ELSEVIER

Expert Systems with Applications 27 (2004) 331–340

Expert Systems with Applications

www.elsevier.com/locate/eswa

Data mining to aid policy making in air pollution management

Sheng-Tun Li[a,*], Li-Yen Shue[b]

[a]Institute of Information Management, National Cheng-Kung University, Tainan, Taiwan, ROC
[b]Department of Information Management, National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan, ROC

Abstract

In the past two decades, the heavy environmental loading has led to the deterioration of air quality in Taiwan. The task of controlling and improving air quality has attracted a great deal of national attention. The Taiwanese government has since set up the National Air Quality Monitoring Network (TAQMN) to monitor nationwide air quality and adopted an array of measures to combat this problem. This study applies data mining to uncover the hidden knowledge of air pollution distribution in the voluminous data retrieved from monitoring stations in TAQMN. The mining process consists of data acquisition from Web sites of 71 data gathering stations nationwide, data pre-processing using multi-scale wavelet transforms, data pattern identification using cluster analysis, and final analysis in mapping the identified clusters to geographical locations. The application of multi-scale wavelet transforms contributes greatly in removing noises and identifying the trend of data. In addition, the proposed two-level self-organization map neural network demonstrates its ability in identifying clusters on the high-dimensional wavelet-transformed space. The identified distribution of suspended particulate PM10 represents a complete, national picture of the present air quality situation, which contrasts the present pollution districts, and could serve as an important reference for government agencies in evaluating present and devising future air pollution policies.

Keywords: Data mining; Air pollution management; Multi-scale analysis; Self-organization neural network; Decision support

1. Introduction

Data mining, also known as knowledge discovery in databases (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), is the process of discovering useful knowledge from large amount of data stored in databases, data warehouses, or other information repositories. It is a hybrid disciplinary (Zhou, 2003) that integrates technologies of databases, statistics, machine learning, signal processing, and high-performance computing. This rapidly emerging technology is motivated by the need for new techniques to help analyze, understand or even visualize the huge amounts of stored data gathered from business and scientific applications. The major data mining functions that are developed in commercial and research communities include summarization, association, classification, prediction and clustering (Zhou, 2003). Data mining has been shown capable of providing a significant competitive advantage to an organization by exploiting the potential knowledge of large databases (Bose & Mahapatra, 2001). Recently, a number of data mining applications and prototypes have been developed for a variety of domains (Liao, 2003; Mitra, Pal, & Mitra, 2002), including marketing, banking, finance, manufacturing, and health care. In addition, data mining has also been applied to other types of scientific data (Abidi, 2001; Read, 2000) such as bioinformatical, astronomical, and medical data.

In general, techniques and functions that are to be applied in a data mining process depend very much on the application domain and the nature of the data available. This creative process generally involves phases of data understanding, data preparation, modeling, and evaluation (Fayyad et al., 1996). Data understanding starts with an initial data collection and proceeds with activities to get familiar with the data, to identify data quality problems, and to discover first insights into the data. Data preparation covers all activities that construct the final dataset to be modeled from the initial raw data. The tasks of this phase may include data cleaning for removing noise and inconsistent data, and data transformation for extracting the embedded features. The modeling phase applies various modeling techniques, determines the optimal values for parameters in models, and finds the one most suitable to meet the objectives. The evaluation phase evaluates the model found in the last stage to

* Corresponding author. Address: No. 1, Ta-Hsueh Road, Tainan 701, Taiwan, ROC. Tel.: + 886-6-2757575x53126, fax: +886-6-2362162.
E-mail address: stli@mail.ncku.edu.tw (S.-T. Li).

0957-4174/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2004.05.015

confirm its validity to fit the problem requirements. No matter which areas data mining is applied to, most of the efforts are directed toward the data preparation phase (Pyle, 1999). In this study of mining air pollution data, our data preparation phase particularly emphasizes the data scale issue.

The purpose of this study is to apply data mining technology to identify the national air quality distribution of Taiwan, whose hourly air quality data are continuously collected and archived through a network of 71 EPA stations. In dealing with voluminous data, we combine both wavelet transform (WT) and self-organization map (SOM) neural networks as our data mining technology. The former is accredited with capability of investigating temporal variation with different scales, and the latter is known to be effective in isolating clusters in high-dimensional space. With both technologies, one can benefit from better understanding and interpretation of the pollution data. The rest of this paper is organized as follows. Section 2 provides a brief review of air pollution management in Taiwan. Section 3 presents the issues of mining air quality data from the EPA Web site and the underlying technologies for dealing with the issues. Section 4 elaborates on the mining procedure that consists of data acquisition, missing-value handling, data transform, modeling, and performance evaluation. Section 5 discusses the mining results and its comparisons with official distribution districts. Section 6 concludes this paper.

2. Air pollution management in Taiwan

The island of Taiwan runs from north to south like a sweet potato and is divided into an eastern seaboard and western seaboard by the mountain range that runs also from north to south. With a total area of 35,873 square kilometers and only 26% of it being plain, her population of more than 21 millions represents one of the densest countries in the world. The western seaboard contains a much wider area of plain than the eastern one and is hence much densely populated and also heavily industrialized. Presently, for every square kilometer, Taiwan has 611 people, 453 vehicles, 2.78 factories. Her present population and vehicle density are about 2 times those of Japan, 3 times those of Germany and British, 22 times those of America, and the factory density, ranging from 2.4 to 69.5 times, is even worse. With such a heavy environmental loading, the industrial and related pollution have, in the past 30 years, caused air quality to deteriorate alarmingly, and the air quality control and improvement task has become an urgent task for successive governments. The EPA (Environmental Protection Administration) of Taiwan was formally established in 1987 and was given the mission to control and improve national air quality. It later set up the National Air Quality Monitoring Network (TAQMN) in 1990 to monitor nationwide air quality.

