Characterizing the spatial distribution of multiple pollutants and populations at risk in Atlanta, Georgia

John L. Pearce, *Medical University of South Carolina*
Lance Waller, *Emory University*
Stefanie Sarnat, *Emory University*
Howard Chang, *Emory University*
Mitchel Klein, *Emory University*
James A. Mulholland, *Georgia Institute of Technology*
Paige Tolbert, *Emory University*

**Journal Title:** Spatial and Spatio-temporal Epidemiology
**Volume:** Volume 18
**Publisher:** Elsevier | 2016-08, Pages 13-23
**Type of Work:** Article | Post-print: After Peer Review
**Publisher DOI:** 10.1016/j.sste.2016.02.002
**Permanent URL:** [https://pid.emory.edu/ark:/25593/s4k92](https://pid.emory.edu/ark:/25593/s4k92)

Final published version: [http://dx.doi.org/10.1016/j.sste.2016.02.002](http://dx.doi.org/10.1016/j.sste.2016.02.002)

**Copyright information:**
© 2016 The Authors. Published by Elsevier Ltd.
This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License ([http://creativecommons.org/licenses/by-nc-nd/4.0/](http://creativecommons.org/licenses/by-nc-nd/4.0/)).

Accessed November 25, 2017 10:13 PM EST
Characterizing the spatial distribution of multiple pollutants and populations at risk in Atlanta, Georgia

John L. Pearce\textsuperscript{a,}\textsuperscript{*}, Lance A. Waller\textsuperscript{b}, Stefanie E. Sarnat\textsuperscript{c}, Howard H. Chang\textsuperscript{b}, Mitch Klein\textsuperscript{c}, James A. Mulholland\textsuperscript{d}, Paige E. Tolbert\textsuperscript{c}

\textsuperscript{a} Department of Public Health Sciences, College of Medicine, Medical University of South Carolina, 135 Cannon Street, Charleston, SC 29422, United States
\textsuperscript{b} Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, United States
\textsuperscript{c} Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, United States
\textsuperscript{d} School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, United States

\section*{Abstract}

\textbf{Background}: Exposure metrics that identify spatial contrasts in multipollutant air quality are needed to better understand multipollutant geographies and health effects from air pollution. Our aim is to improve understanding of: (1) long-term spatial distributions of multiple pollutants; and (2) demographic characteristics of populations residing within areas of differing air quality.

\textbf{Methods}: We obtained average concentrations for ten air pollutants (p = 10) across a 12 km grid (n = 253) covering Atlanta, Georgia for 2002–2008. We apply a self-organizing map (SOM) to our data to derive multipollutant patterns observed across our grid and classify locations under their most similar pattern (i.e., multipollutant spatial type (MST)). Finally, we geographically map classifications to delineate regions of similar multipollutant characteristics and characterize associated demographics.

\textbf{Results}: We found six MSTs well describe our data, with profiles highlighting a range of combinations, from locations experiencing generally clean air to locations experiencing conditions that were relatively dirty. Mapping MSTs highlighted that downtown areas were dominated by primary pollution and that suburban areas experienced relatively higher levels of secondary pollution. Demographics show the largest proportion of the overall population resided in downtown locations experiencing higher levels of primary pollution. Moreover, higher proportions of nonwhites and children in poverty reside in these areas when compared to suburban populations that resided in areas exhibiting relatively lower pollution.

\textbf{Conclusion}: Our approach reveals the nature and spatial distribution of differential pollutant combinations across urban environments and provides helpful insights for identifying spatial exposure and demographic contrasts for future health studies.

\section*{Introduction}

Air quality within urban environments involves a mixture of gaseous and particulate concentrations that are affected by a variety of emission sources, local topographies,
and meteorological conditions. As such, complex spatial patterning can occur in urban air quality making the variability of such phenomena difficult to characterize as different pollutants often exhibit differential spatial patterns (e.g., ozone vs. nitrogen dioxides). This is a concern for health scientists in the field of air pollution epidemiology who need to identify appropriate spatial contrasts in their exposure assessments of air pollution (Marshall et al., 2008; Hajat et al., 2013). Such challenges, in part, have led investigators performing chronic exposure studies to typically focus on one pollutant at a time (Hoek et al., 2013); however, it is well understood that intercorrelations among various pollutants can be problematic for statistical models designed to estimate individual pollutant risk (Tolbert et al., 2007; Jerrett et al., 2013). Therefore, investigations reporting associations between long-term exposure to air pollution and adverse health generally acknowledge that reported associations are likely the result of a pollutant mixture, not the sole effect of the proxy pollutant (Pope et al., 2004; Lee et al., 2009; Hoek et al., 2013).

