Regional air pollution mixtures across the continental US

Weeber J. Requia\textsuperscript{a,}\textsuperscript{*}, Brent A. Coull\textsuperscript{b}, Petros Koutrakis\textsuperscript{a}

\textsuperscript{a} Harvard University, Department of Environmental Health, School of Public Health, Boston, MA, United States
\textsuperscript{b} Harvard University, Department of Biostatistics, School of Public Health, Boston, MA, United States

\textbf{A R T I C L E   I N F O}

\textbf{Keywords:} \\
Air pollution \\
Pollutant mixtures \\
Spatiotemporal analysis \\
Cluster analysis

\textbf{A B S T R A C T}

A limited literature body have estimated regional differences in air pollution mixtures. More comprehensive analyses are necessary to accurately depict differences in air pollution characteristics over space and time. Our objective is to further these efforts by investigating spatial differences of air pollution mixtures across the US. We employed spatially-constrained clustering approach (based on k-means algorithm) to group air pollution monitoring sites that exhibit distinct pollutant profiles or mixtures in the US over 9 years (2008–2016). We accounted for 20 chemical components of PM\textsubscript{2.5}. The resulting clusters of pollution mixtures are characterized and validated based on source emissions represented by land-use information. Our analysis resulted in 27 clusters with different number of sites. For example, the cluster 1 has 14 sites and it covers part of the southeast, including the states of North Carolina, South Carolina, Georgia, and Florida. The southwest has a very prominent cluster with 8 sites (cluster 26), covering part of the Louisiana, Mississippi, Texas, and Arkansas. In the west coast, two clusters were highlighted in our analysis, cluster 3 in California and cluster 7 in Washington and part of Oregon. Both clusters with 5 sites. We estimated that Cu, Se, NO\textsubscript{3}–, Cr, and Ba were the top five species that divided the study area into cluster of sites more effectively. Observing the concentration ratios (concentrations of the species/concentration of PM\textsubscript{2.5}) for some of these clusters, our results show that clusters 3 and 7 in the west coast represent sites with high Na ratios. Cluster 13 in the northwest and part of the Midwest represents sites with high SO\textsubscript{4}²⁻ ratio. The cluster 16 with a single site in northeast has the highest SO\textsubscript{4}²⁻ ratio, representing almost the third quartile of the SO\textsubscript{4}²⁻ ratio. This is one of the few studies focused on spatial patterns analysis to estimate regions that exhibit distinct pollutant mixtures on a large scale. We expect that further investigations can use our findings to analyze the relationship between areas that exhibit distinct pollutant mixtures and the impact of regulations, climate change, and health effects in the US.

\textsuperscript{*}Corresponding author.
\textit{E-mail addresses:} weeberb@gmail.com, wjrequia@hsph.harvard.edu (W.J. Requia).

1. Introduction

Studies have demonstrated significant regional differences in air pollution transport, emission sources, and atmospheric reactions, which lead to considerable spatial heterogeneity of pollutant mixtures (Jeong and Park, 2013; Lee et al., 2012; J. Liu et al., 2009a, b; Ying and Kleeman, 2006). Researchers and environmental agencies highlight the importance of multi-pollutant mixtures in air pollution investigations, showing that the multi-pollutant approach enhances our understanding of the multivariate relationship between pollutants at a given site (Cooper et al., 2012; Oakes et al., 2014). For example, the United States Environmental Protection Agency (US EPA) has adopted multi-pollutant approach to develop policies and plans in order to control and regulate air quality. Scientists from EPA have reported that policies and plans based on multi-pollutant approach are able to achieve greater reductions of particulates and gases at as observed by the monitoring networks, improve air quality regionally, and reduce health risks (Wesson et al., 2010).

More detailed analyses have been conducted using emission inventories and ambient data from monitoring networks for identifying multi-pollutant profiles in air pollution data (Austin et al., 2013, 2012). Cluster framework has been one of the main methods used by these previous studies. Investigators have also used pollution mixture clusters to assess the heterogeneity in the relationship between air pollution and health (Zanobetti et al., 2014). The literature has shown that the interactions among air pollutants may cause different impacts on individuals, including additive, multiplicative, and antagonistic effects (Mauderly et al., 2010). Understanding the effect of mixtures on health, rather than the effect of the individual components, is a crucial step that must be undertaken to enhance our understanding about the health effects or air pollution.

These studies compose the short literature body on regional
differences in air pollution mixtures. More comprehensive analyses are necessary to accurately depict differences in air pollution characteristics over space and time. Our objective is to further these efforts by investigating spatiotemporal regional differences of air pollution mixtures across the US.

We previously developed cluster analysis-based framework to identify distinct groups of days with similar pollutant mixtures (Austin et al., 2012), to cluster days based on weather parameters (Austin et al., 2015), and then to identify spatial patterns in PM$_{2.5}$ composition (Austin et al., 2013). Now, we aim to expand this framework by including a spatial constraints parameter and expanding the study period - over 9 years (2008–2016) to group areas that exhibit distinct pollutant profiles or mixtures in the US. These areas are referred to as air pollution regions. We will then characterize and validate the spatial patterns of pollution mixtures based on emission rates and source emissions influencing air quality.

We hypothesize that our cluster analysis based on air pollutant mixtures will make it possible to minimize within-region variability and maximize between-regional variability of regional mixture profiles. We posit that areas exhibiting similar pollutant mixtures are impacted by similar sources and atmospheric processes, and this can inform development of more targeted and regionally tailored air quality public health management practices.

