Classification of Urban Districts based on Mobile Carbon Monoxide Exposure Using Self Organizing Maps

Andrè Morton

International Center for Water Resources Management

Mc Lin Building, Central State University, Wilberforce, OH 45384

Abstract

The paper elaborates the concepts of exploring Self Organizing Maps (SOM) as a tool for studying few urban 'air districts' in USA with respect to the demography using a single mobile source air pollutant. The methodology is as follows: SOM is employed to cluster the model, *Hazardous Air Pollutant Exposure Model for Mobile Sources*, *Version 3* (HAPEM-MS3) constructed CO emission data. Annual average CO emission computed using HAPEM-MS3 for each selected demographic group for each county (or air district) was tabulated. A MATLAB based code was written using SOM functions for classification. In the future, for each demographic variable, these counties (or air districts) will be clustered into CO emission level groups using SOM based on the demographic groups of the demographic variable. Within each CO emission level group, each of the demographic group comprising the particular demographic variable can be compared to see the degree of exposure to each demographic group. Using ArcGIS, counties will be geographically mapped, and their proximities to the highways will be looked into within each CO emission level group for similarities. Since this study is in progress, only the SOM theory, description of case study data, methodology and short discussion are provided in this paper.

Introduction

Mobile sources that emit pollutants to air such as cars, motorcycles, trains, tractors and planes cause serious health problems in humans and also contribute to the visibility issues to the drivers. This becomes one of the major concerns in designing vehicles, in selecting the type of the fuel used in an automobile and even in transportation design and traffic control. Due the presence of Highway vehicles and construction equipments, the impact of the air pollution due to mobile sources is higher in the urban areas than in the rural areas. Mobile source originated air pollutants include carbon monoxide (CO), nitrogen oxides (NO_x), volatile organic compounds (VOCs), sulfur dioxide (SO_2) and particulate matter (PM). Among these, the latter two are of less concern in US since there is low percentage of use of diesel run vehicles and good maintenance of the vehicles.

In addition to the type of the mobile source air pollutant under consideration, exposure of each of these mobile source pollutants by a human -directly or indirectly- depends on and varies with number of parameters:

Vehicle and fuel related parameters: fuel composition, air pollution controlling features of the motor vehicle and quality of maintenance.

Terrain and Time related parameters: type of the urban area, time of the exposure event and duration of the exposure.

Human related parameters: who the individual exposed to the pollutant is with his demographical identity.

To elaborate how these parameters can influence the exposure, imagine the following contrast scenarios; the exposure to carbon dioxide by a 50 year old Caucasian business executive who annually earns $100 K and travels averagely 15 miles per day in his or her own car for 25 minutes in Denver suburban area in a winter off peak hour traffic may very much be different from the exposure to carbon dioxide by a 8 year old African American school boy who travels 3 miles per day in average in a school bus for 45 minutes in Washington D.C., downtown in a spring peak hour traffic.

Assume a city into number of 'air districts' – with the specific air pollutant concentration in each district is represented by a monitoring station. It can be safely said that even within a certain city, depending on the demographical variations, the exposure to mobile source pollutant may vary from individual to individual. At the same time, one 'air district' or county of one city may share similar pollutant exposure patterns like one 'air district' or county of another city. Understanding the demographical similarities and dissimilarities with respect to air pollutant exposure pattern may help

1. to predict the exposure pattern in a city where enough decision-making demographic information related to air pollutants are not available, and

2. to address the mobile source pollution issue in demographical perspective for controlling purpose

To classify the urban 'air districts' into similar groups and to further study the similarities and dissimilarities with respect to demographical distribution, data mining techniques such as Self Organizing Maps (SOM) may be employed.

The paper elaborates the concepts of exploring SOM as a tool for studying few urban 'air districts' in USA with respect to the demography using a single mobile source air pollutant. Since this study is in progress, only the SOM theory, description of case study data, methodology and short discussion are provided in this paper.

Theory

One of the most popular types of unsupervised artificial neural network, generally known as self organizing map (SOM) was first formulated by Kohonen has been extensively explored in many fields for the purposes of classification and pattern recognition (Manolakos, *et al.*, 2007).

In SOM networks, a competitive learning process is done with inputs but not with desired outputs. These networks are comprised of the input layer with original data and the output layer with the reduced two dimensional data. In such a network each input layer neuron represents an input variable and this input neuron is connected to each of the mapped output layer using a nonlinear projection (Figure 1). The network itself decides the features to group the input data, and this yields the namesake, '*self-organization'*. The high dimensional input data can be reduced to two dimensional during the iterative self-organization process and consequently the input data can be grouped into clusters. Analyzing the extracted relationships among the input variables, the system can be understood (Kohonen, 1990; Kohonen, 2001). A detail discussion on self organizing maps is beyond the scope of this paper. If needed, readers can consult the references given above.

