Chemosphere xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Chemosphere

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

Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach

Ho-Wen Chen [a,1], Ching-Tsan Tsai [b,1], Chin-Wen She [a], Yo-Chen Lin [b], Chow-Feng Chiang [b,*]

[a] Department of Environmental Science and Engineering, Tunghai University, 181 Section 3, Tauchung Port Road, Taichung 407, Taiwan
[b] Department of Public Health, China Medical University, 91 Hsueh-Shih Road, Taichung 404, Taiwan

ARTICLE INFO

Article history:
Received 9 March 2010
Received in revised form 27 July 2010
Accepted 8 August 2010
Available online xxxx

Keywords:
Cluster analysis
High-tech industrial park
Inorganic acid and base
Spatial and temporal background distribution

ABSTRACT

Air pollution data around a monitored site are normally difficult to analyze due to highly inter-related meteorological and topographical factors on top of many complicated atmospheric chemical interactions occurred in local and regional wind fields. The challenge prompts this study to develop a comprehensive data-mining algorithm of cluster analysis followed by meteorological and interspecies correlations to mitigate the inherent data complexity and dissimilarity. This study investigated the background features of acidic and basic air pollutants around a high-tech industrial park in Taiwan. Monthly samplings were taken at 10 sites around the park in a year. The temporal distribution plots show a baseline with two characteristic groups of high and low peaks. Hierarchical cluster analysis confirms that high peaks were primarily associated with low speed south wind in summer for all the chemical species, except for F^-, Cl^-, NH_3 and HF. Crosschecking with the topographical map identifies several major external sources in south and southwest. Further meteorological correlation suggests that HCl is highly positively associated with humidity, while Cl^- is highly negatively associated with temperature, both for most stations. Interestingly, HNO_3 is highly negatively associated with wind speed for most stations and the hotspot was found in summer and around the foothill of Da-Tu Mountain in the northwest, a stagnant pocket on the study site. However, F^- is highly positively associated with wind speed at downwind stations to the prevailing north wind in winter, indicating an internal source from the north. The presence of NH_4^+ stimulates the formation of NO_3^-, SO_4^{-2} ($R = 0.7$), and HNO_3, H_2SO_4, NH_3 ($R = 0.3$–0.4). As H_2SO_4 could be elevated to a level as high as 40% of the regulated standard, species interactions may be a dominate mechanism responsible for the substantial increase in summer from external sources.

1. Introduction

As of 1998, 85 industrial parks have been officially built, operated or planned in Taiwan to ensure the island's continual economic growth. Due to the limited land area in Taiwan, most industrial complexes are inevitably situated adjacent to urban development areas where people reside and work. The Taichung site of Central Taiwan Science Park (CTSP) is one of the most important bases worldwide for the high-tech industries of semiconductors, electronics and electrical peripherals with a total investment of USD 25 billions. The South part of Taichung site is surrounded by the Greater Taichung area residing about 1.5 millions populations. Inorganic acidic (HF, HCl, HNO_3, H_2SO_4) and basic (NH_4OH) chemicals are the concerned pollutants emitted from the cleaning and etching processes (Tsai et al., 2003). However, some of these pollutants may be attributed to external sources such as urban activity, power plant or other combustion processes (Orel and Seinfeld, 1997). In addition, ocean is another potential source, which may compound the complexity in source identification. Accordingly, understanding the temporal and spatial background features of pollutants around the site is critical to addressing the public concerns and searching for the responsible parties in environmental forensics (Gargava and Aggarwal, 1996; Neves, 1996). Many early studies covered a variety of aspects, including formation mechanisms (Baek et al., 2004; Tsai et al., 2004), health effects (Spengler et al., 1990), concentration contours, hot spots and source identification (Danalatos and Glavas, 1999; Bari et al., 2003; Walker et al., 2004; Chittaro et al., 2006; Horng and Cheng, 2008; Zhu et al., 2008; Hosseinlou and Sohrabi, 2009), and meteorological effects (Webb and Vincent, 1999; Jäggi et al., 2006; Pejman et al., 2009). However, little effort was made to develop a comprehensive data analysis approach for complicated air pollution systems like the Taichung site.

* Corresponding author. Tel.: +886 4 22053366; fax: +886 4 22037717.
E-mail address: amur.chiang@gmail.com (C.-F. Chiang).
[1] H.W. Chen and C.T. Tsai have equal contribution to this study.

0045-6535/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chemosphere.2010.08.019

Prior to the full operation, the site was monitored every month for five gaseous species (HCl, HF, HNO_3, H_2SO_4 and NH_3) and five fine particles ionic species (Cl^-, F^-, NO_3^-, SO_4^{-2} and NH^{+4}) at 10 stations around the site for 1 year. The primary objective of this study was to develop a comprehensive data-mining algorithm to explore the background features of acidic and basic pollutants around the Taichung site. Special effort was made to identify potential external sources to the study site under the meteorological and topographical constraints. Accordingly, an algorithm employing hierarchical cluster analysis (HCA) followed by meteorological and interspecies correlation analysis was developed in this study so that monitored data with specified similarity could be simplified into several groups for better insight.

2. Materials and methods

2.1. Study site and sampling stations

The Taichung site of the Central Taiwan Science Park (CTSP) launched its first test run in middle 2005, but was not fully operated until early 2007. Fig. 1a shows a site map of the Taichung site, having a total area of 413 hectares. The east and south of the site are immediately connected to the Greater Taichung Metropolitan area with 1.5 millions of population. About 5 km to the south is the Taichung Industrial Park, a conventional machinery industry complex being heavily operated for about 30 years. About 3 km to the southwest is the Taichung Refuse Incinerator with a daily