2. Materials and methods

2.1. Study design and data

We evaluated spatiotemporal patterns for 20 components of ambient PM$_{2.5}$, including As, Ba, Ca, Cr, Cu, EC, Fe, K, Mn, Na, NH$_4^+$, Ni, NO$_3^-$, OC, Pb, Se, Si, SO$_2^{2-}$, V, and Zn. Other elements obtained as part of the speciation of the filters were considered but were excluded because of the data did not reach the criteria defined in the data preparation (these criteria are detailed below).

Data for this analysis were obtained from the EPA Air Data Monitoring Program - air quality data collected at outdoor monitors across the US (https://www.epa.gov/outdoor-air-quality-data). We accessed data constructed on a daily basis listed by year (data with information on the day, month, and year of monitoring). We considered the period between 2008 and 2016.

We only considered air pollution monitoring stations that have less than 25% missing observations for the elements of interest. Stations that did not meet this criterion were excluded. We also require that each season within the study period has less than 25% missing data. This is to ensure that the site means are not unduly influenced by missing data sets between 2008 and 2016. As result, 108 sites with complete datasets were selected. This 25% missing criteria was applied across all sites, regardless of whether they were sampled every day, 3 days, or 6 days.

Then, for each site, we estimated the average concentration between 2008 and 2016 for those 20 component of ambient PM$_{2.5}$. This final dataset was used in the further stages of this study, as described in the next sections.

2.2. Mixture profiles

To analyze mixtures, we created a new parameter representing species profiles. We considered a mixture that consists of $p$ species ($i = 1, 2, 3, ..., p$) measured during a time period $t$ ($t = 1, 2, 3, ..., m$) at site $j$. The profile of a species $i$ in a site $j$ and period $t$, $f_{ijt}$, is equal to the ratio $C_{i}/P_{jt}$, where, $C_{i}$ and $P_{jt}$ are the concentrations of the species $i$ and PM$_{2.5}$ in site $j$ during the period $t$, respectively. The correlation between profiles of two mixtures indicates the degree of similarity. Therefore, the normalization of species concentrations enabled us to compare mixture characteristics between and within regions. To eliminate differences in the order of magnitude between profile levels, the species fractions were standardized using a z-transformation. The z-transformation is the mean for all values minus the observation value and divided by the standard deviation for all values.

We used PM$_{2.5}$ as a reference pollutant as its concentration will be more reliably determined using different data sources, including ground monitor, satellite, and modeled data. Although our analysis accounts only for ground monitor, future studies can use our framework for satellite and modeled data. Also, PM$_{2.5}$ modifies the sum of species included in this analysis. In addition, we chose PM$_{2.5}$ because it is responsible for a large proportion of air pollution mortality. Thus, future health studies can express mixture toxicity in a risk unit per 1 g/m$^3$ of PM$_{2.5}$. Using PM$_{2.5}$ as a reference pollutant for both the mixture profiles and health effects makes it possible to directly link mixture composition to its toxicity.

The results of the mixture profiles were consolidated in a single dataset (20 profiles for each of the 108 air pollution monitoring stations) that was used in the cluster analysis, as detailed in the next section.

2.3. Spatial patterns analysis

As mentioned above, we estimated spatiotemporal clustering of pollutant mixtures based on the framework we previously developed (Austin et al., 2013). We used the k-means algorithm to partition air pollution monitoring stations into clusters. This algorithm is easily implemented, computationally efficient, and not very sensitive to outliers (Punj and Stewart, 1983). The difference from our previous framework is that in this current analysis we accounted for a spatial constraint parameter.

We employed Spatially Constrained Multivariate Clustering (SCMC) approach, which clusters multivariate vectors of pollution constituents while accounting for spatial autocorrelation among measurements in neighboring sites, to yield air pollution regions. This method estimates clusters using local measures of spatial autocorrelation based on spatial constraints parameter (we describe that in the next section). The use of spatial constraints results in the creation of air pollution regions including contiguous sites, as compared to having sites of the same air pollution region to belong to different geographic areas. We anticipate that this approach maximizes air pollution homogeneity within regions and heterogeneity across regions. Toward to this end, we applied the SCMC method using an algorithm based on unsupervised machine learning process. This algorithm supports the goal in determining the best solution that maximizes both intra-cluster similarity and inter-cluster differences (Assuncao et al., 2006; Duque et al., 2007).

2.3.1. Spatial aspect – modeling spatial relationships

In our study, the spatial aspect (spatial constraints parameter) refers to the conceptualization of spatial relationships (spatial neighboring) among the ambient monitors across the US. We chose the inverse distance method to describe those spatial relationships. This method is recommended by the literature to model variables where the closer two features (monitoring sites) are in space, the more likely they are to interact with each other (Wong et al., 2004). The inverse distance considers that every monitor is potentially a neighbor of every other monitor. We forced the model to estimate the minimum distance that ensures every feature has at least one neighbor. The inverse distance method was applied by generating a spatial weight matrix. This matrix encompasses the conceptualization of the relationships among a set of points.

2.3.2. Optimal number of clusters (Pseudo-F statistic)

In order to determine each cluster, the SCMC approach first identifies a seed site (monitors) randomly, and then assigns all sites to the closest seed site. Then, it computes a mean data center for each cluster of sites, and reassigns each site to the closest center.

