Ecological Modelling 145 (2001) 275–289

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Prediction of ozone peaks by mixture models

C. Ambroise *, Y. Grandvalet

UMR CNRS 6599, Centre de recherches de Royallieu, BP 20259, 60205 Compiègne Cedex, France

Received 19 February 2001; received in revised form 15 June 2001; accepted 29 June 2001

Abstract

This paper applies recent developments of mixture models to the prediction of peak episodes of ozone in Lyon (a major French city). Forecasting for the next day should be available at 14:00 h GMT. One day ahead prediction of such events allow public authorities to warn the population of the danger of outdoor exercises and activities. Compared to a standard discrimination problem, the database has many unusual characteristics among which missing measurements and missing decisions. All this peculiarities are well taken into account by mixture models, which additionally allow class densities to have complex shapes and decision boundaries to be non-linear. The proposed approach is compared to the persistence method, which consists in forecasting the same decision as the day before. The results show that mixture models outperform this simple predictor and offer the advantage of properly handling the missing labels and data. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Pollution; Mixture model

1. Introduction

Urban ozone, the main ingredient of smog, develops when volatile organic compounds (VOCs), nitric oxide (NO) and nitrogen dioxide (NO_2) react chemically in the presence of heat and sunlight.

The primary sources of VOCs and NO_x are cars and industrial facilities. Other sources of VOCs include consumer products such as paints, insecticides, solvents and cleaners. Ozone affects plant-life and human health by causing inflammation and irritation of the respiratory tract, particularly during heavy physical activity. According to medical studies, it can aggravate asthma and chronic lung diseases, it also lowers the resistance to diseases such as cold and pneumonia and damages the lung tissue. Repeated exposure during childhood may result in reduced lung function in adults.

This paper presents an application of mixture models applied to the forecast of ozone peaks in the city of Lyon (France). One day ahead prediction of such events allow public authorities to warn the population of the danger of outdoor exercises and activities in the early morning or in the evening. Voluntary emission reduction is encouraged, and it is also foreseen to take preventive measures of traffic limitation.

* Corresponding author. Tel.: + 33-3-44-234947; fax: + 33-3-44-234477; http://www.hds.utc.fr/ambroise.

E-mail addresses: ambroise@utc.fr (C. Ambroise), yves.grandvalet@utc.fr (Y. Grandvalet).

PII: S0304-3800(01)00399-4

Besides human experts, two types of predictor can be used: determinist air quality models on the one hand and statistical models on the other hand. Three-dimensional air quality models are based on meteorology and chemistry. The current state of pollution and the mathematical description of its evolution (e.g. by partial differential equations) are used to estimate the state at some future time. These models are extremely useful for explaining the phenomenon, but they are expensive to develop and to maintain (Rufeger and Mieth, 1998). Indeed, since ozone formation is a complex process, long studies are required for building a model at a new site, and powerful computers are needed to run in reasonable time. Statistical models can be more or less complex, but are always easier to develop. They are based on the detection of some regular patterns before pollution events and require thus several years of monitoring records.

Many different statistical approaches have been suggested to predict ozone peaks. The simplest one consists in using a combination of threshold values among selected variables to forecast an excess or non-excess of ozone concentration (EPA, 1999). This basic method exhibits a moderate prediction accuracy and needs an expert for selecting the relevant variables. A decision tree is a more sophisticated tool; it automatically produces rules which are easy to interpret, but which may also be less accurate than a black-box models decision rule (Burrows, 1999). Artificial neural networks are increasingly popular in this domain (Garnder and Dorling, 1998; Prybutok et al., 2000; Yang et al., 2000). They require some statistical expertise, but often prove to be superior to more conventional statistical models such as Box-Jenkins ARIMA, multiple linear regression or discrimination.

In this paper, we illustrate the application of mixture models to forecast ozone pollution. This statistical modelling provides an effective handling of missing data. Moreover, recent developments ease its use in discrimination (Ambroise and Govaert, 2000) and improve its robustness to outliers and model misspecification (Ambroise et al., 2000).

Section 2 describes the application from the data analyst point of view. The problems are introduced, and the grounds for choosing mixture models are presented. Section 3 gives the outline of the methodology proposed here for the design of an automatic forecast based on mixture models. The results obtained are reported in Section 4.1, followed by a conclusion on the general applicability of the method in Section 5.

2. Problem description

2.1. Historical data

The urban area of Lyon, France, has more than one million inhabitants. It is situated on the Rhue river, in a narrow valley between low mountains (1000 m) and the alps. Ozone pollution is a real concern since this valley is an important traffic axis where industrial facilities abound.

Five years of historical data (1994–1998) have been made available by the local urban air control authority COPARLY. A first subset of variables describes the current global meteorological situation of the whole urban area and its predicted state for the next day, it comprises:

- the temperature profile (one measurement per hour) for the current day;
- the ground wind speed and wind direction measurements for the current day;
- the forecasted maximum temperature for the following day;
- a local meteorological classification, the Benichou class, measured at midday, its forecast for the current day at midnight, and for the day after at midday.

The Benichou class is a surrogate for the cloud cover information which is not available. No indicator of the vertical mixing nor faithful wind prediction are available at present. The other subset of variables are profiles (one measurement per hour) recorded by each monitoring station:

- seven sensors measuring ozone;
- 17 sensors measuring NO and NO_2.

The monitoring network has evolved in the past years. The number of monitoring stations has increased (from 2 to 7 for ozone, and 14 to 17 for

NO_x). With the present monitoring network, up to 1012 attributes are available each day, but the changes in the monitoring network and occasional sensor failures result in a data set having many missing attributes.

The decision label (polluted/non-polluted event) is also included in the records. It is based on the order of the prefect (a local French state authority), whose simplified statement reads as follows: a day is considered to be polluted if there are hourly exceedings of the 180 μg/m^3 threshold for at least two monitoring stations within 3 hours. In the sequel, the prediction of pollution according to this statement will be referred to as problem 1.

Although ozone values at different monitoring stations are highly correlated, the order statement considers more days to be polluted as the monitoring network grows. In accordance with the air control authority, another problem definition was used to check the predictability of the local ozone concentration, independently of the monitoring network evolution. A day is considered to be polluted if there is an hourly exceeding of the 180 μg/m^3 threshold at a central monitoring station (Saint-Just) operating since 1994. The prediction of this exceeding will be referred as problem 2. For both problems, the pollution forecasts for the next day are required by 14:00 h.