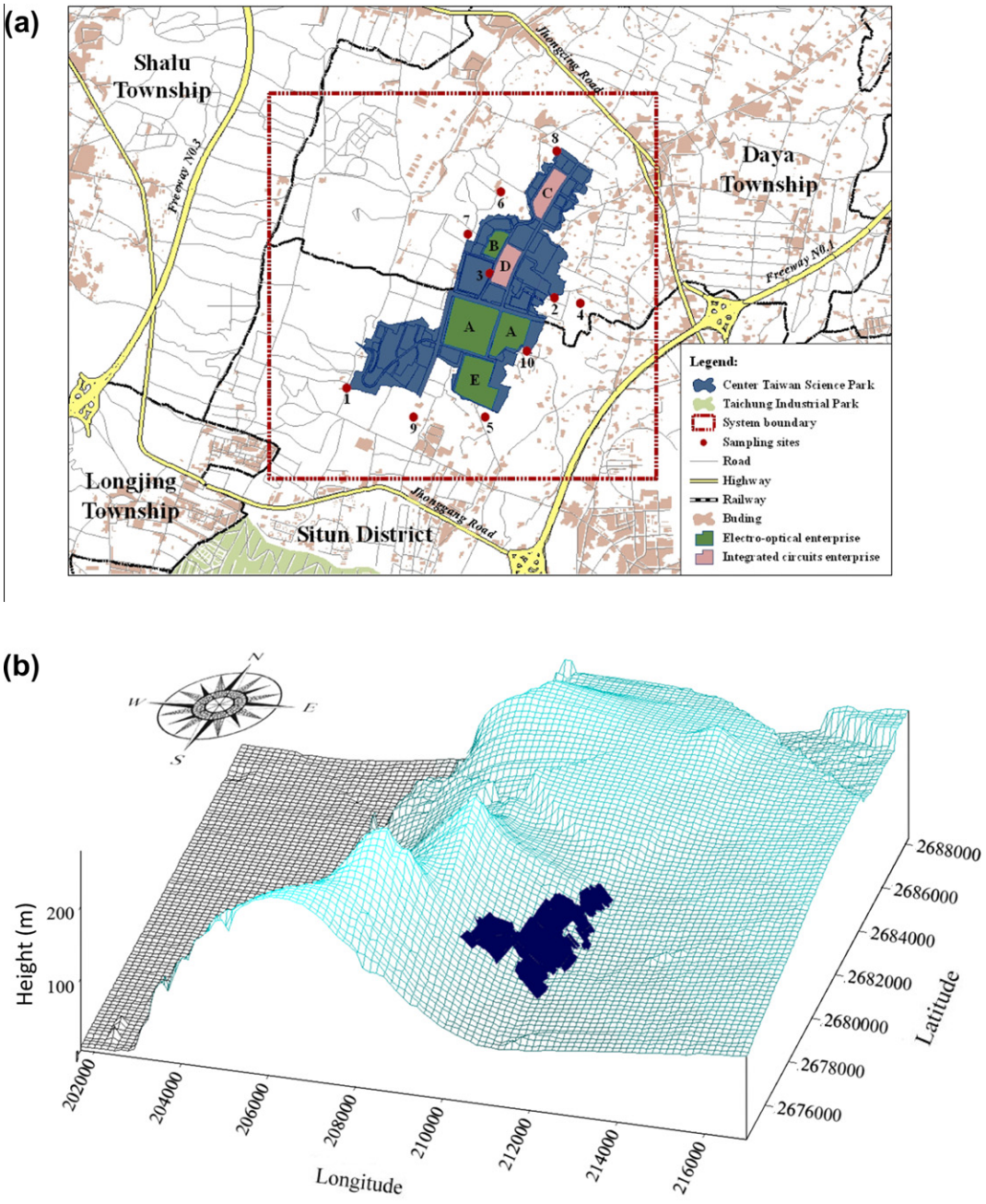


Fig. 1. The Taichung site of the Central Taiwan Scientific Park investigated in this study, showing: (a) the surrounding potential external pollution sources and 10 sampling stations around the site, (b) three-dimensional topographical map of the study site.

Please cite this article in press as: Chen, H.-W., et al. Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. Chemosphere (2010), doi:10.1016/j.chemosphere.2010.08.019

throughput of 900 wet tones. About 2 km to the east is a national highway system. The coast line of the Taiwan Strait and Taichung coal-burned power plant are about 15 km to the east of the study site. The topography of the site is characterized by a low altitude terrain of 50–300 m above the sea level, declining from northwest to southeast as shown in Fig. 1b.

As shown in Fig. 1a, a total of 10 sampling sites distribute around the boundary of the factory lots and evenly covers the entire area on the Taichung site. Stations 6–8 locates in the north, Stations 1, 5, 9 in the south, and 2, 4, 10 in the east. Station 3 is the only site located within the factory complex.

2.2. Sampling device and schedule

In each sampling event, 10 chemical species were monitored, including five ionic species (NH_4^+, F^-, Cl^-, NO_3^- and SO_4^{-2}) on particulate matter and five gaseous species (HF, HCl, NH_3, HNO_3 and H_2SO_4). The air samples were collected by a porous metal denuder which was well reported by Tsai et al. (2003, 2004). The denuder contains three primary structured units: a single-stage impactor with a cutoff aerodynamic diameter of 2.5 μm, two porous metal discs with a pore size of 100 μm with a disc diameter of 4.7 cm, and a three-stage filter pack. As air samples enter into the first stage impactor, coarse particles are removed. The remaining gaseous pollutants are then absorbed by a layer of sodium carbonate (5% w/v) and citric acid (4% w/v) coated on each of the porous metal discs for acidic and basic species. The filter packs consist of a Teflon filter (Gelman Science, 2 μm pore size) to collect fine particles, a 4.7-cm nylon filter (Gelman Science, 1 μm pore size) to collect HNO_3 and HCl, and a 4.7-cm glass fiber filter (AP40, Millipore Inc.) coated with citric acid to collect NH_3 that volatilizes from the collected particles on the first layer of Teflon filter. Particulate NO_3^- and Cl^- concentrations were determined as the sum of those collected on the Teflon filter and the nylon filter, respectively. The flow rate was maintained at 2 L min^{-1}. Samplers were placed at an average height of 1.5–2 m from the ground level. The samplings were taken simultaneously at all 10 sampling sites once every month for a period of 18 h in a day throughout the entire study from March 2005 to February 2006, prior to the full-scaled operation of the site.

2.3. The algorithm of data analysis

A total of 1120 (eight stations × 12 months + two stations × 8 months) valid air data were obtained, each monitoring 10 chemical species for a period of 12 months from March 2005