TAQMN, presently, consists of 71 air quality monitoring stations on the Taiwan Island, and can automatically collect and monitor air quality on an hourly basis (EPA, 2000). In addition to others, each monitor station may collect one or more of the five major types of priority pollutants: PM10 (suspended particulate), SO2 (sulfur dioxides), NO2 (nitrogen dioxide), CO (carbon monoxide), and O3 (ozone), with PM10 and O3 being the main air pollutants. EPA also maintains a Web site for each station for publishing archived and real-time pollutant information and forecasting as well. The locations of these stations are mainly based on population distribution, as shown in Fig. 1, and are distributed among eight areas on the island: T-K (Taipei–Keelung), Ilan, T-H-M (Taoyuan–Hsinchu–Miaoli), T-C (Taichung–Changhua), Nantou, Y-C (Yunlin–Chiai), T-K-P (Tainan–Kaohsiung–Pingtung), and H-T (Hualien–Taitung).

In order to control air quality and reverse the trend, successive governments have adopted the polluter-pay principle and have devised a strategy that consists of both control component and incentive component. One main task of the control component is to map out air quality districts with different ratings and impose different pollution levy for each rating. The incentive component provides financial incentives to encourage industry owners to replace existing equipments with newer ones or install better pollution reduction devices. At the same time, a Pollution Prevention Fund was legislated for pooling all levies together for the common goal. The central government is currently retaining 40% of this fund for national infrastructure use and refunding 60% back to the district governments for local use. This 60% refund from the Pollution Prevention Fund has arisen much interest among local governments, and it was not hard to see the involvement of politics in drawing up the air quality districts for getting the maximum refund. Thus, the ideal process of determining air quality districts, which should be based on air quality, population density, and different types of land utilization, has been greatly compromised during the process. The present air quality control district, as is shown in Fig. 1, is based on the official government administrative districts.

In this study, we intend to investigate the current air quality distribution by applying data mining to the pollution

Fig. 1. Air quality administration districts on Taiwan.

data collected and archived in the Web sites of all 71 stations of the nation. An initial study has revealed the fact that PM10 is the only key pollutant that has been collected and archived by all stations in the past; hence we use the national PM10 data of 1 year for the following study.

3. Issues of mining air quality data and underlying technologies

The nature of air pollutants PM10 is of spatial variability of temporal data, and the result of data mining will decompose a large complex area into several smaller homogeneous regions. Depending on different data scales, the homogeneous regions may vary. Data scale could range from small scale (e.g. hourly, daily, etc.) to large scale (e.g. monthly, seasonal, or annual) (Rainsford & Roddick, 1999). The selection of an appropriate scale is dependent on the application purpose. The purpose of this study is to find out present pollution distribution that may serve as a reference to government pollution control agency, hence it needs to take into consideration both short-term and long-range plan in air pollution management. As a result, we need to consider both small scale and large scale at the same time. Research (Mallat, 1989) has shown that multi-scale wavelet transform can better handle the multi-scale issue in this case than any single scale one, since it provides the capability for investigating the temporal variation with different scales. Another issue involved is the selection of a proper technology in identifying homogeneous regions using clustering, which involves tackling the problem of high dimensionality that is the nature of the transformed temporal data. In literatures, self-organizing map (SOM) neural network has shown to be an effective clustering technology in isolating clusters in a high-dimensional space (Kohonen, 1997). We will briefly discuss these two major technologies in this section.

3.1. Wavelet transform in multi-scale study

Wavelet transform can be regarded as an extension of Fourier transform. Fourier transform has been widely used in transferring signals to the frequency domain for extracting information from the time series. Fourier transform can provide the frequency information of a signal, however, it could not identify the time when the embedded frequency varies with time and thus could not be applied to non-stationary signals. For remedying such a shortcoming, the wavelet transform has been proposed to extract the time-frequency information simultaneously. In particular, the continuous wavelet transform (CWT) provides the capability to investigate the temporal variation with a different scale. Formally, CWT is defined as the convolution of a time series $f(t)$ with a wavelet function $\psi(t)$ shifted in time by a translation parameter τ and a scale parameter s (Mallat, 1989):

$$\mathrm{CWT}_f^{\psi}(\tau, s) = \Phi_f^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int f(t) \cdot \psi^* \left(\frac{t-\tau}{s} \right) \mathrm{d}t \qquad (1)$$

where $*$ is the complex conjugate, s and τ are real numbers and can be varied continuously. Wavelets are a family of functions that are derived from translations and dilations of one basic function, which is referred to as the so-called mother wavelet. In other words, the mother wavelet can be used at any scale and its position can also be shifted continuously over the entire time domain of the signal being analyzed. The transform analysis produces wavelet coefficients that are a function of parameters of scale and position, where scale represents the constant by which the wavelet is uniformly stretched or compressed and position represents the constant by which the onset of the wavelet is shifted.