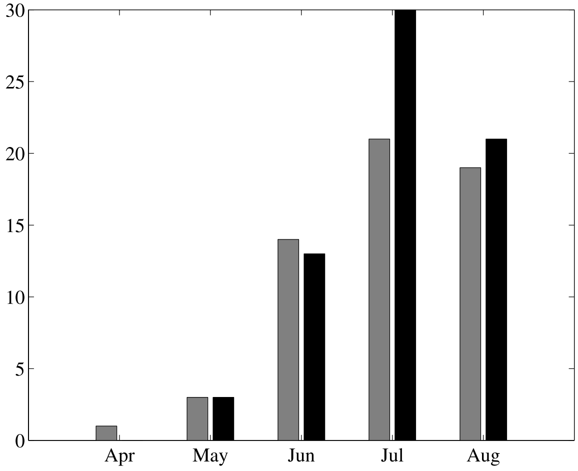


Fig. 1. Polluted events occurrence during 1994–1998, according to the order of the prefect (grey) and to the threshold exceeding on Saint-Just monitoring station (black).

From a data analyst point of view, the two problems differ significantly. Over the whole 1994–1998 period, the two classifications contradict each other 20 times (to be compared with about 60 polluted events): the exceeding of the 180 μg/m^3 threshold at Saint-Just does not cause the prefectural alarm to start off for 15 days, while the order of the prefect detects a polluted event without exceeding at Saint-Just for 5 days. Note that the numerical prediction of ozone would be even more informative, but the prediction of maximum numerical values is difficult. Particularly, it often under-predicts pollution as it is highly biased downwards (EPA, 1999). It is thus likely that the direct prediction of the threshold exceeding is more effective than the prediction based on an estimated maximum value.

The database comprises about 900 days, corresponding to the April–September period of 1994–1998. This reduction is due to the absence of polluted events during winter. The barplot in Fig. 1 displays the number of pollution events per month during this period. There are, respectively 58 and 67 polluted days for problems 1 and 2. They occur mainly during June to August which are the hottest months. May and September are cooler, but the long days in May favor ozone formation. Note that the weekday (not displayed here) is not a discriminant variable for any of the two problems.

For evaluation purposes, the database is split in two parts. The first 4 years are used for training, and the last year is dedicated to testing. The representativeness of test results is questionable since testing is done over a single year with only 11 or 12 pollution episodes, but further data (1999–2000) should become available soon.

2.2. Particularities

Compared to a standard discrimination problem, the data set has many unusual characteristics. Regarding the number of attributes and the sample size:

- there are numerous correlated attributes, and their number grows with the monitoring network;
- the polluted events are rare.

The number of attributes calls for variable selection. With more attributes than cases, most black-box models overfit. The strong correlations between attributes are due to the temporal and spatial dependences between several measurements of the same quantities at close time intervals or geographic distances. Temporal redundancy is used in the two-step variable selection procedure described in Section 3.3. Taking into account geographical redundancy is more difficult since it comes with time shifts. Regarding the unbalanced data set, special attention should be paid to the costs associated with misprediction. This point is detailed in Section 3.2.

Regarding the type of information conveyed by each example:

- some variables are quantitative (e.g. temperatures, ozone concentrations …) and other important ones are qualitative (Benichou classes);
- some data are not reliable, as shown by a few major errors;
- some attributes and class labels are missing;
- some measurements are not available in the 1st years pending the evolution of the monitoring network;
- the sensors and the communication network are subject to failure;
- for problem 2, the class label is missing when the ozone sensor does not provide faithful measurements the day after.

Many discrimination methods, such as nearest neighbors, kernel methods or radial basis functions (Ripley, 1996) classify according to some similarity measure between prototypes and the query example. The definition of a relevant similarity measure with qualitative variables can be difficult. Mixture models can be considered as a prototype method, but, with or without qualitative variables, the similarity is simply estimated by maximizing the likelihood. Outliers can be removed by analysing data prior to learning the model parameters, but an automatic treatment demands robustness, which can be obtained in mixtures by a simple modelling of measurement errors. Last but not least, missing labels are routinely processed by the EM algorithm which is used to optimize mixture parameters. Missing attributes are also taken into account by a slight modification of the original algorithm. All these points are detailed in Section 3.

3. Methodology

In the previous section, the prediction of ozone pollution is described as two discrimination problems. In that situation, the statistical learning theory (Vapnik, 1995) recommends to use an algorithm whose objective is to build a decision rule minimizing the error rate. Mixture models are generative models, based on density estimation. They estimate conditional densities from which a decision rule can be obtained by plugging the estimates in the Bayes rule. This approach is critisized by Vapnik (1995), as the intermediate problem of density estimation is more difficult than the problem of rule inference. However, as pointed out in the previous section, our problems are not in the standard form. We choose to use mixture models as they allow to cope directly with missing data and labels that would otherwise require complex pre-processing in direct rule estimation.

Mixture models are briefly presented. We then discuss the important role of decision costs in measuring the classifier accuracy. Finally, we present the variable selection scheme which is required for automatic rule formation.

3.1. Mixture models

Mixture models are a popular tool for density estimation. They are also used in clustering for modelling different groups within a heterogenous population, and in discrimination for modeling conditional class densities. We first recall the basics of mixture models in the density estimation framework, before explaining their use in discrimination.

3.1.1. Density estimation

In the mixture models setting (Titterington et al., 1985; McLachlan and Basford, 1989), data $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are assumed to arise independently from a random vector with density,

$$f(\mathbf{x};\boldsymbol{\Phi})=\sum_{k=1}^{K}p_k f_k(\mathbf{x};\theta_k), \tag{1}$$

where K is the number of components in the mixture, p_k are the mixing proportions ($0<p_k<1$ for all $k=1,\ldots,K$ and $\Sigma_k p_k=1$), $f_k(\mathbf{x};\theta_k)$ denotes a component, i.e. a probability distribution function parametrized by θ_k, and $\boldsymbol{\Phi}=(p_1,\ldots,p_{K-}1,\theta_1,\ldots,\theta_K)$ gathers all parameters.

