available at www.sciencedirect.com

www.elsevier.com/locate/scitotenv

Forecasting of ozone episode days by cost-sensitive neural network methods

*Che-hui Tsai, Li-chiu Chang, Hsu-cherng Chiang**

Department of Water Resources and Environmental Engineering, Tamkang University, Tamsui, Taipei-hsien, Taiwan

ARTICLE DATA

Article history:
Received 12 September 2008
Received in revised form
2 December 2008
Accepted 3 December 2008
Available online 20 January 2009

Keywords:
Photochemical air pollution
Statistical model
Imbalanced dataset classification

ABSTRACT

Forecasting the occurrence of ozone episode days can be regarded as an imbalanced dataset classification problem. Since the standard artificial neural network (ANN) methods cannot make accurate predictions of such a problem, two cost-sensitive ANN methods, cost-penalty and moving threshold, were used in this study. The models classify each day as episode or non-episode according to the standard of daily maximum 8 h O_3 concentration. The ozone measurements from six monitoring stations in Taiwan were used for model training and performance evaluation. Two different input datasets, regional and single-site, were generated from raw air quality and meteorological observations. According to the numerical experiments, the predictions based on the regional dataset are much better than those obtained from the single-site dataset. Two cost-sensitive ANN methods were evaluated by receiver operating characteristic (ROC) curves. It was found that the results obtained by the two approaches are similar. If the misclassification costs are known, the cost-sensitive method can minimise the total costs. If the misclassification costs are unknown, the cost-sensitive ANN can obtain a better forecast than the standard ANN method when an appropriate cost ratio is used. For clean areas where episode days are very rare, the forecasts are poor for all methods.

1. Introduction

Ground-level ozone is a secondary air pollutant formed by the photochemical reactions in the atmosphere from its precursors, the oxides of nitrogen and volatile organic compounds (Seinfeld and Pandis, 1998). It has caused intensive concern because of its adverse effects on human health and vegetation. A wide variety of operational warning systems have been developed to forecast the next day's O_3 levels (Schlink et al., 2003). On days of expected high O_3, health warnings and calls for community action can be issued.

Many methods exist for forecasting ground-level ozone concentrations (USEPA, 2003). These methods can be classified in two broad categories: statistical and deterministic. The deterministic approaches usually employ a three-dimensional chemical transport model to carry out numerical computations to obtain the concentration distributions of ozone and other photochemical air pollutants. These methods can generate three-dimensional concentration distributions and find the cause–effect relationships between ozone concentration and emission of precursors. They are, however, difficult to develop and operate. In addition, the required emission and meteorological data are difficult to determine accurately in real time. On the other hand, statistical models directly use air quality and meteorological observations to develop empirical relationships between ozone concentration and other environmental parameters. These approaches are simple to develop and have been widely used in short-term forecasting of air quality (Zannetti, 1990).

Statistical methods that have been proposed for ozone forecasting include linear multiple regression, nonlinear regression, classification and regression tree (CART), and ANN. There

* *Corresponding author.*
E-mail address: hcchiang@mail.tku.edu.tw (H. Chiang).

0048-9697/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2008.12.007

is growing interest in using ANNs for air quality forecasts because they are able to simulate strongly nonlinear behaviours and to learn complex and even a priori unknown relationships directly from the training data. The development and application of ANN models for air quality forecasting has made considerable progress in recent years (Gardner and Dorling, 1999; Kolehmainen et al., 2001; Kukkonen et al., 2003; Niska et al., 2004; Wang and Lu, 2006). Some weaknesses still exist, however, when ANN models are applied to air quality forecast. We will focus on two issues in this paper.

The first issue concerns the input variables. The ANN ozone forecast models could be classified as single-site and regional models (Gardner and Dorling, 2000; Thompson et al., 2001). Single-site models use the data collected from one site to carry out the forecast, while regional models use the data collected from multiple sites to represent the regional conditions. Most ANN models developed in previous studies are the single-site model. The formation of ozone in the troposphere is, however, a complex process involving regional transport of ozone and its precursors. Therefore it is reasonable to anticipate that regional models would be superior to single-site models. We will compare the results obtained by these two different approaches and evaluate their performance.

The second issue concerns the training method of ANN models. Many researchers have reported that ANN models can make reasonable predictions when they are used to predict the daily maximum 1 h and/or 8 h concentrations. As a matter of fact, the predictions in the middle range of concentration are fairly good in general. Many ANN models, however, over-predict low values and under-predict high values; see, for example, Comrie (1997) and Niska and colleagues (2004). Generally speaking, the performance of most ANN models in a high ozone episode is rather poor. This bias is caused by non-uniform distribution of training data (Chen, 2006). Like that of other standard statistical models, the ANN model's performance depends largely on the quality of training data. When the data are not homogeneous, results can be biased toward the most commonly sampled regime. From modelling daily maximum ozone concentrations, we have a large amount of data for low and middle concentration events and limited data for high concentration events. ANNs and other machine learning techniques trained to model these problems will be biased towards high frequency values, that is, the middle range. To avoid this bias, a new training algorithm should be considered.

In previous studies, most ANN models were used to predict the absolute value of the ozone concentration. We will focus, however, on predicting the occurrence of the episode day. The aim of this study is to develop a model that can classify each day as episode or non-episode. This is the so-called two-class classification problem in the field of machine learning. There are many classifiers that can be used for such problems, such as decision trees, support machine vectors (SVM), and ANN. Unfortunately, when these methods are applied to ozone episode forecasts, they face a so-called imbalance data set classification problem. This problem occurs when the training data for one class greatly outnumber those of the other class. With highly imbalanced data it is difficult to detect the rare but important event since standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. The results are heavily biased toward the majority class. In most parts of the world, the ratios of the ozone episode days are relatively low, typically less than 20%; in some clean regions the ratio may be lower than 1%. Thus, the forecasting of ozone episode days is a highly imbalanced data classification problem. Some researchers in the field of air quality have tried to find a solution. For example, Lu and Wang (2008) used an SVM approach to predict the occurrence of ozone episode days.

A number of solutions to the class imbalance problem have been proposed. These solutions include many different forms of resampling techniques, adjusting the costs of misclassification, adjusting the decision threshold, and recognition-based learning (Chawla et al., 2004).

The most direct method for dealing with the highly imbalanced classification problem is to use cost-sensitive learning (McCarthy et al., 2005). A cost-sensitive learner can accept cost information from a user and assign different costs to different types of misclassification errors. Maloof (2003) has argued that learning from imbalanced data sets and learning when misclassification costs are unequal and unknown can be handled in a similar manner. Zhou and Liu (2006) also suggest that cost-sensitive learning is a good answer to the two-class imbalance problem. This motivates us to investigate the use of cost-sensitive ANN for the forecasting of ozone episode days.

The aim of this study is to compare the performance of different cost-sensitive ANN algorithms when they are applied to ozone episode forecasts. We will focus on how to forecast correctly the occurrence of episodes rather than predict the absolute concentration level. Ozone observations from six air quality stations in Taiwan with diverse distribution features are employed as test cases from which to draw general conclusions and guidelines on the use of cost-sensitive techniques for air quality forecasts.

2. Methods

2.1. *ANN architecture*

Current researches indicate that ozone can affect human health over long periods of time, not just during a few 1 h peak events. A new standard based on daily maximum 8 h ozone concentration was promulgated in the USA. Therefore, the forecasting models will classify each day as an episode or non-episode according to the standard of the daily maximum 8 h ozone concentration. If the daily maximum 8 h O_3 concentration at a specific day and site exceeds 80 ppb, this site will be regarded as an ozone episode day on that day; otherwise, it is not an episode day.

