Expert Systems with Applications 36 (2009) 9046–9055

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 should be well suited to extracting information from imprecise and non-linear data such as air quality and meteorology. They are useful tools for prediction, function approximation and classification. Despite their theoretical superiority, they have produced only a slight improvement over the linear statistical model in forecast accuracy. Extreme values are represented well if they are present in the data set that the network was trained on, but the network cannot accurately extrapolate values outside the training set. For example, Chaloulakou et al. (2003) examine the possibility of using neural network methods as tools for daily average PM_{10} concentration forecasting. Their results show that, compared with linear regression, root mean square error values are reduced by 8.2–9.4% and false alarm rate values by 7–13%. Other advantages of neural networks include superior learning, noise suppression, non-linear function and parallel computation abilities. One of the major problems with neural networks is that they are not designed with an explanatory capability, the so-called black box approach. In addition, successful implementation of a neural network-based system strongly depends on proper selection of the type of network structure and amount of training data, which are not always available. Recently, the applications of neural networks have become more popular (Chaloulakou et al., 2003; Chelani, Gajghate, & Hasan, 2002; Gardner & Dorling, 1998; Kukkonen et al., 2003; McKendry, 2002; Perez, Trier, & Reyes, 2000). The objective of Thomas and Jacko's work is to develop a reliable model for forecasting hourly $PM_{2.5}$ and CO concentrations at a microscale (adjacent to the expressway). A linear regression model and an neural network model are developed to forecast hourly $PM_{2.5}$ and CO concentrations using the year 2002 traffic, pollutant, and meteorological data. Both models had reasonable accuracy in predicting hourly $PM_{2.5}$ concentration. A major problem for these models is that they are developed specifically for the Borman Expressway, some modifications should be made in order for these models to be used for other expressways and roadways (Thomas & Jacko, 2007).

The model-driven statistical learning methods assume that both operational data and a mathematical model are available. State space model and Bayesian networks belong to this category. In contrast to the black box approaches such as neural networks, a mathematical functional mapping between the drifting parameters and the selected input variables can be established. Moreover, the model-driven statistical learning methods can be adapted to increase accuracy and to address subtle performance problems. Consequently, model-driven methods can significantly outperform data-driven approaches.

Cossentino, Raimondi, and Vitale (2001) employed Bayesian networks to model the temporal series of the particulate matter during the day and the influence that meteorological parameters have upon them. Typical inputs of the networks have been the pollutant concentration at a certain hour and the meteorological parameters at the further hours of the day. The output provided by the networks is the estimate of the probability of reaching a certain pollutant level in the various hours of the day. They concluded that the results are satisfactory and this approach can be profitably used to foresee critical episodes. As indicated by authors, the quality of the results depends on the number of the evidences that are supplied to the network.

Chelani et al. (2001) presented a state space model coupled with Kalman filter to forecast metal concentrations observed at Delhi, India. Wind speed is used as an external input. Compared to an autoregressive model, the state space model gives better predictions. The state space model also provides a way of incorporating model and measurement uncertainty into the forecasts (Harnandez, Martin, & Valero, 1992). However, the prediction may not be accurate for peak forecasting.

The HMM (hidden Markov-model) approach has become increasingly popular and quite effective in some applications such as speech processing and handwritten word recognition. There are two major reasons for this. First, the models have a rich mathematical structure and can form the solid theoretical foundation for a wide variety of applications. Second, the models have many successful applications in practice (Rabiner, 1989). An added benefit of employing HMMs is the ease of model interpretation in comparison with pure "black box" modeling methods such as artificial neural networks (Baruah & Chinnam, 2003). Through the detection of the adulterated words from a blacklist of words frequently used by spammers, Gordillo and Conde (2007) applied HMMs to classify spam mails. In order to forecast financial market behaviour, a fusion model by combining the HMM, ANN and Genetic Algorithms (GA) was proposed. Using ANN, the daily stock prices are transformed to independent sets of values that become input to HMM. The initial parameters of HMM are optimized by GA (Rafiul Hassan, Nath, & Kirley, 2007). However, there is an inherent limitation associated with the HMMs. This limitation is that the state duration of HMM follows an exponential distribution. In other words, HMM does not provide adequate representation of temporal structure for prediction problems. For example, a HMM does not provide adequate representation of the temporal structure of speech and segmental structure of the handwritten word. To overcome the limitations of HMMs in prediction, a novel prediction methodology is developed using a model-driven statistical learning method, called the hidden semi-Markov model (HSMM). A HSMM is constructed by adding a temporal component into the well-defined HMM structures (Guédon, 1999, 2003, 2005; Rabiner, 1989; Schmidler, 2000; Yu & Kobayashi, 2003a, 2003b, 2006; Aydin, Altunbasak, & Borodovsky, 2006; Dong & He, 2007a, 2007b). Instead of holding time distributions attached to states, Guédon (1999) proposed a hidden semi-Markov chain in which the time distributions are attached to transitions. Then, Guédon (2003) further extended previously proposed HSMMs in which the end of a sequence systematically coincides with the exit from a state, that is, the sequence length is not independent of the process. This article defines hidden semi-Markov chains with absorbing states and thus defines the likelihood of a state sequence generated by an underlying semi-Markov chain with a right censoring of the time spent in the last visited state. A new forward–backward algorithm is proposed with complexities that are quadratic in the worst case in time and linear in space, in terms of sequence length. This opens the way to the application of the full machinery of hidden semi-Markov chains to long sequences such as DNA sequences. In order to retain the flexibility of hidden semi-Markov chains for the modeling of short or medium size homogeneous zones along sequences and enable the modeling of long zones with Markovian states at the same time, Guédon (2005) investigated hybrid models that combine Markovian states with implicit geometric state occupancy distributions and semi-Markovian states with explicit state occupancy distributions. The Markovian nature of states is no longer restricted to absorbing states since non-absorbing Markovian states can now be defined. In the context of the application to gene finding, the incorporation of non-absorbing Markovian states is critical since the distributions of the lengths of the longest homogeneous zones are approximately geometric. The underlying assumption in the existing HMM and HSMM models is that there is at least one observation produced per state visit and that observations are exactly the outputs (or "emissions") of states. In some applications, these assumptions are too restrictive. Yu and Kobayashi (2003a) extended the ordinary HMM and HSMM to the model with missing data and multiple observation sequences. They proposed a new and computationally efficient forward–backward algorithm for HSMM with missing observations and multiple observation sequences. The required computational amount for the forward and backward variables is reduced to $O(D)$, where D is the maximum allowed duration in a state. Existing algorithms for estimating the model parameters of a HSMM usually require computations as large as $O((MD^2 + M^2)T)$, where M is the number of states, T is the period of observations used to estimate the model parameters. Because of such computational requirements, these algorithms are not practical to construct a HSMM model with large state space, large explicit state duration and a large amount of measurement data. Yu and Kobayashi (2003b) proposed a new forward–backward algorithm whose computational complexity is only $O((MD + M^2)T)$, a reduction by almost a factor of D when $D > M$. Since the joint probabilities associated with observation sequence often decay exponentially as the sequence length increases, the implementation of the forward–backward algorithms by programming in a real computer would suffer a severe underflow problem. Yu and Kobayashi (2006) redefined the forward–backward variables in terms of posterior probabilities to avoid possible underflows. The existing algorithms for HSMM are not practical for implementation in hardware because of the computational or logic complexity, thus, a forward recursion is used that is symmetric to the backward one and can reduce the number of logic gates required to implement on a field-programmable gate-array (FPGA) chip. HSMM has been widely used in protein secondary structure prediction. Schmidler (2000) believed the intra-segment residue independence and geometric length distributions implied by HMMs to be inappropriate for modeling protein secondary structure, they presented a Bayesian inference-based method for predicting the secondary structure of a protein from its amino acid sequence. Their model is structurally similar to the class of semi-Markov source models described in Rabiner (1989). Aydin et al. (2006) introduced an improved residue dependency model by considering the patterns of statistically significant amino acid correlation at structural segment borders. The results show that new dependency models and training methods bring further improvements to single-sequence protein secondary structure prediction. In this paper, we developed the hidden semi-Markov models by adding the temporal structures into the HMMs and used them for the prediction of high $PM_{2.5}$ concentration values.