Generally, the maximum likelihood estimation of this model cannot be obtained analytically, but learning Φ may be trivial if the component f_k responsible for the existence of each observation $\mathbf{x}_i$ is known. The classical approach to solve this problem is the EM algorithm (McLachlan and Krishnan, 1997), where the fitting problem is addressed by considering a dummy variable $\mathbf{z}_i=(z_{i1},\ldots,z_{iK})$ which flags the component responsible for observation $\mathbf{x}_i$. The 'missing' variable $\mathbf{z}_i$ can be estimated from $\mathbf{x}_i$ and $\boldsymbol{\Phi}$, while parameters Φ can be estimated from all $\mathbf{y}_i=(\mathbf{x}_i,\mathbf{z}_i)$, $i=1,\ldots,n$. This provides an iterative procedure for computing the maximum likelihood estimate of $\boldsymbol{\Phi}$.

Starting from an initial parameter $\boldsymbol{\Phi}^0$, an iteration of the EM algorithm computes the parameter $\boldsymbol{\Phi}^{q+1}$ which maximizes the expectation $Q(\boldsymbol{\Phi}/\boldsymbol{\Phi}^q)=\mathrm{IE}[L(\boldsymbol{\Phi};\mathbf{y})|\mathbf{x},\boldsymbol{\Phi}^q]$, where L is the log-likelihood of $\boldsymbol{\Phi}$. This qth iteration is defined as follows:

- *Expectation step:* Compute the expectation $Q(\boldsymbol{\Phi}/\boldsymbol{\Phi}^q)$ where the unknown quantities are the probabilities

  $$t_k^q(\mathbf{x}_i)=\frac{p_k^q f_k(\mathbf{x}_i;\theta_k^q)}{f(\mathbf{x}_i)},$$

  using $\boldsymbol{\Phi}^q=(p_1^q,\ldots,p_{K\,1}^q,\theta_1^q,\ldots,\theta_K^q)$.
- *Maximization step:* Calculate $\boldsymbol{\Phi}^{q+1}$ which maximizes

  $$Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^q)=\sum_{i=1}^{n}\sum_{k=1}^{K}t_k^q(\mathbf{x}_i)\log p_k f_k(\mathbf{x}_i|\theta_k).$$

This simple re-estimation scheme converges towards a stationary point of the likelihood of $\boldsymbol{\Phi}$.

3.1.2. Estimation of posterior class probabilities

When mixture models are used in discrimination, each class density is modelled by a mixture of simple components. It is thus assumed that the observations arise from a stratified mixture model, which is itself a mixture model. An algorithm dedicated to this type of hierarchical model uses partially known components labels to estimate all component parameters in parallel (Ambroise and Govaert, 2000). It provides a proper handling of missing labels by EM in the context of discrimination.

Robustness with respect to outliers is treated along the same way by adding a special 'noise' component. The latter gathers all atypical observations, and is modelled by a proper or improper uniform distribution. Its robustifying effect is due to the removal of outliers in the estimation of the parameters of class densities components. The observations mixture density function is expressed as:

$$f(\mathbf{x};\boldsymbol{\Phi})=\sum_{k=1}^{K_1}p_k f_k(\mathbf{x};\theta_k)+\sum_{l=1}^{K_2}p_l f_l(\mathbf{x};\theta_l)+p_\eta f_\eta(\mathbf{x}), \tag{2}$$

where

- k indexes the components of the mixture modelling P 'the pollution class';
- ℓ indexes the components of the mixture modelling $\bar{P}$ 'the normal class';
- $f_\eta(\mathbf{x})$ is the density of N the 'noise component'.

When an observation is labelled by P or $\bar{P}$, it is partially labelled, respectively by (P or N) or ($\bar{P}$ or N) to account for possible outliers. Only a subset of components is admitted to be correct. When an observation is not labelled, all components compete to explain it. This processing is dealt with very easily using the EM algorithm with partially known labels (Ambroise and Govaert, 2000).

When all variables are quantitative, the normal and pollution classes are modelled by mixtures of simple continuous distributions (typically Gaussian). Missing attributes are easily handled by assuming independence of variables within (conditionally to) mixture components. This assumption does not mean that variables are independent, as complex dependences can be accurately modelled by a mixture of simple components.

When qualitative variables are included, the mixture is defined analogously on the product space of qualitative × quantitative variables. The observation vector is thus split in two parts $\mathbf{x}_i = (\mathbf{r}_i, \mathbf{s}_i)$ where $\mathbf{r}_i$ and $\mathbf{s}_i$ denote, respectively the qualitative and quantitative variables. Assuming again independence among variables within each component, the densities are written as:

$$f_k(\mathbf{x}_i;\, \theta_k) = f_k(\mathbf{r}_i;\, \theta_{kr}) \cdot f_k(\mathbf{s}_i;\, \theta_{ks}), \tag{3}$$

where $f_k(\mathbf{r}_i;\, \theta_{kr})$ is a product of parameterized multinomial distributions.

Suppose that C of the $C+D$ feature variables in $\mathbf{x}_i = (\mathbf{r}_i, \mathbf{s}_i)$ are categorical, where the cth categorical variable takes on m_c $(c = 1, \ldots, C)$ distinct values. With the assumed multinomial model, let us consider that $\mathbf{r}_i^{cj} = 1$ if the realization of the cth categorical variable in $\mathbf{x}_i$ corresponds to the jth pattern. Let π_{kcj} be the probability that $\mathbf{r}_i^{cj} = 1$ given its membership of the kth component of the mixture.

Suppose that D of the $C+D$ feature variables are quantitative. $\mathbf{s}_i$ is thus a vector in R^D and $f_k(\mathbf{s}_i;\, \theta_{ks})$ is a multivariate Gaussian density with adjustable mean μ_k and covariance matrix $\Sigma_k = \sigma_k I$.

Similarly, when all variables are quantitatives, the density modelling outliers is a uniform continuous distribution over the smallest hyper-rectangle, $\Re$, including all observations.

A binary vector $\mathbf{z}_i \in \{0, 1\}^{(K1+K2+1)}$ indicates the components from which feature vector $\mathbf{x}_i$ may have been generated, where $\mathbf{z}_{ik} = 1$ means that $\mathbf{x}_i$ *may have been* generated by component k whereas $\mathbf{z}_{ik} = 0$ means that $\mathbf{x}_i$ *was not* generated by component k.

Another binary vector $\mathbf{h}_i \in \{0, 1\}^{C+D}$ indicates missing variables, where $\mathbf{h}_{id} = 0$ means that the dth variable of vector $\mathbf{x}_i$ is missing.