In order to improve our understanding of the health effects of long-term exposure to multiple pollutants it is necessary to examine the entire mix of pollutants (Dominici et al., 2010; Vedal and Kaufman, 2011; Levy et al., 2014). However, expanding chronic exposure studies of air pollution to incorporate information on multiple pollutants is expected to be challenging for at least two reasons: (1) measuring/modeling the joint spatial distribution of multiple air pollutants is difficult (Jerrett et al., 2005; Marshall et al., 2008; Riley et al., 2014; Sororian et al., 2014), and (2) characterizing the spatial distribution of multipollutant exposure is complex (Oakes et al., 2014). To further complicate matters, different subgroups within the populations at risk (e.g., those with low socioeconomic status (SES)) may be more intensely exposed to air pollution than others, a situation that may confound estimated associations between air pollution and health (Laurent et al., 2007; Yanosky et al., 2008; Hajat et al., 2013).

Given such challenges, development of approaches that can be useful for investigating the health effects of complex multipollutant exposures are highly desired (Dominici et al., 2010). Recently, many techniques have been presented for characterizing multipollutant exposure (Oakes et al., 2014); however, very few have been applied in spatial settings (Molitor et al., 2011; Austin et al., 2013). Although limited, findings from these studies have noted significant spatial variation in multipollutant exposures within and across cities in the US. Therefore, it is clear more studies are needed to better understand spatial variation of complex exposures as well as heterogeneity in exposure to populations at risk.

In the present study, we use Atlanta, Georgia, as a case study to illustrate a methodological approach for characterizing long-term trends in population exposure to multiple pollutants. Atlanta’s air quality issues are well known and several studies have documented associations with health outcomes including asthma, cardiorespiratory morbidity, and preterm births (Alhanti et al., 2015; Chang et al., 2015; Pearce et al., 2015; Winquist et al., 2015). Moreover, a novel set of spatially and temporally resolved multipollutant data is available for the region (Sororian et al., 2014) that will allow us to more closely examine air pollution exposure across a unique and diverse population (Pooley, 2015). Our general objective is to determine whether and to what extent long-term patterns in multipollutant combinations and populations at risk systematically map onto one another in the Atlanta region. We aim to achieve our objective by addressing the following questions of interest:

1. What types of long-term multipollutant combinations occur at locations within our study?
2. What is the spatial distribution of types of multipollutant combinations across our study region?
3. What demographics are associated with areas differentiated by types of multipollutant combinations?

In answering these questions we hope to improve future epidemiologic studies by increasing our understanding of: (1) the long-term geographic patterns of multipollutant air quality across our study region; and (2) the demographic makeup of populations residing in areas that experience distinct long-term multipollutant exposure.

### 2. Methods

The principal focus of our approach is to identify geographic locations in our study area with similar long-term multipollutant characteristics in order to better understand local, long-term population exposure to ambient multipollutant mixtures. This is achieved in four stages: (1) divide the study area into grid cells, within which it is assumed the spatial distribution of pollution is relatively homogeneous, (2) define a number of multipollutant spatial types that describe the nature of the pollutant attributes of the grid cells, (3) characterize multipollutant spatial types by mapping grid assignments to multipollutant spatial types in the study area, and (4) describe the demographic characteristics of the populations residing in locations corresponding to areas defined by the multipollutant spatial types.