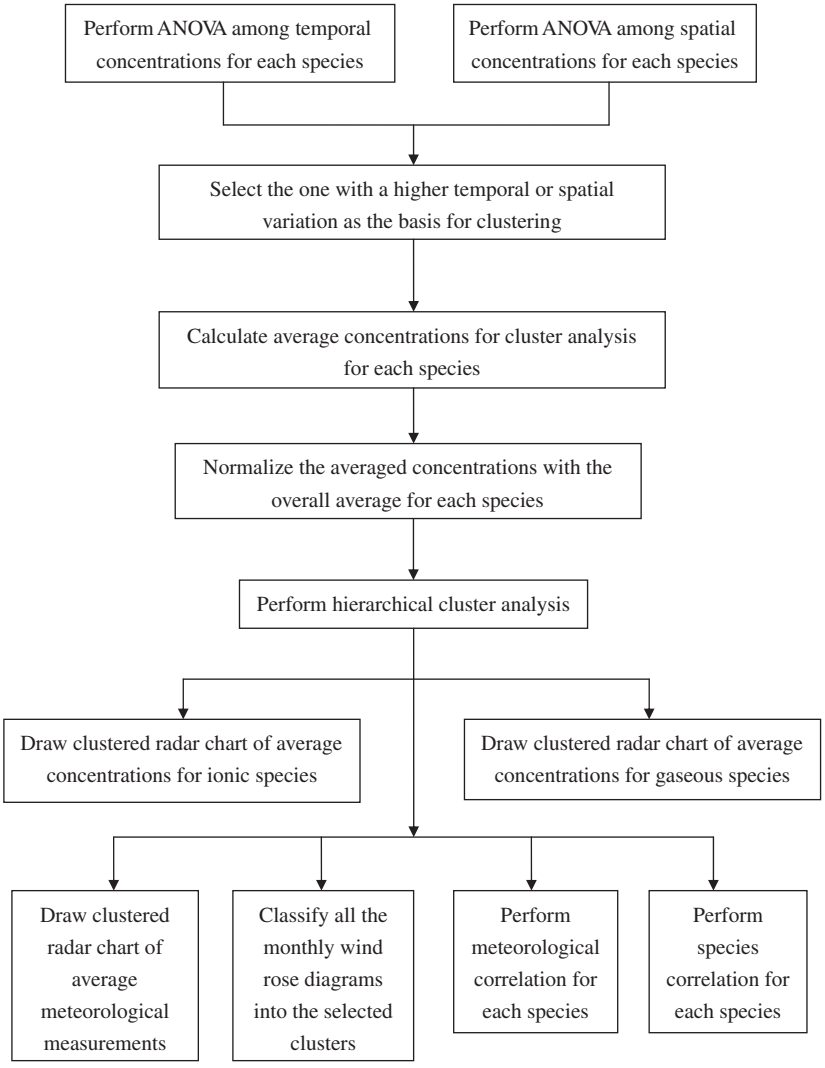


Fig. 2. The algorithm of data mining approach developed in this study, incorporating a preliminary ANOVA and hierarchical cluster analysis followed by a confirmatory meteorological and interspecies correlation.

Please cite this article in press as: Chen, H.-W., et al. Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. Chemosphere (2010), doi:10.1016/j.chemosphere.2010.08.019

to February 2006. The sampling was collected once every month. On each sampling date, four meteorological parameters (wind speed, wind direction, ambient air moisture, and ambient air temperature) were also obtained each hour for 14 h (5 am–7 pm) from a local meteorological station. This gives a total of 762 (four parameters × 14 h × 12 months) meteorological data. The air pollutant and meteorological data were analyzed using a novel hierarchical cluster analysis (HCA) approach developed in this study. For the type of air dispersion concentration data inherent with high dissimilarity like this study, it is advantageous to aggregating measurements into different groups with spatial or temporal similarity (Lee et al., 2001; Reghunath et al., 2002; Banerjee et al., 2009). The HCA algorithm shown in Fig. 2 consists of eight steps and are briefly stated.

1. For each chemical species, use one way ANOVA to determine whether the monitored concentrations are significantly different among the sampling sites and/or among sampling times during the entire course of study.
2. Select the one with a higher temporal or spatial variation as the basis for clustering.
3. If temporal basis is selected, calculate the spatial-averaged concentration for each sampling time and each chemical species, and vice versa. For easy understanding, the temporal basis is assumed in the following steps.
4. Calculate the yearly arithmetic mean of all the station-averaged concentration for each chemical species obtained in Step 3.
5. Divide the monthly station-averaged concentration of each chemical species by the corresponding yearly mean concentration to obtain the corresponding standardized unitless concentration for all the sampling months.
6. Calculate the Euclidean distance (d_{ij}) between the two 10 species-dimensioned points of any two sampling months, $d_{ij} = \sqrt{\sum_{p=1}^{10}(x_{ip} - x_{jp})^2}$ i, j denoting any two sampling months (1–12), p denoting the type of pollutant (1–10). It can be realized that $d_{ij} = d_{ji}$ and $d_{ij} = 0$ as $i = j$. A total of 72

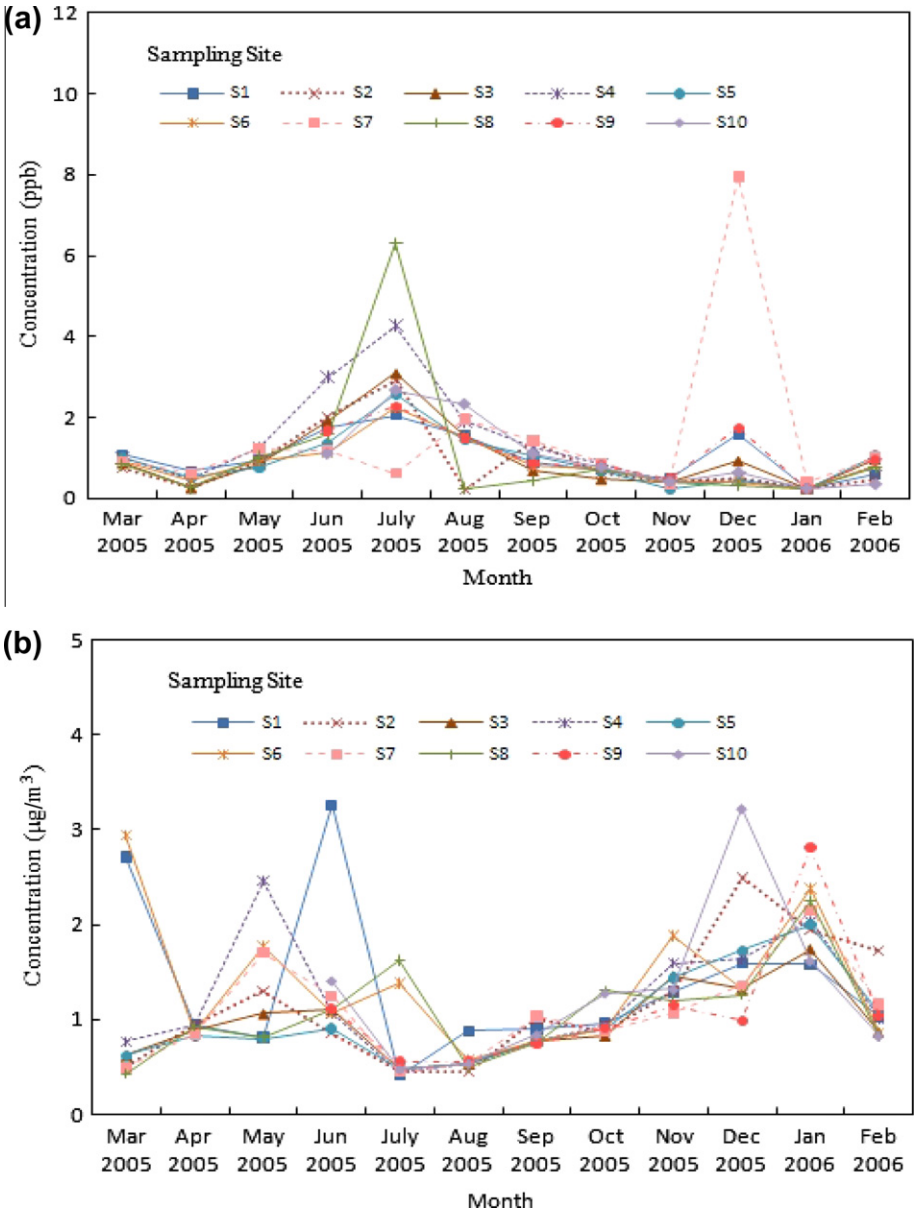


Fig. 3. Results of the temporal distributions of concentrations monitored in each month for the 10 sampling sites in this study, illustrating a pattern of two groups of peaks for: (a) HNO_3 and, (b) Cl^-.

Please cite this article in press as: Chen, H.-W., et al. Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. Chemosphere (2010), doi:10.1016/j.chemosphere.2010.08.019

$(12 \times 12 \div 2)$ d_{ij} are calculated and used as the basis for clustering in this study.