The parameter scale in the wavelet analysis is similar to the scale used in maps. For example, the higher (coarser) scales correspond to non-detailed global views (of the signal), and the lower (finer) scales correspond to detailed views. Scaling, as a mathematical operation, either dilates or compresses a signal. Larger scales correspond to dilated (or stretched out) signals and small scales correspond to compressed signals. In the definition of wavelet transform, the scaling term is used in the denominator. Therefore, scales $s > 1$ dilate the signals whereas scales $s < 1$ compress the signal. The term translation is related to the location of a small window shifted through the signal. Intuitively, a CWT coefficient is an index of how closely the wavelet matches the original signal at a specific time and a particular scale. If the signal has a major component of the frequency similar to the current scale, then at the current scale the wavelet will be close to the signal at the particular location where this frequency component occurs. The CWT coefficient computed at this point in the time-scale plane thus will respond to a relatively large magnitude. With the joint of parameters s and τ, CWT provides a time-frequency representation of a signal.

Wavelet transform could naturally play an important role in data mining since it provides presentations of data that make the mining process more efficient and accurate and it can be incorporated into the kernel of many data mining algorithms (Li, Li, Zhu, & Ogihara, 2002). The study in Li et al. (2002) provides a comprehensive survey of wavelet application in data mining. In particular, the authors show how wavelet transform can be successfully applied to support the creative process of data mining in the phases of data understanding, data preparation, modeling, and evaluation. In this study, we apply CWT to investigate the temporal variation with different scales due to the non-stationary characteristics of air pollution indexes.

3.2. Self-organizing map neural networks

The SOM neural network is one of the most popular unsupervised neural network models, which quantize

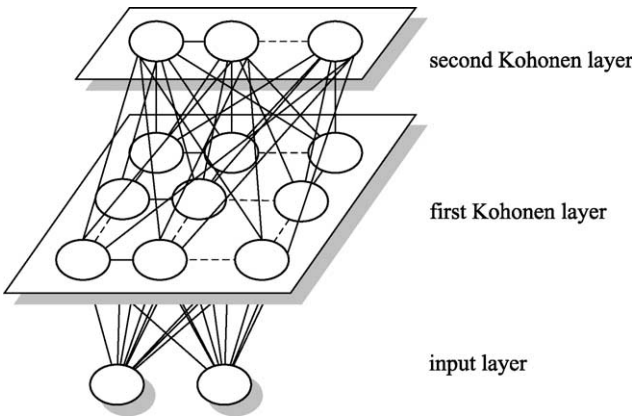

Fig. 2. The architecture of the two-level SOM neural network.

the data space and simultaneously performs a topology-preserving projection from the data space onto a regular two-dimensional grid (Kohonen, 1997). The SOM network can be used for clustering, classification, visualization and modeling. In particular, it has visualization capabilities in providing informative pictures of the data space and in exploring data vectors or whole data sets. The versatile properties of the SOM network make it a valuable tool in data mining and knowledge discovery (Vesanto, 2000).

A basic SOM network is composed of an input layer and a Kohonen layer (see Fig. 2). The input layer contains neurons for each element in the input vector. The Kohonen layer is formed by neurons which are located on a regular, usually two-dimensional grid, and are fully connected with those at the input layer. Each neuron i in the map is represented by an n-dimensional weight or reference vector $\mathbf{w}_i = [w_1, \ldots, w_n]^{\mathrm{T}}$, where n is equal to the number of neurons in the input layer. The neurons in the map are connected to adjacent ones by a neighborhood relation dictating the topological structure of the neurons. When an input vector $\mathbf{x} \in R^n$ is presented to the network, the neurons in the map compete with each other to be the winner (or the best-matching unit, BMU) b, which is the closest to the input vector in terms of some kind of dissimilarity measure such as Euclidean distance,

$$\|\mathbf{x} - \mathbf{w}_b\| = \min_i \{\|\mathbf{x} - \mathbf{w}_i\|\} \tag{2}$$

During training session, weights of neurons that are topographically close in the map within a certain geometric distance are moved toward the input $\mathbf{x}$ using the so-called 'self-organization' learning rule:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \eta h_{bi}(t)[\mathbf{x}(t) - \mathbf{w}_i(t)] \tag{3}$$

where $t = 0, 1, 2, 3, \ldots$ is the time lag, η is a small positive learning rate and $h_{bi}(t)$ is the neighborhood kernel around the BMU b at time t (Kohonen, 1997). In general, $h_{bi}(t)$ can be defined as

$$h_{ci}(t) = h(\|r_c - r_i\|, t), \tag{4}$$

where r_c, $r_i \in R^2$ are the location vectors of neurons c and i, respectively, and when $\|r_c - r_i\|$ increases, h_{ci} decreases to zero gradually. This leads to local relaxation or smoothing effects on the weight vectors of neurons in the neighborhood of the BMU. Therefore, similar input vectors are grouped into a single neuron or neighboring ones in the map when learning is accomplished.

In addition to the ability of exploring internal structures of high-dimensional data, it is of great necessity for a SOM network to delimit regions in the map for supporting cluster analysis. There are a number of different approaches to decide the number of clusters in a trained SOM network. The most popular way is to assign the number of neurons in the map to be equal to the number of expected clusters in the data set (Alam, Booth, Lee, & Thordarson, 2000; Chen, Mangiameli, & West, 1995; Mangiameli, Chen, & West, 1996). However, this approach often results in an overestimated number of clusters for larger maps. Su et al. proposed a sophisticated map transformation algorithm to analyze a unlabeled SOM network to decide about the number of cluster and the locations of the cluster centroids (Su, DeClaris, & Liu, 1997). Other alternatives can be found in Flexer (2001), Kiang (2001) and Vesanto and Alhoniemi, 2000. A comprised approach is the two-level SOM neural network which augments the conventional SOM network by an additional one-dimensional Kohonen layer in which each neuron is connected to the ones in the previous Kohonen layer (Li, 2002; Martín-del-Brío & Medrano, 1995). Fig. 2 shows the architecture of the two-level SOM network. The training process consists of two successive steps in which the weight vectors in the first Kohonen layer are fed into the second Kohonen layer as inputs. Table 1 summarizes the learning algorithm of the proposed two-level SOM network.