The number of seed features selected randomly matched the number
of clusters. To estimate the optimal number of clusters, we applied the Calinski-Harabasz pseudo F-statistic (Calinski and Harabasz, 1974), as presented by the following equations:

\[ CH = \frac{TSS - k \times n}{k - 1} \]

(1)

where, \( CH \) is the Calinski-Harabasz pseudo F-statistic; \( n \) is the number of clusters; \( k \) is the number of sites; and \( TSS \) is defined as follows:

\[ TSS = \sum_{i=1}^{n} \sum_{j=1}^{m} (V_{ij} - \overline{V}_m)^2 \]

(2)

where \( SST \) is a reflection of between-clusters differences (described by Equation (3)) and \( SSE \) is a reflection within-clusters similarity (described by Equation (4)).

\[ SST = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} (V_{ijk} - \overline{V}_m)^2 \]

(3)

\[ SSE = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} (V_{ijk} - \overline{V}_k)^2 \]

(4)

where \( ni \) is the number of sites in cluster \( m \); \( nc \) is the number of clusters; \( nv \) is the number of variables used to group sites; \( V_{ijk} \) is the value of the \( k \)th variable in the \( j \)th site in the \( m \)th cluster; \( \overline{V}_m \) is the mean value of the \( k \)th variable; and \( \overline{V}_k \) is the mean value of the \( k \)th variable in cluster \( m \).

The pseudo F-statistic was performed for 30 simulations (simulating 2 clusters, 3 clusters, 4 clusters, up to 30 clusters). The highest pseudo F-statistic value among these simulations determines the optimal number of clusters.

2.3.3. Analysis of variables to distinguish cluster

The SCMC analysis calculates an \( R^2 \) value for each variable (species fraction). This value indicates the variable that divides the study area into clusters of sites more effectively (Assuncao et al., 2006; Duque et al., 2007). The larger the \( R^2 \) value, the better the discrimination among the sites. In other words, the \( R^2 \) value reflects how much of the variation in the original elemental concentration data was retained after the clustering process. The \( R^2 \) value is calculated as following:

\[ R^2 = \frac{ESS}{TSS} \]

(5)

where \( TSS \) is the total sum of squares, represented by squaring and then summing deviations from the global mean value for a particular variable; and \( ESS \) is the explained sum of squares, represented the same way, except that deviations are calculated by subtracting every value from the mean value for the cluster it belongs to and is then squared and summed.

2.3.4. Membership probabilities

We evaluated cluster membership likelihood using 1000 permutations of random spanning trees and evidence accumulation (Lage et al., 2001; Maravalle et al., 1997). This process was employed by the method Skater (Assuncao et al., 2006; Lage et al., 2001). Cluster analysis of spatial objects has an inherent problem with relational (contiguity) constraints. Lage et al. (2001), Maravalle et al. (1997), and Assuncao et al. (2006) define this as an optimization problem related to the contiguity-constrained clustering (known as clustering problems on Trees). The method Skater reduces the original graph representation of the spatial information of the objects by pruning edges generating a minimal spanning tree. The permutation process performed by the Skater approach assesses the occurrence at which sites of a particular cluster are clustered together under the varying spanning trees. Sites that switch clusters as consequence of variations in the spanning tree will be assigned with low membership probabilities while sites that do not switch clusters are assigned with high membership probabilities. A low membership probability indicates that the site could be classified in a different cluster group, whereas a high membership probability suggests confidence that the site belongs in the cluster group it was included. The variation range of the membership probabilities is between 0 and 1. In other words, the membership probability measures how well our methodology minimized within-region variability and maximized between-regional variability of regional mixture profiles (the closer to 1, the higher confidence that our methodology achieved the purpose).

2.4. Sensitivity analysis

We conducted a sensitivity analysis to examine how sensitive the results were to the following parameters: i) spatial constraints; ii) data completeness; iii) site inclusion, iv) characterization of pollutant mixtures; and v) number of clusters.

First, we changed the spatial constraints method, from the inverse distance approach to two other methods - the spatial weight matrix and the Trimmled Delaunay Triangulation. Then, to test the sensitivity to data completeness, for each site, 20% of the days were randomly excluded and the season means were recalculated. To test how the clusters are subject to which sites are included in the analysis, we only standardized the values using z-transformation. Finally, we repeated the analyses accounting for the dataset without mixture. Here it was considered the measured pollutant reported by the US EPA and we only standardized the values using z-transformation. Finally, we tested the sensitivity to the number of clusters by looking at the trade-offs between the number of clusters and the Pseudo F Statistics.

3. Results

3.1. Summary statistics

Table S.1 in supplemental information shows the descriptive statistics (over the 108 sites) of the 20 components of ambient PM_{2.5} included in the study. OC, SO_{4}^{2-}, and NO_{3}^{-} were the elements with the highest mean, 1.99; 1.62; and 1.27 µg/m³, respectively. The lowest mean was observed for As, Se, V, and Ni, with mean concentration equal to 0.0006; 0.0006; 0.0007, and 0.0010, respectively.