Handling the missing variables as described in Dang (1998), the EM algorithm for estimating the parameters of the proposed hierarchical mixture model, iterates as follows:

E-step: Computation of the posterior probabilities of the components:

$$t_{ik}^{q} = \frac{z_{ik} p_k^q \prod_{c=1}^{C} \prod_{j=1}^{m_c} (\pi_{kcj}^q)^{h_{ic} r_i^{cj}} f_k^{\{o\}}(\mathbf{s}_i;\, \mu_k^q,\, \sigma_k^q)}{\sum_k z_{ik} f_k(\mathbf{x}_i;\, \Phi^q)},$$

where $f_k^{\{o\}}(\mathbf{s}_i)$ is the conditional density of the observed (i.e. non-missing) quantitative variables.

M-step: Maximisation of $Q(\Phi|\Phi^q)$:

$$n_{kd}^{(q+1)} = \sum_i h_{id} t_{ik}^q,$$

$$\pi_{kcj}^{q+1} = \frac{\sum_i t_{ik}^q h_{ic} r_i^{cs}}{\sum_i \sum_{s=1}^{m_c} t_{ik}^q h_{ic} r_i^{cs}},$$

$$p_k^{(q+1)} = \frac{\sum_i t_{ik}^q}{\sum_i \sum_k t_{ik}^q},$$

$$\mu_{kd}^{(q+1)} = \frac{1}{n_{kd}^{(q+1)}} \sum_i t_{ik}^q h_{id} s_{id},$$

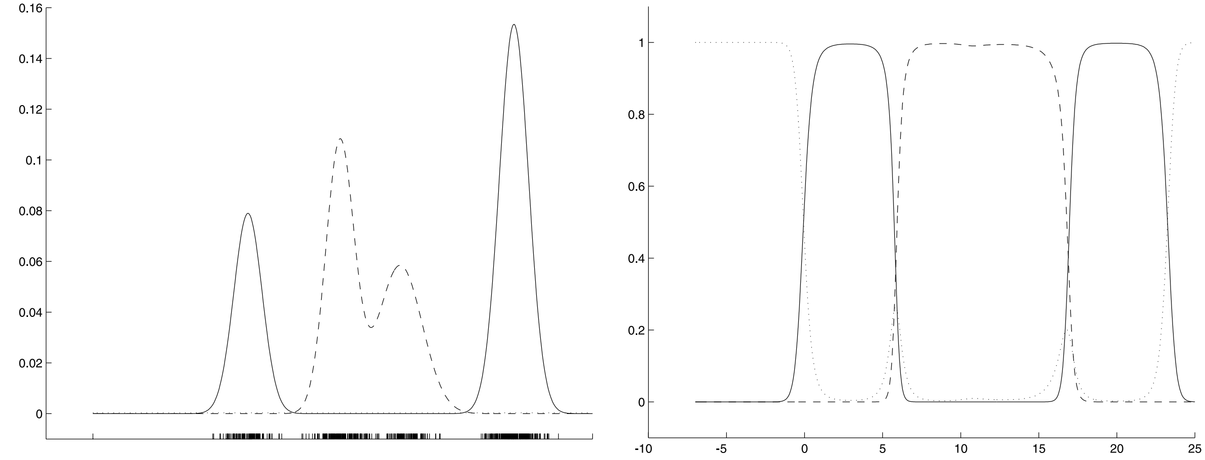
$$(\sigma_k^{(q+1)})^2 = \frac{1}{\sum_d n_{kd}^{(q+1)}} \sum_i \sum_d t_{ik}^q h_{id} (s_{id} - \mu_{kd}^{(q+1)})^2.$$

Once the mixture parameters are identified, the derivation of the posterior class probabilities by Bayes rule is straightforward. For example, the posterior probability of the pollution class ω_P is

$$P(\omega_P|\mathbf{x}) = \frac{\sum_{k=1}^{K_1} p_k f_k(\mathbf{x};\, \theta_k)}{f(\mathbf{x};\, \Phi)}, \tag{4}$$

where $f(\mathbf{x};\, \Phi)$ is computed by summation over the pollution, non-pollution, and outlier classes, as shown in Eq. (2).

This parametrization allows class densities and posterior class probabilities to have complex shapes: a one-dimensional example is given in Fig. 2 which illustrates a simple mixture of two classes with two Gaussian components each. The 'noise component' density is small compared to other class densities (hardly visible on Fig. 2a), but its corresponding posterior probability catches well extreme values which can be considered as outliers.



(a)

(b)

Fig. 2. (a) Hierarchical mixture with three high-level components: dashed, the central class is made of two Gaussian sub-components; solid, the outer class comprises also two Gaussian clusters; dotted, the 'noise component' is made of one uniform component. Ticks on the abscissa represent observations. (b) Respective posterior class densities.

3.2. Discrimination

The Bayesian decision theory recommends the decision associated with smaller risk. The risk R is computed as:

$$R(d|\mathbf{x}) = C(d, \omega_P)\, P(\omega_P|\mathbf{x}) + C(d, \omega_{\bar{P}})\, P(\omega_{\bar{P}}|\mathbf{x}), \tag{5}$$

where $C(d, \omega)$ is the loss incurred when deciding d (predicting a pollution state) when the truth is ω (the pollution state tomorrow).

Once posterior probabilities are estimated, a decision rule is obtained by plugging them in the Bayes risk. In our two-class problems, there are three possible decisions: 'tomorrow is predicted polluted', 'tomorrow is predicted not polluted', 'tomorrow is not predicted', respectively denoted d_P, $d_{\bar{P}}$ and d_0. The last (absence of) decision is often referred to as the reject option. It can be useful when several predictors are available, and that they have to be combined to perform a final prediction. Hereinafter we only consider the two first decisions.

It is thus straightforward to build the decision rule provided $C(d, \omega)$ are available. The losses associated with correct decisions are usually set to zero, but what should be the ones associated with misclassification–classification? In the absence of effective losses provided by the end-user (what is the economic/sociologic costs of misclassification?), the classical solution considers that all errors are equally wrong, so that all misclassification losses should be equal (set to one). For this solution, the risk is equal to the probability of misclassification: $R(d_P|\mathbf{x}) = P(\omega_{\bar{P}}|\mathbf{x})$ and $R(d_{\bar{P}}|\mathbf{x}) = P(\omega_P|\mathbf{x})$.

