Atmospheric Environment 62 (2012) 64–73

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

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

Prediction of ozone levels using a Hidden Markov Model (HMM) with Gamma distribution

Hao Zhang [a], Weidong Zhang [a], Ahmet Palazoglu [b], Wei Sun [a,*]

[a] Beijing Key Lab of Membrane Science and Technology, College of Chemical Engineering, Beijing University of Chemical Technology, 15 Beisanhuan East Road, Beijing 100029, China
[b] Department of Chemical Engineering and Materials Science, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA

HIGHLIGHTS

► A Hidden Markov Model with Gamma distribution is derived in this paper.
► It is applied on ozone prediction in Livermore Valley near San Francisco, CA and Houston Metropolitan Area, TX.
► Results show that HMM-Gamma can predict all exceedances correctly and reduce false alarms significantly.

ARTICLE INFO

Article history:
Received 6 January 2012
Received in revised form
4 August 2012
Accepted 6 August 2012

Keywords:
Ozone prediction
Model-driven models
Hidden Markov models
Wavelets

ABSTRACT

Ground level ozone, generated by the photochemical reaction between nitrogen oxides and volatile hydrocarbons, is harmful to humans and the environment. Prediction and forecasting play an important role in the regulatory policies aimed at the control and reduction of surface ozone. Belonging to the family of model-driven statistical models, Hidden Markov Models (HMMs) provide a rich mathematical structure and perform well in many applications. While conventional HMM applications assume Gaussian distribution for the observation statistics, several key meteorological factors and most ozone precursors exhibit a non-Gaussian distribution, which would weaken the performance of a conventional HMM in modeling ozone exceedances. We propose a method based on a HMM with a Gamma distribution (HMM-Gamma) where each monitoring day is pre-labeled according to its maximum 8-h average ozone concentration and monitoring days are further grouped into zones with different ozone levels. Then, HMMs associated with each zone are trained using air quality monitoring data where the model parameters are estimated by a modified Expectation–Maximization (EM) algorithm. We derive a new re-estimation formula for the model parameters for observation sequences that exhibit a Gamma distribution. The trained HMM-Gamma models are used to predict ozone exceedances in two geographic areas, Livermore Valley near San Francisco, CA and Houston Metropolitan Area, TX. Compared to the conventional HMM (HMM-Gaussian), HMM-Gamma for the ground level ozone in Livermore Valley can reduce false alarms by 77% and HMM-Gamma for that in Houston Metropolitan Area can reduce false alarms by 32%.

1. Introduction

Ground level ozone is produced by the photochemical reaction between nitrogen oxides and volatile hydrocarbons (Jenkin and Clemitshaw, 2000) and is harmful to the human respiratory system (Golden et al., 1978). Exposure to prolonged high levels of ozone can lead to chronic respiratory illnesses such as bronchitis, emphysema, asthma, etc. (Koren et al., 1989; Lefohn et al., 1994; Tilton, 1989). In addition to its effect on the human body, high levels of ozone also cause the reduction of crop output by affecting photosynthesis efficiency of crops (Fuhrer et al., 1997). In densely populated areas of the United States, ground level ozone has been shown to easily reach harmful levels under favorable meteorological conditions (Beaver and Palazoglu, 2006; Prybutok et al., 2000). US EPA sets national ambient air quality standards for ozone and reviews these standards periodically. In 1979, the National Ambient Air Quality Standard (NAAQS) for ozone was set as the hourly average concentration being less than 120 parts per billion (ppb). In 1997, this has been modified to the daily maximum 8-h concentration being limited to 80 ppb; in 2008, further reductions set the threshold as 75 ppb.

[*] Corresponding author. Tel.: +86 10 6444 5826.
E-mail address: sunwei@mail.buct.edu.cn (W. Sun).

1352-2310/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2012.08.008

In the last decade, many urban areas have enacted voluntary ozone control programs, aimed at increasing public awareness and participating in local clean air efforts. A key component of these programs is the prediction of ozone exceedance days and the reduction of human exposure to high ozone concentrations. The benefits to human health and the environment by correctly forecasting high ozone days and issuing health warnings can be significant (Neidell, 2009).

To predict ozone concentrations, various modeling approaches have been proposed in recent years. These can be classified into four categories: (1) chemical transport models (CTMs); (2) empirical models; (3) data-driven statistical models and (4) model-driven statistical learning models.

CTMs comprise numerical models which typically simulate atmospheric chemistry (Jacob, 1999). Examples of CTMs include CMAQ, CAMx, GEOS-Chem, LOTOS-EUROS, MOZART and CLaMS (Brasseur et al., 1998; Byun et al., 1999; Fusco and Logan, 2003; McKenna et al., 2002; Schaap et al., 2008; Tesche et al., 2006). In these models, the meteorological model output and emission inventory are used as inputs and fundamental physical equations are used to simulate transport, diffusion, reaction and depletion of air pollutants. The execution of these complex models can be time-consuming and require vast computing resources. Their accuracy also depends on the scale they are applied and the quality of the emission data (Han et al., 2008; Isukapalli, 1999).

Empirical models are based on field observations of ozone concentration and meteorological conditions. Chang and Rudy (1993) built a semi-empirical model which relates the ratio of non-methane organic gases/nitrogen oxides with the ozone levels to predict ozone concentration. Jimenez and Baldasano (2002) used data from other stations for validation and found the model tending to overestimate the mean values of ozone. Generally, empirical models are embodiments of the observations at certain monitoring stations and would not apply to other locations. Moreover, empirical models do not explicitly account for the mechanisms underlying the actual process and cannot be further generalized.

Data-driven models are developed from data collected from various sources. Through statistical techniques, data-driven models can find expedient relations and knowledge hidden in a data set that cannot be discovered by empirical models. Data-driven models include regression models, clustering, artificial neural networks (ANNs), etc. Among these, ANNs have the most complex mathematical structure and can simulate human learning and pattern recognition. This allows information to be extracted from imprecise and nonlinear data sets using different types of ANNs (Hagan et al., 1996). Many ANN-based models have been developed to forecast ozone concentrations and the result shows that they perform well in forecast accuracy. Extremely high ozone days can be predicted well when the training data set contains sufficient number of ozone exceedance days. However, ANNs are developed with a non-explanatory structure and are black box approaches (Psichogios and Ungar, 1992). Generally, the successful implementation of ANN-based models depend on the proper selection of training data, network structure and connection mode among neural nodes (Foody et al., 1995). In recent years, ANN performance has improved when they are combined with other techniques.

In contrast to black box approaches such as ANNs, model-driven statistical techniques can be used to establish mathematical functional mapping between the patterns and input variables. State-space models, Bayesian networks and HMMs fall into this category. Cheon et al. (2009) developed a Bayesian network to forecast daily ozone levels and showed that the Bayesian network performed better than a decision tree model. An improved state-space model was proposed by Sahu et al. (2009) to forecast next day ozone in the eastern US.