Table 1

<table>
<thead>
<tr>
<th>Species fraction</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Minimum</th>
<th>Maximum</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cu</td>
<td>0.00250</td>
<td>1.004448</td>
<td>-0.13503</td>
<td>7.912261</td>
<td>1</td>
</tr>
<tr>
<td>Se</td>
<td>0.002499</td>
<td>1.004448</td>
<td>-0.141095</td>
<td>7.387605</td>
<td>0.999993</td>
</tr>
<tr>
<td>NO_{3}^{-}</td>
<td>0.002683</td>
<td>1.004422</td>
<td>-0.144867</td>
<td>7.912232</td>
<td>0.99989</td>
</tr>
<tr>
<td>Cr</td>
<td>0.001724</td>
<td>1.004538</td>
<td>-0.126204</td>
<td>10.357268</td>
<td>0.99930</td>
</tr>
<tr>
<td>Ba</td>
<td>0.002213</td>
<td>1.004482</td>
<td>-0.150681</td>
<td>7.379933</td>
<td>0.999881</td>
</tr>
<tr>
<td>V</td>
<td>0.001830</td>
<td>1.004527</td>
<td>-0.145462</td>
<td>8.728342</td>
<td>0.999877</td>
</tr>
<tr>
<td>Mn</td>
<td>0.002927</td>
<td>1.004573</td>
<td>-0.190506</td>
<td>9.369018</td>
<td>0.99801</td>
</tr>
<tr>
<td>K</td>
<td>0.001765</td>
<td>1.004485</td>
<td>-0.229605</td>
<td>7.534327</td>
<td>0.997474</td>
</tr>
<tr>
<td>Ni</td>
<td>-0.013538</td>
<td>0.997743</td>
<td>-0.370325</td>
<td>7.799231</td>
<td>0.98489</td>
</tr>
<tr>
<td>Zn</td>
<td>-0.010823</td>
<td>1.000881</td>
<td>-0.770086</td>
<td>6.825110</td>
<td>0.93903</td>
</tr>
<tr>
<td>Na</td>
<td>-0.080534</td>
<td>0.795406</td>
<td>-0.662960</td>
<td>3.176717</td>
<td>0.89838</td>
</tr>
<tr>
<td>Si</td>
<td>-0.034974</td>
<td>0.923669</td>
<td>-0.804037</td>
<td>3.853498</td>
<td>0.889383</td>
</tr>
<tr>
<td>As</td>
<td>0.0086620</td>
<td>1.002555</td>
<td>-1.548090</td>
<td>6.133993</td>
<td>0.86320</td>
</tr>
<tr>
<td>Fe</td>
<td>-0.010572</td>
<td>0.990227</td>
<td>-1.446661</td>
<td>5.692884</td>
<td>0.851959</td>
</tr>
<tr>
<td>Pb</td>
<td>0.011470</td>
<td>1.000593</td>
<td>-1.133681</td>
<td>5.086068</td>
<td>0.845947</td>
</tr>
<tr>
<td>Ca</td>
<td>-0.020774</td>
<td>0.992678</td>
<td>-0.906632</td>
<td>4.265528</td>
<td>0.803963</td>
</tr>
<tr>
<td>SO_{4}^{2-}</td>
<td>0.004313</td>
<td>0.998129</td>
<td>-2.315519</td>
<td>2.534484</td>
<td>0.802544</td>
</tr>
<tr>
<td>NH_{4}^{+}</td>
<td>0.037589</td>
<td>0.964309</td>
<td>-2.032916</td>
<td>2.407511</td>
<td>0.746411</td>
</tr>
<tr>
<td>EC</td>
<td>0.017339</td>
<td>0.993448</td>
<td>-1.558706</td>
<td>4.693392</td>
<td>0.718052</td>
</tr>
<tr>
<td>OC</td>
<td>0.042189</td>
<td>0.953742</td>
<td>-2.668853</td>
<td>3.128883</td>
<td>0.717184</td>
</tr>
</tbody>
</table>

Notes: the mean, standard deviation, minimum, and maximum values are standardized.
In section 3.3, Table 1, we present the descriptive statistics for the fractions considered in the cluster analysis (concentrations of the species \( i/\text{PM}_{2.5} \) concentration in site \( j \)).

3.2. Number of clusters

Our analysis resulted in 27 clusters. This is the optimal number of clusters that maximizes both intra-cluster similarity and inter-cluster dissimilarity. As described in the methods section, the optimal number of clusters was estimated with the pseudo F-statistic where the highest pseudo F-statistic value represents the optimal number of cluster. For 27 clusters, the estimated pseudo F-statistic value was 28.94. Fig. 1 shows the pseudo F-statistic values for the different simulations (number of clusters, from 2 to 30).

3.3. Effectiveness of each species fraction to divide sites into clusters

Table 1 shows descriptive statistics (considering the whole sample, 108 sites) and the \( R^2 \) value for the 20 fractions (concentrations of the species \( i/\text{PM}_{2.5} \) concentration in site \( j \)) considered in our analysis. Based on the \( R^2 \) value, the results suggest that Cu, Se, NO\(_3^-\), Cr, and Ba were the top five species that divided the study area into cluster of sites more effectively. The larger \( R^2 \) value, the better the discrimination among the sites. EC and OC presented the lowest \( R^2 \) values.

3.4. Cluster membership likelihood

The histogram (plus mean and standard deviation values) of membership probability are presented in Fig. 2. These results can provide a sense of the likelihood of significantly different conclusions (e.g., a particular site could be classified in a different cluster). This is a measure of pollution homogeneity within clusters, and heterogeneity across regions. A low membership probability (lowest value = 0) suggests low confidence in the results, while a high membership probability (highest value = 1) indicates high confidence in the results. In general, our results are relatively significant in terms of confidence. The average membership probability was 0.85, with a standard deviation of 0.18.

3.5. Clusters characteristics

3.5.1. Spatiotemporal distribution and geographic parameters

Fig. 3 presents a map of locations of the clusters and a chart illustrating the distribution of sites number in each cluster. Five clusters (cluster 1, 3, 7, 13, and 26) had more than 4 sites. Our analyses also show that 11 clusters (cluster 2, 8, 9, 10, 15, 16, 18, 21, 23, 24, and 27) presented a single site. Among the clusters with more than 4 sites, the cluster 13 was the one with the highest number of sites, 33 air pollution monitoring stations. This cluster 13 is located in northeast and part of the Midwest (Ohio, Indiana, Illinois, and Wisconsin). The cluster 1 has 14 sites and it covers part of the southeast, including the states of North Carolina, South Carolina, Georgia, and Florida. The southwest has a very prominent cluster with 8 sites (cluster 26), covering part of the Louisiana, Mississippi, Texas, and Arkansas. In the west coast, two clusters were highlighted in our analysis, cluster 3 in California and cluster 7 in Washington and part of Oregon. Both clusters with 5 sites.