This solution is usually not realistic in diagnosis applications where the fault class has low prior probability. For example, let us consider two extreme decision rules predicting systematically pollution and no pollution. None of these rule conveys any information, but according to the error rate, the rule predicting no pollution is much better than the rule predicting pollution because non-polluted events are rare. For the two rules to be identically bad, $C(d, \omega)$ should be inversely proportional to the probability of observing ω. This extreme unbalancing would neither be fair for less rudimentary classifiers, since the features $\mathbf{x}$ hopefully conveys some information related to the class.

In the absence of an objective criteria to compare solutions, it is interesting to display how a predictor performs for a wide range of possible losses. Whatever the misclassification losses may be, for a two-class problem without reject option, the decision rule is a threshold on the ratio $P(\omega_P|\mathbf{x})/P(\omega_{\bar{P}}|\mathbf{x})$. All possibilities can be explored by setting $C(d_{\bar{P}}, \omega_P) = 1$, and letting vary $C(d_P, \omega_{\bar{P}})$ from 0 to ∞ (i.e. from systematic alarm to systematic non-detection). The graph of all these solutions, the so-called Receiver Operating Characteristic ROC curve is always provided hereinafter when posterior probabilities are estimated.

3.3. Variable selection

It is easily shown that the optimal performance of a classifier can only improve as the number of features increases. However, it is more and more difficult to get close to this optimal performance when only a finite sample is available for training: adding unnecessary features makes it more difficult to retrieve the discriminant information. In our application, only 900 examples, described by 1012 features, are given to build a classifier. If all variables are retained, there is an infinity of linear classifiers performing zero error, whatever the label may be. It is likely that none of these perfect discriminator perform well on test data. The discrimination problem has thus to be constrained in some way.

There are two archetypal means of constraining a classifier: by penalization or by subset selection among parameters or variables. In this application, our belief is that many features are useless because of high correlation between variables and low correlation with class labels. We choose thus to select variables from the original set.

The ultimate aim of subset selection is to find the best features, which yield the highest classifier accuracy. This goal is not achievable, first because the true classification cost cannot be computed with a finite training sample, and second because searching for the best subset is an inherently combinatorial procedure. In practise, we have to resort to sub-optimal search strategies based on the minimization of some estimate of classification cost. Moeover, as the classification cost is not well defined in our application, we choose to minimize the cross-entropy. This criterion is the log-likelihood of the estimated posterior probabilities, and is independent of the misclassification losses:

$$E_{\log} = \sum_i \sum_j c_{ij} \log \hat{P}(\omega_j|x_i), \tag{6}$$

where c_{ij} is a dummy variable coding the class label for features $\mathbf{x}_i$ (if $\mathbf{x}_i$ is labelled class j, then $c_{ij} = 1$ and $c_{ik} = 0$ for $k \neq j$, if $\mathbf{x}_i$ is not labelled, then $c_{ik} = 0$ for all k), and $\hat{P}(\omega_i|\mathbf{x}_i)$ is the mixture model estimate of posterior probabilities. The mean cross-entropy is estimated by five-fold cross-validation for each subset of variables.

Regarding computation, all possible subsets cannot be considered. We use a classical forward search. This iterative algorithm starts with an empty set of features, and adds at each step the feature providing the best improvement of the criterion. Termination occurs either when a prespecified number of variables is reached, either when adding any feature results in a criterion increase. The latter solution is used here.

Even with forward search, the selection process is too time consuming due to the huge amount of variables. We thus proceed in two steps:

1. a first forward selection is applied to the subset of all measurements of ozone, NO or NO_2 of one sensor;
2. all the subsets are gathered and a new forward selection is performed among the variables selected in the first step.

The overall search process is acknowledgedly sub-optimal, but it is hoped that the loss is not too important thanks to the high correlations between features. Note finally that it would also be possible to carry out feature selection by using standard criteria of linear separability, but the best set of features depends not only on the problem but also on the classifier.

4. Results

This section illustrates the application of the above methodology to the two discrimination problems described in Section 2. We first present results obtained with the persistence strategy and linear discrimination before evaluating the mixture models based approach.

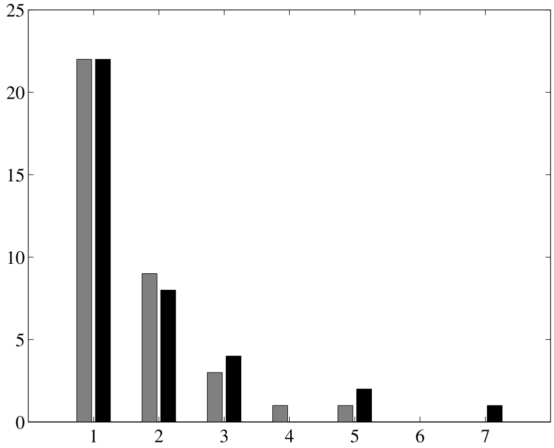


Fig. 3. Length of pollution episodes (in days) during 1994–1998, according to the order of the prefect (grey) and according to the threshold exceeding on Saint-Just monitoring station (black).

4.1. Persistence method

The persistence method consists in forecasting no change in the pollution state. This simple scheme provides a baseline for judging if the improvement in forecast accuracy provided by more complex methods is worth the development efforts. It performs well if pollution occurs on several consecutive days. Fig. 3 shows the length of pollution episodes for both problems. About one-third of pollution events last only 1 day, and long episodes are rare, so that the method is not supposed to be very efficient.

The persistence method is easy to handle, but suffers from at least two severe shortcomings. First, it cannot predict correctly the beginning and end of pollution episodes. Second, it requires an assessed pollution state. Here, the forecast should be done at 14:00 h. For problem 2, the correct decision for the current day can be obtained by 14:00 pm, since the first exceeding of the 18 μg/m^3 threshold of a polluted day always appears before 14:00 h. The pollution state for the next day is predicted identical to the pollution state on the current day. However, for problem 1, the decision for the current cannot always be made by 14:00 h, such that the persistent method has to use the decision of the day before.

Results for both problems are reported on Tables 1 and 2. There is no learning in the persistence method, hence all data are test data. Nevertheless, results are displayed for 1994–1997 and 1998 to allow future comparison with mixture models.

