Contents lists available at ScienceDirect

Expert Systems with Applications

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

$PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining

Ming Dong [a,*], Dong Yang [a], Yan Kuang [b], David He [c], Serap Erdal [d], Donna Kenski [e]

[a] Department of Industrial Engineering and Management, School of Mechanical Engineering, Shanghai Jiao Tong University, 800 Dong-chuan Road, Shanghai 200240, PR China
[b] General Electric (Shanghai) Corporation, 1800 Cai Lun Road, Shanghai 201203, PR China
[c] Department of Mechanical and Industrial Engineering, 842 West Taylor Street, University of Illinois-Chicago, Chicago, IL 60607, USA
[d] Environmental and Occupational Health Sciences, School of Public Health, University of Illinois-Chicago, Chicago, IL 60612, USA
[e] Lake Michigan Air Directors Consortium, 2250 E. Devon Ave., Suite 250, Des Plaines, IL 60018, USA

ARTICLE INFO

Keywords:
Air pollution
Hidden semi-Markov model
$PM_{2.5}$ concentration prediction

ABSTRACT

In this paper, a novel framework and methodology based on hidden semi-Markov models (HSMMs) for high $PM_{2.5}$ concentration value prediction is presented. Due to lack of explicit time structure and its short-term memory of past history, a standard hidden Markov model (HMM) has limited power in modeling the temporal structures of the prediction problems. To overcome the limitations of HMMs in prediction, we develop the HSMMs by adding the temporal structures into the HMMs and use them to predict the concentration levels of $PM_{2.5}$. As a model-driven statistical learning method, HSMM assumes that both data and a mathematical model are available. In contrast to other data-driven statistical prediction models such as neural networks, a mathematical functional mapping between the parameters and the selected input variables can be established in HSMMs. In the proposed framework, states of HSMMs are used to represent the $PM_{2.5}$ concentration levels. The model parameters are estimated through modified forward–backward training algorithm. The re-estimation formulae for model parameters are derived. The trained HSMMs can be used to predict high $PM_{2.5}$ concentration levels. The validation of the proposed framework and methodology is carried out in real world applications: prediction of high $PM_{2.5}$ concentrations at O'Hare airport in Chicago. The results show that the HSMMs provide accurate predictions of high $PM_{2.5}$ concentration levels for the next 24 h.

1. Introduction

Prediction of particulate matter (PM) in the air is an important issue in control and reduction of pollutants in the air. Particulate matter is the term used for a mixture of solid particles and liquid droplets found in the air. In particular, fine particles that are smaller than 2.5 or 10 μm (millionths of a meter) in diameter are defined as $PM_{2.5}$ or PM_{10}. Fine particles (especially, $PM_{2.5}$) harm human health. The US Environmental Protection Agency (EPA) recently promulgated revised standards for PM and established new annual and 24-h fine particulate standards with $PM_{2.5}$ mass as the indicator due to scientific data associating fine particle pollution with significant increases in the risk of death from lung cancer, pulmonary illness (e.g., asthma), and cardiovascular disease (Dockery & Pope, 1994; EPA, 2002; Katsouyanni, 1997; Levy, 2000; Pope, Thurston, & Krewski, 2002). These fine particles are generally emitted from activities such as industrial and residential combustion and from vehicle exhaust. The health effects of exposure to fine particles include: (1) increased premature deaths, primarily in the elderly and those with heart or lung disease, (2) aggravation of respiratory and cardiovascular illness, leading to hospitalizations and emergency room visits, particularly in children, the elderly, and individuals with heart or lung conditions, (3) decreased lung function and symptomatic effects such as those associated with acute bronchitis, particularly in children and asthmatics, (4) new cases of chronic bronchitis and new heart attacks, (5) changes to lung structure and natural defense mechanisms. Fine particles in the air also decrease visibility. The benefits to human health and the environment by reducing fine particles and ozone can be significant. By 2020, the benefits of reductions in fine particles and ozone are estimated to be $113 billion annually (The Clear Skies Act, 2003). By 2010, reductions in fine particles and ozone are estimated to result in substantial early benefits of $54 billion, including 7900 fewer premature deaths, annually (The Clear Skies Act, 2003). Other significant health and environmental benefits include reduced human exposure to mercury, fewer acidified lakes, and reduced nitrogen loads to sensitive ecosystems that cannot currently be quantified and/or monitored but are nevertheless expected to be significant.

* Corresponding author. Tel.: +86 21 34206101.
E-mail address: mdong@sjtu.edu.cn (M. Dong).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.12.017

Predictive models for $PM_{2.5}$ vary from the extremely simple to extremely complex, yet the ability to accurately forecast air quality remains elusive. Much of the variability in PM concentrations is driven by meteorological conditions, which fluctuate on multiple time and spatial scales. Another significant source of variability is changes in the temporal and spatial patterns of emissions activity. Qualitative and quantitative models to forecast PM and ozone were described in a recent EPA document (EPA, 2003). As documented also by Schlink, Pelikan, and Dorling (2003), a particular technique often has good performance in one respect and poor performance in others. The quantitative models are briefly summarized here. Note that, because $PM_{2.5}$ has been regulated only since 1997, and a national measurement program implemented only since 1999, fewer forecasting applications have been developed to date for $PM_{2.5}$ than for ozone. The need for accurate forecasts of $PM_{2.5}$ continues to grow as epidemiological evidence of $PM_{2.5}$'s acute health impacts mounts.

In the past, a number of techniques have been developed for the prediction of PM concentrations. Essentially, approaches for PM prediction can be classified into five categories: (1) empirical models, (2) fuzzy logic-based systems, (3) simulation models, (4) data-driven statistical models, and (5) model-driven statistical learning methods.