As shown in Fig. 1, the ANN architecture used in this study is a fully-connected, feed forward, back-propagation, multiple layer perceptron (MLP). This is a three-layer network; the input and the hidden layer contain multiple units and the output layer only has a single unit. The relationship between the inputs x_i, $i=1,2,\ldots,n_1$, and the output variable $\hat{y}$ is given by (Chaloulakou et al., 2003)

$$\hat{y}=f_2\left(\sum_{m=1}^{n_2} w_{im}^{(2)}\left[f_1\left(\sum_{i=1}^{n_1} w_{im}^{(1)}x_i+b_m^{(1)}\right)\right]+b^{(2)}\right) \quad (1)$$

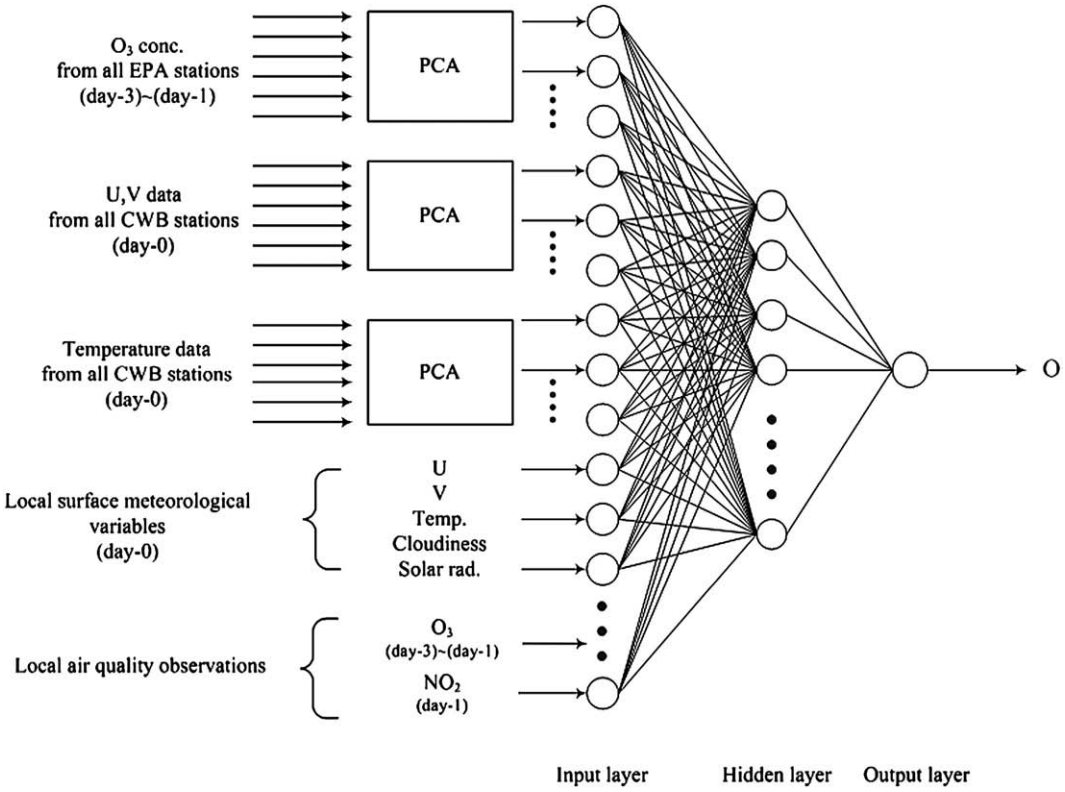

Fig. 1 – The architecture of a multilayer neural network.

where f_1, f_2 are the activation transfer functions for the hidden and output layers, respectively, $w_{im}^{(1)}$ represent the weights from ith input node to mth hidden node, $w_{im}^{(2)}$ denote the weights from the mth hidden node to the output node, $b_m^{(1)}$ and $b^{(2)}$ denote the biases of the mth hidden node and the output node, and n_1 and n_2 denote the number of the input and hidden nodes, respectively. The transfer function used for both the hidden and output layers (f_1 and f_2) is the log-sigmoid function; its output is bounded between zero and one. The target of training data is $y=1$ if the sample is a positive label and $y=0$ otherwise. The example is classified as positive if $\bar{y}>0.5$ for standard ANN.

The ANN was trained (iterative adjustment of weights and biases) by the minimising of an error function to determine the parameters and learn relationships from the presented data. The error function is a measure of discrepancy between the observed and predicted values. Two well-known error functions are OLS (ordinary least square) and LAD (least absolute deviations). For OLS, the error is:

$$Q\left(\underset{\sim}{\beta}\right) = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\left(\underset{\sim}{\beta}\right)\right)^2 \quad (2)$$

while for LAD, the equation becomes:

$$Q\left(\underset{\sim}{\beta}\right) = \sum_{i=1}^{n} \left|y_i - \hat{y}_i\left(\underset{\sim}{\beta}\right)\right| \quad (3)$$

where y_i are observed values, $\hat{y}_i$ are predicted values, β are the unknown parameters, and n is the size of training data.

The choice between different error functions depends on statistical considerations such as the distribution of the target variable as well as the real situation. The standard ANN uses the OLS error function. For cost-sensitive cases, a modified error function will be used which will be discussed later.

The ANN was trained by the traingdx algorithm of the MATLAB software (Demuth and Beale, 2004). This algorithm uses variable learning rates and, although usually slower than the other methods, it can provide more consistent results when early stopping is used. The available data were randomly divided into three subsets: training data, validation data and forecast data. The training data were about half the size of the dataset considered; the remaining data were used for validation and evaluation. The MLP models were then trained to display the relationship between the predictors and daily maximum ozone concentrations using training data. A validation data set checked the performance of the MLP network to determine the epoch at which training should be stopped to avoid over-training. A forecast dataset was used for performance evaluation. Typically, the global minimum is not reached and a good local minimum is treated as an acceptable solution. We trained MLP models 20 times in order to reduce the likelihood of local minima causing problems.

2.2. *Cost-sensitive learning*

Learners can implement cost-sensitive learning in a variety of ways. One common method is adjusting the costs of the various classes so as to counter the class imbalance. This can be done by increasing the penalty associated with misclassifying the positive class relative to the negative class. We refer to this approach as the penalty method later in this paper. The effect of assigning penalties is equivalent to changing the relative data distribution in the two classes, or, in other words, to re-balancing the data.

This method is similar to the standard ANN method except the error function is modified as:

$$Q\left(\underset{\sim}{\beta}\right) = \sum_{i=1}^{n} c\left(y_i - \hat{y}_i\left(\underset{\sim}{\beta}\right)\right)^2 \quad (4)$$

where c is the misclassification cost. For convenience, the costs associated with false-negative and false-positive errors were assigned to λ and unity, respectively. If the misclassification costs are known, λ can be determined and Eq. (4) can train the ANN model by minimising the total costs. If the misclassification costs are unknown, λ is not related to misclassification costs. It can be considered as a parameter used to counter the class imbalance. It will be referred to later as the cost ratio.

