Characterization of Spatially Homogeneous Regions Based on Temporal Patterns of Particulate Matter 2.5 in the Continental United States

COSMOS Technical Report 07-01

Seoung Bum Kim[1], Chivalai Temiyasathit[1], Sun-Kyoung Park[2],

Victoria C.P. Chen[1], Melanie Sattler[3], Armistead G. Russell[4]

[1]Department of Industrial and Manufacturing Systems Engineering
University of Texas at Arlington
Arlington, TX 76019-0017, USA

[2]Department of Transportation
North Central Texas Council of Governments
616 Six Flags Drive P.O. Box 5888
Arlington, TX 76005-5888, USA

[3]Department of Civil Engineering
University of Texas at Arlington
Arlington, TX 76019-0308, USA

[4]Schools of Civil and Environmental Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA

ABSTRACT

Statistical analyses of time-series or spatial data have been widely used to investigate the behavior of ambient air pollutants. Because air pollution data are generally collected in a wide area of interest over a relatively long period, such analyses should take into account both spatial and temporal characteristics. The objective of the present study is twofold: (1) To identify an efficient way to characterize the spatial variations of $PM_{2.5}$ concentrations based solely upon their temporal patterns, and (2) To analyze the temporal and seasonal patterns of $PM_{2.5}$ concentrations in spatially homogenous regions. This study used 24-hour average $PM_{2.5}$ concentrations measured every third day during the period between 2001 and 2005 at 522 monitoring sites in the continental United States. A *k*-means clustering algorithm using the correlation distance was employed to investigate the similarity in patterns between temporal profiles observed at the monitoring sites. A k-means clustering analysis produced six clusters of sites with distinct temporal patterns which were able to identify and characterize spatially homogeneous regions of the United States. The study also presents a rotated principal component analysis (RPCA) that has been used for characterizing spatial patterns of air pollution and discusses the difference between the clustering algorithm and RPCA.

Keywords: air pollution; cluster analysis; $PM_{2.5}$; time series; spatial regions

IMPLICATIONS

The problem of modeling spatial and temporal data is of great practical interest in many different fields. The approaches presented here provide an efficient and objective way to determine spatially homogenous regions of $PM_{2.5}$ mass concentrations based on their temporal patterns over multiple years. The results imply that spatial and temporal patterns are strongly linked, in that spatially homogeneous regions can be characterized solely by their temporal patterns. Furthermore, information about spatial and temporal variations would be useful in improving and evaluating dynamic air quality models.

INTRODUCTION

Statistical analyses of time-series or spatial data have been widely used to investigate the behavior of ambient air pollutants. Because air pollution data are generally collected in a wide area of interest over a relatively long period, such analyses should take into account both spatial and temporal characteristics. In particular, a number of studies have been devoted to characterization of temporal and (or) spatial correlation(s) in air pollution data collected from a number of monitoring sites in an area of interest. Temporal correlation or spatial correlation can be defined as a correlation between the same variables at different times and locations, respectively, and it measures the strength of the relationship of observations. Sometimes, the term "autocorrelation" is used instead of "correlation" to emphasize its characteristic of self-correlation (i.e., correlation of the variable with itself). Therefore, high temporal or spatial correlation implies a strong relationship of observations (e.g., air pollution concentrations) in time or space.

49 This paper focuses on characterizing $PM_{2.5}$, one of the six criteria pollutants
50 identified by the U.S. Environmental Protection Agency under the federal Clean Air Act.[1,
51 2] The other five criteria pollutants include ozone, sulfur dioxide, nitrogen dioxides,
52 carbon monoxide, and lead.[1] $PM_{2.5}$ has the potential to cause adverse health effects in
53 humans, including premature mortality, nose and throat irritation, and lung damage.[3, 4]
54 Furthermore, $PM_{2.5}$ has been known to be associated with visibility impairment, acid
55 deposition, and regional climate change.[5]

56 A number of statistical models have been used to characterize the spatial
57 correlation of $PM_{2.5}$ concentrations. Descriptive statistical analyses that examined daily,
58 seasonal, and spatial trends in mass, composition, and size distributions of 24-hour
59 average $PM_{2.5}$ concentrations at 16 specific sites in several counties over southeast Texas
60 during the period from 2000 to 2001 showed that mass and composition were generally
61 spatially homogeneous, while particle size distributions were not.[**6**] A nonnegative factor
62 analytic model was used to analyze the contribution of meteorology (e.g., temperature,
63 humidity, pressure, and wind speed) and other ambient factors (e.g., ozone concentration)
64 to $PM_{2.5}$ concentrations at 300 monitoring sites in the eastern United States during 2000.[7]
65 Temporal and spatial trends of sulfur dioxide (SO_2), sulfate ($SO_4^=$), nitrogen species, and
66 all major components of $PM_{2.5,}$ were investigated from 1989 to 1995 at 34 rural clean air
67 status and trends network (CASTNet) sites in the eastern United States.[8] In their study, a
68 clustering analysis was performed to group 30 sites adjusted for seasonal effects so that
69 the sites within a cluster had a similar pattern of meteorological factors and ozone levels.
70 A more comprehensive study of spatial and temporal trends of $SO_4^=$ was performed over
71 10 years for 70 monitoring sites in the continental United States.[9] They characterized the

spatial trends of $SO_4^=$ concentrations in summer and winter and quantified the temporal change of the $SO_4^=$ level. A number of studies have been conducted to determine the spatial and temporal patterns of aerosol concentrations for impacting haze and visual effect.[10-12]