#### 2.1. Multipollutant air quality data acquisition

Available data for this study included seven years (2002–2008) of spatially and temporally resolved air pollution concentrations at a twelve kilometer gridded spatial resolution for ten ambient air pollutants obtained for a 31,285 km² study area encompassing Atlanta, Georgia (Sororian et al., 2014). This area contained 253 grid cells (Fig. 1). In brief, data at each grid cell are daily concentration estimates obtained from calibrating gridded output from the Community Multi-scale Air Quality (CMAQ) model against measurements from monitoring sites in the study area – a.k.a. ‘fusion’ data (Sororian et al., 2014). Pollutants available included 1-h maximum carbon monoxide (CO) in ppm, 1-h maximum nitrogen dioxide (NO₂) and nitrous oxides (NOₓ) in ppb, 8-h maximum ozone (O₃) in ppb, 1-h maximum sulfur dioxide (SO₂) in ppb, and five 24-h average PM₂.₅ components in μg/m³: elemental carbon (EC), organic carbon (OC), nitrate (NO₃), ammonium (NH₄), and sulfate (SO₄). See Table 1 for summary statistics of these data.
2.2. Identify spatial profiles that define multipollutant spatial types (MSTs)

To address our first question, we apply an unsupervised learning tool known as the self-organizing map (SOM) to identify the types of multipollutant combinations that occur among the grid cells in our study area (Kohonen, 2001). SOM uses an optimized clustering procedure to identify data-driven profiles that are used to formulate categories and then projects resulting profiles onto a spatially organized array – the ‘map’. We find the SOM algorithm to be appealing for air pollution mixture studies as it has the additional benefit of using the ‘map’ for visualization, a feature we find particularly

Table 1
Summary statistics for air pollution with our study area during the years 2002 to 2008. (SD: standard deviation; CV: SD/mean; n = 646,415 grid cell-days).

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Units</th>
<th>Daily temporal metric</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO</td>
<td>0.39</td>
<td>0.15</td>
<td>0.25</td>
<td>1.01</td>
<td>ppm</td>
<td>1 h max</td>
<td>0.38</td>
</tr>
<tr>
<td>NO₂</td>
<td>10.44</td>
<td>6.40</td>
<td>3.55</td>
<td>35.42</td>
<td>ppb</td>
<td>1 h max</td>
<td>0.61</td>
</tr>
<tr>
<td>NOₓ</td>
<td>0.020</td>
<td>0.015</td>
<td>0.006</td>
<td>0.091</td>
<td>ppm</td>
<td>1 h max</td>
<td>0.76</td>
</tr>
<tr>
<td>O₃</td>
<td>0.043</td>
<td>0.001</td>
<td>0.038</td>
<td>0.045</td>
<td>ppm</td>
<td>8 h max</td>
<td>0.02</td>
</tr>
<tr>
<td>SO₂</td>
<td>8.10</td>
<td>2.90</td>
<td>4.35</td>
<td>28.50</td>
<td>ppb</td>
<td>1 h max</td>
<td>0.36</td>
</tr>
<tr>
<td>EC</td>
<td>0.69</td>
<td>0.24</td>
<td>0.33</td>
<td>1.69</td>
<td>μg/m³</td>
<td>24 h avg</td>
<td>0.35</td>
</tr>
<tr>
<td>OC</td>
<td>2.62</td>
<td>0.29</td>
<td>1.84</td>
<td>3.48</td>
<td>μg/m³</td>
<td>24 h avg</td>
<td>0.31</td>
</tr>
<tr>
<td>NH₄</td>
<td>1.34</td>
<td>0.13</td>
<td>1.06</td>
<td>1.65</td>
<td>μg/m³</td>
<td>24 h avg</td>
<td>0.10</td>
</tr>
<tr>
<td>NO₃</td>
<td>0.60</td>
<td>0.08</td>
<td>0.43</td>
<td>0.78</td>
<td>μg/m³</td>
<td>24 h avg</td>
<td>0.14</td>
</tr>
<tr>
<td>SO₄</td>
<td>4.30</td>
<td>0.22</td>
<td>3.87</td>
<td>5.13</td>
<td>μg/m³</td>
<td>24 h avg</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Fig. 1. Map of study area illustrating air quality modeling fusion grid and population density. Note the white area in the south central portion reflects the location of Hartsfield–Jackson International Airport (no residents).
useful when trying to understand relationships between profiles.