Another popular approach to solving these problems is to bias the classifier so that it pays more attention to the positive instances. For a two-class learning problem, the input pattern is normally assigned to a positive class if $\hat{y} \geq \tau$ and to the negative class if $\hat{y} \geq \tau$ where $\hat{y}$ is the predicted value and τ is a threshold. If the misclassification costs are equal, $\tau = 0.5$. For the threshold moving algorithm, the threshold will be changed by misclassification cost. According to Dorling and colleagues (2003), the use of unequal misclassification costs corresponds to a threshold given by:

$$\tau = C_{fp}/(C_{fn} + C_{fp}) \quad (5)$$

where C_{fn} and C_{fp} are the costs associated with false-negative and false-positive errors, respectively. Since $\lambda = C_{fn}/C_{fp}$, we can show that $\tau = 1/(1+\lambda)$.

2.3. *Experimental data*

The original air quality data used in this study were obtained from EPA, Taiwan (2007). The data consist of hourly average concentrations of ozone and other relevant pollutants collected from 72 monitoring stations in Taiwan from 2000 to 2003. The meteorological data were taken from surface observations collected by the Central Weather Bureau (CWB, 2007). The CWB meteorological data used for ozone forecasts are listed in Table 1. In addition to these observations, another important parameter is solar radiation. An empirical formula suggested by Kasten and Czeplak (1980) was used to determine the solar radiation, R. It is

$$R = R_0\left(1 - 0.75C^{3.4}\right) \quad (6)$$

where C is the fractional cloud cover and R_0 is the clear sky insolation which is calculated as

$$R_0 = 990(\sin\phi) - 30 \quad (7)$$

where ϕ is the solar elevation angle.

Table 1 – Input meteorological parameters

Ave. of surface wind speed from 08:00 to 16:00 LST
Ave. of *u*-components from 08:00 to 16:00 LST
Ave. of *v*-components from 08:00 to 16:00 LST
Max. hourly surface temperatures
Max. surface temperature–Min. surface temperature
Ave. of surface relative humidity from 08:00 to 16:00 LST
Ave. of cloudiness at 11:00 and 14:00 LST
Solar angle at 12:00 LST

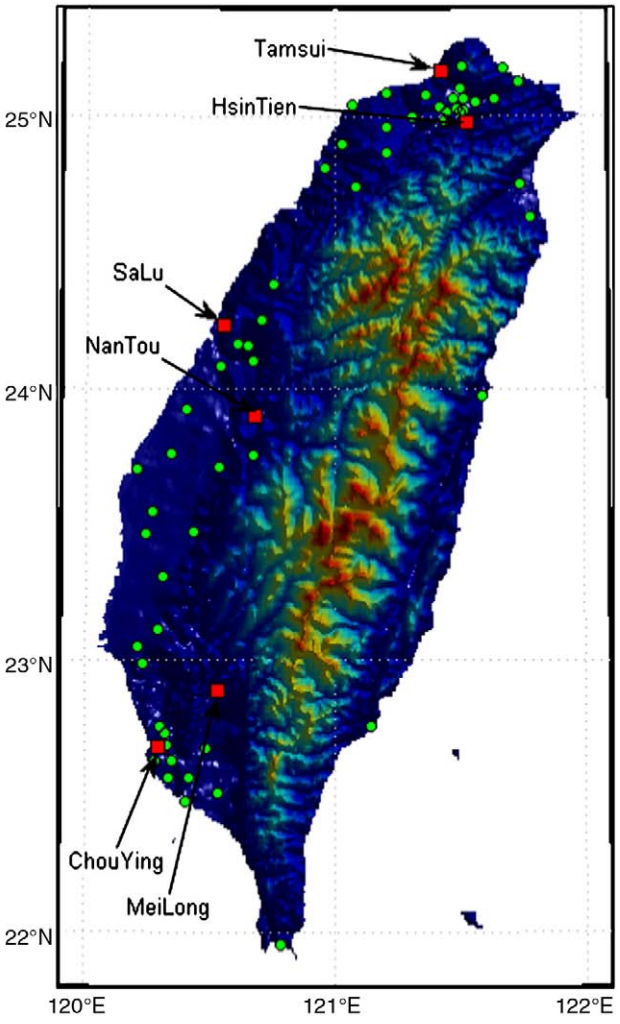

Fig. 2 – The locations of air quality monitoring stations (red squares represent the stations selected for model evaluation and the other air quality monitoring stations are denoted by green circles).

The air quality and meteorological variables are calculated on a daily basis. If more than 6 h of data are missing on one day for any weather variable or for ozone, then the entire day is regarded as missing data. After daily variables were computed, the data quality was examined and the stations with too many missing data were omitted from the analysis. Data from 62 air quality monitoring stations and seventeen surface meteorological stations were kept and used later. At this stage, a fairly small fraction (ranging from 5% to 10%) of missing data was found. The missing values were imputed with the algorithm developed by Schneider (2001). Data imputation allows a consistent and fair model comparison exercise.

Only six air quality stations were selected for prediction and model evaluation. Two stations were selected for three metropolitan areas in Taiwan: one in the coastal area, the other at an inland downwind site. These monitoring stations, shown in Fig. 2, represent different class distributions. The statistics of daily maximum 8 h concentrations for these stations are given in Table 2. The data are listed according to class imbalance (the most imbalanced data sets are listed first). Among these stations,

Table 2 – The locations and statistics of daily maximum 8 h O_3 concentration at selected monitoring stations

Station	Longitude	Latitude	Averaged 8 h O_3 conc.	S.D. of 8 h O_3. conc.	Episodes ratio
	(degree)	(degree)	(ppb)	(ppb)	(%)
TamSui	121.43	25.17	42.0	14.3	1.3
SaLu	120.55	24.23	39.6	17.1	2.8
HsinTien	121.53	24.98	43.2	19.0	4.9
NanTou	120.68	23.90	52.4	21.5	10.2
MeiLong	120.53	22.89	58.2	20.8	15.3
ChouYing	120.29	22.68	56.8	25.3	18.8

the TamSui and SaLu stations are located on the upwind side of major emissions most of the time. The ratios of episode days for these two stations are quite low (less than 3%). HsinTien is a densely populated suburb located on the downwind side of Taipei metropolitan area. The ozone concentrations at HsinTien are influenced by regional contributions as well as local emissions. The concentrations of primary pollutants are fairly high in this station, but the frequency of episode days is low (~5%). This is probably because of the effect of ozone titration. NanTou is located on the downwind side of TaiChung, the largest city in central Taiwan. The high ozone events occurring at NanTou may be caused by the pollutants transported from TaiChung and other cities. MeiLong is a small town located in a valley in southern Taiwan. Ozone episodes occur under specific wind direction that can transport pollutants from nearby sources to this location. ChouYing is located in the centre of Kaohsiung, a major industrial city in Taiwan. It represents a heavily polluted area; the ratio of ozone episode days is about 19%. Many pollutant transport paths will cross this site; hence, the ozone episodes are influenced by regional as well as local conditions.

To account for the persistency of ozone pollution, the daily maximum 8 h ozone concentrations of the previous three days were used for input variables. The forecasts were, however, carried out by use of the meteorological data on the same day. This is based on the assumption that we can carry out a 'perfect' weather forecast. This is the general approach used in the developing stage of a 'perfect prog model' (Wilks, 1995).

It is well-known that different scales of a meteorological system can affect the ground-level ozone concentrations; hence, large-scale meteorological patterns as well as local observations should be considered in the development of an ANN model. To determine large-scale meteorological patterns, the observations from all CWB stations were used. Since the amount of data from all stations was too large, the principal component analysis (PCA) technique was used to reduce and orthogonalise the original input data (Wilks, 1995). These treated variables were then used as new input vectors in the ANN model. The size of the input vectors was reduced by retention of only those components that contributed more than a specified fraction (depending on the variables) of the total variation in the data set.