Table 1
Persistence method for problem 1

Observed	1994–1997 forecasted		1998 forecasted	
	Pollution (%)	Normal (%)	Pollution	Normal
Pollution	9 (1.2)	37 (5.1)	4 (2.2)	8 (4.4)
Normal	37 (5.1)	641 (88.5)	8 (4.4)	161 (89.0)

Results are given in days with the corresponding percentage in parentheses.

Table 2
Persistence method for problem 2

Observed	1994–1997 forecasted			1998 forecasted		
	Pollution (%)	Normal (%)	Missing (%)	Pollution (%)	Normal (%)	Missing (%)
Pollution	23 (3.2)	27 (3.7)	2 (0.3)	7 (3.8)	8 (4.4)	0 (0.0)
Normal	27 (3.7)	598 (82.1)	17 (2.3)	8 (4.4)	157 (86.3)	1 (0.5)
Missing	2 (0.3)	16 (2.2)	16 (2.2)	0 (0.0)	1 (0.5)	0 (0.0)

Results are given in days with the corresponding percentage in parentheses.

Table 3
Linear discrimination for problem 1

Observed	1994–1997 forecasted		1998 forecasted	
	Pollution (%)	Normal (%)	Pollution (%)	Normal (%)
Pollution	26 (3.6)	20 (2.7)	8 (4.4)	4 (2.2)
Normal	41 (5.6)	645 (88.1)	7 (3.8)	164 (89.6)

Results are given in days with the corresponding percentage in parentheses.

Table 4
Linear discrimination for problem 2

Observed	1994–1997 forecasted		1998 forecasted	
	Pollution (%)	Normal (%)	Pollution (%)	Normal (%)
Pollution	36 (4.9)	16 (2.2)	8 (4.4)	7 (3.8)
Normal	41 (5.6)	604 (82.5)	9 (4.9)	158 (86.3)
Missing	6 (0.8)	29 (4.0)	0 (0.0)	1 (0.5)

Results are given in days with the corresponding percentage in parentheses.

The sensor failures are responsible for the missing observations and forecasts for problem 2. The error rates on the whole historical data during the 1994–1998 period for problem 1 and 2 are, respectively 9.9 and 7.7%. This difference is a logical consequence of shorter pollution episodes for problem 1, and also of the increased difficulty to make forecast on the basis of the former decision. The symmetry between false alarms and non-detection arises from the fact that each beginning of a pollution episode is responsible for one non-detection, and the end is responsible for one false alarm.

Notice finally that, in terms of error rates, both persistent models are not better than a systematic non-detection. The error rate of the latter is the proportion of polluted days. On the whole 1994–1998 period, it is as low as 6.4 and 7.6% for problems 1 and 2, respectively. However, we believe that the persistence model, rough as it may be, is still more informative than a model systematically predicting no pollution, and that, as discussed in Section 3.2, the error rate is not an appropriate measure of accuracy in this application. However, as the persistence method provides binary decisions, its solution is not sensitive to misclassification losses. Thus, the ROC curve is not displayed since it is made of just one point.

4.2. Linear discrimination

We use here a version of linear discrimination dealing with missing variables, on the same samples than persistence and mixture models, so that results can be directly compared. Linear discrimination using all variables results in a poor performance since the number of feature is about as great as the number of examples. Consequently, we use a reduced subset of variables selected by means of expert knowledge and exploratory data analysis. The following variables are selected:

- maximum value of the Saint-Just ozone sensor before prediction (between 0:00 and 14:00 h);
- forecasted maximum temperature for the next day;
- wind speed measured at 8:00 h;
- forecast of the Benichou class at 24:00 h on the next day.

Results for the two problems are reported on Tables 3 and 4. Although the answers to problem 1 seem to be slightly more accurate than the one given by persistence, error rates are still higher to the ones obtained by systematic non-detection. For problem 2, there is no sign of improvement over the persistence model.

4.3. Mixture model

Considering the two problems, two different strategies are tested. First we try a fully automatic process where the variables used for classification are automatically selected as explained in Section 3.3 and second we test the classification procedure with the variables selected by experts.

The two forward selection phases estimate 950 of the 1012 input variables as irrelevant for classification problem 1. They are thus discarded from the training sample (years 1994–1997) and the parameters of the mixture models are estimated with the remaining 62 variables.

A mixture of three components for the pollution class and four components for the non-pollution class is chosen. The obtained results are comparable to the performance of the persistence method. We assume that this poor performance is a consequence of an under-selection of the explicative variables. Indeed adding unnecessary features makes it more difficult to retrieve the discriminative information.

Table 5
Mixture model for problem 1

Observed	1994–1997 (learning) forecasted		1998 (test) forecasted	
	Pollution (%)	Normal (%)	Pollution (%)	Normal (%)
Pollution	18 (2.5)	28 (3.8)	6 (3.3)	6 (3.3)
Normal	7 (1.0)	679 (92.8)	1 (0.6)	170 (92.9)

Results are given in days with the corresponding percentage in parentheses. The decision function is built up using equal losses for a false alarm and an absence of detection.

Table 6
Mixture model for problem 1

Observed	1994–1997 (learning) forecasted		1998 (test) forecasted	
	Pollution (%)	Normal (%)	Pollution (%)	Normal (%)
Pollution	27 (3.7)	19 (2.6)	7 (3.8)	5 (2.7)
Normal	22 (3.0)	664 (90.7)	5 (2.7)	166 (90.7)

Results are given in days with the corresponding percentage in parentheses. The decision function is built up using a false alarm loss of one, and an absence of detection loss of two.

Table 7
Mixture model for problem 2

Observed	1994–1997(learning) forecasted		1998 (test) forecasted	
	Pollution (%)	Normal (%)	Pollution (%)	Normal (%)
Pollution	31 (4.2)	21 (2.9)	9 (4.9)	6 (3.3)
Normal	15 (2.0)	630 (86.1)	5 (2.7)	162 (88.5)
Missing	4 (0.5)	31 (4.2)	0 (0.0)	1 (0.5)

Results are given in days with the corresponding percentage in parentheses. The decision function is built up using equal losses for a false alarm and an absence of detection.

