Science of the Total Environment 470–471 (2014) 631–638

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Clustering cities with similar fine particulate matter exposure characteristics based on residential infiltration and in-vehicle commuting factors

Lisa K. Baxter [a,*], Jason D. Sacks [b]

[a] National Exposure Research Laboratory, U.S. Environmental Protection Agency, RTP, NC, United States
[b] National Center for Environmental Assessment, U.S. Environmental Protection Agency, RTP, NC, United States

HIGHLIGHTS

- Cluster analysis grouped 94 cities by similar residential infiltration and commuting factors.
- Cities with older, smaller homes with less central AC grouped together.
- Cities with newer, larger homes more central AC grouped together.
- Cities with newer homes also tended to have longer commuting times and distances.
- Clusters can help group cities with similar exposures $PM_{2.5}$.

ARTICLE INFO

Article history:
Received 17 July 2013
Received in revised form 24 September 2013
Accepted 5 October 2013
Available online 27 October 2013

Editor: Pavlos Kassomenos

Keywords:
Air exchange rates
Cluster analysis
Exposure error
In-vehicle exposure
Infiltration

ABSTRACT

Epidemiological studies have observed between city heterogeneity in $PM_{2.5}$-mortality risk estimates. These differences could potentially be due to the use of central-site monitors as a surrogate for exposure which do not account for an individual's activities or ambient pollutant infiltration to the indoor environment. Therefore, relying solely on central-site monitoring data introduces exposure error in the epidemiological analysis. The amount of exposure error produced by using the central-site monitoring data may differ by city. The objective of this analysis was to cluster cities with similar exposure distributions based on residential infiltration and in-vehicle commuting characteristics.

Factors related to residential infiltration and commuting were developed from the American Housing Survey (AHS) from 2001 to 2005 for 94 Core-Based Statistical Areas (CBSAs). We conducted two separate cluster analyses using a k-means clustering algorithm to cluster CBSAs based on these factors. The first only included residential infiltration factors (i.e. percent of homes with central air conditioning (AC) mean year home was built, and mean home size) while the second incorporated both infiltration and commuting (i.e. mean in-vehicle commuting time and mean in-vehicle commuting distance) factors.

Clustering on residential infiltration factors resulted in 5 clusters, with two having distinct exposure distributions. Cluster 1 consisted of cities with older, smaller homes with less central AC while homes in Cluster 2 cities were newer, larger, and more likely to have central AC. Including commuting factors resulted in 10 clusters. Clusters with shorter in-vehicle commuting times had shorter in-vehicle commuting distances. Cities with newer homes also tended to have longer commuting times and distances.

This is the first study to employ cluster analysis to group cities based on exposure factors. Identifying cities with similar exposure distributions may help explain city-to-city heterogeneity in $PM_{2.5}$ mortality risk estimates.

Published by Elsevier B.V.

1. Introduction

Multi-city population-based epidemiological studies have observed heterogeneity between community- and city-specific $PM_{2.5}$-mortality effect estimates (Dominici et al., 2006; Franklin et al., 2007). One potential reason for these differences is the use of central-site monitors as a surrogate for exposure. This may introduce bias into the observed risk estimates if the central-site monitor–exposure relationship varies by city.

Previous studies have hypothesized and reported higher air pollution risks for cities with higher overall air exchange rates (AERs) or pollutant infiltration efficiencies (Bell and Dominici, 2008; Hodas et al., 2012; Janssen et al., 2002; Levy et al., 2005; Medina-Ramon et al., 2006). A number of factors related to home characteristics can influence the infiltration of ambient air into the home. Some of the most important

* Corresponding author at: 109 T.W. Alexander Drive, MD E205-2, RTP, NC 27711, United States. Tel.: +1 919 541 0671; fax: +1 919 541 4787.
E-mail address: baxter.lisa@epa.gov (L.K. Baxter).

0048-9697/$ – see front matter. Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.scitotenv.2013.10.019

factors include age of construction (Allen et al., 2003; Chan et al., 2005), housing type (i.e., multi- vs. single-family home) (Koenig et al., 2005; Pandian et al., 1993), and central air conditioning (AC) (Johnson and Long, 2005). In addition people may spend time away from their home (e.g. at work) or in other near-source environments (e.g. in vehicles), where the composition and toxicity of pollutants can vary. In-vehicle air pollution measurement studies have also indicated that concentrations of pollutants inside cars and buses are considerably higher than those recorded at nearby central-site monitors (Riediker et al., 2003) and exposure models suggest that even a small amount of time spent in vehicles may contribute significantly to the average daily personal PM exposure (Burke et al., 2001). Recently exposure to traffic pollution while in-vehicle has been shown to result in changes in heart rate variability (Shields et al., 2013). Estimating exposures based on community-average pollution concentrations also does not account for time spent at other locations outside the assigned community, and thus can add bias (Setton et al., 2011).

This analysis continues our attempt to better understand the heterogeneity in $PM_{2.5}$-mortality effect estimates across cities. Our objective is to group cities with similar central-site monitor–exposure relationships by clustering them using a k-means cluster analysis based on residential infiltration and commuting characteristics. Exposure variables related to infiltration and commuting patterns were developed from the American Housing Survey (AHS) from 2001 to 2005 for 94 Core-Based Statistical Areas (CBSAs). It is anticipated that this approach will identify groups of cities with similar exposure characteristics that may explain the heterogeneity in $PM_{2.5}$ mortality risk estimates observed in multi-city epidemiologic studies.