Analyses of spatial and temporal patterns of pollutants can be used to establish representative monitoring sites. A fixed-effect analysis of variance (ANOVA) model was developed to explore spatial and daily variations of pollutant levels and to identify the representativeness of $PM_{2.5}$ monitoring sites in Seattle, Washington.[13] Furthermore, a statistical model was used to quantify the representativeness of existing monitoring sites.[14] Principal components analysis was applied to measure the spatial representativeness of ground level ozone concentrations.[15]

An understanding of spatial correlations of pollutant concentrations would be useful in improving dynamic air quality models. McNair et al.[16] evaluated the performance of the Carnegie/California Institute of Technology (CIT) model and found that spatial inhomogeneity needed to be taken into account in order to develop model performance guidelines. Jun and Stein[17] compared daily $SO_4^=$ levels between observation data and the Community Multiscale Air Quality (CMAQ) model by space-time correlation. The CMAQ model matches the space-time correlation structure of the observed data; however, CMAQ partially captures time-lagged spatial variation of $SO_4^=$ concentrations. Recently, Park et al.[18] investigated effects of spatial variability on the evaluation of the CMAQ model and observed that slight errors in the model were caused by uncertainties due to the different spatial scales between the point-observations and the

94 volume-averaged simulated concentrations. Their recommendation was to use data at
95 spatially representative monitoring sites in model evaluation.

96 The present study seeks to characterize regions of homogenous $PM_{2.5}$
97 concentrations across the continental United States based *solely* upon their temporal
98 patterns over multiple years. Each monitoring site provides a profile that represents the
99 temporal pattern of $PM_{2.5}$ concentrations. Combinations of multiple temporal profiles,
100 each with 609 variables (days), lead to a large number of data points and a situation that
101 poses a great challenge to analytical capabilities. Our first objective was to develop an
102 efficient way to identify homogenous $PM_{2.5}$ concentration regions using these temporal
103 profiles. Our approach yielded groupings of the monitoring sites into spatially
104 homogenous regions. Thus, our second objective was to analyze the temporal and
105 seasonal patterns of $PM_{2.5}$ concentrations that characterize each of the identified spatially
106 homogenous regions.

107 **DATA**

108 Monitoring data were obtained from the Aerometric Information System (AIRS) database
109 in the Environmental Protection Agency's Air Quality System (EPA-AQS)
110 (http://www.epa.gov/ttn/airs/airsaqs/), which contains 24-hour average $PM_{2.5}$ mass
111 concentrations measured every third day from 2001 to 2005 at 522 monitoring sites in the
112 continental United States. At each 24-hour average $PM_{2.5}$ mass monitoring site, 609
113 measurements were recorded between 2001 and 2005. Thus, the $PM_{2.5}$ concentration for
114 monitoring site S_i at time T_j can be represented as follows:

115 $$Z(S_i, T_j) \ for \ i = 1,\ldots I, \ j = 1,\ldots J,$$

where I is the number of monitoring sites (here I=522) and J is the number of time points (here J=609). The database contains a number of missing values. Monitoring sites that had values missing for more than 50% of the observations or more than 10 consecutive missing values were excluded from the study. The database originally contains 1,402 monitoring sites. After excluding those sites, 522 monitoring sites remained. The remaining missing observations in the dataset were replaced with the interpolation of the nearby values, on the assumption that those were the result of measurement errors or instrument malfunctions**.** In addition, we found one observation (October 27, 2003 in California) that had a much higher concentration (239.2 $\mu g/m^3$) than the values in its neighborhood. We considered this as an outlier and replaced it with an interpolated value. The remaining 522 sites include both the urban and rural sites. In the present study, we combined the urban and rural sites in the analysis because we are more interested in analyzing an overall spatial and temporal pattern of $PM_{2.5}$ concentration in the continental U.S. rather than addressing questions related to levels of pollutants around specific commercial, industrial, residential, or agricultural sites. Also, we should point out that $PM_{2.5}$ speciation data can be useful for characterizing the patterns of components of total $PM_{2.5}$ mass concentration. However, because the numbers of monitoring sites where speciation data are available are very limited and the present study seeks to characterize regions of homogenous $PM_{2.5}$ concentrations across the *entire* continental United States (regional scale), we focused on the analysis of total $PM_{2.5}$ mass concentrations.

ANALYTICAL APPROACHES

Interpolation Technique to Impute Missing Observations and Outliers

Missing observations and outliers were replaced with interpolated values using an inverse-distance-squared weighted method.[16] The interpolated value for site S_i at time T_j, $I(S_i, T_j)$ is computed as follows:

$$I(S_i, T_j) = \frac{\sum_{k=1,k\neq i}^{m} Z(S_k, T_j) \cdot \omega_k}{\sum_{k=1,k\neq i}^{m} \omega_k}, \quad (1)$$

where m is the number of monitoring sites and ω_s is calculated as follows:

$$\omega_k = \begin{cases} \frac{1}{r_k^2} & if \quad r_k \leq d\ km \\ 0 & if \quad r_k > d\ km, \end{cases} \quad (2)$$

and r_k is Euclidean distance from site S_i to site S_k at time T_j. Thus, $I(S_i, T_j)$ in (1) is the weighted average $PM_{2.5}$ concentration value observed in the surrounding m sites. The weights are determined by the way that observations in close spatial proximity are given more weight than those that are spatially separated. In this paper, d in (2) was set to 180 km. Based upon our own analysis, using a different d did not lead to significantly different results for interpolation.