Upon completing the training, the first Kohonen layer will contain the topological relations of the input vectors whereas the second Kohonen layer will represent the clusters identified from the input data. An interoperable web-aware data mining system based on the proposed two-level SOM network is constructed by applying RMI and high-level code wrapper mechanisms of Java distributed object computing to address the issues of interoperability in heterogeneous environments. The details of design and implementation

Table 1
Learning algorithm of the two-level SOM network

[Step 1.1] Initialize randomly weights of neurons in the first Kohonen layer, learning rate, and winner neighborhood
[Step 1.2] Determine the winner using Eq. (2)
[Step 1.3] Update the weights of the neighborhood of the winner using Eq. (3)
[Step 1.4] Decrease the neighborhood of the winner appropriately using Eq. (4)
[Step 1.5] Repeat Steps 1.2–1.4 while the learning of Step 1 proceeds
[Step 2.1] Initialize weights of neurons in the second Kohonen layer using ones in the first map. Define the learning rate, and winner neighborhood
[Step 2.2] Iterate Steps 1.2–1.4 when the learning of Step 2 progresses

were presented in (Li, 2002). In this study, we adopt the two-level SOM network approach to mine the air pollution data and identify the number of clusters.

4. Mining procedures

Data mining is an iterative process which involves various steps. A reference model for data mining will greatly help the success of the problem under investigation. In this section, we present the process model used in mining air quality data by adapting the industry standard, the so-called CRISP-DM model (Wirth & Hipp, 2000). Fig. 3 shows the self-explanatory model and the associated steps, which will be discussed in detail in this section.

4.1. Data collection

Each EPA station, presently, posts its hourly air pollution data that includes Pollutant Standard Index (PSI) and various air pollutants on its Web site (see Fig. 4 for example). Hence, the first part of the mining procedure is to find a cost-effective way to acquire the PM10 data that is only a subset of the published Web data. A Java-based Web spider (some called it a crawler or robot) was developed that can automatically visit each EPA Web page, read the pages of interest by traveling through all related hypertext links, and parse and extract the particular table that contains PM10. As a result, the data source for this research contains 71 time series (stations), each of which has 365 times 24 data points and is stored in a local database.

4.2. Missing value handling

Field stations of TAQMN automatically record levels of various pollutants, which are then transmitted back to the monitoring center of EPA via telephone lease lines before becoming official data. Due to problems of instrument malfunction, communication noise, and/or other unknown reasons, around 10% of data are missing during the transmission process. These missing values have to be filled for the classification of clusters to be carried out properly. Several methods have been proposed for tackling the issue of filling the missing values in spatio-temporal data analysis, and they are normally based on the behavior of the known values of the variables. These include normal ratio method (NR), inverse-distance weighting (IDW), optimal interpolation (OI), multiple regression using the least absolute deviation criterion (MLAD), the single best estimator, and the median (MED) of the previous five methods (Eischeid, Baker, Karl, & Diaz, 1995). In this study, we applied the IDW method to estimate the missing air pollution data for its simplicity (NWSRFS, 1996). IDW is a simple distance-weighted 'area average' estimate of the value at the target station. It is based on the assumption that surrounding stations are related to the target station by their proximity to the target station. Thus, the value of a missing data A can be estimated as a weighted average of its surrounding stations. Let A be the original point and the relative coordinates of the nearest eight data from A are $(\Delta x_B, \Delta y_B)$, $(\Delta x_C, \Delta y_C)$,…, and $(\Delta x_I, \Delta y_I)$, the weights are

