Science of the Total Environment 443 (2013) 93–103

Contents lists available at SciVerse ScienceDirect

Science of the Total Environment

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

Prediction of 24-hour-average $PM_{2.5}$ concentrations using a hidden Markov model with different emission distributions in Northern California

Wei Sun [a], Hao Zhang [a], Ahmet Palazoglu [b], Angadh Singh [b], Weidong Zhang [c,*], Shiwei Liu [a]

[a] Beijing Key Laboratory of Membrane Separation Process and Technology, Beijing University of Chemical Technology, Beijing 100029, China
[b] Department of Chemical Engineering and Materials Science, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA
[c] State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology, Beijing 100029, China

HIGHLIGHTS

- A hidden Markov model with different non-Gaussian distributions is developed to match data characteristics.
- The method is applied to the prediction of $PM_{2.5}$ exceedance days in Concord, CA and Sacramento, CA.
- Results show that the HMM can predict most exceedances correctly and reduce false alarms significantly.

ARTICLE INFO

Article history:
Received 18 July 2012
Received in revised form 17 October 2012
Accepted 17 October 2012
Available online 23 November 2012

Keywords:
$PM_{2.5}$
Hidden Markov model
Gamma
Lognormal
GEV distribution

ABSTRACT

Prediction of air pollutant levels plays an important role in the regulatory plans aimed at the control and reduction of airborne pollutants such as fine particulate matter (PM). Deterministic photochemical air quality models, which are commonly used for regulatory management and planning, are computationally intensive and also expensive for routine predictions. Compared to deterministic photochemical air quality models, data-driven statistical models are simpler and may be more accurate. In this paper, hidden Markov models (HMM) are used to forecast daily average $PM_{2.5}$ concentrations 24 h ahead. In conventional HMM applications, observation distributions emitted from certain hidden states are assumed as having Gaussian distributions. However, certain key meteorological factors and most $PM_{2.5}$ precursors exhibit a non-Gaussian distribution in reality, which would degrade the HMM performance significantly. In order to address this problem, in this paper, HMMs with log-normal, Gamma and generalized extreme value (GEV) distributions are developed to predict $PM_{2.5}$ concentration at Concord and Sacramento monitors in Northern California. Results show that HMM with non-Gaussian emission distributions is able to predict $PM_{2.5}$ exceedance days correctly and reduces false alarms dramatically. Compared to HMM with Gaussian distributions, HMM with log-normal distributions can improve the true prediction rate (TPR) by 37.5% and reduce the false alarms by 78% at Concord. And HMM with GEV distribution can improve TPR by 150% and reduce false alarms by 63.62% at Sacramento Del Paso Manor. Comparisons between different distributions used in HMM show that the closer the distribution employed in HMM is to the observation sequence, the better the model prediction performance.

1. Introduction

$PM_{2.5}$ comprises solid particles and liquid droplets whose aerodynamic diameters are 2.5 μm or less, and can be classified into two types by its forming processes: the primary $PM_{2.5}$ which comes from the combustion of fossil fuels, such as transportation, smelting, and metal processing (Jacobson, 2002) and the secondary $PM_{2.5}$ which is generated in reactions among air pollutants, such as ammonia and SO_2 (Ying and Kleeman, 2006). $PM_{2.5}$ has been reported as being the most harmful airborne pollutant, as it can travel deep into the lungs and is made of toxic substances such as heavy metals and carcinogenic organic compounds (Fang et al., 1999; Thomaidis et al., 2003). The large number of deaths and other health problems associated with PM pollution was first demonstrated in the early 1970s and has been reported many times from then on (Brunekreef and Holgate, 2002; Franklin et al., 2006). Exposure to high concentrations of $PM_{2.5}$ leads to increased use of medication and more visits to the doctor or emergency room. In addition to its effect on humans, $PM_{2.5}$ can also lead to growth stunting or mortality in some plant species by clogging stomatal openings of plants and interfering with photosynthesis functions (Bench, 2004). In order to protect people from exposures to high $PM_{2.5}$ concentrations, US EPA sets national ambient air quality standards (NAAQS) for $PM_{2.5}$ based on annual average concentration (15 $\mu g\ m^{-3}$) and 24-h average concentration (35 $\mu g\ m^{-3}$), and review these standards periodically (USEPA, 2006). The World Health Organization (WHO) sets stricter

* Corresponding author at: The college of Chemical Engineering, Beijing University of Chemical Technology, 15 Beisanhuan East Road, Beijing 100029, China. Tel.: +86 10 6442 3628; fax: +86 10 6443 6781.
E-mail address: zhangwd@mail.buct.edu.cn (W. Zhang).