Other approaches for interpolating outliers and missing values include functional, maximum likelihood imputation schemes, and Bayesian modeling. Polynomial functions and splines can be used to interpolate regularly-spaced data. Maximum likelihood or Bayesian modeling, which typically requires high computation, uses an iterative approach

based on model parameter estimation. Examples of this approach include Expectation-Maximization,[19] kriging,[20] radial basis function,[21] and Bayesian hierarchical model.[22, 23]

k-means Clustering Analysis

Clustering analysis systematically partitions the dataset by minimizing within-group variation and maximizing between-group variation, and then assigning a cluster label to each observation.[24] Clustering analysis has been widely used to facilitate the extraction of implicit patterns and to test the validity of the groupings obtained by visualization methods such as principal components analysis. Variation can be measured based on a variety of distance metrics between observations in a dataset. The present study applied a *k*-means clustering algorithm to the set of $PM_{2.5}$ concentrations from each monitoring site in 609 (days) dimensional space. The brief summary of the *k*-means clustering algorithm is as follows: Given *k* seed points, each observation is assigned to one of the *k* seed points close to the observation, which creates *k* clusters. Then, seed points are replaced with the mean of the currently assigned clusters. This procedure is repeated with updated seed points until the assignments do not change. The results of the *k*-means clustering algorithm depend on the distance metrics, the number of clusters (*k*), and the location of seed points.

For the distance metric, the correlation distance that measures the similarity in patterns between the two temporal profiles from each monitoring site was used. More precisely, for the monitoring sites *x* and *y*, the correlation distance between two temporal profiles that consist of a series of *J* time points can be computed as follows:

$$D_{[Z(S_x,T_j),Z(S_y,T_j)]} = \frac{1}{J}\sum_{i=1}^{J}\left(\frac{Z(S_x,T_j)-\overline{Z}_{s_x}}{\sigma_{Z_{s_x}}}\right)\left(\frac{Z(S_y,T_j)-\overline{Z}_{s_y}}{\sigma_{Z_{s_y}}}\right), \quad (3)$$

where, $\bar{Z}_{s_i} = \frac{1}{J}\sum_{j=1}^{J} Z(S_i, T_j)$ and $\sigma_{Z_{S_i}} = \left(\frac{1}{J}\sum_{j=1}^{J}\left(Z(S_i, T_j) - \bar{Z}_{s_i}\right)^2\right)^{1/2}$.

In contrast to Euclidean distance that measures the difference of each time point over the monitoring period, the correlation distance allows us to measure the similarity in shape between the two temporal profiles observed at each monitoring site. In other words, the correlation distance focuses more on an overall pattern rather than scale-difference between the profiles.

To determine the number *k*, a heuristic approach was used based on the assumption that we do not have explicit knowledge of expected $PM_{2.5}$ concentration changes in the continental United States. To be specific, we applied the *k*-means clustering algorithm to our dataset with *k* values ranging from 5 to 15 for 20 replications. We then selected the final *k* so that the average value of the standard deviation of *k* groups (for *k* = 5, 6,…,15) reaches the first minimum. To determine the location of seed points, we used a "sample" method available in MATLAB (MathWorks Inc., Natick, MA).

A previous study applied the *k*-means clustering algorithm with Euclidean distance to SO_2 data from 30 sites in the eastern United States.[8] The study obtained six clusters in which the sites within the cluster had a similar pattern of meteorological factors and ozone levels. The study determined the number *k* based on geographical and climatological characteristics and estimated the location of seed points using the centroid values of each region. In contrast to Holland et al.,[8] our study relied solely on statistical methods to determine the number *k* and the location of seed points. This is a reasonable approach because one of the main purposes of this study is to examine the feasibility of

using only temporal patterns of $PM_{2.5}$ concentrations for characterizing spatial correlations. To facilitate the interpretation of temporal patterns, we applied robust locally weighted polynomial regression (rloess).[25] This basic idea of rloess is to define local subsets of data (within the span) and fit the model locally by giving weight to each data point in a robust manner that can reduce sensitivity to outliers. For more mathematical details, see Cleveland.[26]

A Rotated Principal Components Analysis Technique

A rotated principal components analysis (RPCA) approach has been used to characterize spatio-temporal patterns of air pollution and meteorological fields.[27, 28] We begin with a brief introduction to a traditional PCA approach. PCA is a multivariate data analysis technique primarily for dimensional reduction and visualization. In the atmospheric sciences, PCA has been widely used for determining the important source regions of air pollution[29], and in receptor modeling, which apportions source contributions to air pollution.[30] PCA identifies a lower dimensional space that can explain most of the variability of the original dataset (**X**). The lower dimensional space, represented by the principal components (PCs), is a linear combination of all the original variables. The most important PCs are obtained to maximize the variability of the entire dataset. For example, the i^{th} PC can be expressed as follows:

$$PC_i = \mathbf{x}_1 k_{i1} + \mathbf{x}_2 k_{i2} + ... + \mathbf{x}_N k_{ip} = \mathbf{X}\mathbf{k}_i, \; i = 1, 2, \ldots, p, \tag{4}$$

where p is the total number of variables in the original dataset. A set of coefficients is given by the eigenvector with the corresponding i^{th} largest eigenvalue of the covariance matrix of the original dataset. Because the contribution of each variable to form a PC can be represented by each component of the eigenvector, this vector is often called a

227 "loading vector." For example, k_{i1} in (4) indicates the degree of importance of the first
228 variable in the i^{th} PC domain.

