "motor city" has generally flat topography and is largely suburban with mainly low-rise buildings. Some results may not apply to cities with different configurations.

Conclusions

This paper has examined errors in estimating near-road concentrations of traffic-related air pollutants due to positional errors that result from the use of automated geocoding methods. Using a modern dispersion model and a detailed road network, concentrations were predicted at 160 homes of children participating in an epidemiological study in Detroit, Michigan. Comparing results of automated and manual geocoding methods for home locations, positional errors averaged 35 m and reached 197 m; larger errors are likely in other urban areas. Comparing road edge and road centerline, differences in house-to-highway proximity averaged 23 m and reached 82 m; road curvature and road width were important factors that should be considered in proximity measures used either directly as an exposure metric, or as an input variable to a dispersion or other model. Geocoding errors in home locations produced concentration errors that averaged only 9%; however, much larger errors were found for a subset of homes, for example, up to 54% for annual averages and 87% for daily averages. Positional errors were particularly important for locations near roads. Our results and literature suggest that 5–20% of homes can have geocoding errors exceeding 100 m, a large amount considering the steepness of concentration gradients of traffic-related air pollutants. Whereas the use of automated geocoding methods may be essential in deriving proximity measures and exposures of traffic-related air pollutants, particularly if dealing with thousands of sites, verification of locations and distances from major roads is recommended whenever possible.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The NEXUS study involves a community-based participatory research partnership, and we thank Community Allies Against Asthma (CAAA) and the following organizations: Arab Community Center for Economic and Social Services, Community Health and Social Services Center, Detroit Hispanic Development Corporation, Detroiters Working for Environmental Justice, Friends of Parkside, Detroit Department of Health and Wellness Promotion, Latino Family Services, Southwest Detroit Environmental Vision, Warren/Conner Development Corporation, the University of Michigan School of Medicine, the University of Michigan School of Public Health and an independent community activist. CAAA is an affiliated project of the Detroit Community-Academic Urban Research Center. We also thank Janet Burke, Steve Perry and Dave Heist at the US EPA, Laprisha Berry Vaughn, Sonya Grant, Graciela Menz and other staff at the University of Michigan, and the NEXUS participants and their families who assisted us with the collection of these data. The US Environmental Protection Agency through its Office of Research and Development partially funded the research described here under cooperative agreement

R834117 (University of Michigan). It has been subjected to Agency review and approved for publication. The study was conducted as part of NIEHS grants 5-R01-ESO14677-02 and R01 ES016769-01.