0048-9697/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.scitotenv.2012.10.070

standards for $PM_{2.5}$: annual average concentration is set to be less than 10 μg m^{-3} and 24-h average concentration is set to be less than 25 μg m^{-3} (WHO, 2005).

In densely populated areas of the United States, $PM_{2.5}$ has been shown to easily reach harmful levels under favorable meteorological conditions (Beaver et al., 2010). Although great efforts have been made by EPA and local governments, by 2009 over 126 million people lived in areas exceeding the NAAQS for $PM_{2.5}$ in the United States. Among those, San Francisco Bay Area and Sacramento Valley had been designated as $PM_{2.5}$ nonattainment areas. In both areas $PM_{2.5}$ control programs have been enacted, aimed at increasing public awareness and improving local air quality. A key component of these programs is the prediction of $PM_{2.5}$ exceedance days and the reduction of human exposure to high $PM_{2.5}$ concentrations. The development of a $PM_{2.5}$ prediction method which aims to improve early warning procedures is critical to this effort.

To predict $PM_{2.5}$ concentrations, various modeling approaches have been proposed in recent years. Basically, there are two types of mathematical models: deterministic and statistical models. Deterministic models, also named as chemical transport models (CTMs), focus on the sources and transport of chemical species. Different chemical mechanisms, chemical kinetic expressions, reaction rate coefficients, a number of chemical species and gas phase reactions among them are usually incorporated in these very complex models. Examples of CTMs include CMAQ, GEOS-Chem, LOTOS-EUROS, MOZART and CLaMS (Brasseur et al., 1998; McKenna et al., 2002; Fusco and Logan, 2003; Tesche et al., 2006; Schaap et al., 2008). Their accuracy depends on the scale they are applied and the quality of the emission data (Isukapalli, 1999; Han et al., 2008). Compared to statistical models, they are more complex, computationally intensive and less accurate (Cobourn, 2010).

Statistical models use various techniques to correlate the relationship between air quality and meteorological factors. A large panel of statistical methods for air quality prediction can be found in the literature. However, there are only a few papers for $PM_{2.5}$ prediction using statistical methods.

Among these statistical methods, artificial neural network (ANN) is a popular method for $PM_{2.5}$ prediction. Perez et al. used neural network models to predict hourly $PM_{2.5}$ concentration in Santiago, Chile (Perez et al., 2000). Previous day hourly $PM_{2.5}$ concentration, wind speed and direction, and humidity are used as predictor variables. Prediction errors range from 30% for early hours to 60% for late hours. Ordieres et al. used three types of ANNs to predict daily average $PM_{2.5}$ at El Paso (Texas) and Ciudad Juárez (Chihuahua) and found that radial basis function ANN (RBF-ANN) had the shortest training time and a greater stability during prediction stage (Ordieres et al., 2005). In their work, average levels of $PM_{2.5}$ during the first 8 h of the day, maximum level of $PM_{2.5}$ of the first 8 h of the day, temperature, relative humidity, wind speed, and wind directions were selected as input variables. The highest correlation factor (R^2) was 0.4611 and the average error was about 30%. Voukantsis et al. applied neural networks for forecasting PM_{10} and $PM_{2.5}$ concentrations in Thessaloniki and Helsinki (Voukantsis et al., 2011). CO, NOx and SO_2 were selected as $PM_{2.5}$ precursors. Meteorological factors included temperature, relative humidity and wind fields. The correlation factor of their forecasting result was 0.639 and the relative mean square error was 7.128 μg m^{-3}. Although ANNs can yield good prediction results, due to the lack of theoretical results from the statistical viewpoint as well as the poor interpretability of such black-box models, alternative strategies have been reported recently.