7. Plot the radar chart of average concentration for each clustered group determined in Step 6 by superimposing all the clusters on the same chart for comparison. The radar chart is useful for visual inspection as to whether there is a distinct separation in concentration among the clustered groups for each chemical species.
8. Plot wind roses for each month using the hourly wind data and categorize each month of wind roses in the corresponding cluster as determined in Step 6. The clustered wind rose diagrams are useful in evaluating the effect of clustered wind data on the clustered radar chart of concentrations for each chemical species.
9. Perform correlation analysis between concentration and each meteorological parameter for each species and each station.
10. Perform correlation analysis between the station-averaged concentrations of any two species.

3. Results and discussion

3.1. Background temporal and spatial profiles

To examine whether a long-term temporal or spatial distribution pattern existed for the gaseous and particulate species monitored in this study, their levels versus sampling months were superimposed on the same plot for all the 10 sampling stations for each chemical species. Fig. 3 illustrates the distribution profiles of HNO_3 and Cl^- from March 2005 to February 2006 (12 months). In general, these plots display varied temporal and spatial variations for different chemical species. However, a general pattern of baseline with two characteristic groups of high and low peaks appearing in summer and winter could be distinctly identified for each species. For instance, as shown in Fig. 3a, HNO_3 had a baseline of 1–2 ppb with a summer peak of 4–7 ppb and a winter peak of 3–4 ppb. Nevertheless, an exceptional high summer peak of 7 ppb appeared at Station 8 and a winter peak of 8.5 ppb at Station 7 for HNO_3. Interestingly, the two hotspots were located around the foothill of Du-Tu Mountain to the northwest of the study site. In another instance shown in Fig. 3b, particulate Cl^- also displayed the pattern of two major groups, but with much wider and larger variations among different sampling sites. The other four ionic species also demonstrated the type of distribution. Ionic species absorbed on fine particles was likely associated with long-range transport and hence affected by regional sources. Two major regional sources must be noted in this study. One is the Taichung coal-burned power plant and Taiwan Strait, about 15 km to the west of the study site. Another is the Taichung city and Taichung Industrial Park, about 5 km to the south of the study site. The variation could be also due to the ever changing surface winds (1.5–2 m above ground) over a typical complex terrain like the Taichung site. As the denuder sampling took a total of 14 h, it would be difficult to assure the actual upwind and downwind relationship.

Table 1 shows the mean concentration and percent variation coefficient (CV) of the 10 sampling stations for each sampling month and each species. For such a relatively small site of 413 hectares (2800 m × 1500 m), it would be justified to simplify the analysis by taking an arithmetic mean of the concentrations collected at all the sampling stations each month for each species. For the five gaseous species, the average concentration and the percent of the regulated standard (in parenthesis) was 0.5–2.0 $\mu g\,m^{-3}$ (5–10%) for HF, 1–4 ppb (1–4%) for HCl, and 0.5–3 ppb (1–8%) for HNO_3, 5–20 $\mu g\,m^{-3}$ (10–40%) for H_2SO_4, and 5–20 ppb (0.5–2%) for NH_3, using the factor-surrounding air quality standard (SAQS), 10 $\mu g\,m^{-3}$ for total F^- including HF, 100 ppb for HCl, 40 ppb for NHO_3, 50 for $\mu g\,m^{-3}$ for H_2SO_4, and 1000 ppb for NH_3, currently regulated by Taiwan Environmental Protection Agency (Taiwan EPA, 2007). The above analysis shows that the average background gaseous concentrations in general

Table 1
Results of mean concentrations and variation coefficients (CV) of the 10 sampling sites for each chemical species monitored in this study.

	HF		H_2SO_4		HCl		HNO_3		NH_3	
	Mean[a]	CV%	Mean	CV%	Mean	CV%	Mean	CV%	Mean	CV%
March-2005	0.6	10.0	16.1	17.0	3.1	207.4	0.9	12.2	20.4	33.1
April-2005	0.7	1.4	10.6	50.9	2.8	60.4	0.4	42.5	15.8	43.9
May-2005	0.6	5.0	11.3	34.5	1.7	20.0	1.0	17.0	9.6	31.9
June-2005	2.2	30.5	22.4	33.7	4.3	38.1	1.7	33.5	9.1	19.9
July-2005	1.1	33.6	19.8	42.7	1.2	40.0	2.9	52.1	15.5	73.9
August-2005	0.6	5.0	5.9	49.2	0.8	38.8	1.4	49.3	4.9	63.3
September-2005	1.2	56.7	6.5	38.5	1.4	13.6	1.0	32.0	16.3	29.1
October-2005	0.7	1.4	6.1	36.2	1.3	5.4	0.7	15.7	7.4	30.4
November-2005	1.2	45.8	12.1	51.3	1.2	18.3	0.4	17.5	9.5	22.5
December-2005	1.7	58.8	9.2	49.1	3.0	139.7	1.5	157.7	14.6	46.6
January-06	1.2	38.3	5.4	21.7	1.0	38.0	0.3	16.7	6.7	36.0
February-2006	0.9	40.0	4.0	41.5	1.3	23.9	0.8	30.0	6.8	33.7
	F^-		SO_4^{-2}		Cl^-		NO_3^-		NH_4^+	
	Mean	CV%	Mean	CV%	Mean	CV%	Mean	CV%	Mean	CV%
March-2005	0.5	10.0	7.0	33.3	0.8	91.3	3.3	23.3	7.0	27.4
April-2005	0.6	11.7	5.4	21.1	0.6	5.0	0.6	38.3	3.7	15.1
May-2005	0.5	4.0	6.6	27.0	0.9	45.6	1.0	13.0	3.4	30.6
June-2005	0.6	21.7	3.3	25.8	0.9	53.3	0.8	30.0	3.1	23.6
July-2005	0.5	20.0	10.6	17.6	0.5	60.0	2.0	29.5	8.3	22.2
August-2005	0.5	6.0	10.9	48.6	0.4	20.0	0.9	47.8	4.4	35.5
September-2005	0.5	10.0	2.2	24.6	0.6	11.7	0.9	50.0	2.8	40.4
October-2005	0.5	2.0	2.4	36.3	0.7	17.1	0.5	38.0	2.6	6.9
November-2005	0.5	2.0	5.3	21.5	0.9	17.8	1.0	24.0	4.8	15.0
December-2005	0.7	8.6	2.3	32.2	1.2	38.3	1.6	26.9	3.9	9.0
January-2006	0.5	20.0	1.9	55.3	1.4	18.6	0.5	36.0	1.6	11.3
February-2006	0.5	30.0	7.5	47.6	0.7	25.7	2.2	64.6	4.4	56.8

[a] Unit is ppb for HCl, HNO_3 and HN_3 and is $\mu g\,m^{-3}$ for the rest parameters.