The term ‘prediction’ as used in this paper means establishing the relationship between observed independent variables (predictors, such as meteorological variables) and an observed dependent variable (in this case concentration). When the predictors are forecast by some method, we can “forecast” or “predict” the concentrations.

2. Theoretical background

2.1. Elements of a hidden Markov model

A HMM represents stochastic sequences as Markov chains where the states are not directly observed, but are associated with a probability function. A HMM has the following elements (Rabiner, 1989):

(1) The state transition probability distribution $A = \{a_{ij}\}$ where

$$a_{ij} = P[s_{t+1} = j \mid s_t = i], \quad 1 \leqslant i,j \leqslant N.$$

(2) The observation probability distribution in state i, $B = \{b_i(k)\}$, where

$$b_i(k) = P[v_k \mid s_t = i] \quad 1 \leqslant i \leqslant N,\ 1 \leqslant k \leqslant M.$$

(3) The initial state distribution $\pi = \{\pi_i\}$ where

$$\pi_i = P[s_1 = i], \quad 1 \leqslant i \leqslant N.$$

(4) N, the number of states in the model, i.e., $1,2,\ldots,i,j,\ldots,N$. Although the states are hidden, there is often some physical signal attached to the states of the model. In this work, the state at time t is denoted as s_t.

(5) M, the number of distinct observations for each state. The observation symbols correspond to the physical output of the system being modeled. The individual observation symbols are denoted as $V = \{v_1, v_2, \ldots, v_M\}$.

It can be seen that a complete HMM λ requires the specifications of A, B, π, N and M. For convenience, a compact notation is often used in the literature to indicate the complete parameter set of the model:

$$\lambda = (\pi, A, B).$$

2.2. Durational measure of standard HMMs

The durational behaviour of a HMM is usually characterized by a durational state occupancy distribution (pdf) $P(d)$. For a single state i, the value $P(d)$ is the probability of the event of staying in i for exactly d time units. This event is in fact the joint event of taking the self-loop for $(d-1)$ times and taking the out-going transition (with probability $1 - a_{ii}$) just once. Given the Markovian assumption, and from probability theory, $P(d)$ is simply the product of all the d probabilities:

$$P_i(d) = a_{ii}^{d-1}(1 - a_{ii}). \tag{1}$$

Here, $P_i(d)$ denotes the probability of staying in state i for exactly d time steps, and a_{ii} is the self-loop probability of state i.

It can be seen that this is a geometrically decaying function of d. It has been argued (Russell & Moore, 1985) that this is a source of inaccurate duration modeling with the HMMs since most real-life applications will not obey this function.

2.3. Hidden semi-Markov models (HSMM)

2.3.1. Model description

We begin with a standard discrete-time finite-state Markov model. There are N states in the model: $1,2,\ldots,N$. The state at time t is denoted s_t, where $t = 1,2,\ldots$ is the time index. Given the state s_t at time t, the state s_{t+1} at time $t+1$ is given by the conditional probabilities $P(s_{t+1}|s_t)$, and conditionally independent of earlier states $s_{t-1}, s_{t-2}, \ldots, s_1$. The state transition matrix A specifies the conditional probabilities $A_{ij} = P(s_{t+1} = j | s_t = i)$, for $1 \leqslant i,j \leqslant N$.

Unlike a state in a HMM, a state in a HSMM generates a segment of observations $o_{t1}, \ldots, o_{t2}$ (as opposed to a single observation o_t in the HMM). Let s_t be the hidden state at time t and O be the observation sequence. Characterization of a HSMM is through its parameters. The parameters for a HSMM are defined as: the initial state distribution (denoted by π), the transition model (denoted by A), state duration distribution (denoted by D), and the observation model (denoted by B). Thus, a HSMM can be written as $\lambda = (\pi, A, D, B)$.

2.3.2. State transitions

In this paper, each segment consists of several single states. In the segmental HSMM, there are N states (i.e., $1,2,\ldots,i,j,\ldots,N$), and the transitions between the states are according to the transition matrix A, i.e., $P(i \to j) = a_{ij}$. Suppose a state sequence s has R segments, and let q_r be the time index of the end-point of the rth segment, for $1 \leqslant r \leqslant R$. The segments are as follows:

Time units	$\underline{1,\ldots,q_1}$	$\underline{q_1+1,\ldots,q_2}$	...	$\underline{q_{R-1}+1,\ldots,q_R}$
Observations	$\underline{o_1,\ldots,o_{q_1}}$	$\underline{o_{q_1+1},\ldots,o_{q_2}}$	...	$\underline{o_{q_{R-1}+1},\ldots,o_{q_R}}$
States	$\underline{s_1,\ldots,s_{q_1}}$	$\underline{s_{q_1+1},\ldots,s_{q_2}}$	...	$\underline{s_{q_{R-1}+1},\ldots,s_{q_R}}$
Segments	*1*	*2*	...	*R*

The data points in the rth segment are $o_{(q_{r-1},q_r]} = o_{q_{r-1}+1}\cdots o_{q_r}$, and they have the same state label:

$$s_{q_{r-1}+1} = s_{q_{r-1}+2} = \cdots = s_{q_r}.$$

Let $i = s_{qr}$ be the state label of segment r. Although the segment-state transition $s_{q_{r-1}} \rightarrow s_{q_r}$ is Markov

$$P(s_{q_r} = j \mid s_{q_{r-1}} = i) = a_{ij}$$

the single state transition $s_{t-1} \rightarrow s_t$ is usually not Markov, unless the state distribution is geometric, as implied by a Markov model. This is the reason why the model is called "semi-Markov" (Ferguson, 1980).