2. Methods

2.1. Development of variables

We acquired data from the AHS, available from the Department of Housing and Urban Development's website (Department of Housing and Urban Development, 2001-2005) on community-specific residential infiltration and commuting patterns. The AHS collects data on the Nation's housing, including number of apartments, single-family homes, mobile homes, and vacant housing units; and household characteristics including household income, housing and neighborhood quality, housing costs, heating equipment and fuels, size of housing unit, and recent moves. AHS also collects information on type of transportation (e.g., car, bus, subway) used to commute to work, commuting distance, and commuting time. National data are collected in odd numbered years, and supplemented with data for 47 selected CBSAs about every six years. The national sample covers an average of 55,000 housing units while each metropolitan area sample covers 4100 or more housing units. For this analysis we used the national surveys and any available metropolitan surveys from 2001 to 2005.

Using the housing units sampled in each CBSA as part of the AHS, indicators of AERs were calculated as a means to identify those cities that may have a higher fraction of ambient $PM_{2.5}$ that penetrate indoors. These indicators include percent of home with central air conditioning, average home age, and average square footage of the home for each CBSA. Previous studies have shown that personal and/or indoor concentrations of sulfate (often used as a tracer for PM of ambient origin) are lower and less well correlated with outdoor concentrations for homes with AC than homes without AC (Suh et al., 1994, 1992). This is likely because air conditioned homes typically have lower air exchange rates (AERs) than homes that use open windows for ventilation, suggesting that the fraction of $PM_{2.5}$ from ambient origin that penetrates indoors (i.e. infiltration) is lower in homes with AC than in homes without AC. Other predictors of AER include the year a structure was built, as well as its size (Chan et al., 2005; Sherman and Matson, 2002). Newer homes are generally more tightly sealed with lower AERs due to modern methods for constructing and sealing building envelopes (Chan et al., 2005; Persily et al., 2010). Similarly, larger houses typically have higher AERs compared to smaller houses, since they contain a greater surface area for leaks to develop (Chan et al., 2005).

The mean in-vehicle commuting distance and time were also calculated for each AHS sample subject in each CBSA. Commuting was considered in-vehicle if according to the AHS the mode of transportation was car, truck, van, bus/streetcar, taxicab, or other vehicle. This in-vehicle mode of transportation was then combined with the distance traveled in miles and the time traveled in minutes.

Cluster analysis is based on the distance between points so variables need to be scaled appropriately. If variables are measured on different scales, or units variables within smaller units will lead to a larger range for that variable tending to give that variable a greater effect or "weight" (Han et al., 2012). To help avoid this the variables were standardized prior to performing the cluster analysis. All variables were standardized to a mean of 0 and standard deviation of 1.

2.2. Selection of cites

The total number of CBSAs covered in the national and metropolitan surveys from 2001 to 2005 was 148. The population of the CBSA largely determines the daily number of clinical events, such as mortality and hospitalizations, and thus the statistical power to detect potential adverse health effects of air pollutants, as reflected in the confidence intervals around their effect estimates. Small CBSAs with relatively few daily events will have more uncertainty surrounding their city-specific effect estimates and less statistical power to detect potential adverse health effects of air pollutants. From a previous report we determined that populations of less than 500,000 would not provide enough daily deaths to perform a time-series analysis with sufficient power to detect significant associations between $PM_{2.5}$ and mortality (Baxter et al., 2012). As a result, for the analysis of the 148 CBSAs that are included in AHS, we focused on the 94 CBSAs with a population greater than 500,000 people. Population data for these 94 CBSAs was obtained from the U.S. Census Bureau's website (United States Census Bureau, 2010).

2.3. Cluster analysis

We used a k-means clustering algorithm to cluster CBSAs based on residential infiltration factors and commuting patterns. This iterative algorithm searches for a local solution that minimizes the Euclidean distance between the observations and the cluster centers. The k-means clustering algorithm is somewhat less sensitive to outliers than hierarchical clustering methods (Punj and Stewart, 1983). In a k-means cluster analysis the number of clusters (k) must be assigned a priori based either on pre-existing knowledge of the data or observable characteristics of the data set. For our analysis there was no pre-existing knowledge of the number of unique clusters to specify. We, therefore, calculated the within groups sum of squared errors (SSE) for 14 cluster solutions with k ranging from 2 to 15 to identify an optimal number of clusters.

SSE is defined as the sum of the squared distance between each member of a cluster and its cluster centroid (Kaufman and Rousseeuw, 1990) as shown below.

$$SSE = \sum_{i=1}^{K} \sum_{x \epsilon C_i} dist(c_i, x)^2$$

where K is the number of clusters; x is a city; C_i is the ith cluster; $dist$ is the standard Euclidean distance between two objects of Euclidean space; and c_i is the centroid of cluster C_i. In general, as the number of clusters increases, the SSE should decrease because clusters are, by definition, smaller. A plot of the SSE against a series of sequential cluster levels can provide a useful graphical way to choose an appropriate cluster level.