The methods for the preparation of input data are shown in Fig. 1, which includes:

(1) The previous three days' daily maximum 8 h ozone concentrations from all EPA stations, processed by the PCA technique before input to ANN models.
(2) The wind velocity components (u, v) from all CWB stations, processed by the PCA technique before input to ANN models.
(3) The temperature from all CWB stations, processed by the PCA technique before input to ANN models.
(4) The local meteorological variables, i.e. u, v, temperature, cloudiness, and solar radiations from the nearest CWB station.
(5) The previous day's ozone and NO_2 concentration from the nearest EPA station.

The regional dataset includes all items mentioned above, but the single-site dataset contains only items 4 and 5.

2.4. *Performance evaluation*

The performance of the ozone forecast is evaluated by a confusion matrix as illustrated in Table 3. The columns are the forecast class and the rows are the observed class. In this matrix, TN is the number of non-episode days correctly classified, FP is the number of non-episode days incorrectly classified as episode days (false alarm), FN is the number of episode days incorrectly classified as non-episode days, and TP is the number of the episode days correctly classified. A perfect forecast programme would have values in cells 'TN' and 'TP' only. In the real world, imperfect forecasts result in values in cells 'FP' and 'FN'.

Based on the confusion matrix, several measures can be computed, including probability of detection (POD), false alarm rate (FAR), false-positive rate (FPR) and accuracy. POD represents the fraction of correctly forecast ozone episode days, ranging between [0, 1] with a best value of 1. POD is calculated as:

$$\mathrm{POD} = \mathrm{TP}/(\mathrm{FN} + \mathrm{TP}) \tag{8}$$

FAR is the fraction of false alarms over total forecast positive events, ranging between [0, 1] with a best value of 0. FAR is computed as:

$$\mathrm{FAR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TP}) \tag{9}$$

FPR ranges between [0, 1] with a best value of 0, which is computed as:

$$\mathrm{FPR} = \mathrm{FP}/(\mathrm{TN} + \mathrm{FP}) \tag{10}$$

The accuracy, ranged between [0, 1], is defined as:

$$\mathrm{Accuracy} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}) \tag{11}$$

In order to maintain public confidence in the ozone action advisories, it is desirable that the POD should be reasonably high. At the same time, the FAR and the FPR should be reasonably low. For a highly imbalanced classification problem, accuracy may not a good indicator since the accuracy is generally as high as TN is large.

Table 3 – Contingency table for a two-category forecast

		Forecast	
		<80 ppb	≧80 ppb
observed	<80 ppb	TN	FP
observed	≧80 ppb	FN	TP

Table 4 – The statistics of the forecast using local data and standard ANN method

Station	Observed episodes	Predicted episodes	No. correctly predicted (POD)	No. of false alarms (FAR)	Accuracy
TamSui	7	0	0 (0.000)	0 (0.000)	0.985
SaLu	14	0	0 (0.000)	0 (0.000)	0.969
HsinTien	21	0	0 (0.000)	0 (0.000)	0.954
NanTou	48	22	12 (0.250)	10 (0.455)	0.899
MeiLong	71	35	25 (0.352)	10 (0.286)	0.877
ChouYing	82	44	28 (0.341)	16 (0.364)	0.846

When the operating conditions (misclassification costs and class distribution) are unknown, classifier evaluation requires a method of visualising classifier performance across the full range of possible operating conditions. The receiver operating characteristic (ROC) curve is an ideal graphical method to fulfill this requirement. The horizontal axis of an ROC curve is FPR and the vertical axis is POD. By varying the misclassification cost or threshold parameter, it is possible to draw the ROC curve. Some researchers argue that ROC curves are good indicators of the classifier's performance; for example, Dorling and colleagues (2003) suggested that the area under the ROC curve gives the effectiveness of the classifier. They stated that if nothing is known about the optimal cost ratio, the closer the area to unity the better the classifier. Many researchers suggest that the more the ROC approaches the (0,1) point, the better the classifier will discriminate under various operating conditions.

3. Results and discussion

3.1. *Numerical experiments*

Two ANN algorithms (cost-penalty and moving threshold methods) and two input datasets (single-site data and regional data) were considered; hence, four runs were carried out for each site. Thirty-six cost ratios were considered for each run. They can be represented by $\lambda = 10^{0.1\ k}$, $k = -10, -9, \ldots, 15$. The method will be referred to as the standard method when $\lambda = 1$.

3.2. *Statistics of the forecast from using standard ANN methods*

The statistics of the forecast from using the standard ANN method and single-site input data are summarised in Table 4.

We will examine the results of the TamSui station first. There are 455 days of data included in the forecast dataset and the Tamsui station has seven episode days; unfortunately, none is correctly predicted. These episodes are extremely rare events representing approximately the upper 1.5% of the O_3 distribution. Such extremely rare events are difficult to predict. On the other hand, no alarm was given, so FAR equals zero. The accuracy is very high (0.985), but, as mentioned before, accuracy is not a suitable index for the performance evaluation since it is high when the number of episode days is small. As for the other stations, the results for SaLu and HsinTien are also poor. None of the ozone episode days is correctly predicted. The predictions of NanTou, MeiLong, and ChouYing are somewhat better. Unfortunately, none of the stations can correctly forecast more than 50% of exceedances. Considering the difficulty of forecasts that are at the extreme high end of the O_3 distribution, a POD above 50% and a FAR below 50% are regarded as good outcomes. Although this is a very loose standard, none of the stations performed well in this capacity.

Table 5 is similar to Table 4 except that the regional input data were used instead of single-site input. When we compare these two tables, we find that significant improvement can be obtained if ANN models use a regional dataset. The POD values are increased and the FAR values are decreased when compared with Table 4. The improvements in ChouYing are especially remarkable. The prediction based on regional dataset is 16% higher in POD (correctly detects thirteen more ozone episode days) and 11% lower in FAR (two fewer false alarm days) than standard ANN models that use a single-site dataset. Similar improvement also occurs in NanTou. This is entirely reasonable since photochemical air pollution is a regional problem. The ozone precursors can transport several hours before the maximum ozone concentration is reached. Unfortunately, the improvement in MeiLong is not so significant. The POD increased by 4.2%, but the FAR also increased by 1.4%. This may because MeiLong is located in a valley (see Fig. 2). The effect of large-scale meteorological conditions and ozone distributions may not be so important in this station. According to the results of these two tables, it is concluded that regional input data can improve the performance of ANN in general.

It is worth mentioning that the numbers of the predicted episodes are much smaller than those of observed episodes if standard ANN is used. This is caused by the bias of imbalanced dataset classification. Standard ANN methods have a tendency to treat a minority as noise and therefore disregard it.