The input dataset in our analysis includes the land use classification for each site. This classification was assigned by the EPA and is divided into three groups – rural, suburban and urban/center city areas. We present in Fig. 4 the average mixture concentration rates (specie/PM\(_{2.5}\)) of all specie fractions by cluster and land use type. Our results showed that sites do not necessarily have the same land use classification within a cluster. We can observe that in the clusters 1, 4, 13, and 26, which presented the three land use classes. This may be related to the relationship between regional and local pollution.
Fig. 3. Spatial distribution and number of sites by cluster. Note: the map and the chart has the same color key according to the cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Fig. 4. Average mixture concentration rates (specie/PM$_{2.5}$) by cluster and land use.
3.5.2 Concentration ratios

Fig. 5 shows the distribution of the concentration ratio by all the 27 clusters. These diagnostic ratios permit simple comparisons between cluster types and allow us to interpret the multi-pollutant mixtures according to certain types of pollution regimes (tracer elements of source types). For example, our results show that clusters 2 and 8 represent sites with very high Mn and Cr ratios, respectively.

3.6 Sensitivity of the results

Our results demonstrated some sensitivity to the five parameters tested, spatial constraints, data completeness, site inclusion, approach used to characterize the pollutant mixture, and number of clusters.

For the spatial constraints, the number of clusters changed when we applied the weight matrix approach with $k$ nearest neighbors or when we applied the trimmed delaunay triangulation approach to represent the spatial relationship among sites. Our results showed 2 clusters when we accounted for weight matrix or trimmed Delaunay triangulation. For both spatial constraints approach, the first cluster is composed by 3 sites located in the west coast. The other 105 sites represent the second cluster. $K$ nearest neighbors is based on a prior minimum number of neighbors (we defined 8). This is appropriate for analysis when fixing the spatial scale is less important than fixing the number of neighbors.

Spatial scale is an important aspect in air pollution studies. Regarding the triangulation method, it defines neighbors based on voronoi triangles, which each site is a triangle node. Nodes connected by the triangle edge will be determined as neighbors (Cai et al., 2018; Deng et al., 2011). The limitation of this method is that some grouped triangles are not contiguous over space. Therefore, we recommend the use of the inverse distance approach (used in the primary analysis). The inverse distance method accounts for at least one neighbor and the spatial distribution of the data itself will estimate how many neighbors each site gets. This method make it possible to create air pollution regions including contiguous sites, as compared to having sites of the same air pollution region to belong to different geographic areas. This approach maximize air pollution homogeneity within regions and heterogeneity across regions. Therefore, a more substantiated grouping of regions can inform management of regional air sheds.

Fig. 5. Heatmap of the standardized concentration ratio by cluster of the mixture profiles. Note: Axis-x represents the mixture profiles; axis-y represents the clusters; and the color key distribution across the heatmap was standardized by columns (mixture profiles). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)
The clusters distribution also changed when we excluded 20% of the days (data completeness parameter) and 10% of the sites (site inclusion parameter). By incorporating this change, our results showed a significant difference in the optimal number of clusters. While in the primary analysis we estimated 27 clusters, in this sensitivity analysis we estimated 17 clusters. The sensitivity to data completeness suggests that the site means are influenced by missing data sets between 2008 and 2016. This supports the data treatment that we performed before cluster analysis. In this treatment, we defined the completeness of the original data to be greater than 75% for the sites included in the analysis. This sensitivity was observed when we applied our framework previously (Austin et al., 2013).

The clustering is also subjective to the approach used to characterize the mixture profiles. As we described above, we repeated the analysis without accounting for the mixture profiles (ratio by PM$_{2.5}$ concentration). We repeated this analysis considering only the concentration of individual species. As illustrated by Fig. 6, the number and spatial distribution of clusters changed. For example, in the primary analysis we estimated 27 clusters, while in this sensitivity analysis the number of clusters estimated was 30. The cluster profile also changed with this sensitivity analysis. For example, in the primary analysis we estimated about 26 sites in Northeast reflecting high concentration of SO$_4$$^{2-}$ and OC, whereas in this sensitivity analysis we observed about 38 sites. The results of this sensitivity analysis show the different aspects on the degree of similarity between species and PM$_{2.5}$. As we mentioned before, the correlation between profiles of two mixtures indicates this degree of similarity. Therefore, the normalization (using PM$_{2.5}$ as a reference pollutant) of species concentrations supports the comparison of profile characteristics between and within regions or between time periods.

As shown by Fig. 1, there is a slight difference in the Pseudo F Statistic between approximately 15 clusters to 30 clusters. The biggest inflection point appears around 11 clusters. Given that from the perspective of national air quality management strategies, it would be better to have few numbers of clusters. Therefore, we conducted sensitivity analysis to test how sensitive the results are to three different number of clusters – 10, 15, and 20 clusters. These different number of clusters were based on the tradeoffs between the number of clusters and the Pseudo F Statistics. Fig. 7 shows the spatial distribution of clusters for this sensitivity analysis. We can observe that our primary analysis (27 clusters) is similar to the analysis with 15 and 20 clusters. The results of this sensitivity analysis show that only few sites are assigned to different clusters when we set the model with the parameter number of clusters varying from 15 to 27 clusters.