Besides ANNs, regression models are also used to forecast air quality. Cobourn developed an enhanced $PM_{2.5}$ air quality forecast model based on nonlinear regression and back-trajectory $PM_{2.5}$ concentrations (Cobourn, 2010). In his paper, meteorological data consist of hourly surface temperatures, dew point temperature, wind field, daily minimum and maximum temperatures, along with back-trajectory $PM_{2.5}$ concentrations as predictor variables. Results showed that this model can forecast 88% of exceedance days correctly and the False Alarm Rate (FAR) was 36%. Although the normalized mean absolute error (NMAE) was only 20%, FAR was relatively high.

Compared to nonlinear regression and ANNs mentioned above, hidden Markov models (HMMs) offer a richer mathematical structure and can be used to establish mathematical functional mapping between the hidden patterns and observations. It has been increasingly popular and effective in speech recognition and hand-written word identification since the late 1960s (Rabiner, 1989). In the second half of 1980s, HMMs began to find applications in biological analysis, machine translation, gene prediction, activity recognition and protein folding (Hannenhalli and Russell, 2000; Kato et al., 2003; Cai et al., 2009; Keller et al., 2009; Arias-Londoño et al., 2010; Leu and Adi, 2011). HMMs also have been used for $PM_{2.5}$ prediction at O'Hare airport recently in Chicago (Dong et al., 2009). Meteorological data from 2000 to 2001 is used to train the model and 12 exceedance and 12 non-exceedance days are used as validation data. Results show that the modified HMM can predict high $PM_{2.5}$ concentration levels correctly. In contrast to most previous HMM applications, the prediction of $PM_{2.5}$ focuses on the modeling of $PM_{2.5}$ exceedance days which in fact take a very small fraction of the total number of monitoring days. $PM_{2.5}$ exceedance days are defined as days in which the daily average $PM_{2.5}$ concentration is greater than 35 μg m^{-3} and they would aggregate towards the right tail of the distribution curve. Previous research suggests that some key meteorological factors which affect $PM_{2.5}$ production and depletion also display non-Gaussian distributions (Hubbard and Cobourn, 1998; Cobourn and Hubbard, 1999; Denby et al., 2008). However, in most HMM applications, the hidden state outputs are represented by Gaussian (normal) distributions. As real data seldom distribute normally, there are two ways to address this problem. One is to use multiple Gaussians to mimic the real distribution. As the number of Gaussian distributions increases, the number of parameters to be estimated will also increase rapidly, and so will the corresponding computational load. Also, the number of exceedance days may be too few to estimate the HMM parameters properly. In another words, increasing the number of Gaussians in emission distribution estimation may be impractical. An alternative approach is to use a non-Gaussian distribution to approximate the real distribution. Both approaches are trying to make the model mimic the real situations better, while the latter will not increase the number of parameters to be estimated. O'Connell et al. used a HSMM based on a non-Gaussian sojourn distribution and a Gaussian emission distribution to predict cattle activity and progesterone level. Their result shows that their model can predict 70% of follicular states correctly (O'Connell et al., 2010). This suggests that HMMs with distributions which are closer to the real data behavior may be more suitable for the prediction of high $PM_{2.5}$ level days.

In this paper, continuous multivariate HMMs with different non-Gaussian distributions are developed to predict $PM_{2.5}$ exceedance days at Concord and Sacramento, CA. The remainder of this paper is organized as follows: first, the traditional HMM with a Gaussian distribution is introduced. Then, a modified Expectation–Maximization (EM) algorithm used to estimate the HMMs with non-Gaussian emission distribution parameters is described in detail. Next, data characteristics at Concord and Sacramento are described. This is followed by the results and discussion. Finally, some concluding remarks are presented.

2. Methods

2.1. Hidden Markov model

HMM is a doubly embedded stochastic process, in which one is an underlying Markov chain, a series of hidden states; and the other one is the observation sequence determined by the current hidden state of a given Markov chain (Rabiner, 1989), the outcome of a certain hidden state. One can only see the observations. In most applications,