References

1. Han X, Naeher LP. A review of traffic-related air pollution exposure assessment studies in the developing world. Environ Int. 2006; 32:106–120. [PubMed: 16005066]
2. Grange SK, Salmond JA, Trompetter WJ, Davy PK, Ancelet T. Effect of atmospheric stability on the impact of domestic wood combustion to air quality of a small urban township in winter. Atmos Environ. 2013; 70:28–38.
3. 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. Environ Health Perspect. 1999; 107:761–767. [PubMed: 10464078]
4. Clark NA, Demers PA, Karr CJ, Koehoorn M, Lencar C, Tamburic L, et al. Effect of early life exposure to air pollution on development of childhood asthma. Environ Health Perspect. 2010; 118:284–290. [PubMed: 20123607]
5. Holguin F. Traffic, outdoor air pollution, and asthma. Immunol Allergy Clin N Am. 2008; 28:577–588.
6. Gasana J, Dillikar D, Mendy A, Forno E, Ramos Vieira E. Motor vehicle air pollution and asthma in children: a meta-analysis. Environ Res. 2012; 117:36–45. [PubMed: 22683007]
7. Lindgren A, Stroh E, Nihlen U, Montnemery P, Axmon A, Jakobsson K. Traffic exposure associated with allergic asthma and allergic rhinitis in adults. A cross-sectional study in southern Sweden. Int J Health Geogr. 2009; 8:25. [PubMed: 19419561]
8. McConnell R, Berhane K, Yao L, Jerrett M, Lurmann F, Gilliland F, et al. Traffic, susceptibility, and childhood asthma. Environ Health Perspect. 2006; 114:766–772. [PubMed: 16675435]
9. Gauderman WJ, Vora H, McConnell R, Berhane K, Gilliland F, Thomas D, et al. Effect of exposure to traffic on lung development from 10 to 18 years of age: a cohort study. Lancet. 2007; 369:571–577. [PubMed: 17307103]
10. Bureau USC. Current Housing Reports 2007. Mar 28. 2013 Available from http://www.census.gov/prod/2008pubs/h150-07.pdf
11. Zhu Y, Kuhn T, Mayo P, Hinds WC. Comparison of daytime and nighttime concentration profiles and size distributions of ultrafine particles near a major highway. Environ Sci Technol. 2006; 40:2531–2536. [PubMed: 16683588]
12. Hitchins J, Morawska L, Wolff R, Gilbert D. Concentrations of submicrometre particles from vehicle emissions near a major road. Atmos Environ. 2000; 34:51–59.
13. Karner AA, Eisinger DS, Niemeier DA. Near-roadway air quality: synthesizing the findings from real-world data. Environ Sci Technol. 2010; 44:5334–5344. [PubMed: 20560612]
14. Reponen T, Grinshpun SA, Trakumas S, Martuzevicius D, Wang ZM, LeMasters G, et al. Concentration gradient patterns of aerosol particles near interstate highways in the Greater Cincinnati airshed. J Environ Monit. 2003; 5:557–562. [PubMed: 12948227]
15. Baldauf R, Thoma E, Hays M, Shores R, Kinsey J, Gullett B, et al. Traffic and meteorological impacts on near-road air quality: summary of methods and trends from the Raleigh Near-Road Study. J Air Waste Manag Assoc. 2008; 58:865–878. [PubMed: 18672711]
16. Barzyk TM, George BJ, Vette AF, Williams RW, Croghan CW, Stevens CD. Development of a distance-to-roadway proximity metric to compare near-road pollutant levels to a central site monitor. Atmos Environ. 2009; 43:787–797.
17. Hagler GSW, Baldauf RW, Thoma ED, Long TR, Snow RF, Kinsey JS, et al. Ultrafine particles near a major roadway in Raleigh, North Carolina: downwind attenuation and correlation with traffic-related pollutants. Atmos Environ. 2009; 43:1229–1234.
18. Hu SS, Fruin S, Kozawa K, Mara S, Paulson SE, Winer AM. A wide area of air pollutant impact downwind of a freeway during pre-sunrise hours. Atmos Environ. 2009; 43:2541–2549.
19. Huang YL, Batterman S. Residence location as a measure of environmental exposure: a review of air pollution epidemiology studies. J Expo Anal Environ Epidemiol. 2000; 10:66–85. [PubMed: 10703849]