4. Discussion

This is one of the limited number of studies focusing on spatial patterns analysis to estimate regions that exhibit distinct pollutant mixtures on a large scale (all the US). Our analysis was based on a framework previously developed for a single city (Austin et al., 2013). We adapted this framework for a multi-city study in the US. The challenge here was related to the aspects of cluster analysis that are inherently subjective in selecting the best clustering solution for each location. However, our findings suggest strong confidence (based on the statistical criteria) according to the results of the membership likelihood, which the average membership probability was 0.85, with a standard deviation equal to 0.18. The higher cluster membership is essential in cluster analysis, without it, the clusters are of little use for air pollution studies (Keller et al., 2017).

Our findings show that the spatial variation in air pollution mixtures in the US affect substantially in defining cluster profiles. This is consistent with previous studies that have demonstrated strong spatio-temporal variation in air pollution (Austin et al., 2012; Bell et al., 2007; Li et al., 2017; Querol et al., 2008; Zhang et al., 2015). For example, Austin et al. (2012) observed that seasonal patterns within cluster of pollutant mixtures in Boston. The authors suggest that conditions that lead to the formation of the mixture captured by some clusters occur most often in specific regions, including the northeast. Bell et al. (2007) found distinct regional and seasonal patterns of the PM$_{2.5}$ components in the US. The authors report that the degree of the spatiotemporal variation differs by PM$_{2.5}$ components.

Besides spatiotemporal variation in pollution mixtures, air pollution sources, chemical properties, and geographic parameters were also identified as significant factors in distinguishing one region from another. This is in agreement with the literature that shows the potential of these variables as modifier factors of the air pollution exposure levels (Austin et al., 2013; Bari and Kindzierski, 2016; Keller et al., 2017; Requia et al., 2017; van Donkelaar et al., 2014). In particular, we observed similar influence from these factors when we compared our results with those obtained in the original framework (Austin et al., 2013) – the framework as reference for the present study. Other specific work demonstrates that geographic covariate information increases the precision in exposure assignment when using clusters of air pollution at cohort locations (Keller et al., 2017).

Our findings are also in agreement with previous investigations when we incorporate our results in source apportionment studies considering the regional profiles. This allows us to characterize the clusters into regions with certain types of pollution regimes based on emissions...
Fig. 7. Spatial distribution of clusters (each color represents one cluster) for the sensitivity analysis to examine how sensitive the result are to different number of clusters – 10, 15, and 20 clusters. Note: the clusters are represented by the different colors in the maps.
sources. Therefore, we suggest a characterization into 5 regions, as described below.

4.1. Region 1 - northeast and part of the midwest

The first region is the northeast and part of the Midwest. This region is mostly defined by the cluster 13 (the cluster with the highest number of sites, 33) with a mix of rural, suburban and urban area, plus seven other clusters with few sites. Fig. 8 illustrates this region and the distribution of the mixture ratio of each cluster.

Overall, this region reflects air pollution sites with the higher SO$_4^{2-}$ concentration ratio (except for two clusters, cluster 27 with a single site, and 22 with three sites). Most sulfate aerosol in the atmosphere comes from the photochemical conversion of SO$_2$ (Roberts and Friedlander, 1976). Source apportionment studies have shown that power plants are the main sources of SO$_2$ (Fu et al., 2013; Huang et al., 2012). According to the US Energy Information Administration (EIA), in 2015, about 40% of the SO$_2$ emissions from power plants in the US occurred in New England, Middle Atlantic, and in the East North Central region (EIA, 2015).

The cluster 16 with a single site located in a suburban area in west Pennsylvania (known as Industrial Midwest) exhibited very high (nearly the third quartile) ratio of SO$_4^{2-}$, As, EC, NH$_4^+$, OC, and Pb. This suggests contributions from coal combustion and industrial processes. The high ration of SO$_4^{2-}$ and NH$_4^+$ in this cluster 16 may be due to the chemistry of these elements, which there is an interdependence during the reactions - sulfuric acid and ammonia to form ammonium sulfate. Indeed, sulfate is mostly present in the atmosphere in the form of ammonium sulfate. Studies show that reduction in sulfate will increase the available free ammonium (Ciuraru et al., 2012). Also, ammonia from sources such as fertilizer contributes to the formation of sulfates and nitrates that exist in the atmosphere as ammonium sulfate and ammonium nitrate (Shen et al., 2011). Other clusters located in that Industrial Midwest area, cluster 12 (3 sites), cluster 11 (4 sites), and cluster 21 (1 site) presented similar mixture ratios. The cluster 21 (south Illinois), in particular, highlighted by the very high ratio of Fe and Zn, reflecting emissions from road dust and motor vehicle Fe (Almeida-Silva et al., 2011; Bari and Kindzierski, 2016; Liu et al., 2014).

Specifically in the coastal area (East coast), there is a particular cluster (cluster 17) with 4 sites in urban areas which reflects high ratio of Ni and Na. Regarding the Ni, it suggests that this location is impacted by emissions from ports. Studies have shown contributions from ship emissions to Ni concentrations (Agrawal et al., 2008; Moldanová et al., 2009). Regarding Na, this pollution regime indicates presence of sea salt, the main source of sodium (Bersenkovitsch et al., 2018; Laskin et al., 2003).