Table 5 – The statistics of the forecast using regional data and standard ANN method

Station	Observed episodes	Predicted episodes	No. correctly predicted (POD)	No. of false alarms (FAR)	Accuracy
TamSui	7	0	0 (0.000)	0 (0.000)	0.985
SaLu	14	0	0 (0.000)	0 (0.000)	0.969
HsinTien	21	1	1 (0.048)	0 (0.000)	0.956
NanTou	48	24	15 (0.313)	9 (0.375)	0.908
MeiLong	71	40	28 (0.394)	12 (0.300)	0.879
ChouYing	82	55	41 (0.500)	14 (0.255)	0.879

3.3. *Comparison of different cost-sensitive methods*

Fig. 3 shows the POD v. FAR curves obtained by cost-penalty and moving threshold methods. The regional dataset was used by both methods in order to obtain better results. Each point in this figure corresponds to a different setting of the cost ratio or threshold parameter. The points marked with blue circles represent the results obtained by the cost-penalty method, while those marked with red squares are computed by the moving threshold method. This figure shows the trade-off between POD and FAR since FAR will be increased as POD is increased. Interestingly, the results obtained by the two ANN methods have the same trend and are very close. If, say, we set up some criteria for acceptable model performance, POD should be above 50% and FAR should below 50%; there are some points in NanTou, ChouYing and MeiLong stations that

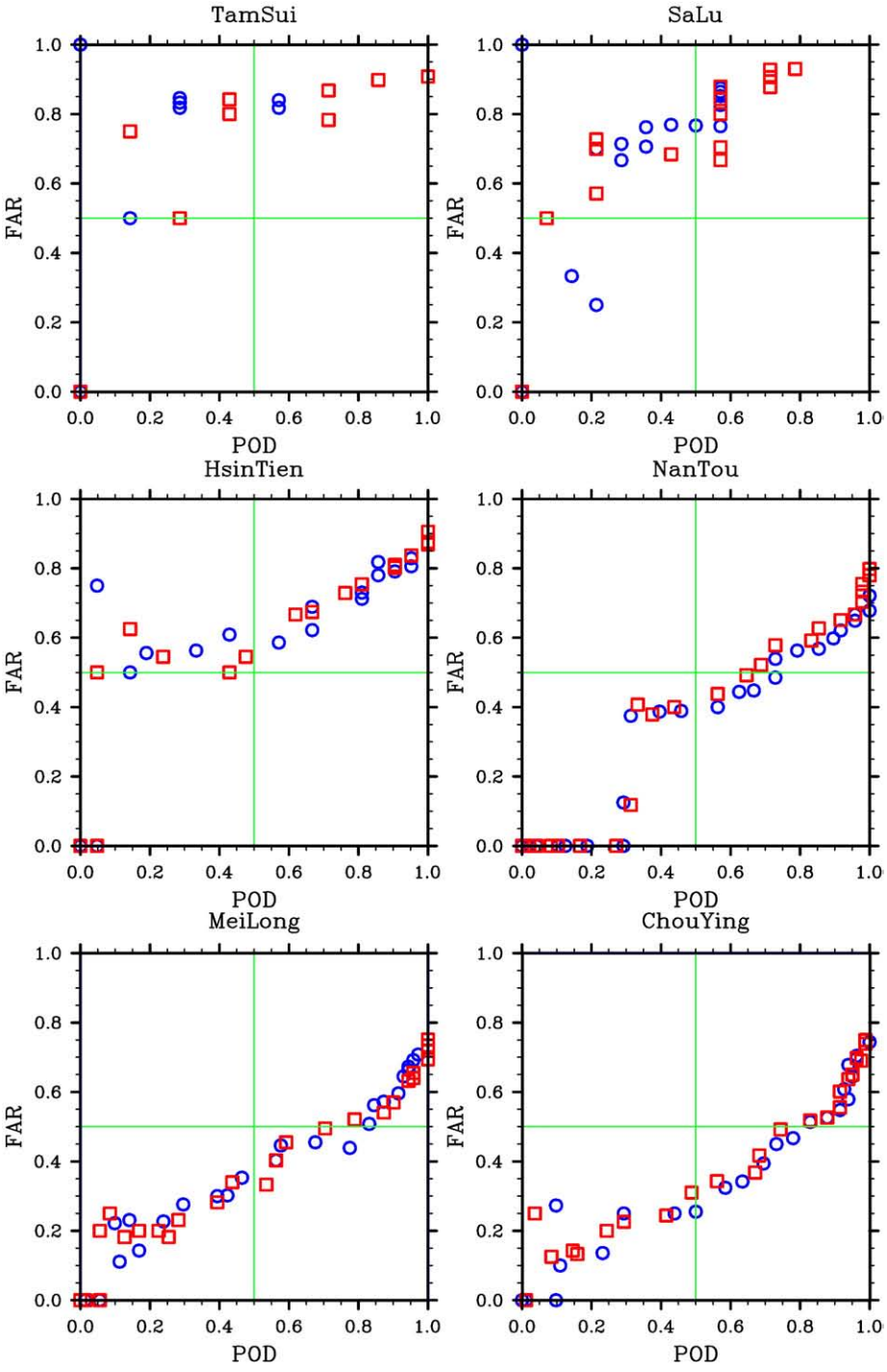

Fig. 3 – Comparison of POD v. FAR curves obtained by cost-penalty and moving threshold methods (○: cost-penalty method, □: moving threshold method).

meet these criteria. In general, these acceptable points occur while the cost ratios range from one to five. For cleaner stations (TamSui, HsinTien, SaLu) the POD will be greater than 50% only when FAR is greater than 50%. This indicates that if the events are rare, the forecasts of cost-sensitive ANN models are rather poor. We should just wait for the event to occur unless we can tolerate the inconveniences caused by numerous false alarms.

Fig. 4 is the computed ROC points obtained by cost-penalty and moving threshold methods. Although there are some fluctuations, the two curves are very close for every station. This suggests that the cost-penalty has the same effect as

