Spatio-Temporal Data Mining for Air Pollution Problems

Seoung Bum Kim
The University of Texas at Arlington, USA

Chivalai Temiyasathit
The University of Texas at Arlington, USA

Sun-Kyoung Park
North Central Texas Council of Governments, USA

Victoria C. P. Chen
The University of Texas at Arlington, USA

INTRODUCTION

Vast amounts of data are being generated to extract implicit patterns of ambient air pollution. 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 temporal and spatial characteristics. Furthermore, combinations of observations from multiple monitoring stations, each with a large number of serially correlated values, lead to a situation that poses a great challenge to analytical and computational capabilities. Data mining methods are efficient for analyzing such large and complicated data. Despite the great potential of applying data mining methods to such complicated air pollution data, the appropriate methods remain premature and insufficient. The major aim of this chapter is to present some data mining methods, along with the real data, as a tool for analyzing the complex behavior of ambient air pollutants.

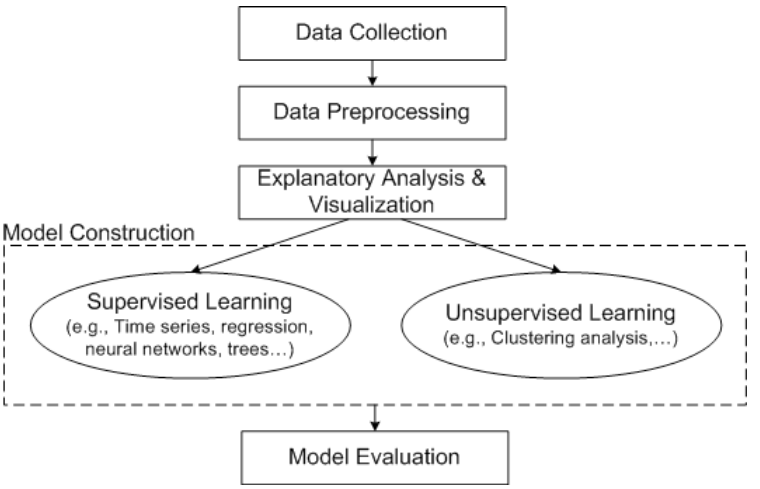
BACKGROUND

In 1990, under the Clean Air Act., the U.S. Environmental Protection Agency (EPA) set the National Ambient Air Quality Standards (NAAQS) for six pollutants, also known as criteria pollutants, which are particulate matter, ozone, sulfur dioxide, nitrogen dioxides, carbon monoxide, and lead (US EPA, 1990). Any exceedance of the NAAQS results in non-attainment of the region for that particular pollutant.

Well-known consequences of air pollution include the green house effect (global warming), stratospheric ozone depletion, tropospheric (ground-level) ozone, and acid rain (Wark, Warner, & Davis, 1998). In this chapter, we present applications on tropospheric ozone and the less publicized air pollution problem of particulate matter. High concentrations of tropospheric ozone affect human health by causing acute respiratory problems, chest pain, coughing, throat irritation, or even asthma (Lippmann, 1989). Ozone also interferes with the ability of plants to produce and store food, damages the leaves of trees, reduces crop yields, and impacts species diversity in ecosystems (Bobbink, 1998; Chameides & Kasibhatla, 1994). Particulate matter is an air contaminant that results from various particle emissions. For example, $PM_{2.5}$ (particulate matter that is 2.5 micrometers or smaller in size) has the potential to cause adverse health effects in humans, including premature mortality, nose and throat irritation, and lung damage (e.g., Pope et al., 2002). Furthermore, $PM_{2.5}$ has been associated with visibility impairment, acid deposition, and regional climate change.

To reduce pollutant concentrations and establish the relevant pollution control program, a clear understanding of the pattern of pollutants in particular regions and time periods is necessary. Data mining techniques can help investigate the behavior of ambient air pollutants and allow us to extract implicit and potentially useful knowledge from complex air quality data. Figure 1 illustrates the five primary stages in the data mining process in air pollution problems: data collection, data preprocessing, explanatory analysis and visualization, model construction, and model evaluation.

Figure 1. Overview of data mining in air pollution problems.



MAIN FOCUS OF CHAPTER

Data Collection

Because air pollution data are generally collected in a wide region of interest over a relatively long time period, the data are composed of both temporal and spatial information. A typical air pollution database consists of pollutant observations $O(S_i, T_j)$, for monitoring site S_i at time T_j for $i=1,2,...,m$, $j=1,2,...,n$, where m and n is the number of monitoring sites and time points, respectively. Since most air pollution data hold these two properties, spatial and temporal variability should be incorporated into the analysis in order to accurately analyze the air pollution characteristics. Table 1 provides a list of publicly accessible databases and their web addresses that contain a variety of air pollution data.

Data Preprocessing

Preprocessing of air pollution data is a crucial task because inadequate preprocessing can result in low-quality data and make it difficult to extract meaningful information from subsequent analyses. The collected air pollution data typically contain a number of potential outliers that are far away from the rest of the observations and missing values possibly due to measurement or instrumental errors. It is necessary to process missing values and outliers in both the time and space domains. Imputing missing values or replacing potential outliers with a sample average is the simplest method because it can be calculated without any pre-specified assumptions or complex mathematical formulas. However, the sample average assumes that each observation is equally important and does not take into account the fact that the data are collected over time and space. A weighted average can be an efficient method to replace the outliers or impute the missing values. One example of using a weighted average is the inverse-distance-squared weighted method (McNair, Harley, & Russell, 1996). This method determines weights based on spatial proximity to the query points. The interpolated value for site S_i at time T_j, $I(S_i, T_j)$ is computed as follows:

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

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

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