20. Nuckols JR, Ward MH, Jarup L. Using geographic information systems for exposure assessment in environmental epidemiology studies. Environ Health Perspect. 2004; 112:1007–1015. [PubMed: 15198921]
21. Peters JM, Avol E, Gauderman WJ, Linn WS, Navidi W, London SJ, et al. A study of twelve Southern California communities with differing levels and types of air pollution. II. Effects on pulmonary function. Am J Respir Crit Care Med. 1999; 159:768–775. [PubMed: 10051249]
22. Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, et al. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. Ann Epidemiol. 2007; 17:464–470. [PubMed: 17448683]
23. Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, et al. Positional accuracy of two methods of geocoding. Epidemiology. 2005; 16:542–547. [PubMed: 15951673]
24. Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, et al. Geocoding in cancer research: a review. Am J Prev Med. 2006; 30:S16. [PubMed: 16458786]
25. Hart TC, Zandbergen PA. Reference data and geocoding quality. Policing. 2013; 36:263–294.
26. Zandbergen PA. Geocoding quality and implications for spatial analysis. Geogr Compass. 2009; 3:647.
27. Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, et al. Accuracy of commercial geocoding: assessment and implications. Epidemiol Perspect Innovat. 2006; 3:8.
28. Karimi HA, Durcik M, Rasdorf W. Evaluation of uncertainties associated with geocoding techniques. Comput Aided Civil Infrastruct Eng. 2004; 19:170–185.
29. Churches T, Christen P, Lim K, Zhu JX. Preparation of name and address data for record linkage using hidden Markov models. BMC Med Inform Decis Mak. 2002; 2:9. [PubMed: 12482326]
30. Laender AHF, Borges KAV, Carvalho JCP, Medeiros CB, de Silva AS, Davis CA Jr. Integrating web data and geographic knowledge into spatial databases. Spatial Databases: technologies, techniques and trends: IGI global. 2005:23–48.
31. Yang DH, Bilaver LM, Hayes O, Goerge R. Improving geocoding practices: evaluation of geocoding tools. J Med Syst. 2004; 28:361–370. [PubMed: 15366241]
32. Cayo M, Talbot T. Positional error in automated geocoding of residential addresses. Int J Health Geogr. 2003; 2:10. [PubMed: 14687425]
33. Zandbergen PA, Green JW. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. Environ Health Perspect. 2007; 115:1363–1370. [PubMed: 17805429]
34. Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, et al. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. Ann Epidemiol. 2007; 17:464–470. [PubMed: 17448683]
35. Vieira VM, Howard GJ, Gallagher LG, Fletcher T. Geocoding rural addresses in a community contaminated by PFOA: a comparison of methods. Environ Health. 2010; 9:18. [PubMed: 20406495]
36. Benson PE. A review of the development and application of the CALINE3 and 4 models. Atmos Environ B Urban Atmos. 1992; 26:379–390.
37. Wu J, Funk TH, Lurmann FW, Winer AM. Improving spatial accuracy of roadway networks and geocoded addresses. Transactions in GIS. 2005; 9:585–601.
38. Jacquemin B, Lepeule J, Boudier A, Arnould C, Benmerad M, Chappaz C, et al. Impact of geocoding methods on associations between long-term exposure to urban air pollution and lung function. Environ Health Perspect. 2013; 121:1054. [PubMed: 23823697]
39. Vette A, Burke J, Norris G, Landis M, Batterman S, Breen M, et al. The Near-Road Exposures and Effects of Urban Air Pollutants Study (NEXUS): study design and methods. Sci Total Environ. 2013; 448:38–47. [PubMed: 23149275]
40. Alistair, A. The Google Maps/Bing Maps Spherical Mercator Projection. 2011. https://alastaira.wordpress.com/2011/01/23/the-google-maps-bing-maps-spherical-mercator-projection/
41. ESRI. ESRI Geocoder Information. Redlands, CA. USA: 2010.
42. USEPA. User's Guide for the AMS/EPA Regulatory Model - Aermod. 2004 Contract No.: EPA-454/B-03-001.

43. Snyder MG, Venkatram A, Heist DK, Perry SG, Petersen WB, Isakov V. RLINE: A line source dispersion model for near-surface releases. Atmos Environ. 2013; 77:748–756.
44. Venkatram A, Snyder MG, Heist DK, Perry SG, Petersen WB, Isakov V. Reformulation of plume spread for near-surface dispersion. Atmos Environ. 2013; 77:846–855.
45. Pouliot G, Pierce T, Denier van der Gon H, Schaap M, Moran M, Nopmongcol U. Comparing emission inventories and model-ready emission datasets between Europe and North America for the AQMEII project. Atmos Environ. 2012; 53:4–14.
46. Wallace HW, Jobson BT, Erickson MH, McCoskey JK, VanReken TM, Lamb BK, et al. Comparison of wintertime CO to NOx ratios to MOVES and MOBILE6.2 on-road emissions inventories. Atmos Environ. 2012; 63:289–297.
47. Cook R, Isakov V, Touma JS, Benjey W, Thurman J, Kinnee E, et al. Resolving local-scale emissions for modeling air quality near roadways. J Air Waste Manage Assoc. 2008; 58:451–461.
48. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology. 2003; 14:408–412. [PubMed: 12843763]
49. Ratcliffe JH. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. Int J Geogr Inform Sci. 2001; 15:473–485.
50. Kravets N, Hadden WC. The accuracy of address coding and the effects of coding errors. Health Place. 2007; 13:293–298. [PubMed: 16162420]
51. Wu J, Jiang C, Liu Z, Houston D, Jaimes G, McConnell R. Performances of different global positioning system devices for time-location tracking in air pollution epidemiological studies. Environ Health Insight. 2010; 4:93–108.
52. Morishita M, Keeler GJ, Wagner JG, Harkema JR. Source identification of ambient PM2.5 during summer inhalation exposure studies in Detroit, MI. Atmos Environ. 2006; 40:3823–3834.
53. Duvall RM, Norris GA, Burke JM, Olson DA, Vedantham R, Williams R. Determining spatial variability in PM2.5 source impacts across Detroit, MI. Atmos Environ. 2012; 47:491–498.
54. MDEQ. Annual Air Quality Report. Michigan Dept. of Environmental Quality; Lansing, MI: 2011. Available: http://www.michigan.gov/documents/deq/deq-aqd-aqe-amu-Annual-2011-Report_390418_7.pdf
55. Hanna SR, Paine R, Heinold D, Kintigh E, Baker D. Uncertainties in air toxics calculated by the dispersion models AERMOD and ISCST3 in the Houston ship channel area. J Appl Meteorol Climatol. 2007; 46:1372–1382.