Please cite this article in press as: Chen, H.-W., et al. Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. Chemosphere (2010), doi:10.1016/j.chemosphere.2010.08.019

contributed less than 10% of the regulated standard, except for the H_2SO_4. For ionic species, their mean levels of concentrations appeared to have less spatial variations than the ones for gaseous species. As also shown in Table 1, the CV among the 10 sampling stations for each sampling month, disregarding a few outliers (>100%), varied from 10% to 50% for all the species. Nevertheless, the temporal distribution plot and spatial Cv calculation was inadequate to generate more concrete findings.

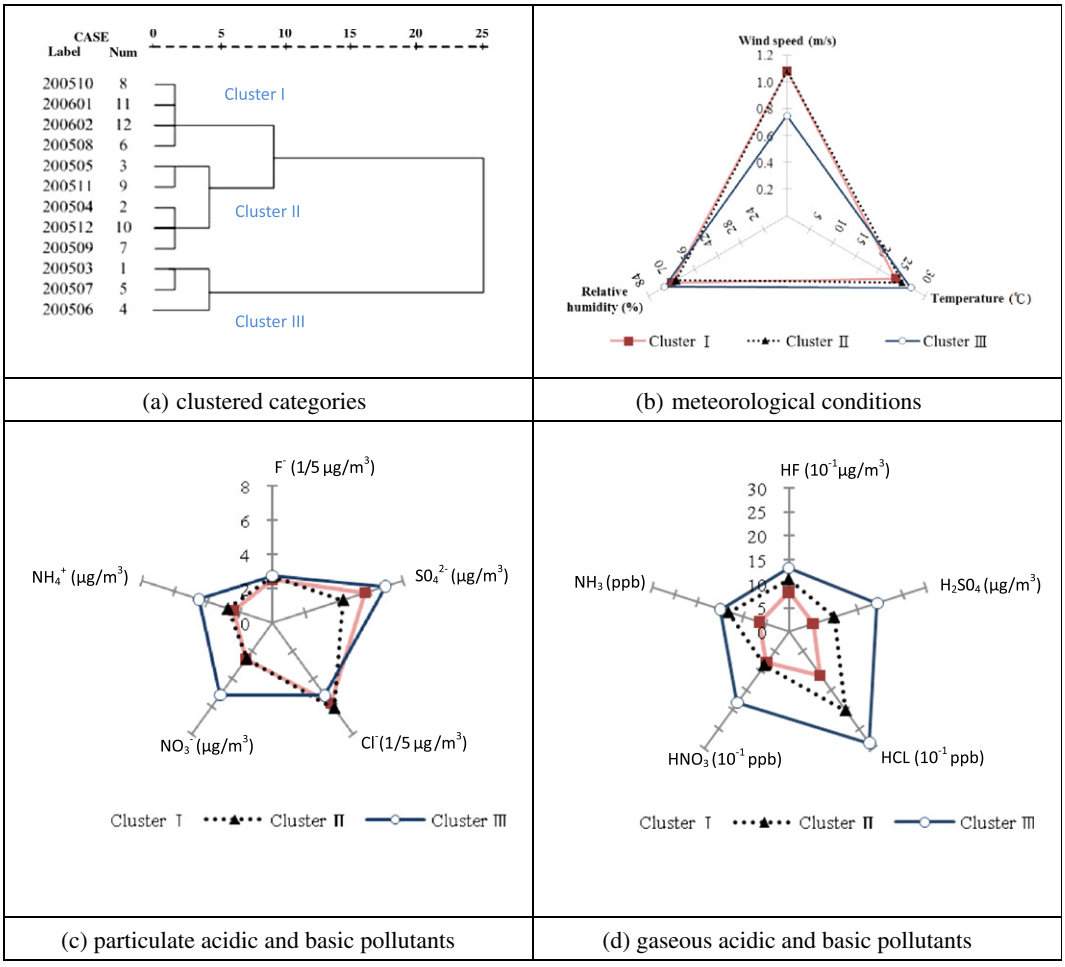
Tsai et al. (2003, 2004) reported the monitoring results of a fully operated Scientific Park in northern Taiwan, having the manufacturing processes similar to this study site. Their monitoring results showed that HF and H_2SO_4 had a concentration as high as 5–10 μg m^{-3} (5–10%) and 30–35 μg m^{-3} (60–70%) around 13 sampling stations in summer at a wind speed of 1.7 m s^{-1}. It should be noted that HF is considered an indicator species to the semiconductor manufacturing process, while H_2SO_4 could be generated in the photochemical reaction associated with the emission of SO_2 from combustion, and favorably occurred in summer in the subtropical area like Taiwan. It is likely, for the study site, that H_2SO_4 is the chemical species most difficult to comply with the regulated standard, particularly in summer, due to the potential transport from Taichung Industrial Park and Taichung City.

3.2. Hierarchical cluster analysis

To begin with the data-mining algorithm proposed in this study (Fig. 2), spatial and temporal variations were evaluated using one way ANOVA approach for each chemical species for 10 sampling sites during 12 sampling months. The dependent variables were the concentrations of each chemical species, and the independent variables were either the sampling month or the sampling station, respectively. The results show that temporal variation in average concentration was significantly different among all the sampling stations ($p < 0.001$), while spatial variation was not among all the sampling months ($p < 0.001$). The relatively small area (413 hectares) of the study site may be part of the reason for the insignificant spatial variation.

To explore features of the temporal variation, all the sampling months were grouped into three statistically meaningful clusters based on the hierarchical cluster analysis (HCA) proposed in this study. Fig. 4a shows the results of three clusters at a Euclidean distance of 5–7, with Cluster I (October, January, February, and August), Cluster II (April, May, November, and December) and Cluster III (March, June, and July). The arithmetic mean of wind speed, relative humidity, and temperature of each clustered month was calculated and plotted in a triangle chart as shown in Fig. 4b. It clearly shows the separation power of the cluster analysis proposed in this study, that the mean wind speed is noticeably lower (about 0.7 m s^{-1}) for Cluster III than that (about 1.1 m s^{-1}) for Clusters I and II. However, the mean of relative humidity and temperature among the three clusters are not so visually different.

By applying the same segregation rule, the 12 months of wind roses were grouped into three clusters in Fig. 5. On the basis of most frequent wind direction, Cluster III of wind roses primarily



(a) clustered categories

(b) meteorological conditions

(c) particulate acidic and basic pollutants

(d) gaseous acidic and basic pollutants

Fig. 4. Results of hierarchical cluster analysis showing: (a) three clusters of sampling months, (b) clustered triangle chart in average number for three meteorological parameters, (c) clustered radar charts in average concentration for five fine particle ionic species, (d) clustered radar charts in average concentration for five gaseous species.