HMMs have been increasingly popular and effective in speech processing and hand-written word recognition since the late 1960s. The HMMs offer a rich mathematical structure and can be used as the theoretical basis for many practical problems. HMMs have found applications in machine translation, gene prediction, alignment of bio-sequences and protein folding (Hannenhalli and Russell, 2000; Kato et al., 2003; Keller et al., 2009; Leu and Adi, 2011). HMMs have been also used for air pollutant prediction recently (Dong et al., 2009). In contrast to other HMM applications, air pollutant prediction is focused on the modeling of exceedance days which take a small fraction of the total number of monitoring days. According to the US EPA regulations, ozone exceedance days are defined as days in which the maximum 8-h average ozone concentration is greater than 75 ppb, most of which reside at the right end of the distribution curve. Previous research suggests that some key meteorological factors which affect the photochemical reactions and air quality data also have a non-Gaussian distribution (Cobourn and Hubbard, 1999; Denby et al., 2008). However, in most HMM applications, the hidden state outputs are represented by Gaussian distributions. As real data seldom has a normal distribution, there are two ways to address this problem. One is to use multiple Gaussians to approximate the real distribution, i.e., more hidden states can be employed in the HMM. As the number of hidden states increase, however, the number of parameters to be estimated will also increase rapidly, and so will the corresponding computational load. Also, the number of ozone exceedance days may be too few to estimate the HMM parameters properly. In other words, increasing the number of Gaussians, i.e., the hidden states, in emission distribution estimation may not be feasible in practice. The other alternative is to use a non-Gaussian distribution that more closely mimics the real distribution. O'Connell et al. (2010) used a HSMM based on a shifted Poisson sojourn distribution and a Gaussian emission distribution to predict cattle activity and progesterone level and their result shows that this model can predict 70% of follicular states. This suggests that HMMs with other distributions may prove to be more suitable for ozone exceedance prediction.

In this paper, a continuous multivariable HMM with a Gamma distribution (HMM-Gamma) is developed to predict ozone exceedance days in Livermore Valley, CA and Houston Metropolitan Area, TX. 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 HMM-Gamma parameters is described in detail. In the next section, data characteristics in Livermore Valley and Houston Metropolitan Area are described. This is followed by the results and discussion. Our concluding remarks are presented in the final section.

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.

How one state transitions to another state in a Hidden Markov Model is depicted in Fig. 1. In the beginning, a hidden state distribution is initialized as π at time t_1. Then, the hidden state moves from the initial state to the next state according to a state-transition probability matrix (A). Each state emits observations according to an emission probability (B), creating an observation sequence. The sequence ends at time t_l where l corresponds to the length of the

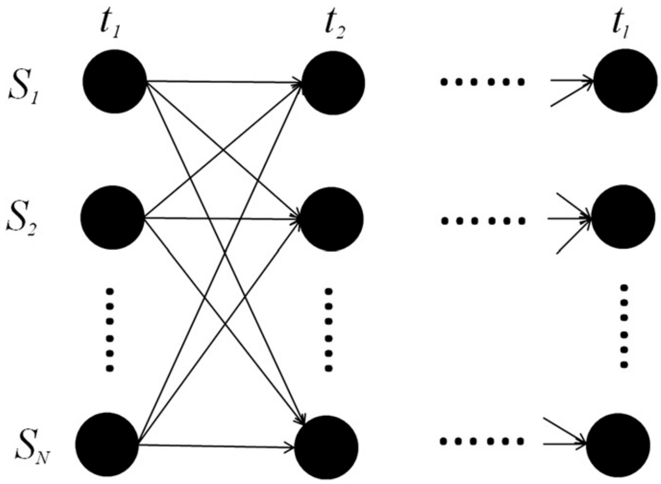

Fig. 1. Illustration of the state transitions. where *S* means states and *t* means time point.

observation sequence. HMM can be described by the following elements:

- π: initial state probability distribution denoted as $\pi = \{\pi(i)\}$; where $\pi(i) = P[q_1 = S_i]$, $1 \le i \le N$.
- *A*: state transition probability matrix denoted as $A = \{a_{ij}\}$; where $a_{ij} = P[q_{t+1} = S_j|q_t = S_i]$, $1 \le i, j \le N$.
- *B*: emission probability distribution in state S_j denoted as $B = \{b_j(k)\}$; where $b_j(k) = P[v_k|q_t = S_j]$, $1 \le j \le N$, $1 \le k \le M$.
- *N*: number of states of the Markov chain. Individual states are denoted as $S = \{S_1, S_2, \ldots S_N\}$, and the state at time *t* is denoted as q_t.

Thus, a HMM model can be described by the specification of *A*, *B*, π, and *N*.

2.2. Hidden Markov Model based on Gamma distribution

In a HMM, observations are made of a random variable $X_t = \{X_{t1}, X_{t2}, \ldots, X_{tk}\}$ that are related to the state S_t while the state itself is not observable. The conditional distribution of the observed variable X_t at the hidden state S_t is referred to as an emission distribution. Generally, the emission distribution is assumed to have a Gaussian distribution or can be characterized by a mixture of Gaussian distributions.

The problem is to use the observation data to adjust the HMM, $\lambda = (\pi, A, B)$, so that the probability of an observation, $P(O|\lambda)$, is maximized. Yet there is no optimal way of estimating the model parameters given any finite observation sequence as training data. Nevertheless, an iterative method, Baum-Welch algorithm, can be used to locally maximize $P(O|\lambda)$ (Baum et al., 1970). This technique was later grouped with a more general class of algorithms named the Expectation–Maximization (EM) algorithm (Dempster et al., 1977).

In the EM algorithm, during the E-step, the probability of being in state S_i at time *t* given the observed sequence ($\gamma_t(i)$) and the probability that the system changes state S_i at time *t* into state S_j at time *t*+1 are calculated by the forward-backward algorithm. The E-step estimation of HMM-Gamma would be similar to HMM-Gaussian. However, during the M-step, estimation of the non-Gaussian emission distribution parameters depends on the chosen distribution. In this paper, an estimation method for the shape and scale parameters of the Gamma distribution is derived.

A local maximum likelihood of the continuous multivariable HMM with a non-Gaussian emission distribution can be found via a modified EM algorithm that follows the steps below:

(1) The probability of being in state S_i at time *t* and state S_j at time *t*+1, given the model and the observation sequences, i.e.,

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j|O,\lambda) \tag{1}$$

From the definition of the forward-backward variables, Eq. (1) can be rewritten as follows:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \tag{2}$$

where $\alpha_t(i)$ is the probability of the partial observation sequence (until time *t*) and state S_i at time *t*; $\beta_{t+1}(j)$ accounts for the remainder of the observation sequence.

(2) The probability of being in state S_i at time *t*, given the observation sequence *O* and the model λ, i.e., is

$$\gamma_t(i) = P(q_t = S_i|O,\lambda) \tag{3}$$

And it also can be written as:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \tag{4}$$

where $\alpha_t(i)$ accounts for $O_1, O_2, \ldots, O_t$ give state S_i at *t*; and $\beta_t(i)$ accounts for the rest of sequence.

When the observations are continuous signals, the probability density function can be described in the following form:

$$b_j(O) = \sum_{m=1}^{M} c_{jm}\Pi\left[O, \mu_{jm}, U_{jm}\right] \quad 1 \le j \le N \tag{5}$$

where c_{jm} is the mixture coefficient for the *m*th mixture in state S_j and Π is the elliptically symmetric density with mean vector μ_{jm} and covariance matrix U_{jm} for the *m*th mixture component in state S_j.