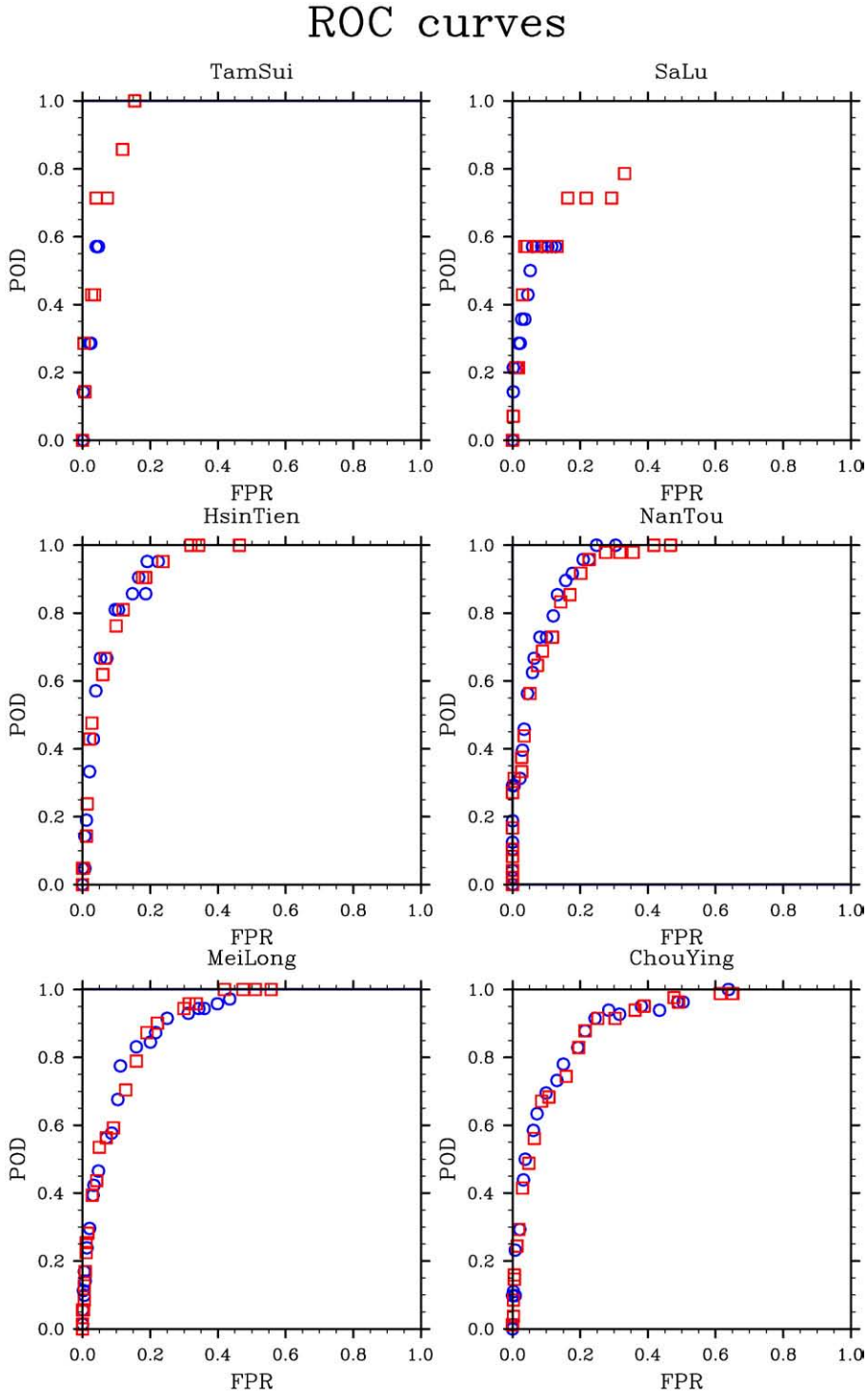

Fig. 4 – Comparison of ROC curves obtained by cost-penalty and moving threshold methods (○: cost-penalty method, □: moving threshold method).

moving the decision threshold. The experiment carried out by Maloof (2003) obtained a similar result.

3.4. *Effects of different input data on the forecast*

Fig. 5 is the POD v. FAR curves obtained by cost-penalty ANN method using different datasets. The red squares are obtained by the regional dataset, while the blue circles represent the results computed by the single-site dataset. One approach dominates another if it has a higher POD and a lower FAR. According to this rule, the forecasting using regional data is better than that using single-site data. The evidence is very obvious in NanTou and ChouYing. Probably, the ratios of episode days are higher in these two stations; therefore, they have enough episodes that can be used to train the ANN models. The relationship between environmental conditions and ozone concentration can be well established in these two sites. On the other hand, since

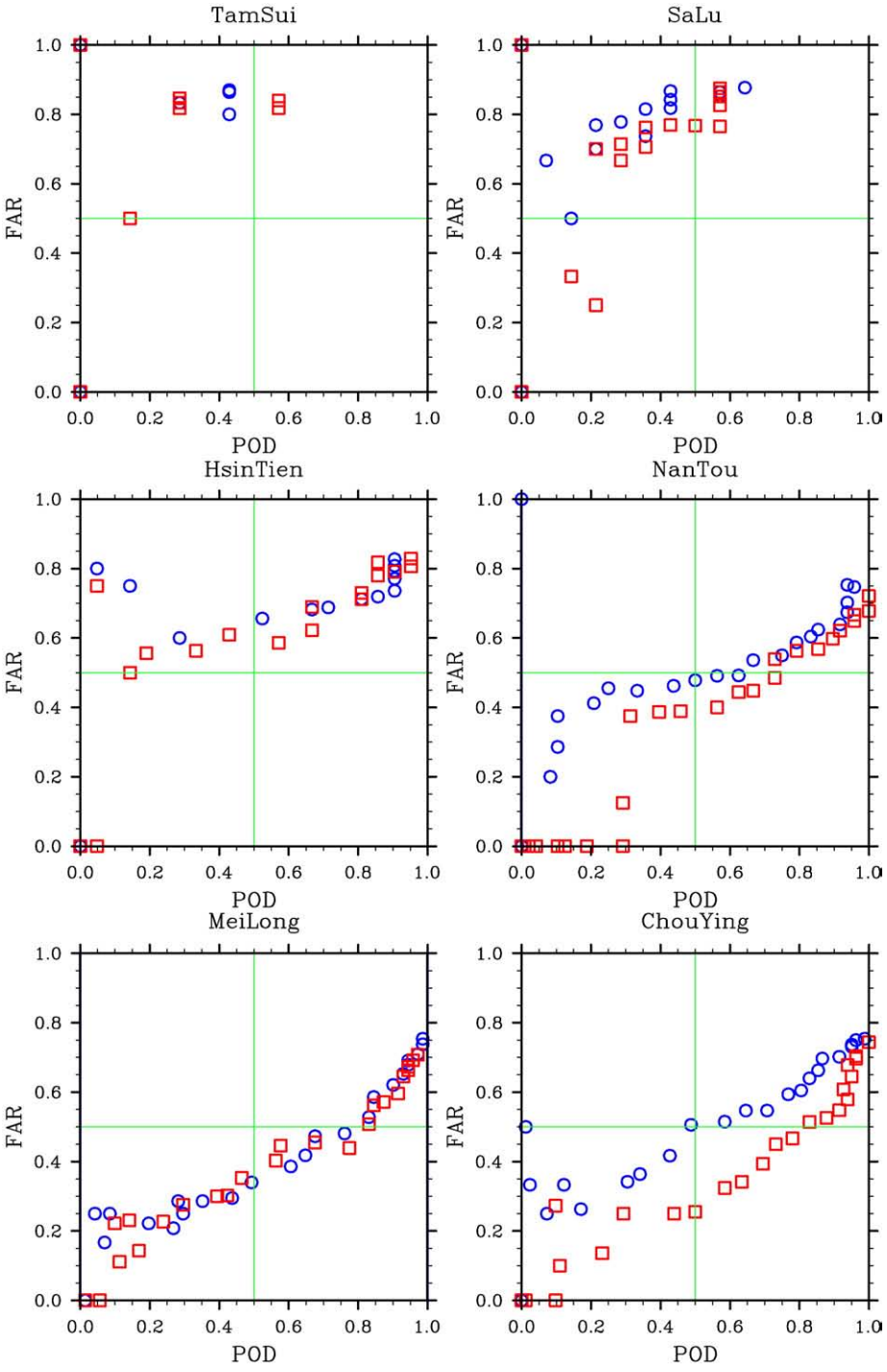

Fig. 5 – Comparison of POD v. FAR curves obtained by cost-penalty method using different input datasets (○: single-site data, □: regional data).

MeiLong is located in a valley, the influences of large-scale meteorological conditions are not so significant. Thus, the POD v. FAR curves obtained from different input datasets are closed in MeiLong.

Fig. 6 shows the ROC curves obtained by the cost-penalty ANN method using different input datasets. Since the ROC curves obtained by using regional data are close to (0,1), this again demonstrates the superior results of regional input data.