Finally, we highlight the high ratio of NH$_4^+$ in this region (except the cluster 22 and 27, more close to the coastal area). This reflects contribution from agricultural area. As we mentioned above, ammonia from sources such as fertilizer contributes to the formation of sulfates and nitrates that exist in the atmosphere as ammonium sulfate and ammonium nitrate (Shen et al., 2011).
4.2. Region 2 - southeast

This region is mostly characterized by the cluster 1, composed by 14 sites located in a mix of rural, suburban and urban areas. In addition, there are three more clusters in this area with single sites, clusters 15, 18, and 23. Fig. 9 illustrates the Southeast region and the distribution of the mixture ratio of each cluster.

The 14 sites grouped as cluster 1 reflect average ratios for most of the pollutants. The exception is for OC, which the ratio is almost the third quartile. Organic aerosols are a complex mixture of chemical compounds formed primarily by incomplete combustion or the oxidation of gas-phase precursors. It can be produced from fossil fuel and biofuel burning and natural biogenic emissions (Kanakidou et al., 2005). Huang et al. (2015) estimated a global emission inventory of OC and found that the Southeast in the US is a region with substantial OC concentration. Huang et al. (2015) also show that in the US more than 90% of the anthropogenic OC comes from oil, gas and coal emissions.

The single site represented by cluster 18 in south Florida reflects very high ratios (above the third quartile) of Ba, Ca, Cu, Fe, K, Na, Ni, NO$_3^-$, Se, Si, V, suggesting significant contribution from a mixture of sources, including road dust (Ba, Ca, Fe, Na and Si), motor vehicle (Cu), wood burning (K), coal combustion (Se), and common sources in coastal areas – ship engine exhaust (Ni) and sea salt (Na), as we mentioned above. In particular about NO$_3^-$, note that this single site in cluster 18 presented extremes values for NH$_4^+$, SO$_4^{2-}$, and NO$_3^-$. This may represent the complex chemistry of this elements, which more ammonia becomes available to react with nitric acid to form ammonium nitrate. When nitric acid and sulfuric acid are present, the reaction between ammonia and sulfuric acid is thermodynamically favored. In addition, sulfuric acid and ammonia are used to form ammonium sulfate. Indeed, sulfate is mostly presented in the atmosphere in the form of ammonium sulfate. Studies show that reduction in sulfate will increase the available free ammonium (Ciuraru et al., 2012).

The other two clusters with single site (cluster 15 and 23) have similar ratios. These clusters are located in a suburban/urban area in Alabama, representing very high ratios (above the third quartile) of As, Ca, Fe, Mn, Pb, and Zn. This represent a mixture of pollution regimes based on emissions sources. The high ratio of As represents significant contribution from coal combustion; Ca, Fe and Mn from road dust; and Pb and Zn from motor vehicle.

4.3. Region 3 - south

This region encompass one cluster with 8 stations in a mix of rural and urban areas (cluster 26), and three clusters with single station (clusters 2 and 8 in an urban area, and cluster 9 in a rural area). The clusters of the south region represents the states of Texas, Louisiana, Mississippi, Arkansas, Oklahoma, and Missouri, which are highlighted in Fig. 10.

Clusters 2 and 8 have very similar ratios. Observing the whole ratio distribution with all 27 clusters estimated in our analysis (Fig. 5), we can see that the clusters 2 and 8 reflect the highest ratio of Ba, Cr, Cu, Ec, K, Mn, Ni, NO$_3^-$, Se, Si, and V. This represent a mixture of local and regional sources, including motor vehicle, road dust, oil/coal combustion, and wood burning.

Cluster 9 is the only one in the region located in a rural area, which
is expected to find substantial contribution from agricultural activities. Similar to the pollution regime observed in the region 1 (Northeast and part of the Midwest), ammonia from fertilizer contributes to the formation of sulfates and nitrates that exist in the atmosphere as ammonium sulfate and ammonium nitrate (Shen et al., 2011). Cluster 9 presented high ratios of NH$_4^+$, NO$_3^-$, and SO$_4^{2-}$. All these ratios were above the third quartile. Cluster 9 also presented very low As and EC ratios. These ratios were lowest one (very close to the minimum value) compared to the clusters in the south region (Fig. 10) and even to the total clusters (Fig. 5). This reflects very little contribution from coal combustion (main As source) and motor vehicles (important EC source plus Zn, Pb, Cu, and Br, which were low for cluster 9 as well).

The cluster with the highest number of stations in the south region (cluster 26) represents ratios with values within the interquartile range, except for the Na, which the ratio was above the third quartile. This suggests some regional contribution from sea salt, since most of the sites in this region are not in the coastal area. A large body of the literature has demonstrated that local air quality can be impacted by pollution from distant sources (e.g., local, regional, and even inter-continental sources) due to the atmospheric transport (Jaffe et al., 1999; Lin et al., 2014; J. Liu et al., 2009a, b; Ngo et al., 2018; Zhang et al., 2017). For example, Lin et al. (2014) estimate that air pollution sources in China contribute 3–10% of annual mean surface sulfate concentrations over the western United States in 2006. Similar to cluster 9, cluster 26 also reflects very little contribution from motor vehicles. The ratios for EC, Zn, Pb, and Cu were in general in the first quartile.

### 4.4. Region 4 – southwest and mountain

This region includes five clusters with similar pollution regime – clusters 4 (three sites), 5 (three sites), 10 (a single site), 14 (two sites), and 24 (a single site) (Fig. 11).