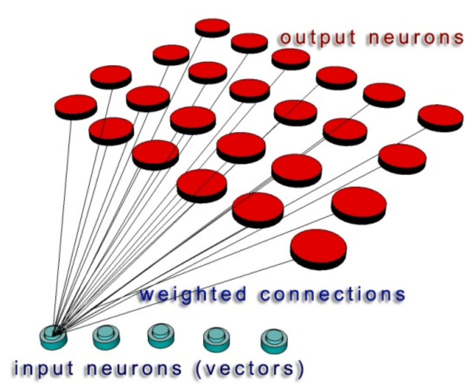

Figure 1: Components of a self organizing map

Case Study

Carbon Monoxide (CO), an odorless, colorless, tasteless gas is a result of incomplete burning of a carbonaceous fuel. It affects the human health significantly by inhibiting the ability of blood carrying oxygen to body tissues by reacting with hemoglobin to form carboxyhemoglobin. Depending on the CO concentration and the duration of exposure, effect varies from slight headaches to death.

Pie chart in Figure 2 provides the US national CO emission to the air in 2007 from various pollutant sources with two mobile sources, highway and off-highway vehicles yielding 68.4 % of the total (EPA, 2008). Citing an earlier study finding provided in the book by Cohn and McVoy, Cooper and Alley quote that in some urban area, this percentage go even to 95% (Cooper and Alley, 1994).

As Shown in Figure 3, according to EPA '*Mobile Source Emissions - Past, Present, and Future'* page, while mobile source carbon monoxide emissions are a little more than half what they were in 1970, the actual emissions today are four times less than seventies due to the controlling technology. And, by 2020, its emission is expected to be less than it is now, and about seven times less due to the future advancement in pollutant controlling technology (EPA, 2007).

The results of a study published by EPA in 1998, '*Analysis of Carbon Monoxide Exposure for Fourteen Cities using HAPEM-MS3*' was selected as the data for elaborating the technique of using SOM to group the urban districts based on the CO exposure (in the unit of $\mu g/m^3$) in various demographic classes (EPA, 1998). Model HAPEM-MS3 which is the abbreviation for *Hazardous Air Pollutant Exposure Model for Mobile Sources*, *Version 3* takes CO monitoring data, time-activity data, microenvironmental data and population data as inputs to yield exposure estimates to ambient carbon monoxide (CO) by demographic group, quarter of the year, and county (or district). Description and discussion of HAPEM-MS3 is beyond the scope of this paper. Interested readers should consult the relevant EPA report.

Although the model is applied to 23 demographic data of 1990 for four quarters of year, we took the demographic data that classify the average annual CO exposure for three different demographic variables for our study; ethnicity, income and age, employment and gender. Data for fourteen cites was studied. The detail demographic variables and their groups are listed in Table 1. Table 2 provides the cities and the numbers of counties (total of 102) and 'air districts' (Total of 118) that are under study; considering the length, the names of the counties are not listed.