2.3.3. State duration distribution

For a segment, the number of data points in the segment $d_i = q_r - q_{r-1}$, i.e., the duration of state i, is according to a state duration distribution $P(d|i)$. In other words, $P(d|i)$ is the probability of duration d in state i.

3. Methods

3.1. HSMM-based framework for PM$_{2.5}$ concentration prediction

The basic idea of using HSMMs to predict the PM$_{2.5}$ concentration levels is shown in Fig. 1. As shown in Fig. 1, past meteorological measurements (observed up to current time t) such as temperature, wind speed, wind direction, and etc. along with the observed PM$_{2.5}$ concentration levels (for example, consider PM$_{2.5}$ concentration level 95%) are used as the training data sets. Through parameter estimation by maximizing the probability of the observation sequence given the model, a trained HSMM can be obtained for each PM$_{2.5}$ concentration level (e.g., $HSMM_{95\%}$). The trained HSMMs will be used for PM$_{2.5}$ prediction.

An essential part of applying HSMM technology to any problem is to decide what the appropriate observations are. In general, the observations could be the raw data or some function or transformation of the data. A transformation of the data is preferable when the result allows for the diminution of the quantity of data which needs to be processed without losing the ability to effectively monitor the HSMM process (Bunks, Mccarthy, & Tarik, 2000).

Each state of the Markov chain associated to a HSMM must have a state process model, which implicitly or explicitly specifies a state occupancy distribution. In the literature, the Gaussian distribution is often used although multi-modal distributions are also used by taking mixtures of Gaussians (Liporace, 1982). Other common choices include mixtures of autoregressive and autoregressive moving-average models. In this paper, a simple multi-dimensional Gaussian distribution is used.

3.2. Inference procedures

To facilitate the computation in the proposed HSMM-based PM$_{2.5}$ concentration prediction framework, in the following, new forward–backward variables is defined and modified forward–backward algorithm is developed. To implement the inference procedures, a *forward variable* $\alpha_t(i)$ is defined as the probability of generating $o_1\ o_2\ \cdots o_t$ and ending in state i:

$$\alpha_t(i) = P(o_1 o_2 \cdots o_t, i \text{ ends at } t \mid \lambda). \tag{2}$$

Assume that $\alpha_{t-d}(i)$ has been determined, the duration in state j is d. Then $o_1 o_2 \cdots o_t$ is generated by a state sequence ending at state j if and only if the following conditions are satisfied: (1) $o_1 o_2 \cdots o_{t-d}$ is generated ending in state i (i.e., $\alpha_{t-d}(i)$); (2) the transition from i to j is chosen (i.e., a_{ij}); (3) the duration in state j is chosen (i.e., $P(d|j)$); and (4) $o_{t-d+1}o_{t-d+2}\ldots o_t$ is emitted in state j. Summing over all states s and all possible state durations d, we get the recurrence relation (Rabiner, 1989):

$$\alpha_t(j) = \sum_{i=1}^{N}\sum_{d=1}^{D}\alpha_{t-d}(i)a_{ij}P(d \mid j)\prod_{s=t-d+1}^{t} b_j(o_s) \tag{3}$$

where D is the maximum duration within any state.

It can be seen that the probability of O given model λ can be written as:

$$P(O \mid \lambda) = \sum_{i=1}^{N}\alpha_T(i). \tag{4}$$

This is the case since, by definition,

$$\alpha_T(i) = P(o_1 o_2 \cdots o_T, q_T = i \mid \lambda) \tag{5}$$

and hence $P(O|\lambda)$ is just the sum of the $\alpha_T(i)$'s.

3.3. Modified forward–backward algorithm for HSMMs

Similar to the forward variable, a *backward variable* $\beta_t(i)$ is defined as the probability of generating $o_1 o_2 \cdots o_t$ and ending in state i:

$$\beta_t(i) = P(o_{t+1}o_{t+2}\cdots o_T \mid i \text{ begins at } t, \lambda).$$

Assume that backward variable $\beta_{t+d}(j)$ has been determined and the duration in state j is d, summing over all states s and all possible state durations d, we have the recurrence relation (see Fig. 2):

$$\beta_t(i) = \sum_{j=1}^{N}\sum_{d=1}^{D}a_{ij}P(d \mid j)\prod_{s=t+1}^{t+d} b_j(O_s)\beta_{t+d}(j). \tag{6}$$

In order to obtain re-estimation formulae for all variables of the HSMM, three more segment-featured forward–backward variables are defined.

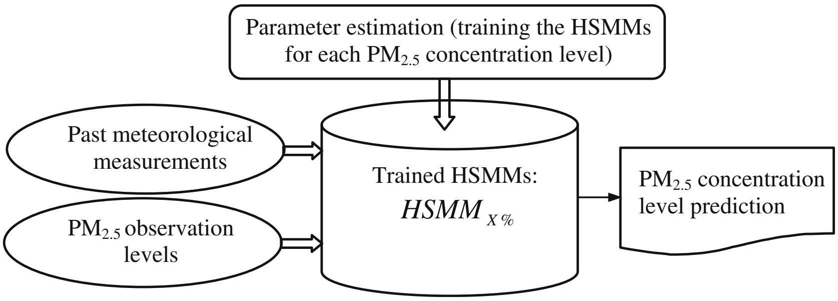
Parameter estimation (training the HSMMs for each PM$_{2.5}$ concentration level)
Past meteorological measurements
PM$_{2.5}$ observation levels
Trained HSMMs: $HSMM_{X\%}$
PM$_{2.5}$ concentration level prediction

Fig. 1. HSMM-based framework for PM$_{2.5}$ concentration prediction.