2.3. SOM algorithm

In order to apply SOM two components must be specified by the user – the input data matrix and the output map (Fig. 2). Here, the input matrix is our multipollutant data set, Z:

$$Z = \begin{bmatrix} Z_{11} & \cdots & Z_{1p} \\ \vdots & \ddots & \vdots \\ Z_{n1} & \cdots & Z_{np} \end{bmatrix}$$

where \( n \) denotes the number of grid cell locations and \( p \) the number of pollutants. Each grid cell is represented by a row \( Z_j \) within \( Z \). The output collection of nodes (i.e., multipollutant profiles) is the “map”, \( M \):

$$M = \begin{bmatrix} m_{1x} & \cdots & m_{xy} \\ \vdots & \ddots & \vdots \\ m_{nx1} & \cdots & m_{nxl} \end{bmatrix}$$

with each profile \( m \) represented as a node at location \((x, y)\) on the map (Fig. 2). Note \( X \times Y \) determines the number of nodes (i.e., number of profiles) and the arrangement (e.g., 1D or 2D) of \( M \). The shape of \( M \) is most commonly rectangular but can be other variations (e.g., hexagonal). Each node \( m \) is associated with a profile defined as vector \( w_m \):

$$w_m = [\mu_{m1}, \mu_{m2}, \ldots, \mu_{mpl}]$$

where \( \mu \) are ‘learned’ coefficient values corresponding to the pollutant concentration values that characterize profile \( m \).

Operationally, SOM implements the following steps. First, given \( M \), map initialization occurs with each \( m \) being assigned a preliminary \( w_m \) from a random selection of \( Z_i \)’s. Then, iterative learning begins where, for each iteration \( t \), the algorithm randomly chooses a grid cell’s profile \( Z_i^{(t)} \) from \( Z \) and then computes a measure of (dis)similarity (in our case the Euclidean distance) between the observation \( Z_i^{(t)} \) and each \( w_m^{(t)} \). Next, SOM provisionally assigns a best matching node \( m^*(t) \) whose \( w_{m^*} \) is most similar to each \( Z_i^{(t)} \). Next, class profile development occurs via the Kohonen learning process:

$$w_m^{(t+1)} = w_m^{(t)} + \alpha(t)N_{m^*}(t)[Z_i^{(t)} - w_m^{(t)}]$$

where \( \alpha \) is the learning rate, \( N_{m^*} \) is a neighborhood function that spatially constrains the neighborhood of \( m^* \) on \( M \), and \( Z \) is the mean of pollutant values on days provisionally assigned to the nodes within the neighborhood set. The learning rate controls the magnitude of updating that occurs for \( t \). The neighborhood function, which activates all nodes up to a certain distance on \( M \) from \( m^* \), forces similarity between neighboring nodes on \( M \). Eq. (4) updates coefficients within a neighborhood of \( m^* \), where the impact of the neighborhood decreases over iterations.

SOM performance is dependent on both \( \alpha \) and \( N \) and thus mappings are sensitive to these parameters. Therefore, in effort to provide guidance we note that \( \alpha \) typically starts as small number and is specified to decrease monotonically (e.g., 0.05–0.01) as iterations increase. Similarly, the range of \( N \) starts large (e.g., 2/3 map size) and decreases to 1.0 over a predetermined termination period (e.g., 1/3 of iterations), after which fine adjustment of the map occurs.

Training continues for the number of user-defined iterations. Kohonen recommends the number of steps be at
least 500 times the number of nodes on the map. Once training is complete, results include final coefficient values for each node’s $w_m$, classification assignments for each day $Z_t$, and coordinates of nodes on $M$. The final step is to visualize the class profiles by plotting the map. For additional details regarding SOM, please refer to the book of Kohonen (2001).

2.4. SOM implementation

Application of SOM requires three steps. First, we calculate long-term means for each pollutant at each CMAQ grid cell during years 2002–2008. Next, we standardize the long-term averages of each pollutant by grid cell by subtracting the overall grid cell mean and dividing by the standard deviation in order to remove the absolute differences between variable magnitudes of different pollutants yet retain ratios between variable amplitudes. We then determined an appropriate number of spatial profiles by assessing (1) the grouping structure of our data, (2) the information retained by resulting classifications, and (3) the area size of the categories in order to better understand the number of potentially exposed. We evaluate grouping structure using principal component analysis (PCA), information retention using regression models where SOM classifications are assessed as a categorical predictor for each pollutant in the profile, and potential area size through evaluation of grid cell class assignments. This information was then used collectively to determine the number of profiles for the SOM algorithm.