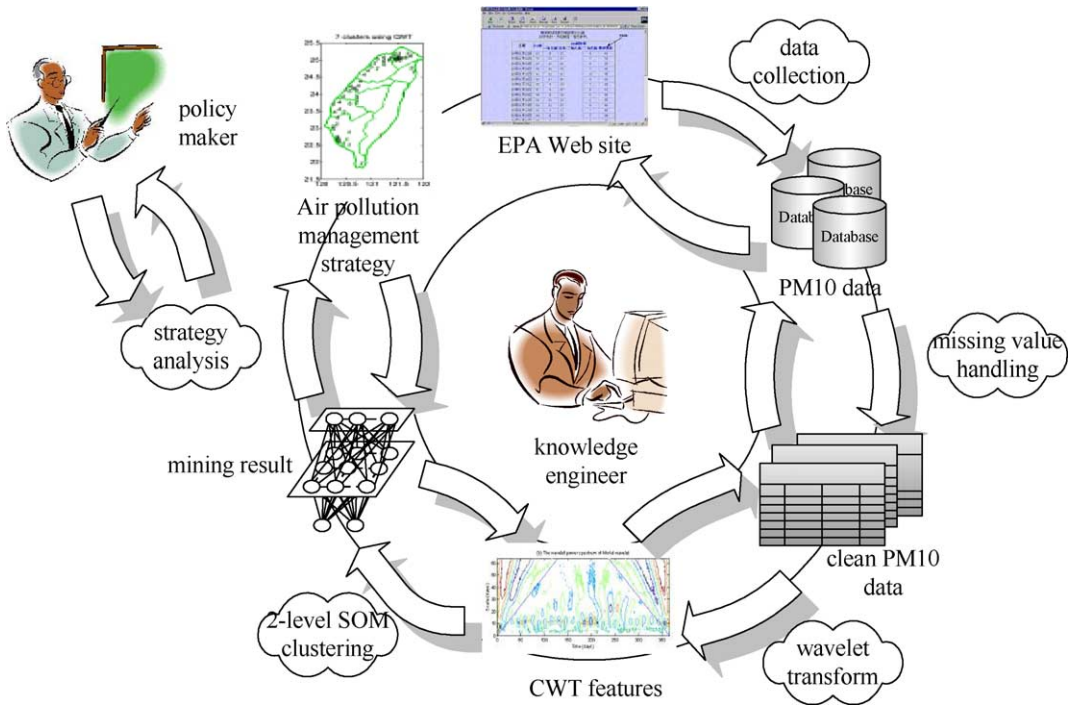

Fig. 3. The data mining process for air pollution.

Fig. 4. Sample of air pollutant data on EPA Web site.

reciprocals of the sums of the squares of Δx and Δy, i.e.

$$D_m^2 = \Delta x_m^2 + \Delta y_m^2 \qquad m = B, \ldots, I \tag{5}$$

$$W_m = \frac{1}{D_m^2} \qquad m = B, \ldots, I \tag{6}$$

The missing value of A is estimated as

$$A_{\text{est}} = \frac{\sum W_m \times P_m}{\sum W_m} \qquad m = B, \ldots, I \tag{7}$$

where P_m is the value of PM10 at station m.

4.3. Data transform

The choice of an appropriate wavelet function is crucial when performing CWT. The wavelet function has to satisfy the admissibility condition (i.e. zero mean) and localization support (i.e. fast decay from its center) in both time and frequency space. There are many wavelet functions proposed in the literature, in which the Morlet wavelet is the most commonly used wavelet for time-scale analysis (Mertins, 1999):

$$\psi(t) = \mathrm{e}^{iw_0 t}\mathrm{e}^{-t^2/2}, \tag{8}$$

where w_0 is a non-dimensional constant. One important feature of the Morlet wavelet is attributed to its best time-frequency localization constrained by the Heisenberg Uncertainty Principle (Rioul & Vetterli, 1991). According to the Heisenberg Uncertainty Principle in quantum mechanics, the higher resolution in the time domain will cause the lower resolution in the frequency resolution, and vice versa. The Morlet wavelet, with a relatively wide shape in the time domain, has a relatively narrow shape in the frequency domain (i.e. can detect fine features) and thus meets the requirements in extracting fine structures from the time series in support of a more accurate cluster analysis.

In order to investigate the behavior of a signal at each scale and translation, the scalogram of the signal is defined as $|\Phi_f^{\psi}(\tau, s)|^2$. The magnitude displayed in the scalogram reflects the correlation between the time series and wavelet function at different scales. A relative high magnitude at a certain time and scale indicates that the high similarity between the shape of the time series and the shape of wavelet function. Fig. 5(a) shows the daily PM10 value averaged over 24 h for a station in a year and Fig. 5(b) displays the corresponding wavelet scalogram using the Morlet wavelet. The x-axis is the time of the data observed in days and the y-axis is the wavelet scale, which represents characteristic modes of variability within the data. The darker the color is, the higher the value of the coefficients. The trapezoid indicates the cone of influence (COI), the region of the wavelet spectrum where the edge effects at both ends of the time series are negligible beyond these two straight lines (Torrence & Compo, 1998). With the wavelet analysis, one can see variations in the frequency of occurrence and amplitude of the major power spectrums. For example, scales 7 through 14 carry more significant features during 180–220 days (summer) whereas scales 18 through 24 do during 240–260 days (post-summer and pre-autumn). Since low scales correspond to high frequencies and high scales to low frequencies, this example indicates the frequency changes rapidly in winter comparing to the slower change in the post-summer and pre-autumn seasons, thus one should pay special attentions to various scales during different periods.

In this study, CWT is performed on the PM10 time series of 71 stations. We chose 1–31 (i.e. including 1-, 2-, 3-,$\cdots$, 30-, and 31-day) scales for generating CWT coefficients, which are then used as inputs to the two-level SOM network for mining.

4.4. Mining and performance evaluation

The mining stage applies a two-level SOM neural network and leads to the clustering of data, as discussed in Section 3.2. The topology of the first-level self-organization map is an 8×8 grid, whereas the second-level is a $1 \times C$ grid, where C is the number of clusters under investigation. In evaluating the performance of clustering results produced by the SOM neural network, the cohesion measure of a cluster (Kiang, Kulkarni, & Tam, 1995) is normally used, which can indicate how close stations are within a cluster. The cohesion is defined in the following:

$$P_i = \frac{\sum_{m \neq n} C_{mn}}{\binom{k_i}{2}} \quad (m, n \in \text{cluster } i), \tag{9}$$

where C_{mn} is the dissimilarity coefficient between stations m and n in cluster i. k_i is the total number of stations within cluster i. Thus, a smaller P_i value indicates a better similarity among stations within the same cluster i.