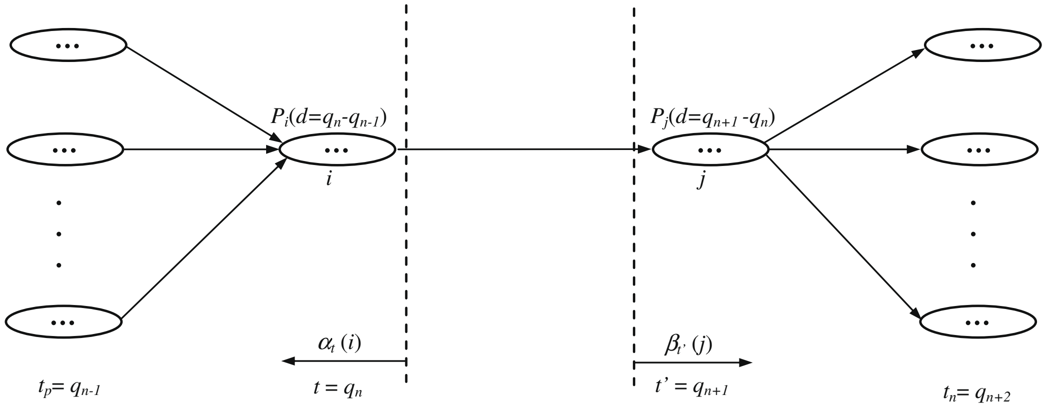

Fig. 2. Illustration of the sequence of operations required for computation of the joint event that the process is in level j at time $t + d$ and level i at time t.

$$\alpha_{t,t'}(i,j) = P(o_1o_2\cdots o_{t'}, t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j \mid \lambda), \tag{7}$$

i.e. the probability of the partial observation sequence, $o_1o_2\cdots o_{t'}$, and state i at time t and state j at time $t'(t' = t + d)$.

$$\phi_{t,t'}(i,j) = \sum_{d=1}^{D}[P(d = t' - t \mid j)\cdot P(O_{t+1}^{t'} \mid t = q_n, s_t = i,\\ t' = q_{n+1}, s_{t'} = j, \lambda)], \tag{8}$$

i.e. the mean value of the probabilities of the system being in state i for d ($d = 1,\ldots,D$) time units and then moving to the next state j. Here, $O_{t+1}^{t'} = o_{t+1}o_{t+2}\cdots o_{t'}$.

$$\xi_{t,t'}(i,j) = P(t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j \mid O_1^T, \lambda), \tag{9}$$

i.e. the probability of the system being in state i for d ($t' = t + d$) time units and then moving to the next state j, given the observation sequence, $o_1o_2\cdots o_T$. Here, $O_1^T = o_1o_2\cdots o_T$.

$\alpha_{t,t'}(i,j)$ can be described, in terms of $\phi_{t,t'}(i,j)$, as follows:

$$\begin{aligned}\alpha_{t,t'}(i,j) &= P(o_1o_2\cdots o_to_{t+1}\cdots o_{t'}, t = q_n, s_t = i, t' = q_{n+1},\\ &\quad s_{t'} = j \mid \lambda) = P(o_1o_2\cdots o_t, t = q_n, s_t = i \mid \lambda)\\ &\qquad \times P(O_{t+1}^{t'}, t' = q_{n+1}, s_{t'} = j \mid O_1^t, t = q_n, s_t = i, \lambda)\\ &= \alpha_t(i)P(O_{t+1}^{t'}, t' = q_{n+1}, s_{t'} = j \mid t = q_n, s_t = i, \lambda)\\ &= \alpha_t(i)P(t' = q_{n+1}, s_{t'} = j \mid t = q_n, s_t = i, \lambda)\\ &\quad \times P(O_{t+1}^{t'} \mid t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j, \lambda) = \alpha_t(i)a_{ij}\\ &\quad \times \sum_{d=1}^{D}P(d = t' - t \mid j)P(O_{t+1}^{t'} \mid t = q_n, s_t = i, t' = q_{n+1}, s_{t'} = j, \lambda)\\ &= \alpha_t(i)a_{ij}\phi_{t,t'}(i,j)\end{aligned} \tag{10}$$

The relationship between $\alpha_t(i)$ and $\alpha_{t,t'}(i,j)$ is given in the following:

$$\begin{aligned}\alpha_{t'}(j) &= P(o_1o_2\cdots o_{t'}, t' = q_{n+1}, s_{t'} = j \mid \lambda)\\ &= \sum_{i=1}^{N}\sum_{d=1}^{D}P(d = t' - t \mid j)P(o_1o_2\cdots o_to_{t+1}\cdots o_{t'}, t = q_n, s_t = i,\\ &\qquad t' = q_{n+1}, s_{t'} = j \mid \lambda) = \sum_{i=1}^{N}\sum_{d=1}^{D}P(d = t' - t \mid j)\alpha_{t,t'}(i,j)\end{aligned} \tag{11}$$

From the definitions of the forward–backward variables, we can derive $\xi_{t,t'}(i,j)$ as follows:

$$\xi_{t,t'}(i,j) = \frac{\sum_{d=1}^{D}\alpha_t(i)a_{ij}\phi_{t,t'}(i,j)\beta_{t'}(j)}{\beta_0(i = \text{“}START\text{”})} \tag{12}$$

The forward–backward algorithm computes the following probabilities:

Forward pass: the forward pass of the algorithm computes $\alpha_t(i), \alpha_{t,t'}(i,j)$ and $\phi_{t,t'}(i,j)$.

Step 1: Initialization ($t = 0$)

$$\alpha_{t=0}(i) = \begin{cases}1, & \text{if } i = \text{“}START\text{”},\\ 0, & \text{otherwise}.\end{cases}$$

Step 2: Forward recursion ($t > 0$). For $t = 1,2,\ldots,T; 1 \leqslant i,j \leqslant N$; and $1 \leqslant d \leqslant D$.

$$\phi_{t,t'}(i,j) = \sum_{d=1}^{D}[P(d = t' - t \mid j)\cdot P(O_{t+1}^{t'} \mid t = q_n, s_t = i,\\ t' = q_{n+1}, s_{t'} = j, \lambda)],$$

$$\alpha_{t,t'}(i,j) = \alpha_t(i)a_{ij}\phi_{t,t'}(i,j),$$

$$\alpha_{t'}(j) = \sum_{i=1}^{N}\sum_{d=1}^{D}P(d = t' - t \mid j)\alpha_{t,t'}(i,j).$$

Backward pass: The backward pass computes $\beta_t(i)$ and $\xi_{t,t'}(i,j)$.

Step 1: Initialization ($t = T$ and $1 \leqslant i, j \leqslant N$)

$$\beta_T(i) = 1.$$

Step 2: Backward recursion ($t < T$). For $t = 1,2,\ldots,T; 1 \leqslant i, j \leqslant N$; and $1 \leqslant d \leqslant D$.

$$\begin{aligned}\beta_t(i) &= \sum_{j=1}^{N}\sum_{d=1}^{D}a_{ij}P(d \mid j)\prod_{s=t+1}^{t+d}b_j(O_s)\beta_{t+d}(j)\\ &= \sum_{j=1}^{N}a_{ij}\phi_{t,t'}(i,j)\beta_{t'}(j)\end{aligned} \tag{13}$$

$$\xi_{t,t'}(i,j) = \sum_{d=1}^{D}\alpha_t(i)a_{ij}\phi_{t,t'}(i,j)\beta_{t'}(j)/\beta_0(i = \text{“}START\text{”})$$

Let D_i be the maximum duration for state i. The total computational complexity for the forward–backward algorithm is $O(N^2LT)$, where $L = \sum_{i=1}^{N}D_i$.