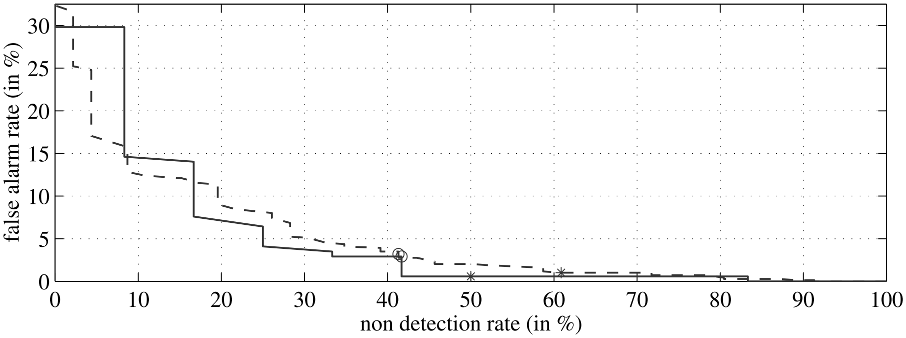


Fig. 4. ROC curves in training (dashed line) and test (solid line) for problem 1. The star represents the operating points for equal misclassification costs, and the circles the operating points when an absence of detection costs twice as much as a false alarm.

The results for the mixture models using expert knowledge (Tables 5–7) can be compared for reference to the persistence method (Tables 1 and 2) and linear discrimination (Tables 3 and 4). Table 5 displays the results obtained when optimizing the error rate. They are significantly improved in comparison with the simpler linear discriminant and the persistence model. The total error rate decreases from 8.8 down to 3.9% for the test set. Note that the results are almost identical on the training set (with a decrease from 10.2 down to 4.8%), and that the improvement is especially noticeable with regard to the false alarm rate.

Table 5 assumes that the two type of errors, absence of detection and false alarm are equally weighted. But, opposed to direct rule estimates like the persistence method, the modelling of posterior probabilities allows some flexibility (see Section 3.2): the relative importance of the two types of errors can be adjusted without reestimation of the model parameters. For example, if an end-user considers that an absence of detection costs twice as much as a false alarm, it leads to a different distribution of errors as presented in Table 6. In that case, the cost specification allows to decrease the non-detection rate from 3.3 down to 2.7% but this has disadvantage to increase the false alarm rate from 0.6 to 2.7% (the error rate on the training set is increased from 4.8 up to 5.6%, but the averaged weighted misclassification cost is decreased from 8.6×10^{-2} down to 8.2×10^{-2}).

Fig. 4 shows the Receiver Operating Characteristic curve, which displays all possible errors produced by varying the misclassification losses. The non-detection rate is plotted versus the false alarm rate. Using ROC curves, the end-user can have a global perception of possible classifier performances, and choose the best operating point according to his subjective criterion. For example, the two operating points corresponding to the decision described by Tables 5 and 6 are, respectively represented in Fig. 4 by a star and a circle.

It is usually expected that test results degrade compared to training set performances, but we observe that, up to the quantification effect due to the small number of polluted events during 1998, the training and test curves are very close. This phenomenon may be explained either by the fact that the pollution episodes that happened during 1998 are easier to forecast than the ones which occurred during 1994–1997, or by the fact that the number of free parameters of the model is small enough to avoid overfitting the training data.

Using the same approach for problem 2 (prediction of ozone concentration exceeding at Saint-Just) leads to qualitatively similar results to those the ones obtained for problem 1. Error rates obtained by mixture models (Table 7) are improved compared to the ones of persistence method (Table 2), even if the absence of decision of the latter is considered as a correct decision. Then, the total error rate decreases from 8.8 down to 6.0% for the test set (and from 7.4 down to 4.9% for the training set).

Notice that, as for linear discrimination, we obtain slightly better results for problem 1, but the difference is small and may not be significant. This absence of significance is confirmed by Fig. 5, which is similar to the one obtained on the first problem. It seems strange that the well defined prediction problem 2 does not provide better results than the prediction of pollution according to the order of the prefect, but it is known to the experts that the one-day-ahead prediction of local fluctuations of ozone tend to be more difficult than the prediction of a global state for a whole urban area.

A better insight into the way the classifier works can be gained by considering couples of variables. Class and posterior densities can be computed for two variables by considering all other as missing. Fig. 6 shows different decision boundaries (Fig. 6b) computed using the same posterior densities for the two classes (Fig. 6a). Notice that the shape of the projected decision boundaries appear quadratic even if the kind of mixture models used for building this boundary is able to produce much more complex shapes. Changing the cost functions does not greatly modify the shapes of the boundary but is more

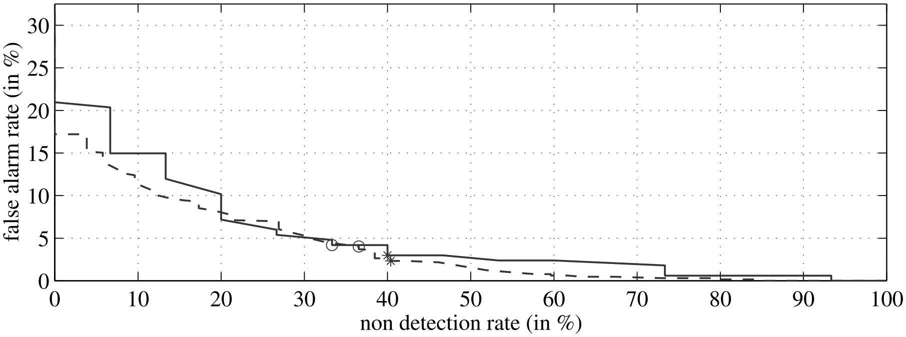


Fig. 5. ROC curves in training (dashed line) and test (solid line) for problem 2. The star represents the operating points for equal misclassification costs, and the circles the operating points when an absence of detection costs twice as much as a false alarm.