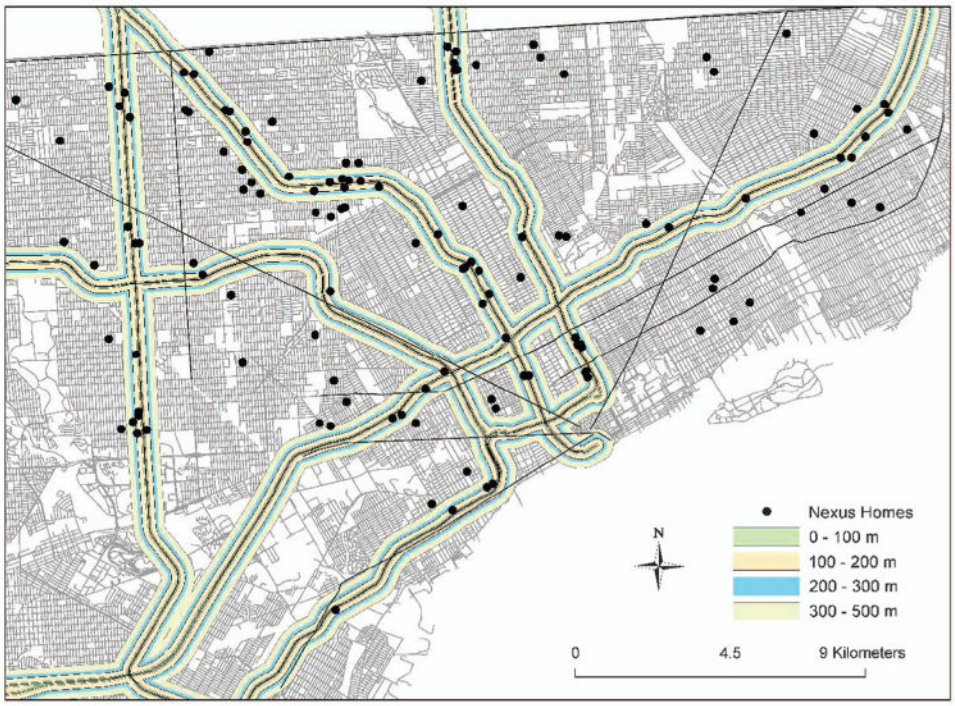

Figure 1.

Roads and buffers along highways in Detroit from which children were recruited to participate in NEXUS. Locations of 160 NEXUS homes are shown as dots.