Once the number of profiles was determined, SOM was applied to our entire data set using parameters described in Pearce et al. (2014). Resulting spatial profiles are referred as multipollutant spatial types (MSTs) and are referenced on the ‘map’ using SOM $[x,y]$ coordinates. It is important to note that our SOM is not a geographic map but rather a projection of resulting profiles onto a two-dimensional grid where locational proximity reflects profile similarity. In short, SOM aims to preserve the topology of the original multidimensional data space, a feature that results in neighboring profiles being more similar and distant profiles being more dissimilar. To enhance interpretation, MST profiles are visualized using barplots with mean centered concentrations on a percentage scale.

2.5. Geographic distribution of multipollutant spatial types

We visualize the spatial distribution of multipollutant combinations across our study area (to address our second question) using color-coded map that differentiates locations based on their assigned MST. The result distinguishes area boundaries among grid cells based on individual cells long-term air quality and serves to identify regions defined by MSTs. The map was spatially referenced using North American Datum 1983 and projected using the Georgia Statewide Lambert Conformal Conic system.

2.6. Population characteristics of multipollutant spatial types

To address our third question, we obtained population data from the US Census Topologically Integrated Geographic Encoding and Referencing (TIGER) products that provide geographic boundary data merged with 2010 Census data and 2008–2012 5-year estimates from the American Community Survey (United States Census Bureau, 2010). Data were collected at the census tract level and variables of interest included total population, child subpopulation (aged < 18 years), nonwhite subpopulation, and the percent of children (aged < 18 years) living in poverty. Poverty statistics presented in the ACS rely on a set of money income thresholds that vary by family size and composition. If the family’s income is less than the federal poverty threshold, then family and all included individuals are considered to be in poverty. The poverty thresholds do not vary geographically and are updated annually to allow for changes in the cost of living. For more detail see: “How Poverty is Calculated in the ACS” (US Census Bureau, 2015).

We then used geographic information systems to match census tracts to the grid cell (Fig. 1) in which their geographic centroids were located. Once matched, we calculate aggregate summaries for demographics under each MST category in order to get region specific population summaries.

3. Results

3.1. Selecting the number of multipollutant spatial profiles

PCA projections suggest at least five primary modes of variation in our data: (1) CO, NO2, NOX, and EC; (2) NOX and NH4; (3) SO4 and OC; (4) SO2; and (5) O3 (Fig. 3). These can generally be described as the building blocks of air quality in Atlanta and likely correspond to variation driven by traffic related pollutants (1), secondary inorganic aerosols (2), secondary organic aerosols and sulfate (3), sulfur dioxide emissions (4), and region-wide ozone levels (5).

Frequency counts of grid cells assigned to each spatial type illustrate an anticipated reduction in the sample size as class number increases (i.e., fewer grid cells in each class when there are more classes) (Fig. 3b). We prefer our SOM analysis to provide categorizations that will be useful for further analysis and thus we have added a reference line of 10%, which shows when our classifications capture ‘rare’ spatial profiles. For example, we see that a SOM classification with eight profiles identifies three spatial types that were observed in less than 10% of the locations.

Results from using SOM classes as categorical predictors of individual pollutant variance show a strong relationship between the number of classes and the explanatory power of the SOM classification (Fig. 3c). The ability of the SOM classification to predict long-term average differs among the pollutants evaluated with NOX and NO2 generally being explained well and SO2 and O3 being explained poorly.

In combination, these results display aspects of the underlying variance structure in the data and illustrate how different partitions of the data can be used to capture features of interest for exposure characterization. For this study, we determined that a partition of the data into six multipollutant categories was appropriate as it reasonably captures the variation of our pollutants (Fig. 3c) and has the benefit of samples sizes that identify both typical and rarer combinations in the data (Fig. 3b).
3.2. Spatial profiles for multipollutant spatial types

To begin, we present a $3 \times 2$ SOM characterizing ambient air quality using six categories of locations reflecting the range of multipollutant combinations modeled at locations in our study area. Each category defines a spatial profile describing a multipollutant spatial type (MST) and is referenced using SOM $[x,y]$ coordinates. Furthermore, relative concentrations of pollutants for the MST profiles are visualized using barplots with mean centered values on a percentage scale (Fig. 4) and actual concentrations are presented in Table 2.