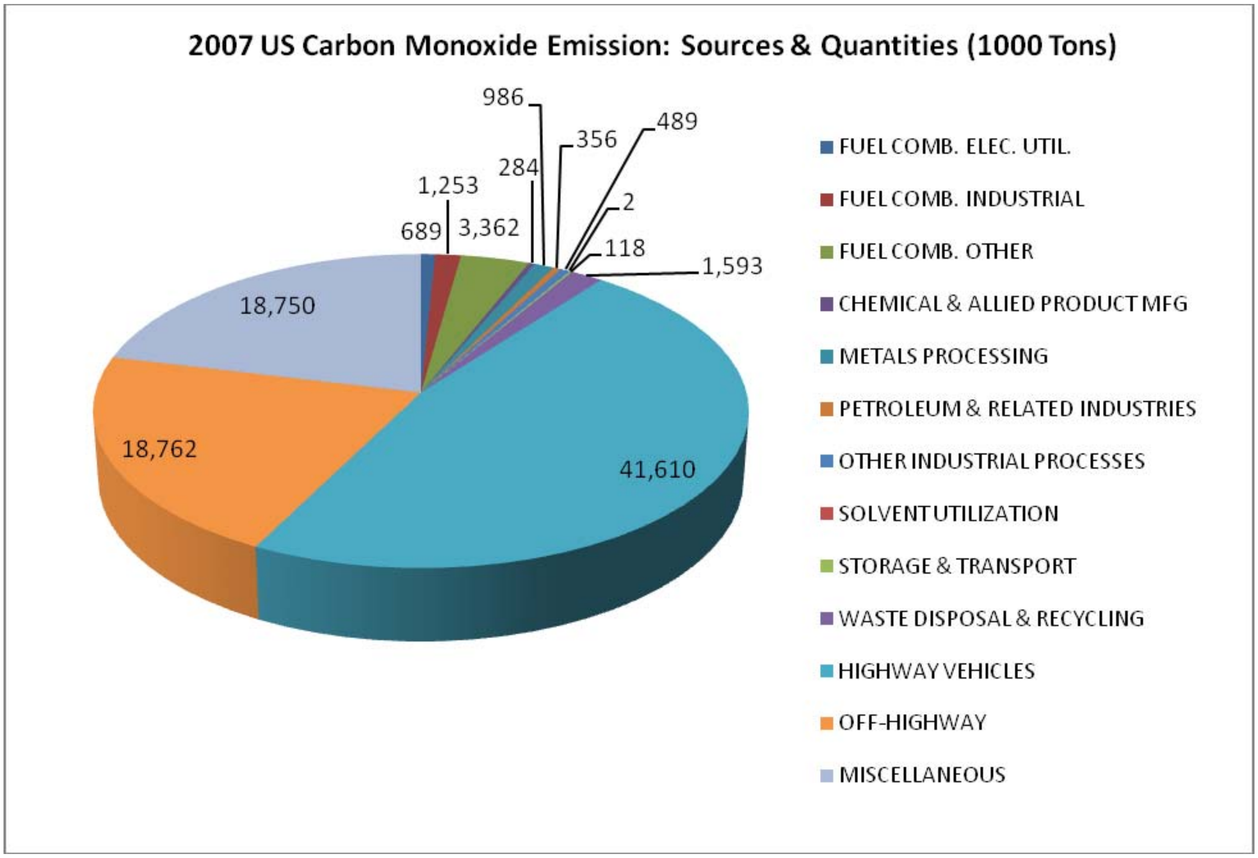

Figure 2: US Nation Carbon Monoxide Emission 2007: Sources & Quantities (Data source: EPA, 2008)

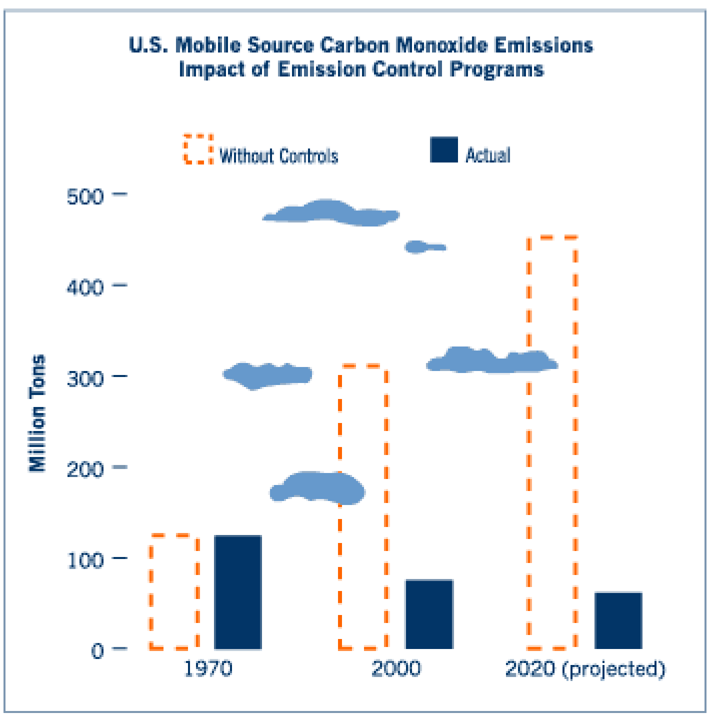

Figure 3: US Mobile Source Carbon Monoxide Emissions: Without Control and Actual (EPA, 2007)

Demographic Variable	Demographic Groups
Ethnicity	Caucasians African Americans Hispanic
Household Income	Less than $10K $10K-$25K $25K-$50K $50K-$75K Greater than $75k
Age_Employment_Gender	Children, 0 to 17 Nonworking men, 18 to 44 Working men, 18 to 44 Nonworking women, 18 to 44 Working women, 18 to 44 Nonworking men, 45 to 64 Working men, 45 to 64 Nonworking women, 45 to 64 Working women, 45 to 64 Men, 65+ Women, 65+