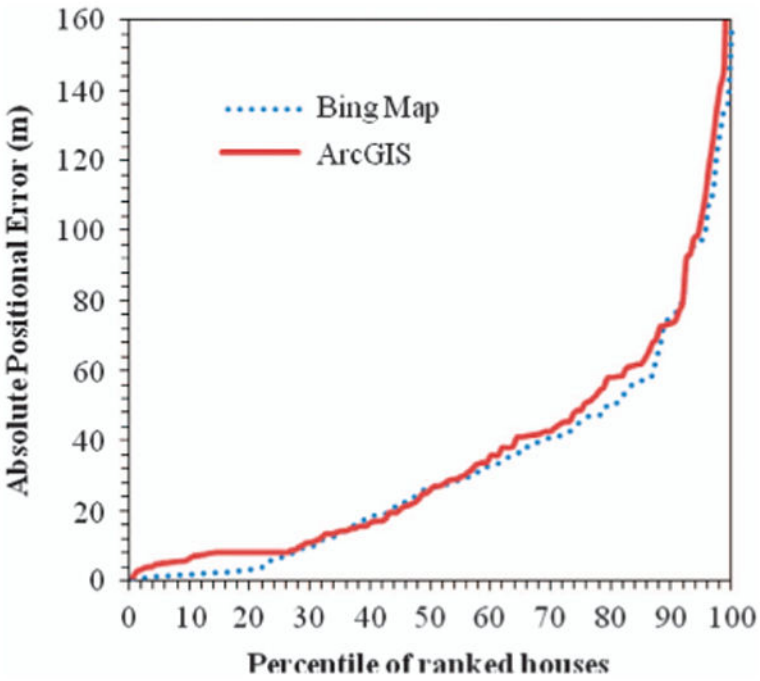

Figure 2.
Cumulative distribution of absolute positional errors for 160 NEXUS homes.

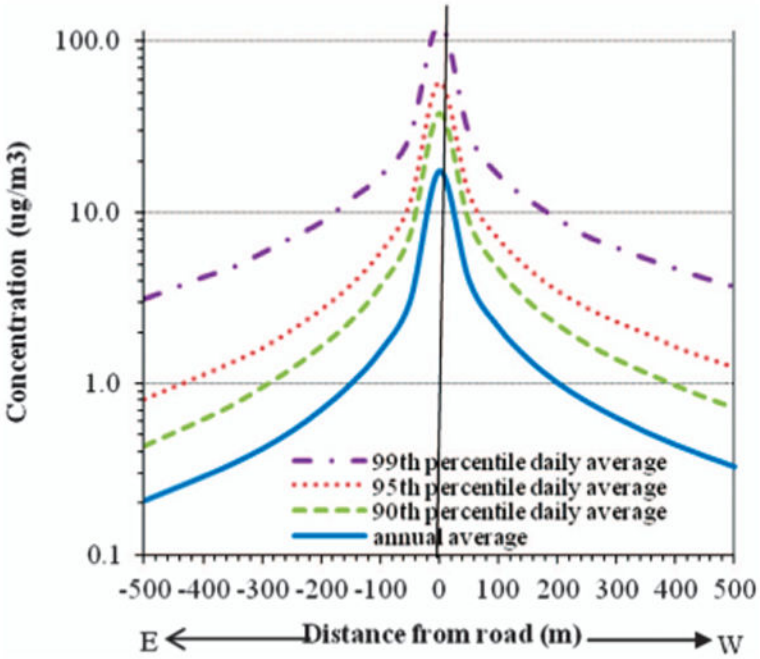

Figure 3.

Crossroad concentrations of $PM_{2.5}$ for test case using prediction from RLINE dispersion model and 2010 Detroit meteorology. Results show annual and 24-h averages.

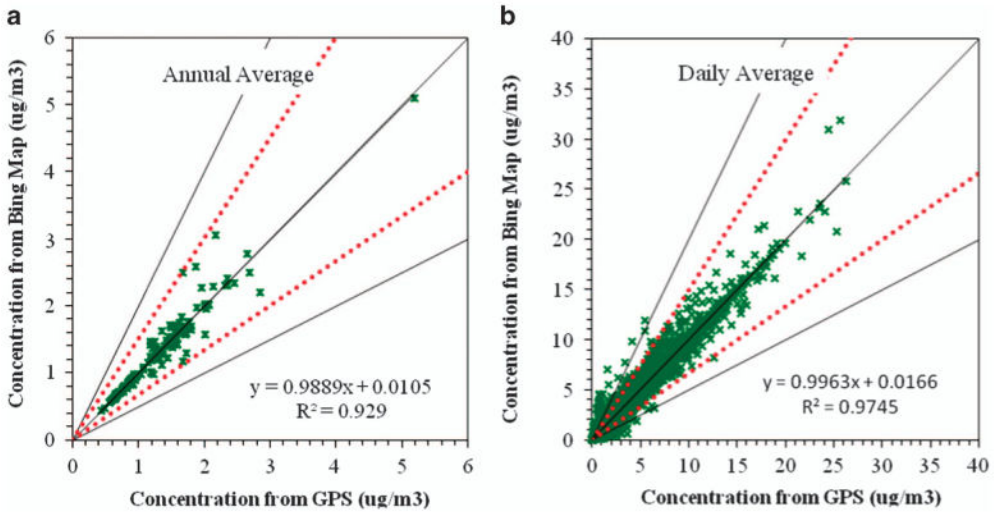

Figure 4.