229 The basic idea of RPCA is to rotate the loading vectors of the traditional PCA
230 approach to facilitate the spatial interpretation. Among the many options for rotation, a
231 varimax rotation method has been widely used.[28] The varimax rotation maximizes the
232 sums of the variances of the squared components in each loading vector of the traditional
233 PCA.[28]

234
235 **RESULTS**
236
237 **Spatial Patterns of $PM_{2.5}$ Concentrations**

238 The *k*-means clustering algorithm using the correlation distance was performed on the
239 dataset of 522 monitoring sites, each of which had 609 time points. Based on the heuristic
240 method described in previous section, the optimal number for *k* is six. The results of six-
241 means clustering analysis on temporal profiles are displayed on the U.S. map (Fig. 1). It
242 is seen that the monitoring sites in close spatial proximity are grouped together,
243 demonstrating the identification of spatially homogeneous regions solely based on the
244 temporal patterns of $PM_{2.5}$ concentrations. To further characterize the spatial regions, the
245 clustered sites can be grouped according to the following ad-hoc categories chosen by
246 geographical locations, with the number of monitoring sites in each cluster indicated in
247 parentheses: (i) Central (68); (ii) Florida & Gulf Coast (44); (iii) Midwest (103); (iv)
248 Northeast (104); (v) Southeast (111); and (vi) West (92). Table 1 shows a list of states in
249 the United States in each clustered region.

250 Main factor analysis that compares the mean $PM_{2.5}$ concentrations for each
251 clustered region showed that mean $PM_{2.5}$ concentrations vary regionally from year to year

252 although the degree of difference was not significant (Fig. 2). In general the highest mean
253 $PM_{2.5}$ concentrations occurred at sites in the Midwest, followed by the Southeast and the
254 Northeast (Fig. 2). This may be because of the high SO_2 emissions generated within the
255 Ohio River Valley in the Midwest region.[9, 31] The mean $PM_{2.5}$ concentration in the
256 Midwest in 2001 (15.02 $\mu g/m^3$) and 2005 (15.56 $\mu g/m^3$), in particular, exceeds the annual
257 federal standard of 15 $\mu g/m^3$ (Fig. 2). Lower mean concentrations are observed in the
258 West, Florida & Gulf Coast, and Central. It appears from Fig. 2 that the mean $PM_{2.5}$
259 concentrations have a downward trend from 2001 to 2004 but increase in 2005, except
260 for the West, which exhibits a decreasing trend over the time period from 2001 to 2005.
261

262 **Comparison with Rotated Principal Components Analysis**

263 A RPCA approach was applied to the same dataset used in *k*-means clustering analysis. A
264 set of ordered eigenvalue-eigenvector pairs was computed from a 522 by 522 covariance
265 matrix containing the pair-wise covariance of the 522 monitoring sites. Usually, only a
266 small number of PCs is needed to explain the variability in the original dataset. There is
267 no definitive answer to determine an appropriate number of PCs to retain.[32] One popular
268 method is to use the property that the proportion of variability explained by each PC can
269 be expressed by the eigenvalues. For example, the proportion of variability explained by
270 the i^{th} PC ($V(PC_i)$) can be calculated from the following equation:

271
$$V(PC_i) = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}, \qquad (5)$$

272 where λ_i is the i^{th} eigenvalue, and p is the total number of original variables. The idea of
273 this method is to plot the ordered $V(PC)$ against its rank and determine an appropriate

number of PCs. This graphical method is rather subjective since the decision involves a visual inspection. The general recommendation is to find an elbow in the plot. In the present study, we found that the elbow point was observed around five, six, and seven PCs. Of these, we decided to retain the six PCs in order to ensure the comparability to the six clusters obtained from the clustering analysis in previous section. Note that six PCs accounted for 65% of the variability of the entire dataset. A varimax rotation of the six PCs was performed. The components in the loading vectors of each of the six rotated PCs were displayed by contour plots on U.S. maps (Fig 3). The regions with higher loading values were highlighted. The first RPCA loading contour plot identified the monitoring sites in the Midwest. The second, third, fourth, fifth, and sixth RPCA loading contour plots identified the monitoring sites in the Northeast, Southern California, Southeast, West, and Central, respectively.

It is somewhat difficult to make a direct comparison between RPCA and *k*-means clustering analysis because of their different ways of determining the spatial groups of homogeneous $PM_{2.5}$ concentrations. RPCA relies on a graphical interpretation of the contour plot of RPCA loadings, while *k*-means clustering analysis assigns a group label to each monitoring site. Note that Fig. 1 is a plot of group labels from *k*-means clustering analysis. Nevertheless, identified homogeneous regions from RPCA and *k*-means clustering analysis seem similar. The main difference is that RPCA did not identify the sites in the Florida & Gulf Coast as a separate group but identified sites in Southern California.

Both the RPCA and *k*-means clustering analysis are unsupervised learning techniques, in that they depend only on input variables (explanatory variables) but do not

297 take into account the information from the response variable. However, from the
298 mathematical point of view, RPCA and *k*-means clustering are different. RPCA
299 identifies a new coordinate system that maximizes the variability of the original dataset
300 through an orthogonal linear transformation, while *k*-means clustering analysis does not
301 use any transformation processes but iteratively partitions the observations by minimizing
302 within-group distances and maximizing between-group distances, then assigning a cluster
303 label to each observation.