The coefficient of mixture density could be calculated by the following equations:

$$\bar{c}_{jm} = \frac{\sum_{t=1}^{T}\gamma_t(j,m)}{\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t(j,m)} \tag{6}$$

$$\bar{\mu}_{jm} = \frac{\sum_{t=1}^{T}\gamma_t(j,m)\cdot O_t}{\sum_{t=1}^{T}\gamma_t(j,m)} \tag{7}$$

$$\overline{U}_{jm} = \frac{\sum_{t=1}^{T}\gamma_t(j,m)\cdot\left(O_t - \mu_{jm}\right)\left(O_t - \mu_{jm}\right)'}{\sum_{t=1}^{T}\gamma_t(j,m)} \tag{8}$$

Thus, the probability of being in state S_j at time *t* with the *k*th mixture component for continuous observations is expressed as:

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(i)}\right]\left[\frac{c_{jk}\Pi\left[O_t, \mu_{jk}, U_{jk}\right]}{\sum_{m=1}^{M}c_{jm}\Pi\left[O, \mu_{jm}, U_{jm}\right]}\right] \tag{9}$$

$\gamma_t(i)$ and $\xi_t(i,j)$ can be related by summing over j, giving

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{10}$$

(3) In the Expectation step, $\gamma_t(i)$ and $\xi_t(i,j)$ is calculated via the forward-backward algorithm which has the complexity $O(J^2T)$.
(4) Based on the Expectation step, HMM parameters are estimated during the Maximization step:

$$\overline{\pi}_i = \gamma_1(i) \tag{11}$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{12}$$

Estimates for the parameters of the emission distribution are dependent on the choice of the distribution. In our model, X_t are assumed to be Gamma distributed given $S_t = i$, that is $X_t|S_t = i \sim \Gamma(k,\theta)$. In this case, the likelihood for Gamma distribution can be maximized with its parameters by solving,

$$\log\left(\widehat{k}_i\right) - \psi\left(\widehat{k}_i\right) = \log(\overline{x}_t) - \overline{\log(x_t)} \tag{13}$$

where $\psi()$ is the digamma function, $\overline{x}_t = \sum_{t=1}^{T-1} \gamma_\mathrm{t}(i)x_t \Big/ \sum_{t=1}^{T-1} \gamma_t(i)$, $\overline{\log(x_t)} = \sum_{t=1}^{T-1} \gamma_t(i)\log(x_t)/\sum_{t=1}^{T-1} \gamma_t(i)$. This equation can be solved using Newton's method (Choi and Wette, 1969).

The scale parameter θ is then estimated as follows:

$$\widehat{\theta}_i = \overline{x}_t/\widehat{k}_i \tag{14}$$

Both E-step and M-step are repeated iteratively until convergence. Equations (7) and (8) are used to estimate the mean and variance of the Gaussian distribution in HMM-Gaussian. In HMM-Gamma, the shape and scale parameters are estimated by (13) and (14), instead.

2.3. HMM with multiple observation sequences

To solve the training problem of HMMs with multiple observation sequences, Li et al. (2000) presented a formal treatment in which the multiple observation probability is expressed as a combination of individual observation probabilities. Given a set of K observation sequences $O = \{O^{(1)}, O^{(2)}, \ldots, O^{(K)}\}$, where $O^{(k)} = o_1^{(k)} o_2^{(k)} \ldots o_{T_k}^{(k)}$, with $l \le k \le K$, are independent observation sequences, with T_k being the length of the observation sequence k, the multiple observation probability is expressed as:

$$P(O|\lambda) = \Pi_{k=1}^{K} P\left(O^{(k)} \middle| \lambda\right) \tag{15}$$

and the updating parameters are:

$$\overline{a_{ij}} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \alpha_t^k(i) a_{ij} b_j\left(O_{t+1}^{(k)}\right) \beta_{t+1}^k(j)}{\sum_{k=1}^{K} \sum_{t=1}^{T_k-1} \alpha_t^k(i)\beta_t^k(i)} \tag{16}$$

and

$$\overline{b_j(l)} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{\substack{t=1 \\ s.tO_t = v_t}}^{T_k-1} \alpha_t^k(i)\beta_t^k(i)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i)\beta_t^k(i)} \tag{17}$$

where P_k is the probability of the sequence k; $\alpha_t^k(i) = P(o_1 o_2 \ldots o_t, q_t = i|.\lambda)$ is the probability of the partial observation sequence $o_1 o_2 \ldots o_t$ of individual sequence k until time t and q_t is state i at t given the model λ; and $\beta_t^k(i) = P(o_{t+1} o_{t+2} \ldots o_T, q_t = i|.\lambda)$ is the probability of partial observation sequence from t+1 to T given state q_t and the model λ. This method can be applied to both HMM-Gaussian and HMM-Gamma with multiple observation sequences.

3. Study area and data characteristics

In this paper, we consider ozone concentrations measured at the Livermore monitoring station near the Bay Area of San Francisco (SFBA), CA and the Houston Deer Park monitoring station which is part of the Houston Metropolitan Area, TX. Both SFBA and Houston were designated as non-attainment areas for many years. Recently, great strides have been made by the Bay Area Air Quality Management District (BAAQMD) and Texas Commission on Environmental Quality (TECQ) to reduce the number of ozone exceedance days. Livermore is dominated by a Mediterranean climate which hardly brings rainfall during summer months and Houston is dominated by a subtropical climate which brings a lot of moisture during summers. Thus, the ozone dynamics of Livermore is expected to be quite different from Houston.

3.1. Bay Area of San Francisco, CA

The geography of the SFBA is shown in Fig. 2. In summer, clean marine air is usually driven by the high pressure off the coastline through the mouth of San Francisco Bay. The marine flow is blocked and deflected by the Coastal Range which is parallel to the seashore and some can pass through the Delta into the Central Valley. The I-

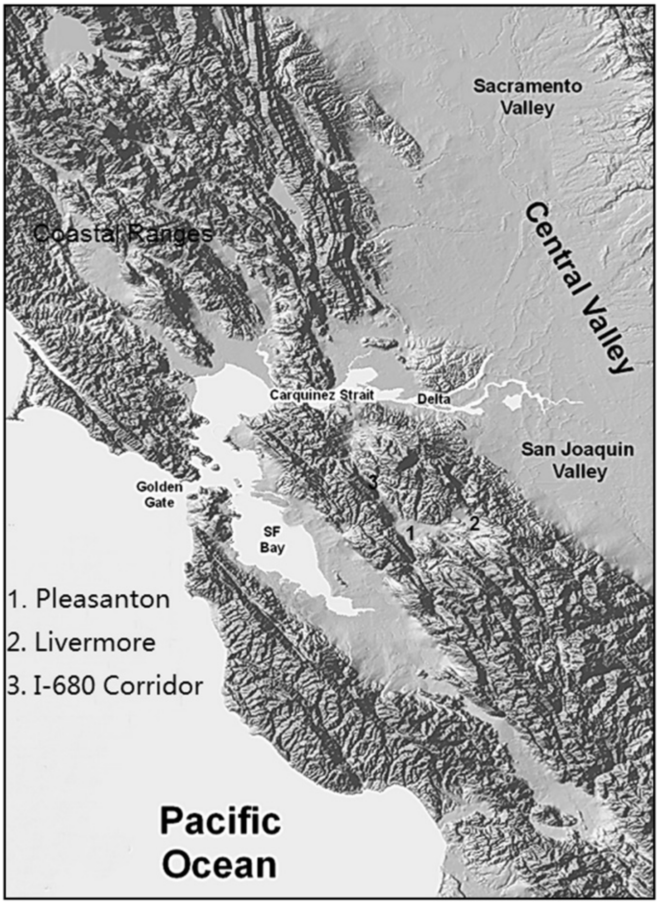

Fig. 2. The geography of Bay Area of San Francisco, CA.