Please cite this article in press as: Chen, H.-W., et al. Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. Chemosphere (2010), doi:10.1016/j.chemosphere.2010.08.019

selects for south and southeast wind in summer at a low speed of 2–4 m s^{-1} (June and July), with an exception for March at a high speed of 6–7 m s^{-1} (March). High speed (4–7 m s^{-1}) of northeast (October, January, February) and northwest (February and August) winter winds dominated for Cluster I. However, Cluster II shows no obvious pattern in wind speed and direction. Surprisingly, it was

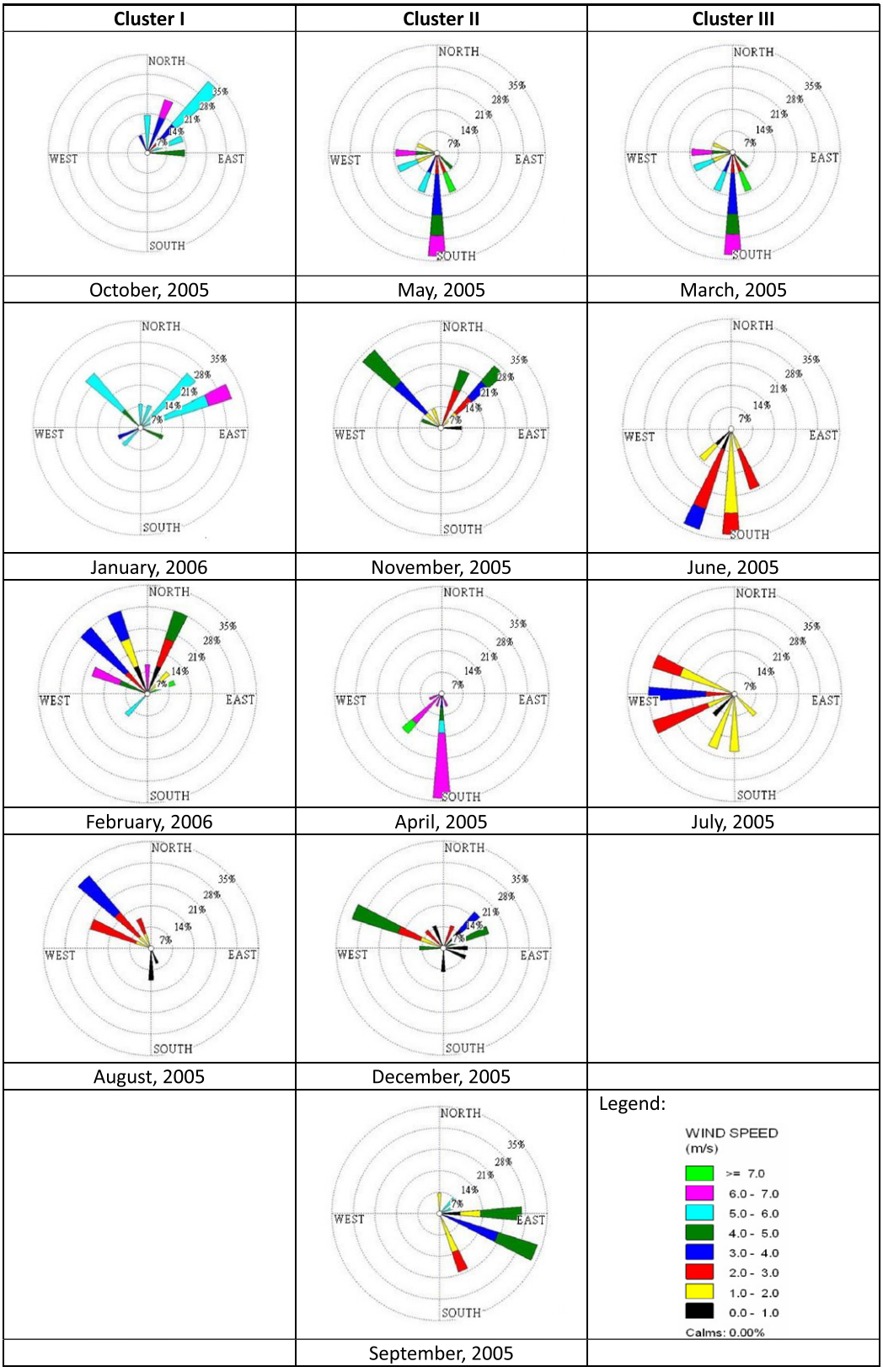


Fig. 5. Results of three clusters of wind rose diagrams, showing high northern winds (>3 m s^{-1}) prevailed in winter for Cluster I, low southern winds (<3 m s^{-1}) dominated in summer for Cluster III, and other types of winds for Cluster II.

Please cite this article in press as: Chen, H.-W., et al. Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. Chemosphere (2010), doi:10.1016/j.chemosphere.2010.08.019

also noted that the high speed (6–7 m s^{-1}) south wind of April was grouped in Cluster II, while the low speed (2–4 m s^{-1}) northeast wind of July was classified in Cluster III. As the cluster analysis is based on the concentrations monitored around the site for each species, the imprecise separation for wind roses reflects the complicate nature of an air dispersion system like the study site. Many local and regional factors other than wind roses may also affect the background concentrations of the study site.

To further explore the association between clustered meteorological parameters and pollutant concentrations, two concentration-clustered radar charts were made, one for the five ionic species in Fig. 4c and another for the five gaseous species in Fig. 4d. Interestingly, Cluster III was clearly separated from the other two clusters for NO_3^-, SO_4^{-2}, and NH_4^+ in Fig. 4c, and for HNO_3, H_2SO_4 and HCl in Fig. 4d. Several interesting points can be noted after crosschecking between the clustered radar charts and wind roses. The elevated levels of NO_3^-, SO_4^{-2}, and NH_4^+ in Fig. 4c and HNO_3, H_2SO_4 and HCl in Fig. 3d were primarily associated with southern summer low wind in cluster III (Fig. 5), suggesting external sources existed in the open southern area to the study site as displayed in the 3-D topography map of Fig. 1b. The average levels of NO_3^-, SO_4^{-2}, and NH_4^+ in Cluster III were significantly increased by 50–100% from those in the other two clusters. However, the levels of F^- and Cl^- in Fig. 4c and HF and NH_3 in Fig. 4d remained relatively stable in all clusters, suggesting that they were little influenced by wind speed and direction ands likely associated with internal sources. These analyses lead to a hypothesis that Taichung Industrial Park and Taichung City southern to the study site could be a significant external source contributing to the background concentrations of NO_3^-, SO_4^{-2}, NH_4^+, HNO_3, H_2SO_4 and HCl. On the other hand, the levels of NH_3, HCl and H_2SO_4 were distinctly lower in Cluster I than those in the other two clusters shown in Fig. 4d, and are associated with the dispersion driven by northern high speed winds of Cluster I in Fig. 5.

Although the use of clustering analysis may oversimplify the complicated nature of the air system, it does help to look into the dominate factors affecting the levels of chemical species under study when a clustering basis is properly used. Theoretically the separating power of the proposed HCA can be enhanced as the monitoring frequency increases in a year, i.e., from once to twice a month (Lee et al., 2001; Reghunath et al., 2002; Banerjee et al., 2009). With the aid of clustered radar charts, clustered wind roses and three-dimensional topography map, the HCA algorithm proposed in this study offers a comprehensive approach to explore the features of a complicate air dispersion system like the Taichung site, to which external pollutant sources may play an important role in regional transport.