The most typical grid cells in our study area are characterized by the MST profiles on the bottom row of Fig. 4. The most common, MST [2,1], identified that 33% of the grid cells in our study area experienced below average concentrations for all pollutants. The second most frequent, MST [3,1], identifies that 26% of locations experienced conditions with above average NH$_4$ and NO$_3$ in combination with below average concentrations for all other pollutants. MST [1,1] captures conditions that experienced well below average concentrations for all pollutants and covered 19% of locations in our study area.

![PCA Plot of Long-term Multipollutant Profiles for 12km Grid](image_url)

**Fig. 3.** Graphical and statistical evaluation measures used to guide number of spatial profiles for self-organizing map (SOM) analysis. Panel (a) presents a principal components analysis projection of our multipollutant data. The gray points represent the scores for each location along the first two principal components and the dark arrows indicate the corresponding loading vectors for each pollutant. Panel (b) displays the distribution of frequency assignments for each spatial type. Gray points reflect observed frequencies and trend line reflects the mean. Panel (c) displays the distribution of adjusted $R^2$ values from simple linear regression models fit to each pollutant using multipollutant spatial types as a categorical predictor in the model. Each pollutant has a unique symbol and the trend line reflects the mean $R^2$.

**Table 2**

Air pollutant and geographic summary statistics (mean (SD)) for grid cells assigned to each multipollutant spatial type.

| SOM  | Area ($km^2$) | Area % CO | NO$_2$ | NO$_x$ | O$_3$ | SO$_2$ | EC | OC | NH$_4$ | NO$_3$ | SO$_4$
|------|---------------|-----------|--------|--------|------|--------|----|----|--------|--------|--------
| [1,1]| 2284.1        | 7.7       | 0.27 (0.01) | 5.42 (2.48) | 0.01 (0) | 0.04 (0) | 6.37 (1.17) | 0.47 (0.07) | 2.34 (0.25) | 1.2 (0.09) | 0.52 (0.06) | 4.05 (0.09)
| [1,2]| 1774.1        | 6.0       | 0.4 (0.06) | 11.7 (2.92) | 0.02 (0.01) | 0.04 (0) | 14.85 (4.66) | 0.76 (0.12) | 2.76 (0.15) | 1.4 (0.1) | 0.6 (0.05) | 4.73 (0.15)
| [2,1]| 11680.3       | 39.2      | 0.34 (0.05) | 8.44 (2.26) | 0.01 (0) | 0.04 (0) | 8.55 (1.67) | 0.63 (0.1) | 2.66 (0.19) | 1.27 (0.08) | 0.55 (0.05) | 4.31 (0.12)
| [2,2]| 4316.3        | 14.5      | 0.61 (0.09) | 20.39 (3.19) | 0.04 (0.01) | 0.04 (0) | 9.22 (1.7) | 1.06 (0.12) | 2.97 (0.11) | 1.39 (0.06) | 0.62 (0.04) | 4.46 (0.13)
| [3,1]| 8247.6        | 27.7      | 0.36 (0.05) | 8.77 (2.49) | 0.02 (0) | 0.04 (0) | 6.09 (1.03) | 0.61 (0.11) | 2.5 (0.2) | 1.48 (0.07) | 0.7 (0.05) | 4.23 (0.13)
| [3,2]| 1479.9        | 5.0       | 0.86 (0.11) | 30.44 (2.71) | 0.07 (0.01) | 0.04 (0) | 11.69 (1.8) | 1.39 (0.13) | 3.18 (0.16) | 1.47 (0.05) | 0.65 (0.02) | 4.63 (0.13)
In the upper row of the map we find spatial types that were less common and indicative of grid cells experiencing higher levels of long-term average pollution. In the upper left, MST [1,2] covers 6% of grid cells in the study region that experienced the highest long-term concentrations of SO\textsubscript{2} in conjunction with slightly above average concentrations for all other pollutants except O\textsubscript{3} and NO\textsubscript{3}. MST [2,2] and MST [3,2] covered a combined 16% of locations with similar profiles exhibiting higher than average concentrations for all pollutants (in particular primary pollutants) except O\textsubscript{3}, which was slightly below average. However, MST [3,2] presents concentrations that are 1.2 times higher overall than MST [2,2]. The identified spatial profiles captured a range of combinations present in the data for locations in our study area, from conditions where all pollutants measured relatively low to conditions with high concentrations of secondary or primary pollutants or both.