680 Corridor allows the cool, clean marine air flow into the Livermore Valley. During the ozone season, the Livermore Valley is usually under high temperature conditions, with relatively low wind speeds and a stable boundary layer. These conditions often trigger high ozone levels in the Livermore Valley due to the precursors brought by the sea breeze.

Meteorological data which consist of wind speed, relative humidity and solar radiation monitored at Livermore from 2000 to 2009 are available at BAAQMD's website (http://www.baaqmd.gov/). Because the temperature data at Livermore is unavailable, data at the nearby Pleasanton site is used instead. Air quality data collected at the Livermore monitor include ozone, NO, NO_2, total hydrocarbon (THC) and non-methane hydrocarbon (NMHC) and can be downloaded from the California Air Resources Board (CARB) website (http://www.arb.ca.gov/html/ds.htm). Ozone, NO and NO_2 are recorded from 2000 to 2009. THC data is available from 2000 to 2005 and NMHC data is available from 2005 to 2009. Compared to NO_x, concentrations of VOCs are much higher; the mean VOC concentration is above 2000 ppb and the mean NO_x concentration is 32 ppb. In this paper, data from 2000 to 2007 are used to train the HMMs and data from 2008 to 2009 are used for validation.

3.2. Houston Metropolitan Area

Houston is the largest city in Texas, and the fourth-largest city in the United States. Most of Houston is located on the Gulf coastal plain and the east of downtown is the Houston Ship Channel. The Ship Channel continues past Galveston and then into the Gulf of Mexico. Houston's climate is classified as humid subtropical, with prevailing winds from the south and southeast during the most of the year bringing heat from deserts of Mexico and moisture from the Gulf of Mexico. During summer months, temperatures could easily reach 32 °C. With flourishing industries of energy, manufacturing, aeronautics and transportation, a lot of ozone precursors are generated locally. The geography of the Houston Metropolitan Area is shown in Fig. 3.

Data used in this paper is from the monitoring site at Houston Deer Park maintained by TCEQ (http://www.tceq.texas.gov/agency/data/ozone_data.html). Deer Park is to the east of Houston downtown and adjacent to an industrial park and during summer months, sea breeze which blows northwesterly carries a lot of ozone precursors. Thus, ozone concentrations may build to significant levels under certain weather conditions. This monitoring site is activated at 1996 and collects data on wind, solar radiation, temperature, relative humidity, dew point, ozone, PM_{10}, $PM_{2.5}$, nitrogen oxides, CO and CO_2.

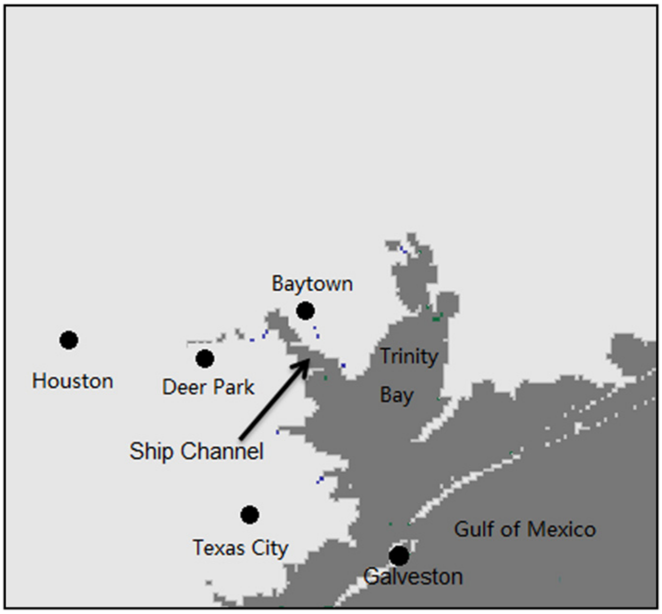

Fig. 3. The geography of Houston Metropolitan Area.

4. Results and discussion

4.1. Forecasting methodology

Before the training step, observation samples are pre-labeled according to the maximum 8-h average ozone level of the target day. Then, specific HMM models are built for different levels of ozone. In this paper, two ozone zones are considered for Livermore Valley, one for ozone concentrations above 75 ppb and the other for below 75 ppb. In Houston, three ozone zones are defined. The exceedance level is defined as ozone concentrations above 75 ppb. The non-exceedance level is split into two parts, one from 0 to 57 and the other is from 58 to 74 ppb. Thus, corresponding HMMs are trained for each concentration zone.

There is no standard method to determine whether an HMM is trained properly or not because the training involves finding parameters that maximize the probability of an observation sequence and iterations may terminate at a local maximum. To address this, simulated annealing method (SA) and genetic algorithm (GA) are often employed in the training step to improve HMM performance (Chau et al., 1997; Paul, 1985). When the training step is completed, the HMMs can then be used for classification.

The details of this method are discussed below:

(1) *Data preparation*. Before training, the available observation sequences are organized into temporal windows which cover 72 h. The rationale behind choosing the length of the window will be discussed in the next step. Short gaps (missing observations) in the data could be interpolated across time. Larger gaps which are more than 2 h and less than 8 h are imputed by the method of Schneider (Schneider, 2001). The imputation result for gaps which larger than 8 samples is not reliable, thus, windows which lack data for more than 8 h are deleted from the training data. As a result, the training data for Livermore Valley and Houston contain 976 and 992 windows, respectively.

(2) *Selection of variables and the length of observation sequence*. As stated before, in the presence of ultraviolet rays, photochemical reactions involving nitrogen oxides and VOCs will produce ground level ozone. Further, the ozone concentration itself also affects the rate of the photochemical reaction. Meteorological factors, including temperature, solar radiation, relative humidity and wind speed are important input variables which have a direct influence on ozone production, accumulation and depletion. Compared to the nitrogen oxide concentration in the boundary layer, VOC concentration is relatively high so that it would not significantly affect the photochemical reaction rate (Dodge, 1977). Because the data quality of measured VOCs in the Livermore Valley is poor, they are omitted from the input variable list. To enable a comparison with the Livermore site, the same variables are also used to train and validate HMM models in the Houston Metropolitan Area. Although most ozone is generated at the current day, the air mass which dominates the local weather regime usually persists for 1–3 days (Beaver et al., 2010). Therefore, we chose observation sequences to cover 72 h (3 days). The variable list is shown in Table 1.

Table 1
The list of input variables for ozone prediction.