Empirical models are developed by field experts and validated by data sets of the studied area. Generally, method performance depends on the variable under study, the geographic location, and the underlying assumptions of the methods. Therefore an empirical method is "best" only for specific situations. Fuller, Carslaw, and Lodge (2002) devised an empirical model to predict concentrations of PM_{10} at background and roadside locations in London. The model accurately predicts daily mean PM_{10} across a range of sites from curbside to rural. Predictions of future PM_{10} can be made using the expected reductions in non-primary PM_{10} and site specific annual mean NO_X predicted from emission inventories and dispersion modeling. However, the model has a limited geographical extent covering London and its immediate surrounding area. The model performance depends on a consistent relationship between PM_{10} and NO_X emissions.

The fuzzy logic approach makes it possible to deal with problems affected by uncertainty and to obtain reliable models for non-linear phenomena whose characterization is based on rough and poor data. However, like rule-based systems, the determination of a fuzzy model knowledge base is obtained by the contribution of experts of the field. Raimondi, Rando, Vitale, and Calcara (1997a, 1997b) proposed a fuzzy logic-based model for predicting ground level pollution. The procedure consists of two different phases. The first phase concerns prediction of meteorological and emission variables (model input) and is implemented through fuzzy prediction of time series. The second phase of modeling concerns the determination (using fuzzy inference methods) of the predicted meteorological classes, each of which contributes in determining model output (i.e. prediction of air pollutant concentration).

In recent years, the use of three-dimensional high frequency mesoscale data sets derived from dynamical models to drive air quality simulation models has been growing. Three-dimensional air quality models have been employed to forecast pollutant concentrations. These models use meteorological model output such as the Penn State Mesoscale Model (MM5) and emissions model output for the forecasting period, then apply a mathematical model to simulate transport, diffusion, reactions, and deposition of air pollutants over the geographical area of interest, from urban scale to national scale. These models are extremely complex to set up and require enormous computing resources. They are capable of predicting air quality in areas where no monitoring data exist, but accuracy is limited by the scale at which they are applied – small scale meteorological and emissions variability may not be represented in the models. Emissions data are notoriously uncertain. Performance of these models for ozone has been reasonably good, but to date their ability to model $PM_{2.5}$ has been poor, due in part to the reasons above but also to the complexities of $PM_{2.5}$ atmospheric chemistry (Baker, 2004).

Data-driven statistical models are developed from collected input/output data. Data-driven statistical models can process a wide variety of data types and exploit the nuances in the data that cannot be discovered by rule-based or fuzzy logic-based systems. Therefore, they are potentially superior to the rule-based systems. Data-driven statistical models include Classification and Regression Tree analysis (CART), regression models, clustering techniques, and neural networks.

CART is based on binary recursive partitioning. Each predictor variable is examined (whether it is a continuous or discrete variable) and the data set is split into two groups based on the value of that predictor that maximizes the dissimilarity between groups. The tree is 'grown' by exhaustively searching the predictor variables at each branch for the best split. Typical predictors include meteorological conditions (especially temperature, wind speed, wind direction) and also air quality conditions. Seasonal or activity data can be incorporated as well. For PM, these models generally account for about 60% of the variability in the data and for ozone, about 80%.

Regression equations have a long history of use as forecasting tools in multiple disciplines. Like CART, multiple predictors are typically incorporated into a regression model that seeks to predict pollutant concentrations. Regression models are most useful and accurate for predicting mean concentrations and less dependable for the extreme values that are generally of most interest when forecasting concentrations for the purpose of warning the public about health risks. Regression models have the advantage of simple computation and easy implementation. However, regression models are based on the assumption of normally distributed data; air quality and meteorological data are generally log-normally distributed. Transformations of the data can improve model performance. Many of the relationships between PM and meteorological variables are curvilinear, which requires additional transformations of the predictor variables. Due to the nature of linear relationship, regression models may not provide accurate predictions in some complex situations. Researchers have applied regression models into different areas such as: downtown area of Santiago, Chile; Ontario, Canada; Taiwan, China; Delhi, India; Maryland, USA and about 100 Canadian sites (Burrows, Montpetit, & Pudykiewicz, 1997; Chaloulakou, Grivas, & Spyrellis, 2003; Chelani, Gajghate, Tamhane, & Hasan, 2001; Fraser & Yap, 1997; Lu & Fang, 2003; Ojha, Coutinho, & Kumar, 2002; Rizzo, Scheff, & Ramakrishnan, 2002; Walsh & Sherwell, 2002).

The main purpose of clustering technique is to identify distinct classes among the data. It can be used for spatial classification of ambient air quality data, in the absence of the huge data sets needed for more sophisticated space–time modeling (Surneet, Veena, & Patil, 2002). However, the analysis is based on grossly-average-level data, not intensive daily data. The clustering algorithm developed by Sanchez, Pascual, Ramos, and Perez (1990) has been applied to PM concentrations recorded at each sampler point, and different pollution levels have been obtained in each of them. This algorithm has revealed a satisfactory relationship between PM concentrations and the identified meteorological types. However, the clustering technique such as *k*-MEANS is very sensitive to the presence of noise and cannot classify outliers. Good quality clustering algorithms are usually expensive. For example, the exact solution of *k*-MEDOIDS (*p-median*) clustering algorithm is NP-hard (Estivill-Castro & Houle, 2001).

Artificial neural networks (ANN) are computer programs that attempt to simulate human learning and pattern recognition, and