304 RPCA renders a graphical result, efficient in facilitating the visualization of a
305 high-dimensional space. However, similar to other graphical methods, the interpretation
306 of RPCA results can be subjective, with different analyzers drawing different conclusions.
307 On the other hand, *k*-means clustering analysis provides a group label for each
308 observation, and thus, the interpretation of results is more objective than RPCA. However,
309 the *k*-means clustering results may vary with different choices of the starting means. No
310 consensus exists about which is the better method (RPCA or clustering analysis) to
311 satisfy all conditions. We believe that visualization methods, such as RPCA, can elicit the
312 natural groupings of the observations, and clustering analysis can test the validity of the
313 groupings obtained by RPCA. The following section discusses temporal and seasonal
314 patterns of $PM_{2.5}$ concentrations according to *k*-means clustering results.

315

316 **Temporal and Seasonal Patterns of $PM_{2.5}$ Concentrations**

317 The smoothed temporal pattern of each spatially homogeneous region identified via six-
318 means clustering analysis over a time period from 2001 to 2005 is summarized using
319 mean, median, 25^{th} percentile, and 75^{th} percentile profiles (Fig. 4). The rloess method
320 with a span of 0.05 was used for smoothing the original time patterns. The similarity

between the 25^{th-} and 75^{th-} percentile profiles confirms that there are no significant outliers in the dataset. A distinct temporal pattern was observed in each region. For ease of interpretation of temporal patterns and to explore seasonal variations, we defined the four seasons in a standard way: spring (March, April, May), summer (June, July, August), fall (September, October, November), and winter (December, January, February). Fig. 2 shows the comparison of mean $PM_{2.5}$ concentrations for the four seasons. It can be seen that the highest mean concentration value was observed in summer, followed by winter for the period between 2001 and 2005. In particular, in 2002 and 2003, the mean concentrations in summer exceed the annual federal standard of 15 $\mu g/m^3$. The lowest mean concentration was observed in spring, except 2001. The results from Tukey's pair-wise comparisons test showed that the mean concentrations in every season were significantly different from each other (p-value < 0.01).

It is important to observe from the box plots shown in Fig. 5 that $PM_{2.5}$ concentrations between regions and seasons have interaction effects in that each clustered region differs in each of the four seasons (Fig. 5). In the box plots, the lines in the middle of the boxes represent the median, and the distance between the top and bottom of the boxes represents the range from the 25^{th} to the 75^{th} percentiles (i.e., interquartile range). The plus sign at the top of the plot is an observation that is more than 1.5 times the interquartile range away from the top or from the bottom of the box.

According to Fig. 5, the West region has the highest level of $PM_{2.5}$ in winter, likely because of the increase in NO_3^- and organic carbon during winter months. Major sources of NOx include transportation, industrial operations, electricity production, and non-industrial fuel burning. Quasi-equilibrium favors the particulate species under cool,

344 moist conditions.[33, 34] This significant increase in the level of NO_3^- in the western United
345 States in winter likely offsets the slight seasonal reduction of $SO_4^=$. A major source of
346 organic carbon during wintertime in the western United States includes fireplace
347 burning.[35]

348 $PM_{2.5}$ concentrations tend to be higher in summer in many parts of the nation's
349 northeastern and southeastern sections (Fig. 5). Sulfate is produced from sulfur dioxide,
350 which is prevalent in the East because of the relatively abundant coal-fired power
351 plants.[35] Higher insolation and humidity during summer months enhance both
352 homogeneous and heterogeneous reactions that produce secondary sulfate particles, one
353 of the major components in $PM_{2.5}$ mass concentrations.[36, 37]

354 Midwest, Central, and Florida & Gulf Coast show comparable $PM_{2.5}$ levels during
355 the four seasons, although the Midwest tends to show higher within-season variability
356 than the Central and Florida & Gulf Coast regions.

357 To be able to predict $PM_{2.5}$ concentration as a function of time in each clustered
358 region, time-series models were developed using the mean of smoothed time-series data
359 (see Fig. 4). The original time series shows a yearly or seasonal trend that causes a non-
360 stationary time series. We subtracted the mean of each time series and used differencing
361 to remove these trends and make the series stationary. To determine the time-series
362 model, we used the Box-Jenkins graphical approach,[38] which relies on the patterns of the
363 autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Fig. 6
364 shows ACF and PACF of the time-series data in each spatially homogeneous region.
365 ACF slowly decays with either an exponential curve or sine waves, while PACF has a
366 large value for the first or second lag and becomes small (close to zero) for higher order

lags. These patterns suggest that a first-order or second-order autoregressive (AR) model might be a good choice.[38] Table 2 summarizes time-series models with the estimated parameters for each clustered region. AR models consider a linear combination of past values and a Gaussian white noise term. AR(1) and AR(2) models are of the forms $Y_t = \phi_1 Y_{t-1} + Z_t$ and $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + Z_t$, respectively. Y_t is the $PM_{2.5}$ concentration at time *t*, ϕs are the parameters of the model, and Z_t is a Gaussian white noise series with mean zero and variance σ_{WN}^2. The parameters of the AR models can be estimated by the maximum likelihood estimation technique, available in many standard computer packages. In the present study, we used S-PLUS 6 (Insightful Corporation, Seattle, WA). To test the adequacy of the time-series model derived, the autocorrelation functions of the estimated residual values (e.g., $Y_t - \hat{\phi}_1 Y_{t-1}$ or $Y_t - \hat{\phi}_1 Y_{t-1} - \hat{\phi}_2 Y_{t-2}$) were generated (Fig. 7). Results show that only a few points out of 40 fall outside the bound, indicating that our derived time-series models fit the data well.