Variable name	Unit	Time covered
Wind speed	m s^{-1}	T − 48:T + 24
Relative humidity	–	T − 48:T + 24
Solar radiation	W m^{-2}	T − 48:T + 24
Temperature	°C	T − 48:T + 24
Ozone	ppb	T − 72:T
NO	ppb	T − 72:T
NO_2	ppb	T − 72:T

(3) *Wavelet decomposition*. Although observation sequences cover three days, 72 nodes in one observation sequence are still too many for a HMM application. Wavelet decomposition can be used to reduce the size of the original sequence and extract data features in the time domain. During wavelet decomposition, a time-series signal is decomposed into an approximation part which represents the low frequency component of the signal and a detail part which captures the high frequency features. Thus, wavelet decomposition summarizes the original observation sequence consisting of a large number of samples with a very few wavelet coefficients. The low frequency part can be used to capture the underlying trend of the meteorological factors and pollutant variations (Sun et al., 2003). In this paper, the Haar wavelet basis function is used to decompose the observation sequences at three decomposition levels resulting in 9 wavelet coefficients for each variable.

(4) *Probability calculation by using trained HMMs for each ozone level*. First, HMMs corresponding to different ozone levels are built for prediction. Then, using the forward-backward algorithm, probability that the observed sequence is produced by the given HMM can be calculated in a straightforward manner.

(5) *Sequence labeling*. At this step, we determine the level to which the given observation sequence belongs by the calculated probability. The HMM which generates the maximum log-likelihood most likely produces the given observation sequence. The procedure is depicted in Fig. 4.

Table 2
Ozone concentration levels for the studied areas.

Location	Level	Ozone value range	Percentage days
Livermore	Level 1	$0 < O_3 < 75$	95.70%
Valley	Level 2	$75 \le O_3 < 120$	4.30%
Houston	Level 1	$0 < O_3 < 75$	86.19%
Metropolitan	Level 2	$75 \le O_3 < 130$	13.81%

4.2. Prediction results

As the aim of air quality prediction is to aid in correctly forecasting days with the harmful levels of pollutant concentrations, the focus of this paper is the days in which maximum 8-h average ozone concentration is 75 ppb and above. Thus, ozone concentrations are classified into two categories as shown in Table 2.

Table 2 shows that the ozone exceedance days comprise a small fraction of the total number of days in the monitoring period. Due to sustained higher temperatures in the long summer months and heavy industrial emissions, there are more exceedance days in the Houston Metropolitan Area than in the Livermore Valley. The average ozone concentration in the Houston Metropolitan Area is also higher than that in the Livermore Valley. Ozone density distributions in Livermore Valley and Houston Metropolitan Area are shown in Fig. 5. In Livermore Valley, the ozone concentration in most days is around 40 ppb and ozone exceedance days take up a very small fraction of the total data set. In contrast to the Livermore Valley, ozone samples distribute more homogenously and the distribution curve shows two peaks in the Houston Metropolitan Area. Houston Deer Park is adjacent to an oil refining industrial park and the Ship Channel; thus, ozone precursors are emitted into the boundary layer directly and can accumulate to high levels. When high-pressure air mass dominates, the stagnant boundary layer in the Houston Metropolitan Area favors the accumulation of local ozone. This may explain why ozone levels higher than 60 ppb in Houston Metropolitan Area is much more frequent than in the Livermore Valley.

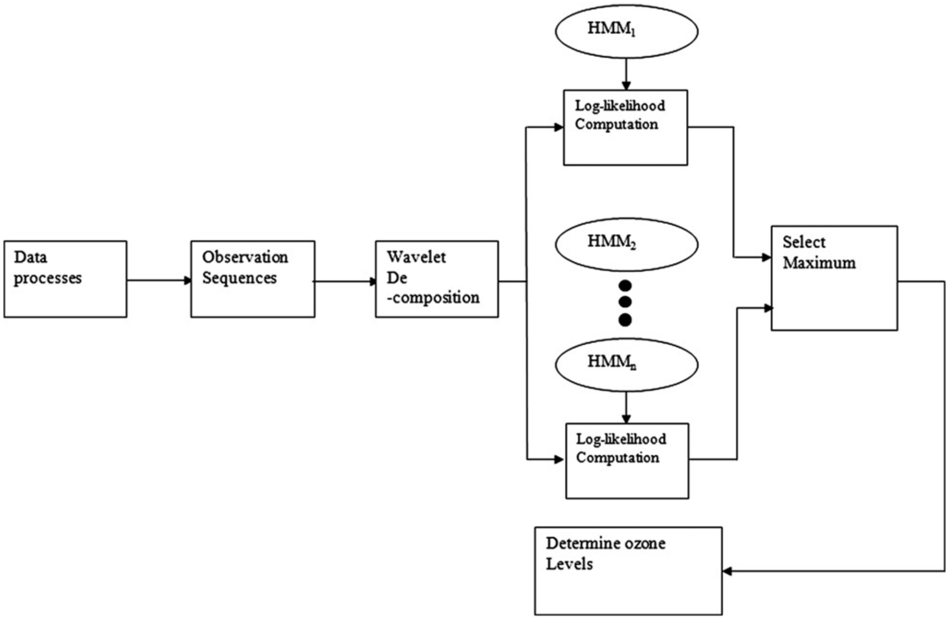

Fig. 4. Block diagram for the ozone prediction method based on HMMs.

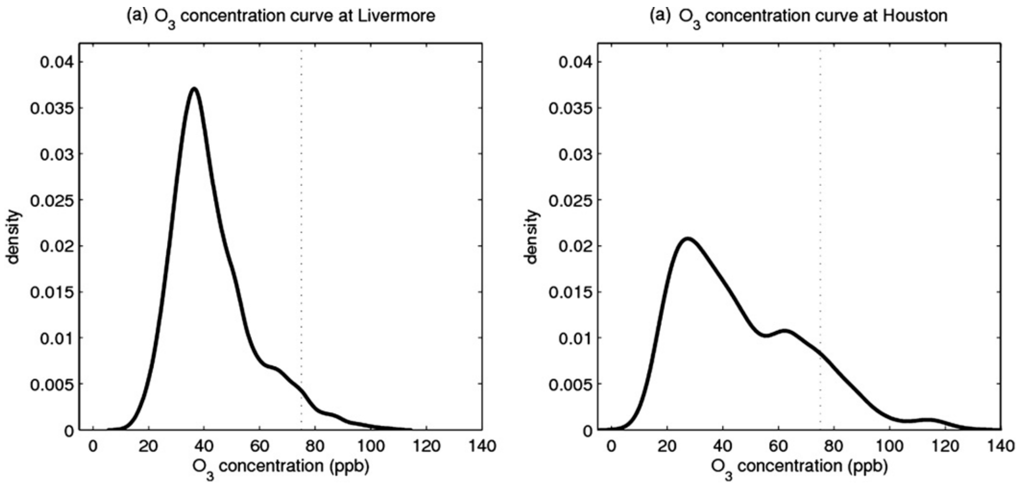

Fig. 5. Ozone distribution curves for Livermore (left) and Houston (right).

At first, a HMM-Gaussian is tested for prediction of ozone in the Livermore Valley and the result shows that all actual exceedance days are correctly identified. The model, however, is too sensitive for such exceedance days resulting in many false alarms. Then, a HMM-Gamma is constructed with the same input variables. The result shows that the HMM-Gamma can reduce false alarms. These two HMMs are also applied to predict ozone exceedance days in the Houston Metropolitan Area. Results show that the HMM-Gamma can also reduce the false alarms compared to the HMM-Gaussian. However, HMM-Gamma for Houston does not perform as well as the one for the Livermore Valley in terms of the number of false alarms due to the significant departure of the real emission distribution from an ideal Gamma distribution. Comparison results are listed in Table 3.