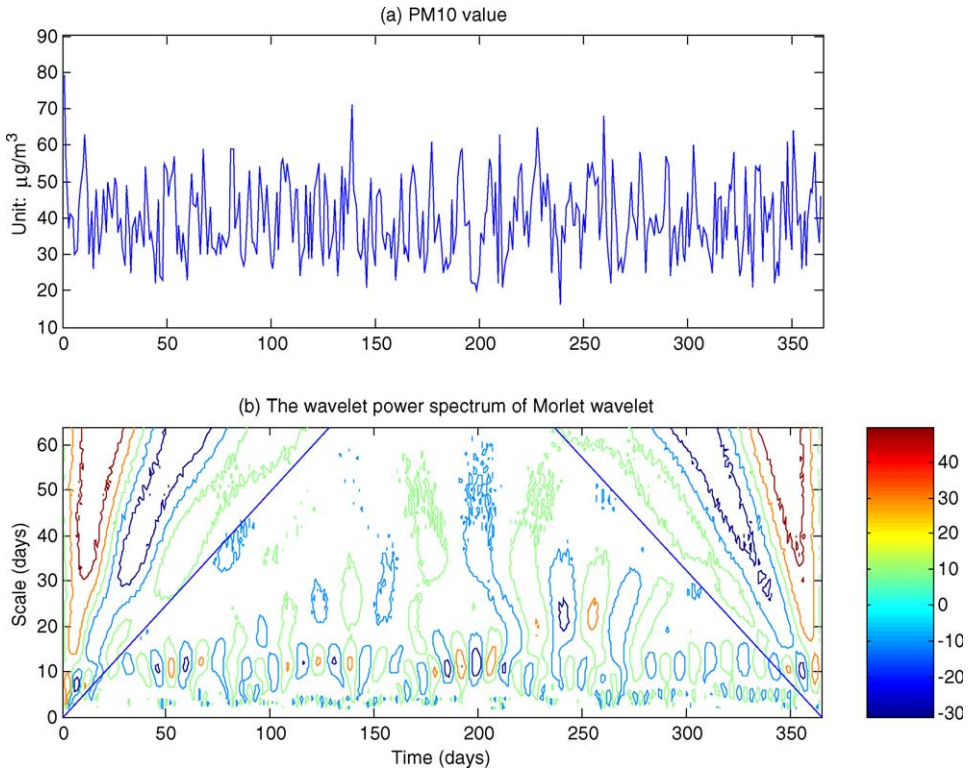

Fig. 5. The scalogram of PM10 time series using the Morlet wavelet.

In addition to the cohesion measure, the variance is often used to measure the dispersion of stations in a cluster; it indicates how evenly the similarities among stations within a cluster are distributed around the mean P_i. The variance of cluster i is defined in the following:

$$V_i = \sum_{m \neq n} \frac{(P_i - C_{mn})^2}{\binom{k_i}{2} - 1} \tag{10}$$

The overall quality of a clustering distribution can then be measured by the average cohesion, P, and the variance, V, which are defined as follows:

$$P = \frac{\sum_{i=1}^{r} (k_i - 1) P_i}{k - r} \tag{11}$$

$$V = \sum_{i=1}^{r} \left(\frac{k_i - 1}{k - r} \right) V_i \tag{12}$$

where k is the total number of stations, and r is the number of clusters. A lower value of P indicates better cohesion among clusters of a mining result. Similarly, a lower value of V indicates more uniformly distributed around the cluster means.

In the following, we conducted a preliminary study to compare the cohesion (P) and its variance (V) of CWT and those of single scale: daily, weekly, bi-weekly, monthly, and seasonal. With the cluster number set to 7, results are shown in Table 2. It is obvious from the table that, within the single-scale range, finer scales (such as daily, weekly, etc.) could result in larger average cohesion and variance, since finer scales provide finer views and focus on detailed matching. Conversely, coaster scales (such as seasonal and monthly) emphasize more on the analogous trends embedded in signals and thus reduce dissimilarities among clusters. The table also shows that the average cohesion and variance of CWT are comparatively larger than single-scale cases, it is because CWT has to process continuous scales that cover embedded features from smoothed to detailed characteristics. These features are required in understanding pollution distribution in both short term as well as longer term time frame, thus, CWT is applied as the method for classifying clusters.

We have to determine a good number of clusters for the CWT to work with. The selection of a good number of clusters is problem-dependent, and is, in general, based on the principle of minimizing the interdistance and maximizing the intradistance among clusters. However, due to

Table 2
The comparison of average cohesion and variance of different scales

Data	P	V
Daily	0.5089	0.2670
Weekly	0.3457	0.0155
Bi-weekly	0.2882	0.0114
Monthly	0.2343	0.0078
Seasonly	0.1465	0.0038
CWT	0.7393	0.0550

Table 3
The average cohesion and variance of CWT based on different clusters

Clusters	*P*	*V*
4	0.7816	0.0865
5	0.7739	0.0826
6	0.7513	0.0640
7	0.7393	0.0550
8	0.7389	0.0582

the missing data factor that was filled using estimated value, the scores that were computed from minimizing the interdistance could be distorted and may not reflect its true values. Hence, we investigated 4–8 clusters, and consulted results with domain experts. Table 3 shows the average cohesion and variance of CWT for different number of clusters.

It can be seen that the *P* value decreases as the number of clusters increases, because, for a given data set, a larger number of clusters will lead to fewer stations in each cluster, and therefore the stations within a cluster are more likely to be closer to each other. One would notice that the decrement of *P* value from 7-cluster to 8-cluster seems insignificant comparing with all previous ones, while the *V* value has reached its minimal at the 7-cluster case. Our domain expert decided that 7 is the number of clusters to be used in the following research.