3.4. Parameter re-estimation for HSMM-based framework

In this paper, we address a more general case in which the state labeling is not available for the training data. Therefore, re-estimation equations are used in the following:

3.4.1. Initial state distribution

The re-estimation formula for initial state distribution is the probability that state i was the first state, given O

$$\bar{\pi}_i = \frac{\pi_i\left[\sum_{d=1}^{D}\beta_d(i)P(d \mid i)b_j(O_1^d)\right]}{P(O \mid \lambda)}. \tag{14}$$

3.4.2. State transition probabilities

The re-estimation formula of state transition probabilities is the ratio of the expected number of transitions from state i to state j, to the expected number of transitions from state i

$$\begin{aligned}\bar{a}_{ij} &= \frac{\sum_{t=1}^{T}\alpha_t(i)a_{ij}\sum_{d=1}^{D}P(d \mid j)b_j(O_{t+1}^{t'})\beta_{t'}(j)}{\sum_{t=1}^{T}\sum_{j=1}^{N}\alpha_t(i)a_{ij}\sum_{d=1}^{D}P(d \mid j)b_j(O_{t+1}^{t'})\beta_{t'}(j)}\\ &= \frac{\sum_{t=1}^{T}\xi_{t,t'}(i,j)}{\sum_{t=1}^{T}\sum_{j=1}^{N}\xi_{t,t'}(i,j)}.\end{aligned} \tag{15}$$

3.4.3. Observation distributions

The re-estimation formula for segmental observation distributions is the expected number of times that observation $o_t = v_k$ occurred in state i, normalized by the expected number of times that any observation occurred in state i. Since $\alpha_t(i)$ accounts for the partial observation sequence $o_1o_2 \ldots o_t$ and state i at t, while $\beta_t(i)$ accounts for the partial observation sequence $o_to_{t+1} \ldots o_T$, given state i at t. Therefore, the re-estimation of segmental observation distributions can be calculated as follows:

$$\bar{b}_i(k) = \frac{\sum^{T}_{\substack{t=1\\ \text{s.t. } O_t=k}}\left[\sum_{\tau<t}\left(\sum_{j=1}^{N}\alpha_\tau(j)a_{ji}\right)\cdot\left(\sum_{d=1}^{D}\beta_{\tau+d}(i)P(d \mid i)\prod_{s=t+1}^{t+d}b_i(O_s)\right)-\sum_{\tau<t}\alpha_\tau(i)\cdot\beta_\tau(i)\right]}{\sum_{k=1}^{M}\sum^{T}_{\substack{t=1\\ \text{s.t. } O_t=k}}\left[\sum_{\tau<t}\left(\sum_{j=1}^{N}\alpha_\tau(j)a_{ji}\right)\cdot\left(\sum_{d=1}^{D}\beta_{\tau+d}(i)P(d \mid i)\prod_{s=t+1}^{t+d}b_i(O_s)\right)-\sum_{\tau<t}\alpha_\tau(i)\cdot\beta_\tau(i)\right]} \tag{16}$$

3.4.4. State duration probability distributions

The re-estimation problem is more difficult for HSMMs than for the standard HMM. One proposal to alleviate some of these problems is to use a parametric state duration density instead of the non-parametric duration density. A parametric model requires far less training and generalizes better results. In this study, Gaussian distribution is adopted. Gaussian distributions have many convenient properties, so random variates with unknown distributions are often assumed to be Gaussian. It is often a good approximation due to a surprising result known as the central limit theorem. This theorem states that the mean of any set of variates with any distribution having a finite mean and variance tends to the Gaussian distribution. In this paper, state duration probabilities are estimated directly from the training data.

The mean $\mu(j)$ and the variance $\sigma^2(j)$ of duration of state i are determined by

$$\mu(j) = \frac{\sum_{d=1}^{D}\left\{\sum_{t=1}^{T-d}\left(\sum_{t=1}^{N}\alpha_t(i)a_{ij}\right)P(d \mid j)b_j(O_{t+1}^{t+d})\beta_{t+d}(j)\right\}d}{\sum_{d=1}^{D}\sum_{t=1}^{T-d}\left(\sum_{t=1}^{N}\alpha_t(i)a_{ij}\right)P(d \mid j)b_j(O_{t+1}^{t+d})\beta_{t+d}(j)} \tag{17}$$

$$\sigma^2(j) = \frac{\sum_{d=1}^{D}\left\{\sum_{t=1}^{T-d}\left(\sum_{t=1}^{N}\alpha_t(i)a_{ij}\right)P(d \mid j)b_j(O_{t+1}^{t+d})\beta_{t+d}(j)\right\}d^2}{\sum_{d=1}^{D}\sum_{t=1}^{T-d}\left(\sum_{t=1}^{N}\alpha_t(i)a_{ij}\right)P(d \mid j)b_j(O_{t+1}^{t+d})\beta_{t+d}(j)} - \mu^2(j) \tag{18}$$

In this paper, vector quantization (VQ) is used to discretize the continuous data. And k-means algorithm is used to train a codebook for vector quantization. The basic procedure is as follows: arbitrarily choose m vectors as the initial centroids (the mean of each cluster). The distance of each vector from these centroids is found and each vector is associated with a cluster. The mean of the vectors is determined. This procedure goes on until the improvement is not substantial.

3.5. PM$_{2.5}$ concentration prediction using HSMM

In real applications, a basic problems associated with HSMMs is classification (i.e., PM concentration level prediction in this paper). That is, given the observation sequence $O = o_1o_2 \ldots o_T$, and a HSMM λ, what is the probability of the observation sequence given the model, i.e., $P(O|\lambda)$. Therefore, for each level of PM concentration, there is a corresponding HSMM. For example, in this paper, we consider two levels of PM concentration (i.e., L_1 and L_2). Therefore, two corresponding HSMMs (i.e. HSMM$_{L_1}$ and HSMM$_{L_2}$) should be trained for each concentration level, respectively. In general, for prediction, the goal is to develop trained HSMMs to recognize N different PM$_{2.5}$ concentration levels of a studied area for a given observation sequence of meteorological measurements. That is, the task is to develop HSMM-based prediction models for classifying the levels of PM$_{2.5}$ concentration. Given these N groups of observation sequences, N different HSMMs (i.e., $HSMM_1$, $HSMM_2, \ldots, HSMM_N$) are modeled (or trained) for characterization of each group, which corresponds to a PM$_{2.5}$ concentration level. Once these HSMMs are trained, the next step is to calculate the probability of the observation sequence given these models, i.e., $P(O|HSMM_1)$, $P(O|HSMM_2), \ldots, P(O|HSMM_N)$. In other words, the procedure for classification of PM$_{2.5}$ concentration levels given an observation sequence of meteorological measurements is: each of the N trained HSMMs is presented with the same sequence. According to the value of highest log-likelihood (i.e. the probability of the observation sequence given a HSMM model), the sequence can be classified (see Fig. 3, in which P_1, P_2 and P_N represent the probabilities of the observation sequence given the HSMM models, i.e. $P_1 = P(O|HSMM_1)$, $P_2 = P(O|HSMM_2), \ldots, P_N = P(O|HSMM_N)$).