As noted before, Table 3 shows that all HMMs can predict ozone exceedance days correctly. However, HMM-Gaussian cannot distinguish these two levels accurately. Compared to HMM-Gaussian, HMM-Gamma can reduce false alarms by 77.42% in the Livermore Valley with only 7 non-exceedance days being labeled as exceedance days from 2008 to 2009. It is noted that the ozone concentration of most false alarm days is above 65 ppb. The same conclusion can be drawn from the study in the Houston Metropolitan Area. Compared to HMM-Gaussian, HMM-Gamma for Houston reduces the false alarm rate from 11.48% to 7.79%.

Fig. 6 shows the difference between a real emission distribution and an ideal Gamma distribution of the ozone level, on ozone exceedance days and normal ozone days at the Livermore site, respectively. We can see in Fig. 6b that the Gamma distribution can model ozone samples in Livermore quite well.

There are more samples that range from 60 ppb to 75 ppb in the Houston Metropolitan Area. As shown in Fig. 7, neither Gamma distribution nor Gaussian distribution can model this data properly. This is the key reason why many days in which maximum 8-h ozone concentration are from 65 to 75 ppb are misclassified as exceedance days.

Table 3
Performance of different HMMs.

	HMM for Livermore Valley		HMM for Houston metropolitan	
	HMM-Gaussian	HMM-Gamma	HMM-Gaussian	HMM-Gamma
TPR	100% (11/11)	100% (11/11)	100% (9/9)	100% (9/9)
FAR	11.65% (31/266)	2.63% (7/266)	11.48% (28/244)	7.79%(19/244)

Next, as the real distribution curve of Houston has two peaks below 75 ppb, we split the non-exceedance days into 2 subcategories. One category varies from 0 ppb to 57 ppb and the other from 58 ppb to 75 ppb. Thus, the whole data set is

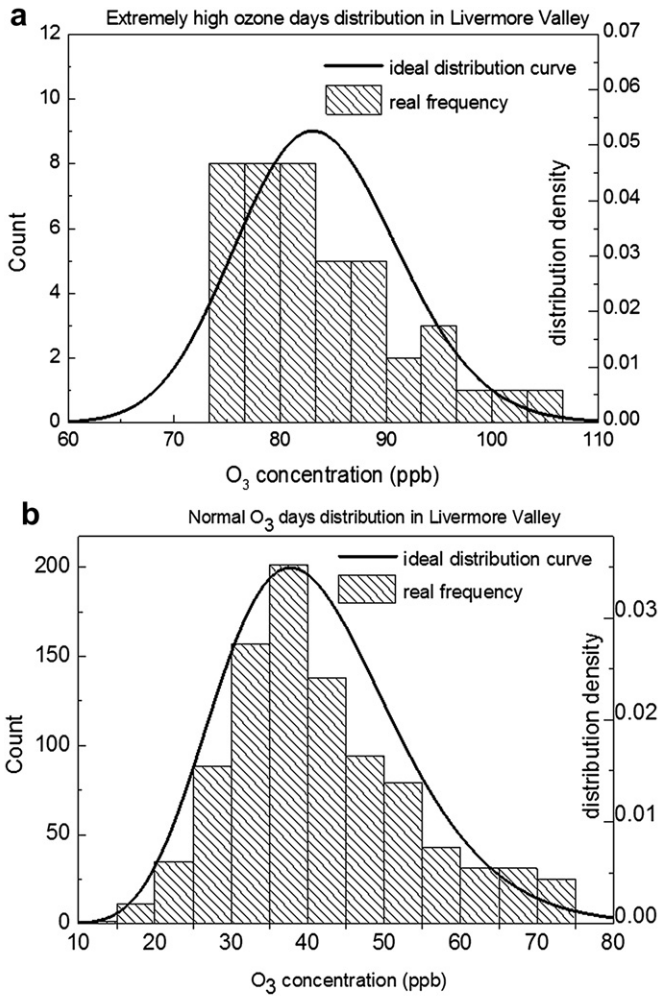

Fig. 6. Ozone distribution for (a) exceedance and (b) non-exceedance days in the Livermore Valley.

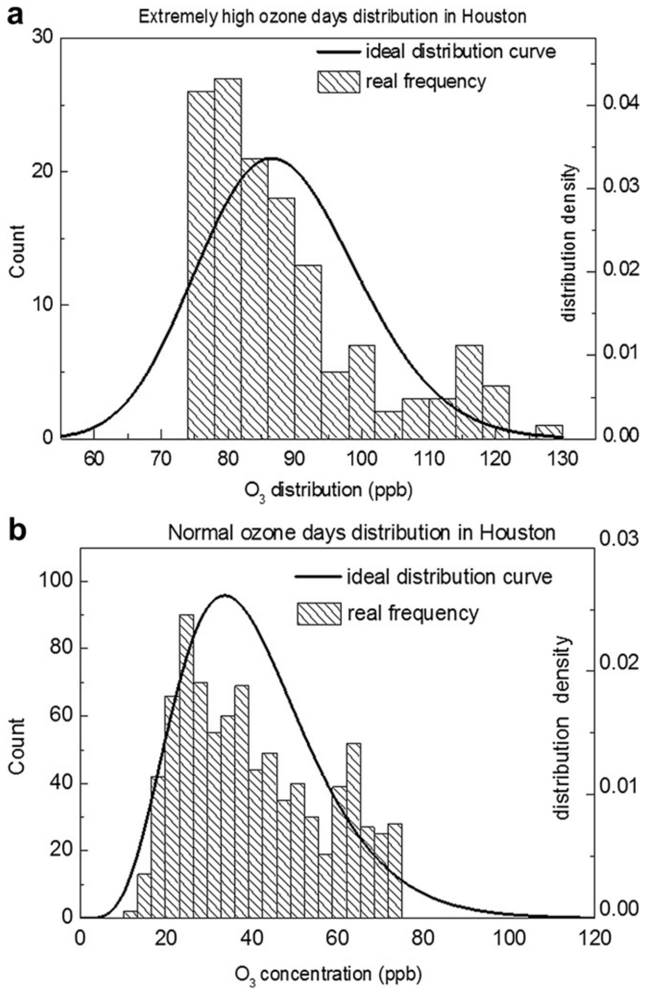

Fig. 7. Ozone distribution for (a) exceedance and (b) non-exceedance days in Houston.

decomposed into three parts. The first part from 0 ppb to 57 ppb contains 676 samples, the second one from 58 to 75 ppb includes 179 samples and the exceedance part contains 137 samples. The statistical details are given in Table 4. The distribution curve for each level is shown in Fig. 8.