5. Mining discovery

Table 4 shows the statistics of mining results using 7 clusters, which includes the number of stations covered by each cluster, mean and standard deviation of PM10 distribution in each cluster. (The cluster identification numbers shown in the case study have been ranked intentionally for a clearer description). It is very clear that cluster 7, with mean 33.9 and Std 9.8, represents the best quality area, while cluster 1 is the worst area with mean 59.5 and Std 24.9. For the rest, one can see that there is little difference between clusters 5 and 6, while the remaining clusters 2, 3, and 4 vary only a little.

Fig. 6 plots the location distribution of each cluster to indicate the geographical distribution of the mining results, in which the background is the official administration

Table 4
Statistics of cluster distribution (unit: $\mu g/m^3$)

Cluster	Stations	Mean	Std
1	15	59.5	24.9
2	5	53.8	21.8
3	5	51.2	20.9
4	13	49.5	17.6
5	4	43.8	15.5
6	20	43.3	14.9
7	9	33.9	9.8

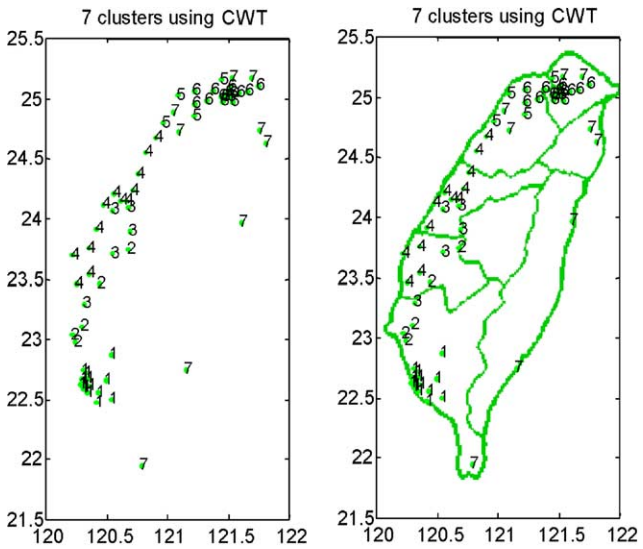

Fig. 6. The location distribution of each cluster.

districts that forms the basis of present government pollution control boundary, as shown in Fig. 1. Geographically, most low pollution areas of cluster 7 falls on the eastern region, which consists of regions of Ilan and H-T and is separated from the industrialized western region by the central mountain range. Its mean value is almost 50% less than that of cluster 1 of the heavily industrialized Kaohsiung area, a part of the K-T-P region, and its Std is more than 50% less. All major high pollution areas fall on the western seaboard, and the pollution situation generally lessens as it moves toward north. The T-K region at the top of the map consists of clusters 5, 6, and 7, and has the second best air quality of the nation because of the almost non-existence of heavy industry and its mountainous area. The next region T-H-M has a much wider area and is characterized by its rigid mountainous terrain at the north, the internationally known Hi-tech development area is situated in this region. However, as the rigid terrain disappears toward the south, more industrial settings were established, as a result, this region spans across clusters 4, 5, 6, and 7. The next two regions, T-C and Y-C, both belong to the same western plain and are mainly dominated by the same industrial settings of clusters 3 and 4. The worst region is T-K-P, its heavy industry and petrochemical industry have contributed heavily towards air pollution, and resulted in many cluster 1s and 2s and a high standard deviation as well, in addition to some cluster 4s. The most surprising of all is the fact that there is a cluster 7 at the very tip of this heavily polluted region, which is clearly out of the character of the T-K-P region, and instead it should be a part of the H-T region.

It is clear that, except the eastern region, every administration district consists of at least two different clusters; some even have four clusters. The T-K-P region, in particular, is made up of four clusters and spans over seven cluster levels from 1 to 7. Based on these findings, we feel

strongly that the effectiveness of pollution control policy, which is entirely based on present administration districts, may be further improved by taking into consideration the actual distribution of pollutants. It is particularly true that the utilization of the 60% refund by the local government should be guided by the locality of seriousness of pollution, which may be best described in zones rather than present official districts.

6. Conclusions

The heavy environmental loading has led to the deterioration of air quality in Taiwan in the past two decades. The task of controlling and improving air quality has attracted a great deal of national attention. The government has since adopted an array of measures to combat this problem. This study applies data mining to identify the national PM10 pollutant distribution, with data retrieved from 71 monitoring stations of the nation. The mining results are presented to contrast the present pollution districts, which could serve as an important reference for the policy maker in formulating future policies.

In carrying out this study, we first retrieved relevant PM10 data of 1 year from archived information of 71 stations, and then filled in all missing data. The data is of spatio-temporal nature, hence, we paid particular attention in investigating the scale issue of time series data, and decided to apply continuous wavelet transform, so that mining results may be applicable for either short or long term reference purposes. SOM neural network was applied to identify clusters in such a high-dimensional space. The results confirm that regions determined from the wavelet transform approach can reduce the local small regions using the small scale input data and improve the over-smoothed regions using one large scale input data. Most important of all, the results clearly indicate the distribution of national PM10 pollutant through 7 clusters and their individual severity. Mapping the findings onto the present air quality districts, one is shocked to learn that there are from 2 to 4 clusters in a district, and a district could span from 2 to 7 cluster levels. Based on these findings, we feel strongly that the effectiveness of present pollution control policy, which is entirely based on convenience of administration, may be further improved by taking into consideration the grouping of pollutants, which may be best described in zones.