Comparison of Annual $PM_{2.5}$ Level of Each Spatially Homogeneous Region with the Federal Standard

Annual mean $PM_{2.5}$ concentrations for each clustered region were compared with the annual federal standard of 15.0 μg/m^3 (Fig. 8). The x-*axis* shows the percent reduction in total $PM_{2.5}$ required to meet the standard. For example, in 2005, in the Central region, 61 of 68 sites (89.7 percent) satisfied the federal standard, which corresponds to the *y-axis* value when the *x-axis* value of the plot is zero (Fig. 8). It also shows that all sites in the Central region will satisfy the federal standard if an 18 percent reduction in total $PM_{2.5}$ is achieved for all sites in the region. The same analysis was performed for the other five clustered regions. The results showed that in 2005, 97.7 percent (Florida & Gulf Coast),

392 39.8 percent (Midwest), 65.4 percent (Northeast), 64.9 percent (Southeast), and 87.0
393 percent (West) of sites met the federal standard. To achieve the federal standard for all
394 sites in each clustered region in 2005 would require pollutant (total $PM_{2.5}$) reductions, by
395 region, of 1 percent (Florida & Gulf Coast), 24 percent (Midwest), 31 percent (Northeast),
396 26 percent (Southeast), and 22 percent (West).
397 An overall pattern of pollutant reductions required in each clustered region seems
398 similar over a period from 2001 to 2005. One clear pattern that emerged is that there were
399 a relatively large proportion of nonattainment sites in 2001 and 2005 compared to 2002,
400 2003, and 2004.
401 Interestingly, the regions with a large proportion of nonattainment sites did not
402 always require large amounts of pollutant reduction to satisfy the federal standard. A
403 comparison of the Midwest and Northeast regions in 2005 provides a good example. In
404 the Midwest region, only 39.82 percent of sites met the federal standard, but 65.38
405 percent in the Northeast met the standard. However, more efforts seemed to be required
406 in order to achieve the federal standard for all sites in the Northeast than in the Midwest
407 region. This implies that the number of sites exceeding the federal standard does not
408 correlate directly with the percent of pollutant reduction required. These results indicate
409 that different pollutant management programs should be applied to specific times and
410 regions. Overall, this analysis discusses percent reductions in total $PM_{2.5}$ required to
411 meet the federal standard based on the clustering results. However, the current analysis
412 does not provide clear recommendations about how to achieve those reductions in $PM_{2.5}$.
413
414

CONCLUSIONS

The present study examines the temporal patterns of $PM_{2.5}$ concentrations over the period from 2001 to 2005 across the continental U.S., so as to characterize spatially homogeneous regions. The *k*-means clustering algorithm using the correlation distance enabled us to measure the similarity of overall temporal patterns among 522 monitoring sites. We believe *k*-means clustering analysis can be useful as an alternate approach to test the validity of the groupings obtained by visualization methods, such as RPCA, which has been used for characterizing spatial patterns in air pollution and meteorological fields. The *k*-means clustering analysis grouped the sites in close spatial proximity. More precisely, the analysis resulted in six spatial regions that exhibit homogenous temporal $PM_{2.5}$ concentration patterns over multiple years: Central, Florida & Gulf Coast, Midwest, Northeast, Southeast, and West. In each spatially homogenous region, distinct temporal patterns were observed. In general, higher $PM_{2.5}$ concentrations occur in winter in the western part of the United States, but in summer in the northeastern and southeastern regions. These results are generally consistent with other existing studies indicating the higher levels of NO_3^- and organic carbon in the west during winter and $SO_4^=$ in the east during summer. The results also indicate that $PM_{2.5}$ concentrations vary from year to year. This may due to meteorological variations or consequences of major human- or nature-related activities. To obtain more understanding of the observed time-series patterns, we fit time-series models based on the Box-Jenkins' graphical approach. Time-series models with mean-centered and differenced data provided AR(1) or AR(2) model for each of six clustered (homogenous) regions. Residual analysis confirmed the adequacy of the derived models. These time series models can be used to predict the

440 future $PM_{2.5}$ mass concentrations in a regional scale. Finally, we showed the amounts of
441 pollutant reduction required to meet the federal standard for all sites in each clustered
442 region from 2001 to 2005.

443
444 **Acknowledgments**
445
446 We thank the referees for the constructive comments and suggestions, which greatly
447 improved the quality of the paper.