Then, three HMM-Gaussian and HMM-Gamma models are trained for each ozone level. Prediction results are shown in Table 5. Compared to HMM-Gaussian trained for 2 zones, HMM-Gaussian with 3 zones eliminates 15 false alarms. HMM-Gamma with 3 zones can eliminate 7 false alarms compared to the HMM-Gamma with 2 zones. Among 4 HMM models, HMM-Gamma corresponding to 3 zones performs the best. It is noteworthy that the performance difference between HMM-Gamma and HMM-Gaussian for 3 zones is almost negligible. It is clear that the number of zones (model granularity) has more impact on the prediction performance than the choice of the emission distributions for Houston. Due to the specific nature of the ozone distribution, three ozone zones can model the ozone days more accurately.

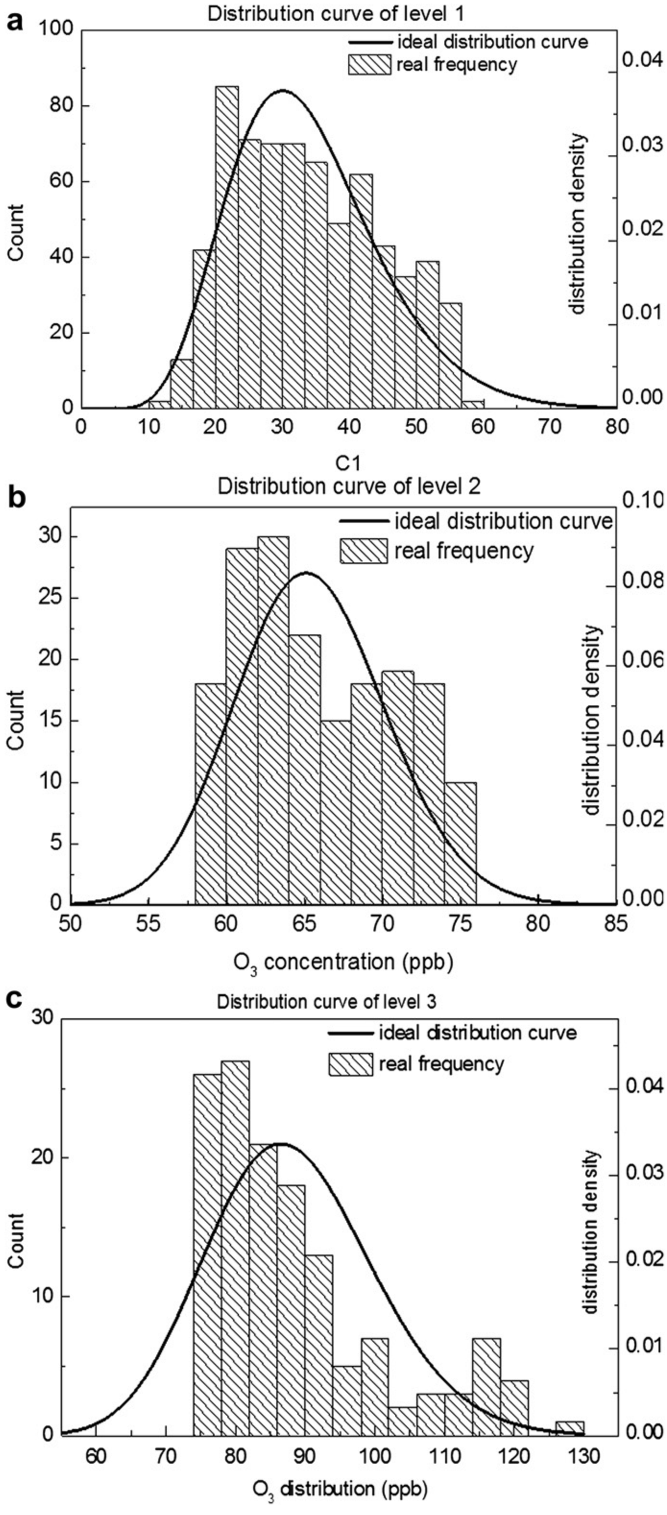

Fig. 8. Distribution curve for each ozone level in Houston.

Table 4
Details of different ozone concentration levels in Houston.

Location	Level	Ozone value range	Number of samples	Percentage days	Mean value
Houston	Level 1	$0 < O_3 < 58$	676	95.70%	33.6 ppb
	Level 2	$58 \leq O_3 < 75$	179	4.30%	65.5 ppb
	Level 3	$75 \leq O_3 < 120$	137	86.19%	88.1 ppb

Table 5
Comparison of different HMMs prediction results in Houston.

Houston	HMM for two levels		HMM for three levels	
	HMM-Gaussian	HMM-Gamma	HMM-Gaussian	HMM-Gamma
TPR	100% (9/9)	100% (9/9)	100% (9/9)	100% (9/9)
FAR	11.48% (28/244)	7.79% (19/244)	5.33% (13/244)	4.92%(12/244)

Another way of improving the performance of a HMM is to increase the number of Gaussian distributions, i.e., the number of hidden states. A mixture model may capture the real emission distribution better and help reduce false alarms. However, increasing the number of Gaussian distributions would result in an increase in the number of HMM parameters to be estimated. It is noted that the training data of Livermore only have 42 ozone exceedance days. Too few training data makes the estimation procedure unreliable. Therefore, increasing the number of hidden states of HMMs is unadvisable. Another possibility is to use more ozone zones (data granularity). This option shares the same drawback as increasing the number of hidden states in that the data sets often do not contain a sufficient number of days to reliably estimate model parameters. On the other hand, the Livermore Valley results showed that using a proper emission distribution can improve HMM performance significantly. Therefore, a more viable option would be to seek ways to better capture the actual emission distribution of the observed data.

5. Conclusions

A HMM with Gamma distribution (HMM-Gamma) is proposed for ozone zone identification to overcome the shortcomings of the conventional HMM (HMM-Gaussian) that assumes a Gaussian behavior for observed variables. The results for the Livermore Valley and Houston Metropolitan Area show that HMM-Gamma can reduce false alarms with all ozone exceedance days being predicted correctly. The Houston Metropolitan Area results further show that the improvement of HMM-Gamma performance is not as significant compared to the Livermore Valley, owing to the fact that the emission distribution used in the HMM deviates significantly from the real distribution. When the data set is large enough, this problem could be solved by dividing the data set into more subsets that capture the distribution with more granularities. However, this would be infeasible when applied to real situations where the data size and quality are often compromised.

Although we have shown that HMM-Gamma exhibits better prediction performance than HMM-Gaussian for ozone exceedance prediction in the Livermore Valley and Houston Metropolitan Areas, there is still room for improvement. For example, at the step where the sequence length is chosen, other wavelet basis functions can be considered for sequence compression. Furthermore, HMMs with a more suitable distribution could yield better prediction results.

Acknowledgement

This work was partially supported by a grant from National Natural Science Foundation (No. 21076011) and Program for New Century Excellent Talents in University (NCET-10-0210). Also, the authors would like to thank anonymous reviewers for their useful comments which have helped to improve the quality of the manuscript.

References

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics 41, 164–171.

Beaver, S., Palazoglu, A., 2006. A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area. Atmospheric Environment 40, 713–725.

Beaver, S., Palazoglu, A., Singh, A., Soong, S.T., Tanrikulu, S., 2010. Identification of weather patterns impacting 24-h average fine particulate matter pollution. Atmospheric Environment 44, 1761–1771.