One limitation of the current study is the fact that the SOM network can only provide the capability for hard clustering, meaning that data can only be assigned to one and only one cluster. However, most of the environment science data have spatial transition characteristics in a time period, which means there could be a transition zone between adjacent regions. This suggests a potential need of clustering models that could provide a data item 'membership degree' of a cluster. The fuzzy logic-based approaches such as fuzzy SOM, fuzzy c-Means, and rotated principal component analysis (RPCA) approach can be potential candidates for future works on this aspect.

Acknowledgements

The authors would like to thank Dr Jeng Jong Pan for his valuable advises on continuous wavelet transform. This work was supported in part by NSC87-2213-E-327-001, Taiwan, ROC.

References

Abidi, S. S. R. (2001). Knowledge management in healthcare: Towards knowledge-driven decision-support services. *International Journal of Medical Informatics*, *63*, 5–18.

Alam, A., Booth, D., Lee, K., & Thordarson, T. (2000). The use of fuzzy clustering algorithm and self-organizing neural networks for identifying potentially failing banks: An experimental study. *Expert Systems with Applications*, *18*(3), 185–199.

Bose, I., & Mahapatra, R. K. (2001). Business data mining—A machine learning perspective. *Information and Management*, *39*(3), 211–225.

Chen, S., Mangiameli, P., & West, D. (1995). The comparative ability of self-organizing neural networks to define cluster structure. *Omega, International Journal Management Science*, *23*(3), 271–279.

Eischeid, J. K., Baker, C. B., Karl, T., & Diaz, H. F. (1995). The quality control of long-term climatological data using objective data analysis. *Journal of Applied Meteorology*, *34*, 2777–2795.

EPA, (2000). *Air quality protection in 25 years 1975–2000, Taiwan.*

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34.

Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization. *Intelligent Data Analysis*, *5*(3), 373–384.

Kiang, M. Y. (2001). Extending the Kohonen self-organizing map networks for clustering analysis. *Computational Statistics and Data Analysis*, *38*(2), 161–180.

Kiang, M. Y., Kulkarni, U. R., & Tam, K. Y. (1995). Self-organizing map network as an interactive clustering tool—An application to group technology. *Decision Support Systems*, *15*(4), 351–374.

Kohonen, T. (1997). *Self-organizing maps*. Berlin: Springer.

Li, S.-T. (2002). A web-aware interoperable data mining system. *Expert Systems with Applications*, *22*(2), 135–146.

Li, T., Li, Q., Zhu, S., & Ogihara, M. (2002). A survey on wavelet applications in data mining. *ACM SIGKDD Explorations*, *4*(2), 49–68.

Liao, S.-H. (2003). Knowledge management technologies and applications—Literature review from 1995–2002. *Expert Systems with Applications*, *25*(2), 155–164.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Pattern Analysis and Machine Intelligence*, *11*(7), 674–693.

Mangiameli, P., Chen, S., & West, D. (1996). A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research*, *93*(2), 402–417.

Martín-del-Brío, B., & Medrano, N. (1995). Feature map architectures for pattern recognition: Techniques for automatic region selection. In D. W. Pearson, N. C. Steele, & R. F. Albrecht (Eds.), *Artificial neural nets and genetic algorithms* (pp. 124–127).

Mertins, A. (1999). *Signal analysis-wavelets, filter banks, time-frequency transforms and applications*. New York: Wiley.

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, *13*(1), 3–14.

National Weather Service River Forecast System (NWSRFS), (1996). *National Oceanic and Atmospheric Administration, USA*.

Pyle, D. (1999). *Data preparation for data mining*. Los Altos, CA: Morgan Kaufmann.

Rainsford, C. P., & Roddick, J. F. (1999). Database issues in knowledge discovery and data mining. *Australian Journal of Information Systems*, *6*(2), 101–108.

Read, B. J (2000). *Data mining and science?* CLRC Rutherford Appleton Laboratory, Chilton, Didcot, Oxon, UK. http://www.ercim.org/publication/ws-proceedings/12th-EDRG/EDRG12_Re.pdf.

Rioul, O., & Vetterli, M. (1991). Wavelets and signal processing. *IEEE Signal Processing Magazine*, *Oct*, 14–38.

Su, M.-C., DeClaris, N., & Liu, T.-K. (1997). Application of neural networks in cluster analysis. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 1–6.

Torrence, C., & Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, *79*(1), 61–78.

Vesanto, J (2000). *Using SOM in Data Mining*. Licentiate's thesis. Helsinki University of Technology, Finland.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*(3), 586–600.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 29–39.

Zhou, Z.-H. (2003). Three perspectives of data mining. *Artificial Intelligence*, *143*(1), 139–146.

Sheng-Tun Li received his BS and MS degrees in computer engineering from Tamkang University, Taiwan, and the PhD degree in computer science from University of Houston, University Park, USA.He is currently an associate professor of Institute of Information Management at National Cheng Kung University, Taiwan. Prior to joining NCKU, he was on the faculty of Department of Information Management at National Kaohsiung First University of Science and Technology, Taiwan, during 1996–2003. His research interests include intelligent decision support systems, data mining, knowledge management, soft computing, and Java interoperability computing. Dr Li is a member of the IEEE/CS and ACM.

Li-Yen Shue graduated from the National Chiao Tung University, Taiwan, finished his PhD from Industrial Engineering Department of Texas Tech University in 1976. He taught in National Tsing Hua University and National Sun Yet-Sen University in Taiwan, and was with Information System Department of Wollongong University Australia. Dr Shue is currently a Professor in the Department of Information Management of the National Kaohsiung First University of Science and Technology, Taiwan. His research interests include decision science, system simulation, production system scheduling, and knowledge management.