Scatterplots contrasting annual average (left) and 24-h maximum (right) $PM_{2.5}$ concentrations using GPS and automated geocoding home locations. Concentrations predicted using RLINE dispersion model and 2010 Detroit meteorology. Plots show 160 homes; dashed red lines indicate a factor of 1.5 agreement; black lines indicate a factor of 2 agreement.

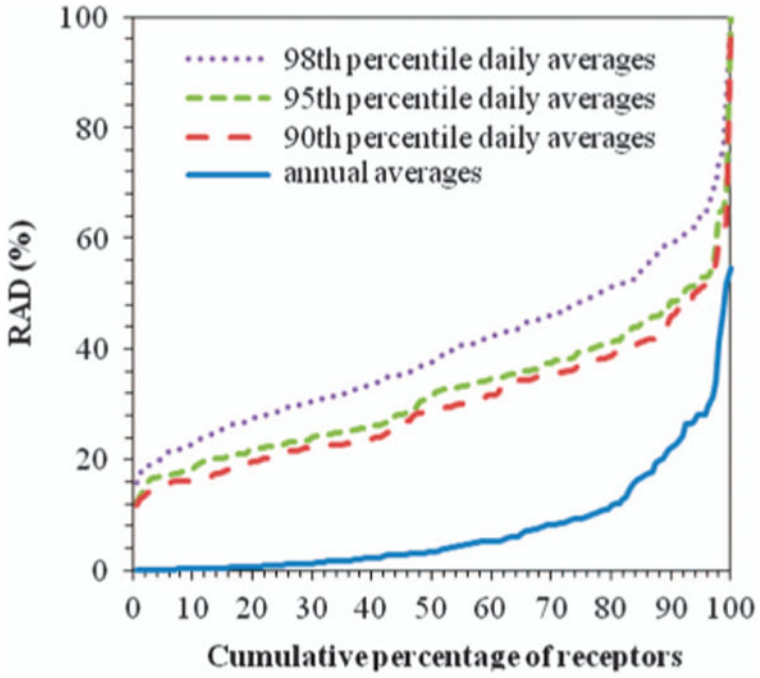

Figure 5.
Relative absolute differences in annual average and 90th, 95th and 98th percentile 24-h concentrations at NEXUS receptors.

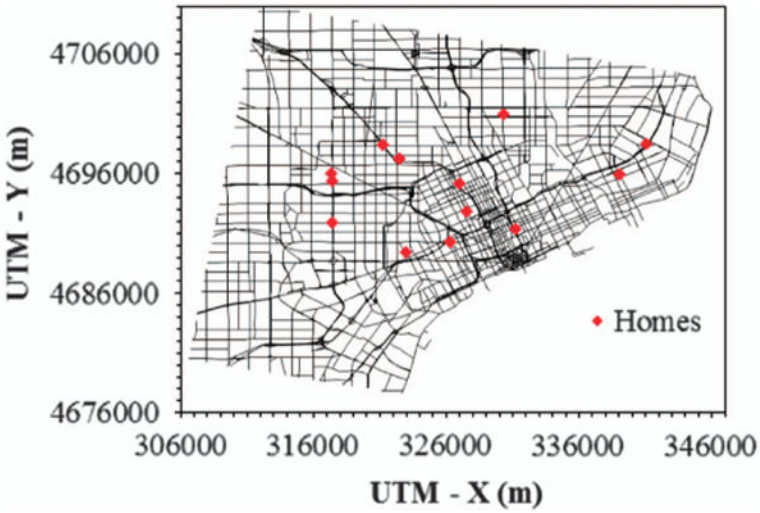

Figure 6.
Location of homes with relative annual average $PM_{2.5}$ concentration errors exceeding 25%.