3.5. *Optimal cost ratio*

The cost ratio (λ) is the most important parameter for cost-sensitivity methods. Now, we will discuss how to determine

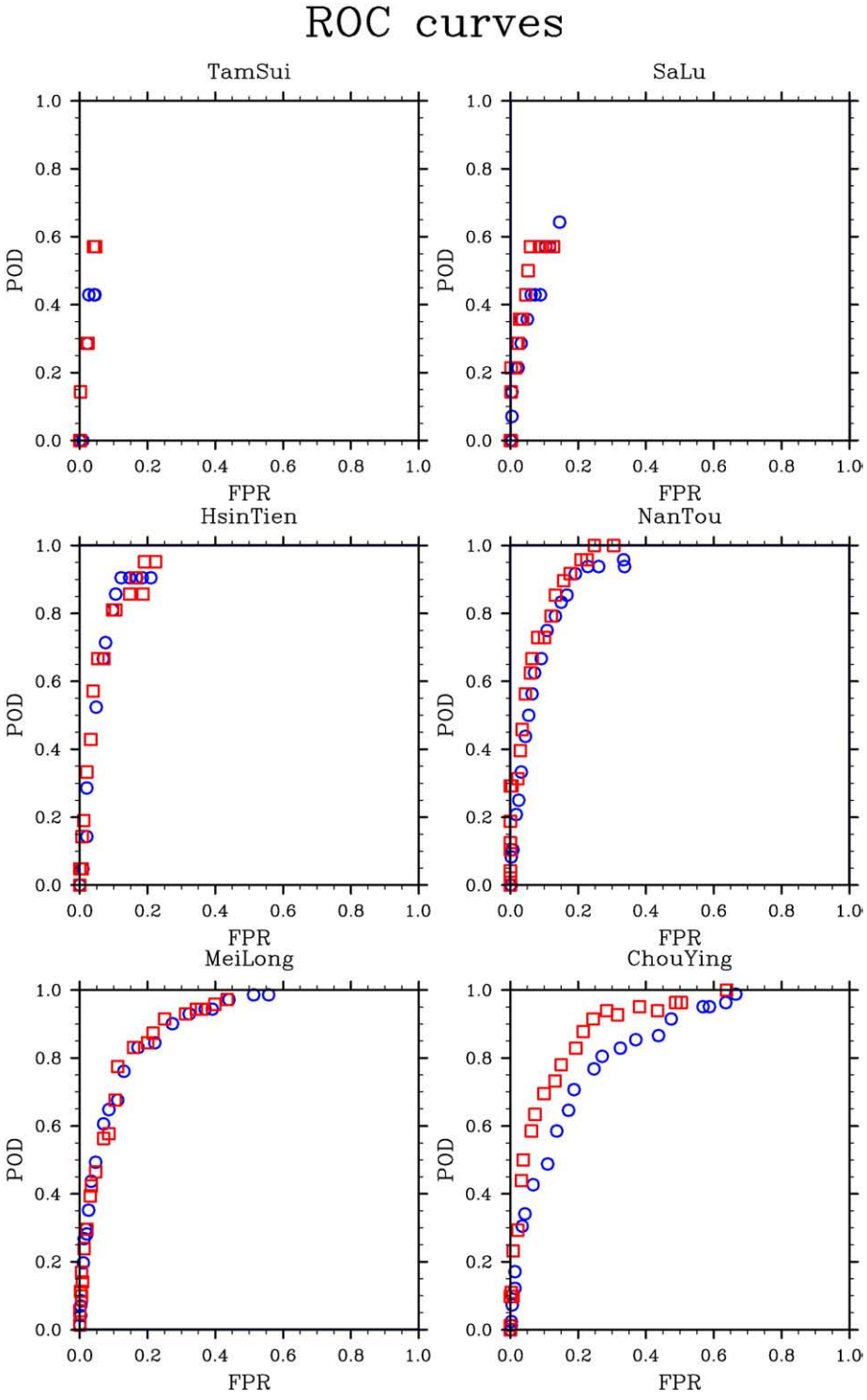

Fig. 6 – Comparison of ROC curves obtained by cost-penalty method using different input datasets (○: single-site data, □: regional data).

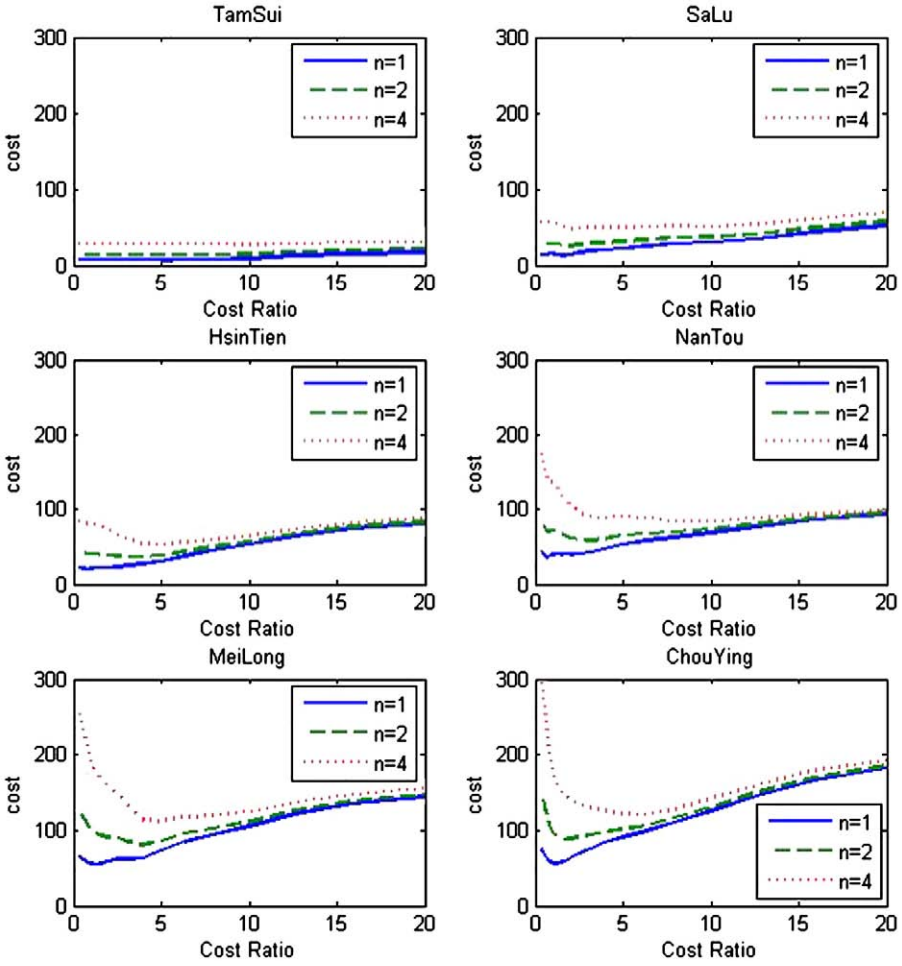

Fig. 7 – Total misclassification costs for different cost ratios.

the optimal cost ratio. Two different situations will be considered. First, when misclassification costs are known or can be assumed the best metric to evaluate overall classifier performance is total cost. The formula for total cost is shown below (McCarthy et al., 2005):

$$\text{Total Cost} = (\text{FN} \times \text{CFN}) + (\text{FP} \times \text{CFP}) \quad (12)$$

where CFN and CFP are the misclassification cost associated with false-negative and false-positive error, respectively. From Eq. (12), the total misclassification costs for six stations at various cost ratios are shown in Fig. 7. The three curves shown in each plot represent the total misclassification costs at $n=1$, 2, and 4, where $n=\text{CFN/CFP}$. Please note that n and λ are different. The cost ratio λ is only a parameter for cost-sensitive ANN models, and n is the 'real' misclassification cost ratio. Setting different values of the cost ratio (λ) would produce significant differences of the total misclassification cost among the different misclassification cost ratio (n). The results show that the total misclassification costs would be close to the same value when the cost ratio is large. Just as we expected, the minimum costs occur when cost ratio equals n. This indicates that the cost-sensitive ANN method using n as the cost ratio can achieve the lowest misclassification cost.