3.3. Geographic distribution of multipollutant spatial types

Mapping the locational assignments of each MST illustrates an approach for characterizing the spatial distribution of the types of ambient air quality mixtures found in our study area (Fig. 5). Results reveal strong contiguity of the classification assignments and indicate a tendency for multipollutant combinations to regionalize across the study area. Central locations representing the urban core in our study area were assigned to MST [3,2] and [2,2], indicating that 20% of the study area experienced relatively high levels of primary pollution. Given the proximity to major interstate highways, these conditions are likely reflective of areas that experienced high traffic volume. Moving away from downtown we can see that 28% of the study region, primarily in the upper northeastern corner, is dominated by above average long-term levels of NH\textsubscript{4} and NO\textsubscript{3} (MST [3,1]). Moving to the west of downtown we see a small collection of disjointed areas dominated by MST...
3.4. Population characteristics of multipollutant spatial types

Demographic summaries indicate that the largest proportion of the population (Table 3) resides in locations with air quality defined by MST [2,2], suggesting that a substantial segment of the Atlanta population experienced relatively high long-term levels of primary pollution (CO, NO₂, NOₓ, and EC) during the study period. The second most populated air quality region is MST [3,2], which is a high pollution region that encompasses the downtown area. Population demographics (percent children, percent nonwhite, and percent of children below poverty level) in each MST follow trends in total population (as expected). However, considering the composition of the population in each MST, MSTs [2,2] and [3,2] have higher population densities, higher proportions of nonwhite residents (47% and 60%) and MST [3,2] has the highest proportion of children living in poverty (6% of total population; 24% of child population) compared to other MSTs (≤5% of total population; ≤17% of child population).

4. Discussion

In this study, we deconstructed complex air pollution data into six multipollutant spatial types (MSTs) that represent long-term patterns in air quality at locations across our study area. Overall, the identified MSTs captured a range of air quality scenarios across our locations as profiles included conditions dominated by: relatively low
levels of pollution, relatively high concentrations of single pollutants, and relatively high levels of multiple pollutants (Fig. 3). We found that the spatial contrasts were most evident for primary pollutants — in particular oxides of nitrogen (NO\textsubscript{X} and NO\textsubscript{2}) and to a somewhat lesser degree CO, EC, and SO\textsubscript{2}. These results agree well with other spatial studies of ambient air pollution in Atlanta and those that have relied on oxides of nitrogen to capture spatial variation in ambient air pollution (Briggs et al., 1997; Wade et al., 2006). Our results also identified very little spatial variation for secondary pollutants, i.e., O\textsubscript{3}, OC, NH\textsubscript{4}, NO\textsubscript{3}, and SO\textsubscript{4}. Such results indicate that primary pollutants (particularly traffic related) may be most useful in identifying spatial contrasts for long term health effects studies of multipollutant mixtures. It is important to note that the data used in this study are the result of a data fusion between ambient air monitoring data and CMAQ model estimates and thus MST profiles reflect a blend of observations and expected concentrations based on the geographic distribution emissions and meteorology in the region (Sororian et al., 2014).

Mapping the geographic assignments of the MSTs reveals strong patterns in the spatial distribution of multipollutant air quality and resulted in the identification of clearly delineated multipollutant regions in our study area (Fig. 5). With these results we found a general pattern of air quality slowly shifting from locations dominated by higher concentrations of primary pollutants to locations dominated by secondary pollution as one moves further away from downtown, however, certain areas outside the downtown area did experience elevated long-term NH\textsubscript{4} and NO\textsubscript{3} or SO\textsubscript{2}. As such, it is clear that the strongest multipollutant contrasts are found between central urban locations and peripheral suburban areas.