Table 1: Demographic variables and their groups used in the case study (Data Source: EPA, 1998)

City	Relevant States	Number of Counties	Number of Air Districts
Baltimore	MD	7	4
Boston	MA,NH	9	5
Chicago	IL,IN	6	9
Denver	CO	6	9
Houston	TX	6	5
Los Angeles	CA	4	16
Minneapolis / St.Paul	MN,WI	11	5
New York City	NY,NJ,CT	18	14
Philadelphia	PA,NJ,DE	12	13
Phoenix	AZ	2	8
San Francisco	CA	7	10
Spokane	WA	1	5
St. Louis	MO,IL	4	8
Washington D.C.	DC,VA,MD	9	7

Table 2: Case study: cities, related states, number of counties & number of air districts (EPA, 1998)

Methodology

SOM is employed to cluster the HAPEM-MS3 constructed CO emission data described the previous section. Annual average CO emission computed using HAPEM-MS3 for each selected demographic group for each county (or air district) has already been tabulated. A MATLAB based code has been written using SOM functions for classification. For each demographic variable, these counties (or air districts) will be clustered into CO emission level groups using SOM based on the demographic groups of the demographic variable. Within each CO emission level group, each of the demographic group comprising the particular demographic variable can be compared to see the degree of exposure to each demographic group. Using ArcGIS, counties are geographically mapped, and their proximities to the highways are looked into within each CO emission level group for similarities.

Discussion

The results that would be obtained with the methodology described may only show which counties (or air districts) have the similar CO emission levels with respect to different demographic variables such as ethnicity, household income and age, employment and gender. The concept and the methodology provided for the case study in progress is a part of a much larger research that includes the other demographic groups as well as other mobile source pollutants (Smith and Kandiah, 2009; Kandiah and Meade, 2009). In the expanded study, the distribution patterns of other demographic variables within each of the CO emission level groups can be compared with the distribution patterns of the demographic groups used in clustering. This will help to understand the underlying relationships better. In addition, clustering the counties with respect to the all mobile source pollutants will help to rank the counties in terms of mobile source related exposure. Classical statistical tools such as principal component analysis and canonical analysis can be incorporated with SOM to extract more information from the available exposure data. Also, it should be noted that there are more accurate mathematical models that could replace HAPEM-MS3 are available. Hence new studies incorporating such models will improve the accuracy.

Results of this study will be helpful in identifying the hot spot counties and understanding the underlying relationships between the air pollutant exposure and demographic variables that would help in decision making in controlling the mobile source emission & pollution.

Acknowledgement

I wish to thank Dr. Ramanitharan Kandiah for mentoring me in this research and International Center for Water Resources Management, Central State University for allowing me using its facilities for my research.

References

Cooper, C. David and Alley, F. C. 1994. *Air Pollution Control: A Design Approach*. Waveland Approach. Second Edition. 694 pg.

EPA. 1998. Analysis of Carbon Monoxide Exposure for Fourteen Cities using HAPEM-MS3: Final Report. National Exposure Research Laboratory, U.S. Environmental Protection Agency. 1998 March. 75 pg.

EPA. 2007. *Mobile Source Emissions - Past, Present, and Future.* Last updated on Monday, July 9th, 2007. http://www.epa.gov/OMS/invntory/overview/results/allmobile.htm

EPA. 2008. 1970 - *2007 Average annual emissions, all criteria pollutants in MS Excel - August 2008*. Last updated on Tuesday, December 2nd, 2008.
http://www.epa.gov/ttn/chief/trends/trends06/nationaltier1upto2007basedon2005v1.xls

Kandiah, Ramanitharan and Meade, Wilbert. 2009. *A Methodology to Estimate the Annual Average Daily On-Road Mobile Source Pollutant Emissions*. 5th Annual Dayton Engineering Sciences Symposium; Wright State University, Dayton, OH. October 26, 2009.

Kohonen, T. 1990. The self-organizing map. *Proceedings of the IEEE*. **78**(9): 1464-1480.

Kohonen, T. 2001. *Self-Organizing Maps.* Springer, Berlin. 2001.

Manolakos, Elias, Virani, Hardik and Novotny, Vladimir. 2007. Extracting knowledge on the links between the water body stressors and biotic integrity. *Water Research.* **41**(18): 4041-50.

Smith, Bartina and Kandiah, Ramanitharan. 2009. *Air Quality Monitoring*. Presented at Statewide Student Research Conference; Cleveland State University, Cleveland, OH. March 27-29, 2009.