REFERENCES

1. U.S. EPA *Clean Air Act*. http://www.epa.gov/oar/caa/caa.txt
2. U.S. EPA *Federal Register, National Ambient Air Quality Standards for Particulate Matter; Proposed Rule, 40 CFR Part 50*. http://www.epa.gov/ttn/amtic/files/ambient/pm25/50frnoticejan2006.pdf
3. Pope, C. A.; Burnett, R.; Thun, N. J.; Calle, E. E.; Krewskik, D.; Ito, K.; Thurston, G. D., Lung cancer, cardiopulmonary mortality, and long term exposure to fine particulate air pollution. *Journal of the American Medical Association* **2002,** 287, 1132-1141.
4. Schwartz, J.; Dockery, D. W.; Nwas, L. M., Is daily mortality associated specifically with fine particles? *Journal of the Air and Waste Management Association* **1996,** 46, 927-939.
5. Shendriker, A. D.; Steinmetz, W. K., Integrating nepholometer measurements for air-borne fine particulate matter ($PM_{2.5}$) mass concentration. *Atmospheric Environment* **2003,** 37, 1383-1392.
6. Russell, M.; Allen., D. T.; Collins, D. R.; Fraser, M. P., Daily, Seasonal, and Spatial Trends in $PM_{2.5}$ Mass and Composition in Southeast Texas. *Aerosol Science and Technology* **2004,** 38, (S1), 14-26.
7. Paatero, P.; Hopke, P. K.; Hoppenstock, J.; Berly, S. I., Advance Factor Analysis of Spatial Distributions of $PM_{2.5}$ in the Eastern United States. *Environmental Science & Technology* **2003,** 37, 2460-2476.
8. Holland, D. M.; Principe, P. P.; Sickles, J. E., Trends in atmospheric sulfur and nitrogen species in the eastern United States for 1989-1995. *Atmospheric Environment* **1999,** 33, 37-49.
9. Malm, W. C.; Schichtel, B. A.; Ames, R. B.; Gebhart, K. A., A 10-year spatial and temporal trend of sulfate across the United States. *Journal of Geophysical Research (Atmospheres)* **2002,** 107, (D22), ACH11.1-ACH11.20.
10. Farber, R. J.; Murray, L. C.; Moran, W. A., Exploring Spatial Patterns of Particulate Sulfur and OMH from the Project MOHAVE Summer Intensive Regional

Network Using Analyses of Variance Techniques and Meteorological Parameters as Sort Determinants. *Journal of Air and Waste Management Association* **2000,** 50, 724-732.
11. Gebhart, K. A.; Malm, W. C., Spatial and Temporal Patterns in Particle Data Measured During the MOHAVE Study. *Journal of Air and Waste Management Association* **1997,** 47, 119-135.
12. Malm, W. C., Characteristics and Origins of Haze in the continental United States. *Earth Science Reviews* **1992,** 33, 1-36.
13. Goswami, E.; Larson, T.; Lurnley, T.; Liu, L. J. S., Spatial Characteristics of Fine Particulate Matter: Identifying Representative Monitoring Location in Seattle, Washington. *Journal of the Air & Waste Management Association* **2002,** 52, 324-333.
14. Chan, C.-C.; Hwang, J.-S., Site representativeness of urban air monitoring stations. *Journal of the Air & Waste Management Association* **1996,** 46, (8), 755-760.
15. Tilmes, S.; Zimmermann, J., Investigation on the spatial scales of the variability in measured near-ground ozone mixing ratios *Geophysical Research Letters* **1998,** 25, (20), 3827-3830.
16. McNair, L. A.; Harley, R. A.; Russell, A. G., Spatial inhomogeneity in pollutant concentrations, and their implications for air quality model evaluation. *Atmospheric Environment* **1996,** 30, 4291-4301.
17. Jun, M.; Stein, M. L., Statistical comparison of observed and CMAQ modeled daily sulfate levels. *Atmospheric Environment* **2004,** 38, 4427-4436.
18. Park, S.-K.; Cobb, C. E.; Wade, K.; Mulholland, J.; Hu, Y.; Russell, A., Uncertainty in Air Quality Model Evaluation for Particulate Matter due to Spatial Variation in Pollutant Concentrations. *Atmospheric Environment* **in press**.
19. Schafer, J. L., *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC: Boca Raton, FL, 1997.
20. Stein, M. L., *Interpolation of Spatial Data: Some Theory for Kriging*. Springer: 2006.
21. Phillips, S. B.; Finkelstein, P. L., Comparison of spatial patterns of pollutant distribution with CMAQ predictions. *Atmospheric Environment* **2006,** 40, 4999-5009.
22. Swall, J. L.; Davis, J. M., A Bayesian statistical approach for the evaluation of CMAQ. *Atmospheric Environment* **2006,** 40, 4883-4893.
23. Riccio, A.; Barone, G.; Chianese, E.; Giunta, G., A hierarchical Bayesian approach to the spatio-temporal modeling of air quality data. *Atmospheric Environment* **2006,** 40, 554-556.
24. Gordon, A. D., *Classification*. Chapman & Hall/CRC: New York, 1999.
25. Fox, J., *Nonparametric simple regression: smoothing scatterplots*. SAGE: Thousand Oaks, CA, 2000.
26. Cleveland, W. S., Robust locally weighted regression and smoothing scatter plot. *Journal of the American Statistical Association* **1979,** 74, 829-836.
27. Eder, B. K.; David, J. M.; Bloomfield, P., A Characterization of the Spatiotemporal Variability of Non-Urban Ozone Concentration over the Eastern United State. *Atmospheric Environment* **1993,** 27A, (16), 2645-2668.
28. Lehman, J.; Swinton, K.; Bortnick, S.; Hamilton, C.; Baldridge, E.; Eder, B.; Cox, B., Spatio-temporal characterization of tropospheric ozone across the eastern United States. *Atmospheric Environment* **2004,** 38, 4357-4369.