Analysis of the demographics associated with our MST regions showed that the largest proportion of the population, along with the largest proportions of our subpopulations of interest (children, nonwhite, and children in poverty), resided in locations where air quality was generally dominated by relatively high long-term levels of primary pollution. This finding agrees well with other studies that have shown that higher exposure to air pollution occurs in communities with higher proportions of poverty and minorities (Molitor et al., 2011; Hajat et al., 2013).

We also illustrate how an unsupervised learning tool (SOM) can be paired with geographic information systems to identify regions experiencing similar multipollutant air quality within an intraurban environment from complex data. The SOM approach has the attractive ability to deconstruct complex data into an interpretable collection of categories that can be visualized on an array revealing associations (the SOM ‘map’) and in a geographic context (the traditional map) to promote further understanding of interclass relationships across the study area. For example, looking at the organization of profiles on our SOM (Fig. 4), we are able to generally infer that residents assigned to MST [3,2] experience air quality similar to residents of MST [2,2] but very different air quality than residents assigned to MST [1,1]. This is because proximity reflects similarity of profiles on the SOM.
A limitation of the approach presented here is our use of mean values at each location to represent local long-term pollution over the study period. Alternative measures such as the maximum or variance, may be more useful in identifying locations that experience the most extreme conditions or the most variable. Another potential shortcoming is our inclusion of pollutants that demonstrate limited spatial variation across our study area. For example, ozone concentrations were found to be quite similar across our MSTs and thus it is likely that O_3 played a limited role in the formation of our spatial profiles. Nevertheless, we chose to include O_3 in this study to assess the approach and to identify which pollutants in our available data would be most appropriate for developing spatial profiles; O_3 is also an important health-relevant component of the air quality mixture in the Atlanta area (Strickland et al., 2010; Pearce et al., 2015; Winquist, 2015). Another potential limitation of the work was the spatial resolution of the 12 km data. While this improves spatial coverage considerably over air monitoring network data, it will be interesting to explore finer scale resolutions for identifying intraburban spatial contrasts as such data become more widely available. Finally, our choice of six multipollutant spatial types was somewhat subjective; nevertheless, optimal statistical methods (i.e., clustering statistics) for identifying groups in data may not be optimal for defining pollutant-health associations.

The natural next step from this work would be to apply SOM to generate a multipollutant exposure metric for a long-term health effects study of multiple air pollutants. Beyond application in a health study, several areas in the development of our spatial exposure metrics could be refined. One area of interest is the evaluation of the importance of geographic scale for spatial studies of multiple pollutants. For example, the development of a multipollutant metric for a study of the southeastern US might include pollutants that are different than a study of downtown Atlanta due to the differences in the size of the study domains and the nature in which pollutants may vary within them. Another area of interest, which is currently under investigation in time-series mixture studies (Bobb, 2015), involves variable selection with a goal of including only pollutants with reasonable spatial variability or pollutants with strong health associations. Another area of continued work involves the use of demographic data to guide the formation of multipollutant regions minimizing potential confounding. For example, the associations between poverty and poorer air quality seen here suggest that separation of an air pollution effect from a poverty effect in an epidemiological study may be difficult.

5. Conclusion

The method presented in this paper can be used to both elucidate the nature in which combinations of pollutants vary across geographic space and to explore associations with populations at risk of exposure. This approach can be useful for multiple purposes, including the development of epidemiologic studies of the long-term health effects of air pollutant mixtures.

Acknowledgments

This publication was made possible, in part, by US Environmental Protection Agency Grant R834799. USEPA does not endorse the purchase of any commercial products or services mentioned in the publication. Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award number K99/RO1ES023475. The content is solely the responsibility of the authors and does not necessarily represent the official views of the USEPA or NIH. We thank Anda Olsen for her comments on the writing, and the research team at the Southeastern Center for Air Pollution Epidemiology (SCAPE: http://www.scape.gatech.edu/) for their comments and reviews of this paper.

References


United States Census Bureau. TIGER/Line with selected demographic and economic data; 2010.