In this paper, only high concentration values are predicted, therefore, each group contains different PM levels (i.e. different PM levels are aggregated into a group) and the groups are disjoint.

4. Experimental data

The input data for the HSMM presented here consists of PM$_{2.5}$ and meteorological data for the 2000–2001 period. PM$_{2.5}$ measurements were obtained from the EPA Air Quality System for all monitoring sites in Cook County, Illinois (the Chicago metropolitan area) as shown in Fig. 4. The daily observations from the 12 stations were averaged together to provide one representative measurement for the area. Meteorological observations were obtained from the National Weather Service for the monitor site at O'Hare Airport. All meteorological data were 1-hour average values, which were then used to determine mean and maximum values for each day. The variables used as inputs to the HSMM for PM$_{2.5}$ prediction are listed in Table 1.

4.1. Concentration data

The 24-h average hourly concentrations of PM$_{2.5}$ were available for the stations at O'Hare airport in Chicago during the selected years.

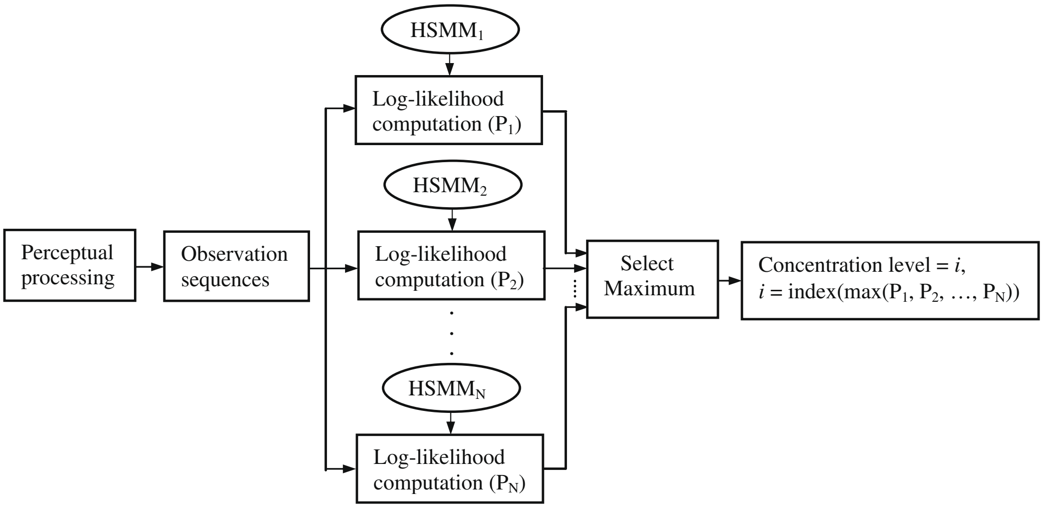
HSMM$_1$
Log-likelihood computation (P$_1$)
HSMM$_2$
Log-likelihood computation (P$_2$)
HSMM$_N$
Log-likelihood computation (P$_N$)
Perceptual processing
Observation sequences
Select Maximum
Concentration level = i,
i = index(max(P$_1$, P$_2$, …, P$_N$))

Fig. 3. Block diagram for a HSMM-based classifier. (The perceptual processing phase produces the processed input data and already trained HSMMs are used for indexing them into a particular class.)

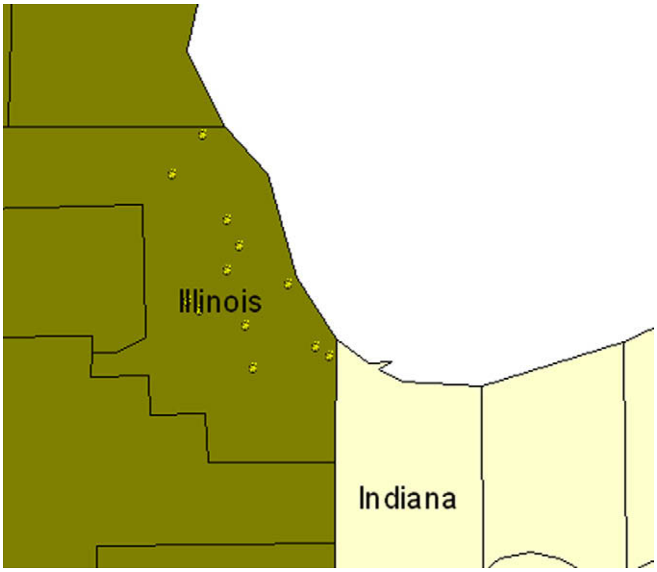

Fig. 4. EPA air quality system for all monitoring sites in Cook County, Illinois (the Chicago metropolitan area).

Table 1
The list of variables used as inputs of HSMMs for $PM_{2.5}$ prediction. (The output time $T+24$ is the time, for which the forecast applies. The input data regarding the concentrations corresponds to the time T.)

Meteorological variables	Unit	Output time
Temperature (mean)	K	$T+24$
Temperature (max)	K	$T+24$
Pressure (mean)	kPa	$T+24$
Pressure (max)	kPa	$T+24$
Cloudiness	0-8	$T+24$
Wind speed (mean)	m/s	$T+24$
Wind speed (max)	m/s	$T+24$
Solar radiation	W/m^2	$T+24$
Dewpoint (mean)	K	$T+24$
Dewpoint (max)	K	$T+24$
Humidity (mean)	%	$T+24$
Humidity (max)	%	$T+24$

4.2. Meteorological data

We have selected the pre-processed meteorological data computed for the location of O'Hare airport in Chicago to be used in this study, as it is best representative for the whole of the urban area. The pre-processed meteorological data is based on the normalization of the data from the stations. A single time series of meteorological data was assumed to be spatially representative for the whole of the study area.

The variables used as inputs of the HSMMs for $PM_{2.5}$ prediction are listed in Table 1.

5. Results

Because accurately predicting high $PM_{2.5}$ concentrations is of most value from a public health standpoint, this case study focused on predicting those high concentrations. Concentrations above 40 μg/m^3 are described by EPA as 'unhealthy for sensitive groups' and as such are of most interest. Forecasts of $PM_{2.5}$ concentrations are routinely made by states so that these sensitive groups can take actions to reduce their exposure. In this study, $PM_{2.5}$ concentrations were classified into two categories ($PM_{\min}$ denotes the minimum $PM_{2.5}$ value and $PM_{\max}$ denotes the maximum $PM_{2.5}$ value).