29. Malm, W. C.; Gebhart, K. A.; Henry, R. C., An Investigation of the Dominant Source Regions of Fine Sulfur in the Western United States and Their Areas of Influence. *Atmospheric Environment* **1990,** 24A, 3047-3060.
30. Henry, R. C.; Wang, Y. J.; Gebhart, K. A., The Relationship Between Empirical Orthogonal Functions and Sources of Air Pollution. *Atmospheric Environment* **1991,** 25A, 503-509.
31. Baumgardner, R. E.; Isil, S. S.; Bowser, J. J.; Fitzgerald, K. M., Measurement of rural sulfur dioxide and particle sulfate: Analysis of CASTNet data, 1987 through 1996. *Journal of Air and Waste Management Association* **1999,** 49, 1266-1279.
32. Johnson, R. A.; Wichern, D. W., *Applied Multivariate Statistical Analysis*. Prentice Hall: Upper Saddle River, NJ, 2002.
33. Seinfeld, J.; Pandis, S., *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. A Wiley-Interscience Publication: New York, NY, 2000.
34. McMurry, P.; Shepherd, M.; Vickery, J., *Particulate matter science for policy makers*. Cambridge University Press: New York, NY, 2004.
35. U.S. EPA *2002 National Emissions Inventory Data & Documentation*. http://www.epa.gov/ttn/chief/net/2002inventory.html
36. Malm, W. C.; Schichtel, B. A.; Pitchford, M. L.; Ashbaugh, L. L.; Eldred, R. A., Spatial and monthly trends in speciated fine particle concentration in the United States. *Journal of Geophysical Research* **2004,** 109.
37. Malm, W. C.; Sisler, J. F.; Huffman, D.; Eldred, R. A.; Cahill, T. A., Spatial and Seasonal Trends in Particle Concentration and Optical Extinction in the United States. *Journal of Geophysical Research* **1994,** 99, 1347-1370.
38. Box, G. E. P.; Jenkins, G. M.; Reinsel, G. C., *Time Series Analysis: Forecasting & Control*. 3rd Edition ed.; Prentice Hall: 1994.

About the Authors

Seoung Bum Kim is an Assistant Professor in the Department of Industrial and Manufacturing Systems Engineering at the University of Texas at Arlington. Chivalai Temiyasathit is a Ph.D. student in the Department of Industrial and Manufacturing Systems Engineering at the University of Texas at Arlington. Sun-Kyoung Park is a transportation planner at the North Central Texas Council of Governments. Victoria C.P. Chen is an Associate Professor in the Department of Industrial and Manufacturing Systems Engineering at the University of Texas at Arlington. Melanie Sattler is an Assistant Professor in the Department of Civil and

Environmental Engineering at the University of Texas at Arlington. Armistead G. Russell is the Georgia Power Professor of Environmental Engineering in the School of Civil and Environmental Engineering at the Georgia Institute of Technology. Address correspondence to: Seoung Bum Kim, 500 W. First Street, Box 1901, 420K Woolf Hall, Arlington, TX 76019-0017, Voice: 1-817-272-3150

List of Figure Captions

1. Figure 1. *k*-means clustering results for the continental United States.

2. Figure 2. A design plot to compare the yearly mean values of $PM_{2.5}$ concentrations by region and season from 2001 to 2005.

3. Figure 3. Contour plots of loadings from each of six RPCA.

4. Figure 4. Smoothed mean, median, 25th percentile, and 75th percentile temporal profiles for each clustered region.

5. Figure 5. Box plots of the seasonal mean $PM_{2.5}$ concentrations in each region over the four seasons from 2001 to 2005.

6. Figure 6. Autocorrelation and partial autocorrelation functions of the mean of smoothed time-series data (from 2001 to 2005) for each clustered region.

7. Figure 7. Autocorrelation of the residuals from time-series models.

8. Figure 8. Percentage of sites meeting the federal standard for annual $PM_{2.5}$ levels.

Table 1. A list of states in the United States in each clustered region.

Clustered Region	Number of states	States
Central	12	North Dakota, South Dakota*, Nebraska*, Kansas, Oklahoma, New Mexico*, Texas*, Minnesota, Iowa*, Missouri, Arkansas*, Illinois*
Florida & Gulf Coast	6	Texas*, Louisiana*, Alabama*, Georgia*, South Carolina*, Florida
Midwest	9	Iowa*, Wisconsin, Illinois*, Indiana, Michigan, Ohio*, New York*, Pennsylvania*, Maine
Northeast	15	Ohio*, West Virginia*, Virginia*, Pennsylvania*, New Jersey, Delaware, Maryland, Connecticut, New York*, Massachusetts, Rhode Island, Vermont, New Hampshire, Maine, Montana*
Southeast	11	Arkansas*, Louisiana*, Tennessee, Mississippi , Alabama*, Georgia*, South Carolina*, Virginia*, West Virginia*, Kentucky, California*
West	14	Washington, Oregon, California*, Nevada, Idaho, Montana*, Wyoming, Utah, Arizona, Colorado, New Mexico*, Texas*, South Dakota*, Nebraska*

* Sites in these states are split into more than one clustered region.

Table 2. Time-series models with the estimated parameters in each clustered region.

Clustered Region	Time-Series Model	$\hat{\phi}_1$	$\hat{\phi}_2$
Central	AR(2)	1.750	-0.775
Florida & Gulf Coast	AR(1)	0.783	-
Midwest	AR(2)	1.733	-0.757
Northeast	AR(2)	1.749	-0.777
Southeast	AR(2)	1.271	-0.434
West	AR(2)	1.796	-0.829