It is encouraging that we can use cost-sensitive ANN models to minimise the total misclassification cost. Although this is straightforward mathematically, it is difficult to determine the misclassification costs in real air quality management systems. The costs must be formulated in financial terms; it is not always possible to derive the true misclassification costs. In addition, different end-users are likely to have different, possibly contradictory, views on the correct set of misclassification costs (Dorling et al., 2003).

Now, we will discuss how to determine an appropriate λ value when the misclassification costs are unknown. The POD vs. FAR curves reveal the inherent trade-off between performance in the positive and negative examples. We could

Table 6 – The statistics of the forecast using regional data and cost-penalty ANN

Station	λ	Observed episodes	Predicted episodes	No. correctly predicted (POD)	No. of false alarms (FAR)
TamSui	12.59	7	11	2 (0.286)	9 (0.818)
SaLu	3.98	14	14	4 (0.286)	10 (0.714)
HsinTien	3.16	21	23	9 (0.429)	14 (0.609)
NanTou	2.00	48	45	27 (0.563)	18 (0.400)
MeiLong	2.00	71	67	40 (0.563)	27 (0.403)
ChouYing	1.59	82	79	52 (0.634)	27 (0.342)

choose a point on these curves and make whatever trade-off we thought appropriate. For example, an air quality manager can use the curves to choose the λ value so that POD will be greater than 0.5 and FAR will less than 0.5. Here, we will propose another ad hoc method to determine the λ values that suits the general interest. As we have noted in Tables 4 and 5, the number of predicted episodes is too small when compared with the number of observed episodes. We may select a λ so that these two numbers become closer. According to this principle, suitable λ were selected for each station. These values were used to carry out the forecast by the cost-penalty ANN method, and the results are shown in Table 6. When compared with Table 5, the number of predicted episode days, as well as the number of correctly predicted episodes, was increased. This is a noteworthy improvement. The price we have to pay for this progress, however, is a slight increase of the false alarm rate. Nevertheless, it is interesting that the POD values in most stations are greater than 0.5 and the FAR values are less than 0.5 if appropriate λ values are used.

4. Conclusions

This study addressed the imbalance classification problem applied to ozone episode day forecast. The models classify each day as episode or non-episode based on the standard of daily maximum 8 h O_3 concentration. Six air quality stations in Taiwan, with diverse distribution features, were employed as test cases.

Two different input datasets, regional and single-site, were generated from raw air quality and meteorological observations. According to the numerical experiments, the predictions based on the regional dataset indeed improved the forecasting accuracy. If the amount of regional data is too large, the PCA technique can be used to reduce and orthogonalise the original input data.

Two ANN models were evaluated by receiver operating characteristic (ROC) curves. The results obtained by the two approaches were found to be similar. In other words, the cost-penalty method has the same effect as moving the decision threshold.

If the misclassification costs can be evaluated, the cost-sensitive method can minimise the total costs. If the misclassification costs are unknown, the cost-sensitive method can obtain a better forecast if a proper cost ratio is used. For clean areas where episodes are very rare, the cost-sensitive ANN models can do little. We should just wait for the episode to occur unless we can tolerate the inconveniences caused by a large number of false alarms.

REFERENCES

Central Weather Bureau, Taiwan, 2007. http://www.cwb.gov.tw.

Chaloulakou A, Saisana M, Spyrellis N. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. Science of the Total Environment 2003;313:1–13.

Chawla N, Japkowicz N, Kolcz A. Editorial: special issues on learning from imbalanced data sets. SIGKDD Explorations 2004;6:1–6.

Chen, W-t., 2006. Forecasting of ground-level ozone concentrations in Taiwan by artificial neural network. Master's Thesis, Tamkang University, Tamsui, Taiwan (in Chinese).

Comrie AC. Comparing neural networks and regression models for ozone forecasting; J. Air & Waste Management Association 1997;47:653–63.

Demuth H, Beale M. Neural network toolbox user's guide, Ver. 4. Natick, MA: The Mathworks; 2005.

Dorling S, Foxall R, Mandic D, Cawley G. Maximum likelihood cost functions for neural network models of air quality data. Atmospheric Environment 2003;37:3435–43.

EPA, Taiwan, 2007. http://www.epa.gov.tw.

Gardner MW, Dorling SR. Neural network modeling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmospheric Environment 1999;33:709–19.

Gardner MW, Dorling SR. Meteorologically adjusted trends in UK daily maximum surface ozone concentrations. Atmospheric Environment 2000;34:171–6.

Kasten F, Czeplak G. Solar and terrestrial radiation dependent on the amount and type of cloud. Solar Energy 1980;24:177–89.

Kolehmainen M, Martikainen H, Ruuskanen J. Neural networks and periodic components used in air quality forecasting. Atmospheric Environment 2001;35:815–25.

Kukkonen J, Partanen L, Karppinen A, Ruuskanen J, Junninen H, Kolehmainen M, Niska H, Dorling S, Chatterton T, Foxall R, Cawley G. Extensive evaluation of neural networks models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmospheric Environment 2003;37:4539–50.

Lu W-z, Wang D. Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. Science of the Total Environment 2008;395:109–16.

Maloof M. Learning when data sets are imbalanced and when costs are unequal and unknown. Proc. of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II; 2003. p. 73–80.

McCarthy K, Zabar B, Weiss G. Does cost-sensitive learning beat sampling for classifying rare classes? Proc. of the ACM SIGKDD First International Workshop on Utility-Based Data Mining. ACM Press; 2005. p. 69–75.

Niska H, Hiltunen T, Karppinen A, Ruuskanen J, Kolehmainen M. Evolving the neural network model for forecasting air pollution time series. Engineering Applications of Artificial Intelligence 2004;17:159–67.

Schlink U, Dorling S, Pelikan E, Nunnari G, Cawley G, Junninen H, Greig A, Foxall R, Eben K, Chatterton T, Vondracek J, Richter M, Dostal M, Bertucco L, Kolehmainen M, Doyle M. A rigorous intercomparison of ground–level ozone predictions. Atmospheric Environment 2003;37:3237–53.

Schneider T. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate 2001;14:853–71.

Seinfeld JH, Pandis SN. Atmospheric chemistry and physics: from air pollution to climate change. New York: John Wiley; 1998.

Thompson ML, Reynolds J, Cox LH, Guttorp P, Sampson PD. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmospheric Environment 2001;35:617–30.

USEPA. Guidelines for developing an air quality (ozone and PM2.5) forecasting program. Research Triangle Park, NC: EPA-456/R-03-002; 2003.

Wang D, Lu W-z. Forecasting of ozone level in time series using MLP model with a novel hybrid training algorithm. Atmospheric Environment 2006;40:913–24.

Wilks DS. Statistical methods in the atmospheric sciences. San Diego: Academic Press; 1995.

Zannetti P. Air pollution modeling: theories, computational methods, and available software. New York: Van Nostrand Reinhold; 1990.

Zhou Z-h, Liu X-y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering 2006;18:63–77.