(1) Level 1: ($PM_{\min}$, 40];
(2) Level 2: (40, $PM_{\max}$] (see Table 2).

Since, there are two $PM_{2.5}$ concentration levels, two HSMMs will be trained for them: $HSMM_{L_1}$ and $HSMM_{L_2}$. If we use $HSMM_{L_1}$ to test the two $PM_{2.5}$ concentration levels, the following results will be obtained (see Table 3). If $HSMM_{L_2}$ is used to test the two $PM_{2.5}$ concentration levels, the corresponding results are provided in Table 4.

Table 2
$PM_{2.5}$ concentration levels at the studied area.

Level	PM value range	Percentage among total number of data points (%)
Level 1	$2.8 < PM \leqslant 40.0$	98.35
Level 2	$40.0 < PM \leqslant 48.44$	1.65

Table 3
Results of $PM_{2.5}$ concentration prediction using HSMM$_{L_1}$.

Testing on concentration level 1 using HSMM$_{L1}$ (log-likelihood values)	Testing on concentration level 2 using HSMM$_{L1}$ (log likelihood values)	Prediction Results (√: correct; ×: wrong)
67.6556	72.8767	√
67.3849	74.9864	√
67.4442	124.6524	√
68.3434	78.2085	√
67.3528	73.3247	√
67.2945	75.4933	√
66.4298	72.4356	√
68.4943	71.5884	√
68.5344	79.5543	√
68.5740	79.1543	√
68.0535	129.4341	√
67.5335	128.4431	√

Table 4
Results of $PM_{2.5}$ concentration prediction using HSMM$_{L_2}$.

Testing on concentration level 1 using HSMM$_{L2}$ (log-likelihood values)	Testing on concentration level 2 using HSMM$_{L2}$ (log-likelihood values)	Prediction Results (√: correct; ×: wrong)
−68.1449	2.5753	√
−22.1231	1.2442	√
−38.5335	1.9586	√
−25.8644	1.2345	√
−18.0245	3.0485	√
−95.5927	5.5943	√
−58.8720	4.7424	√
−26.5294	3.9844	√
−26.0284	5.5884	√
−120.3905	6.8709	√
−106.4028	4.4858	√
−99.5940	3.8575	√

Table 5
Results of $PM_{2.5}$ concentration prediction using HSMMs.

PM level	PM range	Input parameters	Prediction accuracy (%)
Level 1	$2.80 < PM \leqslant 40.00$	Solar radiation, cloudiness, temperature, pressure, humidity, wind speed, dewpoint	100
Level 2	$40.0 < PM \leqslant 48.44$		100

In Tables 3 and 4, the log-likelihood value is defined as logarithm of the probability that the HSMM model generated the observation sequence. The larger the log-likelihood score, the larger probability of the observation sequence given the model. Take the first row of Table 3 as an example, the trained HSMM$_{L_1}$ gives a log-likelihood value of −68.7966 for a data point from concentration level 1, and a log-likelihood value of −74.6634 for a data point from concentration level 2. So this result shows that, for HSMM$_{L_1}$, the concentration level 1 has larger probability. In addition, the magnitude of the log-likelihood value is also important. The larger difference between two log-likelihood values, the more powerful classification capability.

The results of the study are summarized in Table 5.

As we can see from Table 5, the HSMMs provide accurate predictions of $PM_{2.5}$ concentration levels for the next 24 h.

6. Conclusions

This investigation presents a HSMM-based framework and methodology for the prediction of $PM_{2.5}$ concentration levels. A HMM is a probabilistic function of a Markov chain and strictly controlled by the property of the Markov chain. This property says that the current state of the system depends only on the previous one. Therefore, a HMM normally has a short-term memory of the past history and this short-term memory of a HMM limits its power in prediction of future events. The proposed HSMM-based framework and methodology overcomes this problem by adding a temporal component to the HMM structures. In the HSMM case, the conditional independence between the past and the future is only ensured when the process moves from one state to another distinct state (this property holds at each time step in the classical Markovian case). Since a HSMM is equipped with a temporal structure, it can be used to predict pollutant concentrations in air monitoring applications.

The evaluation of our proposed approach is carried out in a real world application: the prediction of $PM_{2.5}$ concentrations in Chicago. The results show that the classification accuracy of $PM_{2.5}$ concentrations is indeed very promising and HSMMs are able to provide accurate predictions of extreme concentration levels 24 h in advance.

References

Aydin, Z., Altunbasak, Y., & Borodovsky, M. (2006). Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. *BMC Bioinformatics, 7*, 178.

Baker, K. (2004). $PM_{2.5}$ model performance statistics. In *Presented to EPA model performance meeting*, Research Triangle Park, NC.

Baruah, P., & Chinnam, R. B. (2003). HMMs for diagnostics and prognostics in machining processes. In: *Proc. of the 57th society for machine failure prevention technology conference. Virginia Beach, VA, April 14–18).*

Bunks, C., Mccarthy, D., & Tarik, A. (2000). Condition based maintenance of machines using hidden Markov models. *Mechanical Systems and Signal Processing, 14*(4), 597–612.

Burrows, W. R., Montpetit, J., Pudykiewicz, J. (1997). A non-linear regression procedure to produce statistical air-quality forecast models. In *Proceedings of the Air & Waste Management Association's annual meeting & exhibition, 97-TP2B.04*.

Chaloulakou, A., Grivas, G., & Spyrellis, N. (2003). Neural network and multiple regression models for PM_{10} prediction in Athens: A comparative assessment. *Journal of the Air and Waste Management Association, 53*(10), 1183–1190.

Chelani, A. B., Gajghate, D. G., Tamhane, S. M., & Hasan, M. Z. (2001). Statistical modeling of ambient air pollutants in Delhi. *Water, Air, and Soil Pollution, 132*(3–4), 315–331.

Chelani, A. B., Gajghate, D. G., & Hasan, M. Z. (2002). Prediction of ambient PM_{10} and toxic metals using artificial neural networks. *Journal of the Air and Waste Management Association, 52*(7), 805–810.

Cossentino, M., Raimondi, F. M., Vitale, M. C. (2001). Bayesian models of the PM_{10} atmospheric urban pollution. In *Ninth international conference on modeling, monitoring and management of air pollution: Air Pollution IX, September 12–14 2001, Ancona, Italy* (pp. 143–152).

Dockery & Pope (1994). Acute respiratory effects of particulate air pollution. *Annual Review of Public Health, 15*, 107–132.