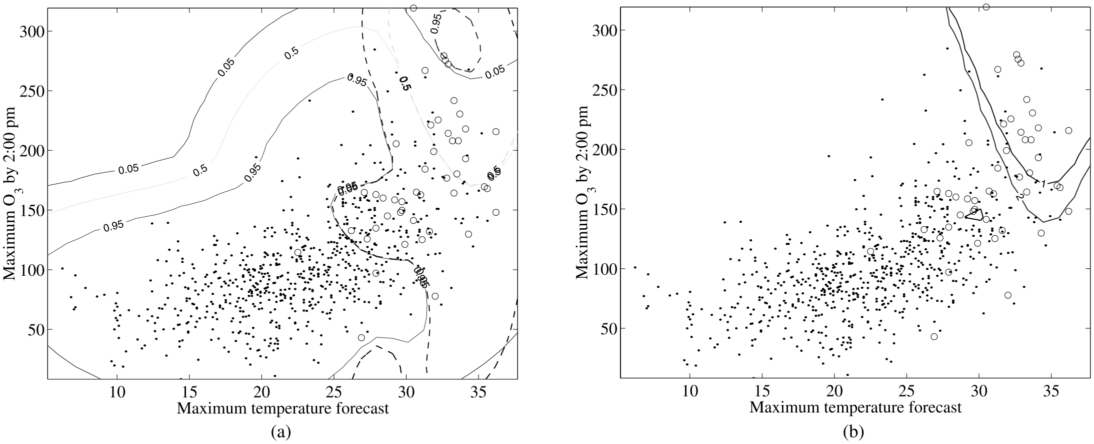


Fig. 6. (a) Projection of the posterior class densities of pollution (solid lines) and non-pollution (dashed lines) in the plane determined by the sole knowledge of forecasted temperature and measures of ozone concentration at Saint-Just the day before the forecasting. (b) Decision boundary for different cost functions. The lower curve corresponds to a false alarm loss of one, and an absence of detection loss of two. The higher curve represents the decision boundary for equal losses.

related to a translation of the decision boundary. Considering this translation allows to understand why a decrease of the absence of detection rate implies an increase of the false alarm rate.

5. Conclusion

This paper presents the application of mixture models to two prediction problems of ozone peaks in the city of Lyon. In both problems, the forecast for the day after is required at 14:00 h, and ozone pollution is defined by exceedings of the 180 $\mu g/m^3$ threshold.

The first problem consists in forecasting pollution according to the order of the prefect, which aims at rendering the pollution state of the whole urban area. The second problem considers only a local pollution state, at one sensor location.

The results are difficult to summarize because there is no obvious criterion for judging of the classifiers accuracy since no objective misclassification loss exists. Moreover, once misclassification losses are fixed, the results seem to be highly variable. The last point is illustrated by the important difference in error rates between 1994–1997 and 1998 for the persistence method for which there is no learning.

We can nevertheless conclude that automatic variable selection is not successful in discarding enough irrelevant features. This failure is due to the very high difficulty of the task. The number of examples in the training set (about 700) is too low compared to the number of initial features (1012) so that some spurious correlation appear between some features and the correct prediction by pure chance. The problem is even more difficult than the above figures suggest, as the information content of many examples is very low: about one-third of examples can be classified as non-polluted without doubt, by only looking at the temperature forecast.

As a consequence of the failure of automatic variable selection, the mixture models are doing rather poorly. The selection procedure should be regarded as a means to select a first subset of potential 'good' variables which may pass the examination by a data analyst. For example, the wind speed at 8:00 h was selected in that way. According to the air quality expert, it has no direct connection with tomorrow's ozone concentration. However, its analysis shows that it is more correlated to tomorrow's wind speed than any other variable, and tomorrow's wind speed is directly connected to ozone pollution.

In both problems, and for any misclassification losses, mixture models with 'expert features' always do better than systematic alarm, systematic non-detection, the persistence model and linear discrimination. The results may however appear disappointing, but we believe that the slightness of improvements is due to the lack of relevant meteorological predictions in the features.

If not reliable for a fully automated prediction, mixture models are well adapted when relevant features have been selected. The approach offers the advantage of properly handling the missing labels and data. This handling is beneficial for learning, since incomplete examples can enter the training set, and is essential for predicting in the absence of some features.

Besides wind speed, the most informative variables are the forecasted temperature and the forecasted Benichou class, which is a surrogate for the cloud cover forecast. Present ground level ozone concentrations also improve the prediction, but all other available variables seem to be detrimental for our models. The present surface air quality plays a marginal role in tomorrow's air quality, and meteorological conditions explain most of the day-to-day changes in ozone concentration. Indeed the two meteorological forecasts are the two main discriminatory features. This observation is backed-up by previous studies using different models in other cities EPA (1999). We thus conclude that, for expecting significant improvements in prediction, efforts should rather be directed towards better meteorological predictions (regarding cloud cover, surface wind and vertical mixing) than towards more complex classifiers.

Acknowledgements

This work was initiated in the work team 'Prévision des pics de pollution' organized by CRIN. The authors would also like to acknowledge partial funding from the GdR ISIS within project 'Méthodes ensemblistes pour la surveillance de l'Environnement'.

References

Ambroise, C., Govaert, G., July 2000. EM algorithm for partially known labels. In: IFCS 2000.

Ambroise, C., Govaert, G., Denoeux, T., 2000. Développement d'un logiciel temps réel pour le contrôle par émission accoustique des appareils à pression. Tech. Rep. 1.8.1333, CETIM.

Burrows, W., 1999. Combining classification and regression trees and the neuro-fuzzy inference system for environmental data modeling. In: 18th International Conference of the North American Fuzzy Information Processing Society, NAFIPS, IEEE.

Dang, V., 1998. Classification de données spatiales: modèles probabilistes et critères de partitionnement. Ph.D. thesis, Université de Technologie de Compiègne.

EPA, July 1999. Guideline for developing an ozone forecasting program. Tech. Rep. EPA-454/R-9-009, United States Environmental Protection Agency.

Garnder, M., Dorling, S., 1998. Artificial neural networks (the multilayer perceptron) a review of applications in the atmospheric sciences. Atmospheric Environment 32, 2627–2636.

McLachlan, G., Basford, K., 1989. Mixture Models. Inference and Applications to Clustering. Wiley.

McLachlan, G., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley.

Prybutok, V., Junsub, Y., 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 (1), 31–40.

Ripley, B., 1996. Pattern Recognition and Neural Networks. Cambridge University Press.

Rufeger, W., Mieth, P., 1998. The dymos system and its application to urban areas. Environmental Modelling and Software 13 (3–4), 287–294.

Titterington, D., Smith, A., Makov, U., 1985. Statistical Analysis of Finite Mixture Distributions. Marcel Dekker.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer Series in Statistics. Springer, New York.

Yang, P., Zhou, X., Bian, J., 2000. A nonlinear regional prediction experiment on a short-range climatic process of the atmospheric ozone. Journal of Geophysical Research 105 (D10), 12 253–12 258.