Brasseur, G., Hauglustaine, D., Walters, S., Rasch, R., Müller, J.F., Granier, C., Tie, X., 1998. MOZART, a global chemical transport model for ozone and related chemical tracers 1. Model description. Journal of Geophysical Research 103, 28265–28289.

Byun, D.W., Ching, J., Research, U.S.E.P.A.O.o., Development, Division, N.E.R.L.A.M, 1999. Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. US Environmental Protection Agency, Office of Research and Development, Washington, DC.

Chang, T.Y., Rudy, S.J., 1993. Ozone-precursor relationships: a modeling study of semiempirical relationships. Environmental Science & Technology 27, 2213–2219.

Chau, C., Kwong, S., Diu, C., Fahrner, W., 1997. Optimization of HMM by a Genetic Algorithm, IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, Munich, Germany, pp. 1727–1730.

Cheon, S.-P., Kim, S., Lee, S.-Y., Lee, C.-B., 2009. Bayesian networks based rare event prediction with sensor data. Knowledge-Based Systems 22, 336–343.

Choi, S., Wette, R., 1969. Maximum likelihood estimation of the parameters of the gamma distribution and their bias. Technometrics, 683–690.

Cobourn, W.G., Hubbard, M.C., 1999. An enhanced ozone forecasting model using air mass trajectory analysis. Atmospheric Environment 33, 4663–4674.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1–38.

Denby, B., Schaap, M., Segers, A., Builtjes, P., Horálek, J., 2008. Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale. Atmospheric Environment 42, 7122–7134.

Dodge, M., 1977. Combined Use of Modeling Techniques and Smog Chamber Data to Derive Ozone-precursor Relationships, pp. 881–889.

Dong, M., Yang, D., Kuang, Y., He, D., Erdal, S., Kenski, D., 2009. PM2. 5 concentration prediction using hidden semi-Markov model-based times series data mining. Expert Systems with Applications 36, 9046–9055.

Foody, G.M., McCulloch, M.B., Yates, W.B., 1995. Classification of remotely sensed data by an artificial neural network: issues related to training data characteristics. Photogrammetric Engineering and Remote Sensing 61, 391–401.

Fuhrer, J., Skärby, L., Ashmore, M., 1997. Critical levels for ozone effects on vegetation in Europe. Environmental Pollution 97, 91–106.

Fusco, A.C., Logan, J.A., 2003. Analysis of 1970–1995 trends in tropospheric ozone at Northern Hemisphere midlatitudes with the GEOS-CHEM model. J. Geophys. Res 108, 1988–1997.

Golden, J., Nadel, J., Boushey, H., 1978. Bronchial hyperirritability in healthy subjects after exposure to ozone. American Review of Respiratory Diseases (United States) 118.

Hagan, M.T., Demuth, H.B., Beale, M.H., University of Colorado, B, 1996. Neural Network Design. PWS Pub.

Han, Z., Ueda, H., An, J., 2008. Evaluation and intercomparison of meteorological predictions by five MM5-PBL parameterizations in combination with three land-surface models. Atmospheric Environment 42, 233–249.

Hannenhalli, S.S., Russell, R.B., 2000. Analysis and prediction of functional sub-types from protein sequence alignments. Journal of Molecular Biology 303, 61–76.

Isukapalli, S.S., 1999. Uncertainty Analysis of Transport-transformation Models. Rutgers, The State University of New Jersey.

Jacob, D.J., 1999. Introduction to Atmospheric Chemistry, first ed. Princeton Univ Pr.

Jenkin, M.E., Clemitshaw, K.C., 2000. Ozone and other secondary photochemical pollutants: chemical processes governing their formation in the planetary boundary layer. Atmospheric Environment 34, 2499–2527.

Jimenez, P., Baldasano, J., 2002. Validation of an empirical ozone model. Advances in Air Pollution Series, 1–12.

Kato, R., Noguchi, H., Honda, H., Kobayashi, T., 2003. Hidden Markov model-based approach as the first screening of binding peptides that interact with MHC class II molecules. Enzyme and Microbial Technology 33, 472–481.

Keller, A., Schleicher, T., Schultz, J., Müller, T., Dandekar, T., Wolf, M., 2009. 5.8S–28S rRNA interaction and HMM-based ITS2 annotation. Gene 430, 50–57.

Koren, H., Devlin, R., Graham, D., Mann, R., McGee, M., Horstman, D., Kozumbo, W., Becker, S., House, D., McDonnell, W., 1989. Ozone-induced inflammation in the lower airways of human subjects. American Review of Respiratory Diseases (United States) 139.

Lefohn, A., Edwards, P., Adams, M., 1994. The characterization of ozone exposures in rural West Virginia and Virginia. Air & Waste: Journal of the Air & Waste Management Association 44, 1276.

Leu, S.-S., Adi, T.J.W., 2011. Probabilistic prediction of tunnel geology using a Hybrid-Neural-HMM. Engineering Applications of Artificial Intelligence 24, 658–665.

Li, X., Parizeau, M., Plamondon, R., 2000. Training Hidden Markov Models with multiple observations-a combinatorial method. Pattern Analysis and Machine Intelligence, IEEE Transactions on 22, 371–377.

McKenna, D.S., Grooß, J.U., Günther, G., Konopka, P., Müller, R., Carver, G., Sasano, Y., 2002. A new chemical Lagrangian model of the Stratosphere (CLaMS): 2. Formulation of chemistry scheme and initialization. J. Geophys. Res 107, 4256.

Neidell, M., 2009. Air quality warnings and outdoor activities: evidence from Southern California using a regression discontinuity design. Journal of Epidemiology and Community Health 64, 921–926.

O'Connell, J., Tøgersen, F.A., Friggens, N.C., Løvendahl, P., Højsgaard, S., 2010. Combining cattle activity and progesterone measurements using hidden semi-Markov Models. Journal of Agricultural, Biological, and Environmental Statistics 16, 1–16.

Paul, D., 1985. Training of HMM Recognizers by Simulated Annealing. IEEE. pp. 13–16.

Prybutok, V.R., Yi, J., Mitchell, D., 2000. Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. European Journal of Operational Research 122, 31–40.

Psichogios, D.C., Ungar, L.H., 1992. A hybrid neural network-first principles approach to process modeling. AIChE Journal 38, 1499–1511.

Rabiner, L.R., 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77, 257–286.

Sahu, S.K., Yip, S., Holland, D.M., 2009. Improved space–time forecasting of next day ozone concentrations in the eastern US. Atmospheric Environment 43, 494–501.

Schaap, M., Timmermans, R.M.A., Roemer, M., Boersen, G., Builtjes, P., Sauter, F., Velders, G., Beck, J., 2008. The LOTOS EUROS model: description, validation and latest developments. International Journal of Environment and Pollution 32, 270–290.

Schneider, T., 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate 14, 853–871.

Sun, W., Palazoğlu, A., Romagnoli, J.A., 2003. Detecting abnormal process trends by wavelet-domain Hidden Markov Models. AIChE Journal 49, 140–150.

Tesche, T., Morris, R., Tonnesen, G., McNally, D., Boylan, J., Brewer, P., 2006. CMAQ/CAMx annual 2002 performance evaluation over the eastern US. Atmospheric Environment 40, 4906–4919.

Tilton, B.E., 1989. Health effects of tropospheric ozone. Environmental Science and Technology (United States) 23.