All these clusters reflect little contribution from power plants (low ratios of SO$_4^{2-}$), including coal combustion (low ratios of As and Se). In contrast, these clusters suggest high contribution from road dust (high ratios of Ca and Si). Overall, these clusters in the southwest/mountain region are differentiated by the ratio of EC, Fe, Na, NH$_4^+$, OC, Pb and Zn. This is related to the geographical distribution of some tracer elements of source types. For example, the ratios of carbon particles (EC and OC) tend to be higher in the west coast. This reflects the significant EC and OC emissions from wildfire, most strongly in California, Nevada, and Arizona (Bendix and Commons, 2017; Doerr and Santín, 2016; Marlon et al., 2012). Some studies have defined the chemical characteristics of PM$_{2.5}$ as the outcome in the wildfire-related air pollution models (Gunsch et al., 2018; Jaffe et al., 2008; Spracklen et al., 2009). These studies have shown that the relationship between particle components and wildfire varies significantly over space and time depending on the chemical characteristics of PM$_{2.5}$ and geographical characteristics, including weather parameters (McClure and Jaffe, 2018; Spracklen et al., 2007). Among those numerous chemical components of PM$_{2.5}$, particulate carbon (including EC and OC) have been the most indicated as trace elements of wildfires (McClure and Jaffe, 2018). In our analysis, the cluster 14 with high ratio of EC and OC is located in the state of Nevada (the station in this state is very close to California).
and Arizona.

4.5. Region 5 – west coast

Finally, the last region that we suggest according to our cluster analysis cover the west coast of the US, which can be considered a coastal region. This region encompass 5 clusters – cluster 3 with 5 sites, cluster 7 with 4 sites, cluster 19 with 4 sites, cluster 20 with 2 sites, and cluster 25 with 2 cluster (Fig. 12).

These clusters have similar pollution regime (as we identified in the region 4 as well) for most of the element fractions. For example, all the 5 clusters presented high values of Na, OC, and EC (for this element the clusters 3, 19, and 25 had values within the interquartile range). As we discussed above, Na and OC are indicators of marine aerosols and wildfire emissions, respectively. Both sources are significant in the west coast (Bendix and Commons, 2017; Doerr and Santín, 2016; Hand et al., 2012).

Most clusters presented low rate of $\text{SO}_4^{2-}$ suggesting little contribution from power plants (similar as we found for the region 4). Coal combustion, main source of As, is contributing only to the clusters 7 and 25. On the other hands, clusters 7, 19, and 25 reflect low rates of $\text{NH}_4^+$, while the clusters 3 and 20 reflect average rates.

5. Conclusions

We propose an innovative approach to classify regions in the US based on the clusters of air pollutant mixtures. We observed a strong influence of the relationship between regional and local pollution in clustering air pollution mixture. First, we found that sites spatially close are assigned to different clusters. Then, when we categorized the results based on land use, we detected that sites do not necessarily have the same land use class within a cluster. These results may be related to whether the cluster profile was influenced by regional pollution versus local pollution. We suggest that the concentration measured at specific air pollution monitoring station will have differing amounts of measurement error, depending on the spatial heterogeneity of a given pollutant across the study region.

We suggest that this study can benefit researchers, policy makers, and local communities to create future strategies related to air pollution and environmental health. The important and imminent policy implications is that our approach supports more targeted and regionally air quality management practices by minimizing within-region variability and maximize between-regional variability of regional mixture profiles. This is based on the concept that regions with similar pollutant mixtures are impacted by similar air pollution sources and atmospheric processes. We expect that further investigations can use our findings to analyze the relationship between areas that exhibit distinct pollutant mixtures and the impact of regulations, climate change, and health effects in the US.

Finally, given that differences in the $\text{PM}_{2.5}$ constituents explain the varying effect size of the association between $\text{PM}_{2.5}$ and health (Achilleos et al., 2017; Dai et al., 2014; Zanobetti et al., 2009), we suggest that our study can support further investigations to assess the health effects of $\text{PM}_{2.5}$ components. For example, in the U.S., previous
studies have shown that health impacts for PM$_{2.5}$ mass are higher when the PM$_{2.5}$ content of Br, Cr, Ni, or Na was higher (Franklin et al., 2008; Zanobetti et al., 2009). In our study, we estimated that the clusters 2 and 8 (located in the region 3 – South) reflect high ratio of these PM$_{2.5}$ content. Bell et al. (2009) estimated that regions in the U.S. with high concentration of EC, V, or Ni had higher risk of hospitalizations associated with short-term exposure to PM$_{2.5}$. In our analysis, we observed that the region 1 (in the northeast and part of the Midwest) includes the cluster 17 with 4 sites that reflect high ratio of EC, V, and Ni. In Boston, (Zanobetti et al. 2014) found that cluster characterized by high concentrations of the elements related to primary traffic pollution and oil combustion emissions has significant association of PM$_{2.5}$ with daily deaths. Zanobetti et al. (2014) found a 3.7% increase (95%CI: 0.4, 7.1) in total mortality, per 10 μg/m$^3$ increase in the same day average of PM$_{2.5}$. In our study, clusters suggesting high contribution from traffic and oil combustion are significant in the regions 1 (especially in the coastal area - East coast) and 5 (west coast). Therefore, we suggest that taking our findings together, further investigations can assess the health effects of PM$_{2.5}$ components by accounting for effect modification and mediation of effects of spatial patterns via air pollution mixtures on health.

Acknowledgement

This work was supported by the US Environmental Protection Agency (grant RD-834798 and RD-835872). The contents of this report are solely the responsibility of the grantee and do not necessarily represent the official views of the US Environmental Protection Agency. Further, the agency does not endorse the purchase of any commercial products or services mentioned in the publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2019.06.006.

References