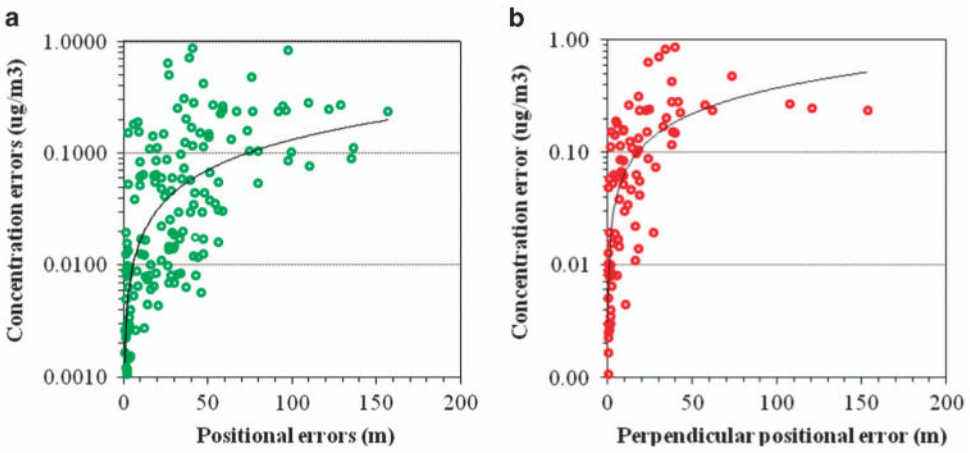

Figure 7.

Dependency of absolute annual average concentration errors at 160 homes on total positional errors (left) and perpendicular errors to the major roads for residences within 500 m (right). Power law model shown as solid line.

Table 1

Summary of positional errors (m).

					Percentiles					
Error	**Geocoder**	**Mean**	**SD**	**Min**	**25th**	**50th**	**75th**	**95th**	**Max**	**No. of homes**
Errors in geocoding homes										
	Bing Map	32	32	1	7	26	45	97	157	160
	ArcGIS	35	35	1	8	26	48	100	196	160
Difference between home-to-road edge versus home-to-road link										
	GPS	23	17	1	7	21	34	49	72	82

Table 2

Relative difference and absolute concentration errors for both annual and maximum 24-h averages because of home geocoding errors.

					Percentiles					
Averages	**Distance bin**	**Mean**	**SD**	**Min**	**25th**	**50th**	**75th**	**95th**	**98th**	**Max**
Absolute concentration error (μg/m^3)										
	0–100 m	0.16	0.21	0.00	0.05	0.09	0.16	0.67	0.75	0.84
Annual average	100–200 m	0.12	0.15	0.00	0.01	0.05	0.20	0.28	0.37	0.88
	200–500 m	0.06	0.09	0.00	0.01	0.02	0.04	0.21	0.25	0.27
	>500m	0.04	0.08	0.00	0.01	0.01	0.04	0.20	0.27	0.51
	All	0.09	0.15	0.00	0.01	0.03	0.11	0.29	0.62	0.88
	0–100 m	0.16	0.34	0.00	0.02	0.06	0.15	0.65	1.13	6.79
Maximum 24-h average	100–200 m	0.12	0.25	0.00	0.01	0.03	0.13	0.48	0.84	5.93
	200–500 m	0.06	0.20	0.00	0.00	0.01	0.03	0.28	0.51	6.27
	>500m	0.04	0.13	0.00	0.00	0.01	0.02	0.19	0.42	2.88
	All	0.09	0.24	0.00	0.00	0.02	0.08	0.40	0.76	6.79
Relative error (%)										
	0–100 m	10	13	0	3	5	10	35	53	54
	100–200 m	10	11	0	1	5	17	27	33	47
	200–500 m	6	9	0	2	2	4	21	25	28
Annual average	>500m	5	8	0	1	2	8	22	32	41
	All	8	10	0	1	3	9	28	40	54
	0–100 m	10	14	0	2	6	11	45	58	87
Maximum 24-h average	100–200 m	10	12	0	1	4	14	37	49	83
	200–500 m	6	11	0	1	2	4	36	45	73
	>500m	5	9	0	1	2	5	25	41	77
	All	8	12	0	1	3	9	33	49	87