Dong, M., & He, D. (2007a). Hidden semi-Markov model based methodology for multi-sensor equipment health diagnosis and prognosis. *European Journal of Operational Research, 178*(3), 858–878.

Dong, M., & He, D. (2007b). A segmental hidden semi-Markov model (HSMM)-based diagnostics and prognostics framework and methodology. *Mechanical Systems and Signal Processing, 21*(5), 2248–2266.

EPA (2002). *Third external review draft. Air quality criteria for particulate matter* (Vol. 1). Office of Research and Development. EPA/600/P-99/002aC.

EPA (2003). *Guidelines for developing an air quality forecasting program.* Environmental Protection Agency Report. EPA-456/R-03-002.

Estivill-Castro, V., & Houle, M. E. (2001). Robust distance-based clustering with applications to spatial data mining. *Algorithmica – Special Issue on Algorithms for Geographic Information, 30*(2), 216–242.

Ferguson, J. D. (1980). Variable duration models for speech. In *Proc. Symposium on the application of hidden Markov models to text and speech* (pp. 143–179).

Fraser, D., & Yap, D. (1997). Daily and continuous monitoring of PM_{10}/$PM_{2.5}$ in Ontario, Canada. In: *Proceedings of the Air & Waste Management Association's annual meeting & exhibition, 97-WP96.08*.

Fuller, G. W., Carslaw, D. C., & Lodge, H. W. (2002). An empirical approach for the prediction of daily mean PM_{10} concentrations. *Atmospheric Environment, 36*(9), 1431–1441.

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) – A review of applications in the atmospheric sciences. *Atmospheric Environment, 32*, 2627–2636.

Gordillo, J., & Conde, E. (2007). An HMM for detecting spam mail. *Expert Systems with Applications, 33*, 667–682.

Guédon, Y. (1999). Computational methods for discrete hidden semi-Markov chains. *Applied Stochastic Models in Business and Industry, 15*(3), 195–224.

Guédon, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics, 12*(3), 604–639.

Guédon, Y. (2005). Hidden hybrid Markov/semi-Markov chains. *Computational Statistics & Data Analysis, 49*(3), 663–688.

Harnandez, E., Martin, F., & Valero, F. (1992). Statistical forecast models for daily air particulate iron and lead concentrations for Madrid, Spain. *Atmospheric Environment, 26B*(1), 107–116.

Katsouyanni et al. (1997). Short term effects of ambient sulfur dioxide and particulate matter on mortality in 12 European cities: Results from time series data from the APHEA project. *British Medical Journal, 314*, 1658–1663.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., et al. (2003). Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modeling system and measurements in central Helsinki. *Atmospheric Environment, 37*(32), 4539–4550.

Levy et al. (2000). Estimating the mortality impacts of particulate matter: What can be learned from between-study variability? *Environmental Health Perspectives, 108*(2), 109–117.

Liporace, L. A. (1982). Maximum likelihood estimation for multivariate observations of markov sources. *IEEE Transactions on Information Theory, IT-28*, 729–734.

Lu, H. C., & Fang, G. C. (2003). Predicting the exceedances of a critical PM_{10} concentration – A case study in Taiwan. *Atmospheric Environment, 37*(25), 3491–3499.

McKendry, I. G. (2002). Evaluation of artificial neural networks for fine particulate pollution (PM_{10} and $PM_{2.5}$) forecasting. *Journal of the Air and Waste Management Association, 52*(9), 1096–1101.

Ojha, S., Coutinho, J., & Kumar, A. (2002). Developing systems to forecast ozone and particulate matter levels. *Environmental Progress, 21*(2), J7–J12.

Perez, P., Trier, A., & Reyes, J. (2000). Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile. *Atmospheric Environment, 34*(8), 1189–1196.

Pope Thurston & Krewski (2002). Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *Journal of the American Medical Association, 287*(9), 1132–1141.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Rafiul Hassan, M., Nath, B., & Kirley, M. (2007). A fusion model of HMM, ANN and GA for stock market forecasting. *Expert Systems with Applications, 33*, 171–180.

Raimondi, F. M., Rando, F., Vitale, M. C., Calcara, A. M. V., 1997. Short-time fuzzy DAP predictor for air pollution due to vehicular traffic. In *Proceedings of the first international conference on measurements and modeling in environmental pollution (MMEP 97)*. Computational Mechanics Publications

Raimondi, F. M., Rando, F., Vitale, M. C., Calcara, A. M. V., 1997. A Short-term air pollution predictor for urban areas with complex orography: Application to the town of Palermo. In Proceedings of the first international conference on measurements and modeling in environmental pollution (MMEP 97). Southampton and Boston: Computational Mechanics Publications.

Rizzo, M., Scheff, P., & Ramakrishnan, V. (2002). Defining the photochemical contribution to particulate matter in urban areas using time-series analysis. *Journal of the Air and Waste Management Association, 52*(5), 593–605.

Russell, M. J., & Moore, R. K. (1985). Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition. In: *Proceedings of ICASSP'85. Tampa, Florida* (pp. 5–8).

Sanchez, M. L., Pascual, D., Ramos, C., & Perez, I. (1990). Forecasting particulate pollutant concentrations in a city from meteorological variables and regional weather patterns. *Atmospheric Environment, 24A*(6 pt 1), 1509–1519.

Schlink, U., Pelikan, E., Dorling, S., et al. (2003). A rigorous intercomparison of ground-level ozone predictions. *Atmospheric Environment, 37*(23), 3237–3253.

Schmidler et al. (2000). Bayesian segmentation of protein secondary structure. *Journal of Computational Biology, 7*, 233–248.

Surneet, S., Veena, J., & Patil, R. S. (2002). Determining spatial patterns in Delhi's ambient air quality data using cluster analysis. In *East-West Center working paper: Environment series*.

The Clear Skies Act (2003). Technical support package, section B: Human health and environmental benefits.

Thomas, S., & Jacko, R. B. (2007). Model for forecasting expressway $PM_{2.5}$ concentration – Application of regression and neural network models. *The Journal of Air and Waste Management Association, 57*(4), 480–488.

Walsh, K., & Sherwell, J. (2002). Estimation of ambient $PM_{2.5}$ concentrations in Maryland and verification by measured values. *Journal of the Air and Waste Management Association, 52*(10), 1161–1175.

Yu, S.-Z., & Kobayashi, H. (2003a). A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking. *Signal Processing, 83*, 235–250.

Yu, S.-Z., & Kobayashi, H. (2003b). An efficient Forward–Backward algorithm for an explicit duration hidden Markov model. *IEEE Signal Processing Letters, 10*(1), 11–14.

Yu, S.-Z., & Kobayashi, H. (2006). Practical implementation of an efficient forward–backward algorithm for an explicit-duration hidden Markov model. *IEEE Transactions on Signal Processing, 54*(5), 1947–1951.