3.3. Meteorological and interspecies correlation

Taiwan is in a subtropical area, the relative humidity and ambient temperature monitored were as high as 60–80% and 29–31 °C in summer and as low as 40–50% and 16–18 °C in winter. In order to examine the association between the monthly monitored concentrations for each station and each of the three metrological parameters (relative humidity, temperature, and wind speed) collected on the same sampling date, a univariate correlation analysis was performed. The results of correlation coefficients (R) were given in Table 2 for the five gaseous species (HF, H_2SO_4, HCl, HNO_3, and NH_3) and the five corresponding ionic species (F^-, SO_4^{-2}, Cl^-, NO_3^-, NH_4^+). It can be seen that HCl were highly positively associated ($R > 0.6$) with relative humidity (RH) at a significance level (α) of 0.05 or 0.01 for seven of the 10 sampling stations. Interestingly, the three excluded stations were located at the foothill of Da-Tu Mountain in the northeast of the study site, a stagnant area favorable for the development of a local hotspot. HF also demonstrated the similar correlation but only for three sampling stations. It was noted that Station 3, inside the factory complex, showed a very strong correlation of 0.8 with relative humidity for both of HF and HCl.

On the other hand, the concentration of F^- is positively ($R = 0.6$) associated with the wind speed at Stations 1, 3, 4, 5 and 10. It

Table 2
Results of correlation coefficients between pollutants concentrations and each of the three meteorological data measured in the corresponding month for each sampling site and each chemical species.

No.	HF			H_2SO_4			HCl			HNO_3			NH_3		
	RH[a]	Temp[a]	W.S[a]	RH	Temp	W.S	RH	Temp	W.S	RH	Temp	W.S	RH	Temp	W.S
S1	0.505	−0.489	−0.458	0.428	0.366	−0.074	0.637*	0.039	0.262	0.494	0.447	−0.755**	−0.096	0.368	0.237
S2	0.341	−0.227	−0.566	0.316	0.271	−0.407	0.704*	0.125	−0.252	0.139	0.466	−0.52	0.187	−0.004	0.143
S3	0.837**	−0.145	−0.451	0.234	0.219	−0.572	0.792**	0.099	−0.207	0.197	0.476	−0.726**	0.128	0.226	0.362
S4	0.446	−0.083	−0.544	0.401	0.344	−0.49	0.763**	0.117	−0.326	0.187	0.594*	−0.642*	0.164	0.073	0.275
S5	0.501	−0.131	−0.465	−0.032	0.192	−0.318	0.700*	0.219	0.101	0.042	0.652*	−0.588*	0.152	0.178	−0.243
S6	−0.483	−0.132	−0.244	0.186	0.275	−0.387	0.082	0.021	0.12	−0.015	0.687*	−0.601*	−0.35	0.376	0.056
S7	0.752**	−0.508	−0.33	0.531	−0.187	−0.309	0.528	−0.456	−0.267	0.492	−0.334	−0.328	−0.043	0.180	−0.416
S8	0.214	−0.403	−0.359	0.389	0.065	0.274	0.39	0.327	0.253	−0.103	0.387	−0.414	−0.197	−0.046	0.197
S9	0.649	−0.122	−0.114	0.53	0.323	−0.662	0.858**	−0.049	−0.384	0.475	0.519	−0.818**	0.163	−0.198	−0.205
S10	0.671*	−0.209	−0.311	0.105	0.409	−0.596	0.751*	0.020	−0.598	−0.009	0.801**	−0.651	0.295	0.126	−0.394
No.	F^-			SO_4^{-2}			Cl^-			NO_3^-			NH_4^+		
	RH	Temp	W.S	RH	Temp	W.S	RH	Temp	W.S	RH	Temp	W.S	RH	Temp	W.S
S1	−0.132	−0.402	0.813**	−0.261	0.550	−0.271	−0.141	−0.301	0.641*	−0.131	−0.011	0.061	−0.091	0.233	−0.161
S2	−0.051	−0.272	0.391	−0.022	0.371	−0.101	0.161	−0.801*	0.122	−0.242	−0.051	0.084	−0.124	0.213	−0.061
S3	−0.334	−0.075	0.912**	−0.432	0.481	−0.222	0.252	−0.832*	0.031	−0.174	0.032	−0.121	−0.531	0.321	−0.112
S4	−0.035	−0.302	0.771**	−0.273	0.591*	−0.312	0.382	-0.511	−0.061	−0.124	−0.011	−0.171	−0.302	0.282	−0.252
S5	−0.262	−0.384	0.801**	−0.422	0.431	−0.112	0.201	−0.951*	0.032	−0.141	−0.204	−0.022	−0.491	0.271	−0.151
S6	−0.071	−0.412	0.572	−0.392	0.482	−0.384	0.282	−0.442	−0.252	−0.344	0.142	−0.213	−0.574	0.171	−0.252
S7	0.471	−0.535	0.343	−0.291	0.432	−0.341	0.343	−0.612*	0.071	−0.222	0.024	−0.371	−0.282	0.332	−0.404
S8	−0.425	−0.012	0.334	0.291	0.442	−0.323	0.131	−0.571	−0.092	0.113	0.023	−0.012	−0.184	0.301	−0.144
S9	0.224	−0.531	0.444	−0.491	0.551	−0.341	0.711*	−0.791*	−0.081	−0.144	0.151	−0.204	−0.513	0.432	−0.243
S10	−0.381	−0.273	0.811**	−0.401	0.482	−0.223	0.111	−0.731*	0.502	−0.561	0.282	0.161	−0.604	0.501	−0.131

[a] RH denotes relative humidity, Temp denotes temperature, W.S denotes wind speed.
* significance level $\alpha = 0.05$.
** significance level $\alpha = 0.01$.

Please cite this article in press as: Chen, H.-W., et al. Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. Chemosphere (2010), doi:10.1016/j.chemosphere.2010.08.019

Table 3
Results of linear correlation between two mean concentrations of all 10 sampling stations for any two species during the entire sampling period of 12 months.

	HF	H_2SO_4	HCl	HNO_3	NH_3	F^-	SO_4^{-2}	Cl^-	NO_3^-	NH_4^+
HF	1									
H_2SO_4	0.356**	1								
HCl	0.385**	0.416**	1							
HNO_3	0.387**	0.443**	0.391**	1						
NH_3	0.112	0.187*	0.172	0.107	1					
F^-	0.445**	0.040	0.193*	0.135	0.074	1				
SO_4^{-2}	−0.284**	0.195*	−0.090	0.301**	0.013	−0.248**	1			
Cl^-	0.301**	−0.059	0.260**	−0.168	−0.057	0.309**	−0.318**	1		
NO_3^-	−0.089	0.230*	0.057	0.154	0.356**	−0.037	0.438**	−0.009	1	
NH_4^+	−0.140	0.403**	−0.052	0.379**	0.324**	−0.088	0.679**	−0.218*	0.716**	1

* Significance level $\alpha = 0.05$.
** Significance level $\alpha = 0.01$.

should be noted that Stations 3 and 4 are situated at the upwind to the prevailing north wind in winter. For this reason, F^- could be very likely associated with internal sources as Ta-Du Mountain might block external sources in winter. However Cl^- is little associated with the wind velocity and negatively associated ($R = -0.6$) with the ambient temperature for most stations. The concentration of HNO_3 was positively correlated ($R = 0.6$) with temperature while negatively correlated ($R = -0.6$) with the wind speed, both at a significance level of 0.05 or 0.01. Further analysis shows that HNO_3 levels increased when the wind velocity was less than 2 m s^{-1} and/or the temperature is higher than 28 °C.

To further investigate the chemical speciation as often occurred in the ambient environment of a urban city, a correlation was also performed among different chemical species as shown in Table 3. Ammonium (NH_4^+) was highly positively associated ($R = 0.7$) with the two oxidized form of acidic anions (SO_4^{-2} and NO_3^-) at a significance level of 0.01. In a theoretical simulation study, Orel and Seinfeld (1997) pointed that the formation H_2SO_4 and HNO_3 from SO_3 and NO_2, respectively, was involved in two subsequent steps, formation of acids in gaseous phase followed by heterogeneous condensation with water vapor on the existing particles as a surface site for catalysis (Danalatos and Glavas, 1999; Baek et al., 2004). They also concluded that NH_4^+ plays an important role in the formation of SO_4^{-2} and NO_3^- from their gaseous forms. This demonstrates that particle concentration and species competition may play an important part to contribute the background concentration of the elated species. As many factors may affect the concentration profiles monitored in this study, the results of the statistical analysis poses a challenge for developing a more advanced data-mining technique.

4. Conclusions

To explore the background features of acidic and basic air pollutants around a new high-tech industrial park, this paper presents a novel analytical algorithm of data mining followed by meteorological and interspecies correlation. By clustering air-monitoring data and applying the segregation rule to the corresponding meteorological data, inherent data complexity and dissimilarity can be mitigated and key influencing factors can be identified. The proposed algorithm is also valuable in dealing with the problem of source identification for a complicate air dispersion system like the Taichung site in this study. This study concludes that, at the presence of NH_4^+, several background peaks, such as NO_3^-, SO_4^{-2}, HNO_3, H_2SO_4 and HN_3, occurred in summer when the prevailing wind velocity is slow (<2 m s^{-1}), are likely associated with the external sources from Taichung urban city and Taichung Industrial Park. The results of this study will serve as a basis for the air quality planning and management for the Taichung site.

Acknowledgement

The authors would like to thank the National Science Council of the Republic of China for financially supporting this study under Contract No. NSC95-2221-E-324-011.

References

Baek, B.H., Aneja, V.P., Tong, Q., 2004. Chemical coupling between ammonia, acid gases, and fine particles. Environ. Pollut. 129, 89–98.
Banerjee, D., Chakraborty, S.K., Bhattacharyya, S., Gangopadhyay, A., 2009. Appraisal and mapping the spatial-temporal distribution of urban road traffic noise. Int. J. Environ. Sci. Technol. 6 (2), 325–335.
Bari, A., Ferraro, V., Wilson, L.R., Luttinger, D., Husain, L., 2003. Measurements of gaseous HONO, HNO_3, SO_2, HCl, NH_3, particulate sulfate and $PM_{2.5}$ in New York, NY. Atmos. Environ. 37, 2825–2835.
Chittaro, P.M., Usseglio, P., Fryer, B.J., Sale, P.F., 2006. Spatial variation in otolith chemistry of Lutjanus apodus at Turneffe Atoll, Belize. Estuar. Coast. Shelf Sci. 67, 673–680.
Danalatos, D., Glavas, S., 1999. Gas phase nitric acid, ammonia and related particulate matter at Mediterranean costal site, Patras, Greece. Atmos. Environ. 33, 3417–3425.
Gargava, P., Aggarwal, A.L., 1996. Industrial emissions in a coastal region of India: prediction of impact on air environment. Environ. Int. 22, 361–367.
Hosseinlou, H.M., Sohrabi, M., 2009. Predicting and identifying traffic hot spots applying neuro-fuzzy systems in intercity roads. Int. J. Environ. Sci. Technol. 6 (2), 309–314.
Horng, C.L., Cheng, M.T., 2008. Distribution of $PM_{2.5}$, acidic and basic gases near highway in central Taiwan. Atmos. Res. 88, 1–12.
Jäggi, M., Ammann, C., Neftel, A., Fuhrer, J., 2006. Environmental control of profiles of ozone concentration in a grassland canopy. Atmos. Environ. 40 (28), 5496–5507.
Lee, J.Y., Chen, J.Y., Lee, K.K., Lee, S.Y., Lee, M.H., 2001. Statistical evaluation of geochemical parameter distribution in a ground water system contaminated with petroleum hydrocarbons. J. Environ. Qual. 30, 1548–1563.
Neves, N.M.S., 1996. Air monitoring at Camacari petrochemical complex. Water Sci. Technol. 33 (3), 9–16.
Orel, A.E., Seinfeld, J.H., 1997. Nitrate formation in atmospheric aerosols. Environ. Sci. Technol. 11 (10), 1000–1007.
Pejman, A.H., Nabi Bidhendi, G.R., Karbassi, A.R., Mehrdadi, N., Esmaeili Bidhendi, M., 2009. Evaluation of spatial and seasonal variations in surface water quality using multivariate statistical techniques. Int. J. Environ. Sci. Technol. 6 (3), 467–476.
Reghunath, R., Murthy, T.R.S., Raghavan, B.R., 2002. The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India. Water Res. 36 (10), 2437–2442.
Spengler, J.D., Brauer, M., Koutrakis, P., 1990. Acid air and health. Environ. Sci. Technol. 24, 946–956.
EPA Taiwan, 2007. Emission Standard for Stationary Airborne Pollutants Sources. Taipei, Taiwan.
Tsai, C.J., Aggarwal, S.G., Chang, C.T., Hung, I.F., 2003. Concentration profiles of acidic and basic air pollutants around industrial park of Taiwan. Water Air Soil Pollut. 151, 287–304.
Tsai, C.J., Chang, C.T., Liu, T.W., Huang, C.C., Chien, C.L., Chein, H.M., 2004. Emission characteristics and control efficiency of acidic and basic gases and aerosols from packed towers. Atmos. Environ. 38 (4), 643–646.

Please cite this article in press as: Chen, H.-W., et al. Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. Chemosphere (2010), doi:10.1016/j.chemosphere.2010.08.019

Walker, J.T., Whitall, D.R., Robarge, W., Paerl, H.W., 2004. Ambient ammonia and ammonium aerosol across a region of variable ammonia emission density. Atmos. Environ. 38, 1235–1246.

Webb, M.P., Vincent, C.E., 1999. Comparison of time-averaged acoustic backscatter concentration profile measurements with existing predictive models. Mar. Geol. 162 (1), 71–90.

Zhu, X., Zhihua, F., Wu, X., Meng, Q., Wang, S.W., Tang, X., Ohman-Strickland, P., Georgopoulos, P., Zhang, J., Bonanno, L., Held, J., Lioy, P., 2008. Spatial variation of volatile organic compounds in a "Hot Spot" for air pollution. Atmos. Environ. 42 (32), 7329–7338.

Please cite this article in press as: Chen, H.-W., et al. Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. Chemosphere (2010), doi:10.1016/j.chemosphere.2010.08.019