Cluster analysis applied to atmospheric PM_{10} concentration data for determination of sources and spatial patterns in ambient air-quality of Kathmandu Valley

D. Giri[1], V. K. Murthy[2,*], P. R. Adhikary[3] and S. N. Khanal[4]

[1,3]Department of Mathematical Sciences and
[4]Department of Environmental Sciences and Engineering, Kathmandu University, Dhulikhel, Kavre, P.O. Box 6250, Kathmandu, Nepal
[2]Department of Chemistry – Science and Humanities, P.E.S. Institute of Technology, 100 Ft Ring Road, BSK III Stage, Bangalore 560 085, India

The statistical cluster analysis of particulate matter concentration, measured as PM_{10} in a network of air-quality monitoring stations located in Kathmandu Valley is presented. PM_{10} data for the period 2003–05 has been analysed. Spatial classification was attempted on the basis of ambient air-quality data. A hierarchical agglomeration schedule using linkages between groups by the Euclidean distance metric was used and Ward's strategy was followed to unite two clusters. Two distinct clusters were formed irrespective of the season. The pre-monsoon cluster of monitoring sites reflected similar rural characteristics. The other cluster comprising air-monitoring sites was typified in having urbanized areas with densely populated commercial characteristics with high motor vehicular traffic. The cluster characteristics of monitoring sites during the monsoon, post-monsoon and winter seasons reflected city-centric and distant area characteristics. Cluster analysis reiterates the fact that it is the nature of sources that matters, on the overall air quality of the area. Results of both cluster analysis and non-parametric tests suggest that in the Kathmandu Valley there is difference in pollutant levels across land-use types and topographical characters.

Keywords: Air-monitoring network, cluster analysis, emission source, particulate matter.

KATHMANDU Valley situated in the central part of the Himalayan kingdom is the largest urban centre in Nepal comprising three districts, namely Kathmandu, Lalitpur and Bhaktapur. The valley is an oval-shaped, flat-bottomed basin, improperly ventilated. It is influenced by the Indian monsoon (Bay of Bengal). There are four climatic seasons: March–May (pre-monsoon/spring), June–August (monsoon/summer), September–November (post-monsoon/autumn) and December–February (winter). The general pattern is characterized by a windy, hot and humid climate in March–May, followed by a well-defined monsoon approximately from June to the middle of September. Autumn is characterized by pleasant temperatures and occasional short bursts of rain, while the period December to February is dry, but can be cold, especially at night; winter temperatures drop to below freezing point, with a high level of snowfall in the mountains.

Due to rapid but uncontrolled urbanization and factors such as traffic movement, emissions from brick-kilns, waste disposal and biomass burning, environmental pollution has been constantly increasing. This is adversely affecting land, water, air and biological systems. Though long-term data on pollution are lacking, available information reveals that the nature and extent of air pollution has a serious dimension in the Kathmandu Valley area.

Cluster analysis is a relatively reliable and simplest among the available statistical methods for classification of objects in a dataset. The aim was to establish a set of clusters such that cases within a cluster are more similar to each other than they are to cases in other clusters. The present study deals with the particulate matter concentration (PM_{10}) observed during a period of three years at six air-monitoring sites of the ambient air-quality network in Kathmandu Valley.

The classification procedure involved delineation of homogeneous clusters. The criterion of homogeneity was observation of minimum and maximum cluster variance within and between clusters. This was achieved with the application of hierarchical clustering technique. A hierarchical agglomeration technique is the most common method to form clusters. In agglomerative clustering, each object is initially placed in its own group. The two 'closest' groups are combined into a single group. Each cycle of the process combines two closest groups into a single group, thereby reducing the number of groups by one. The purpose is to form each possible number of groups n, $n - 1$,..., 1, in a manner to minimize the loss of information.

The first step was to combine two clusters whose fusion yielded the least increase in the sum of squares within clusters distance from each individual to the centroid of its present cluster n, resulting in $n - 1$ groups. The next step was to examine the $n - 1$ groups to determine if a third member should be linked with the first pair or another pair be made in order to secure the optimum value of the objective function for $n - 2$ groups. This process continued till all the stations were clustered in one group. The underlying mathematics of this method was relatively simple, but a large number of computations were needed, which can put a heavy demand on the computer.

Cluster analysis has been used to design and analyse air-quality monitoring networks. The most common application of cluster analysis for aerometric data relates to source identification. The characteristics of acid precipitation collected in Taipei by performing cluster analysis using the Euclidean distance and Ward's method to link clusters has been reported. Three potential sources of water-

*For correspondence. (e-mail: profvkm55@yahoo.com)

soluble ions in rainwater were identified[1]. Cluster analysis was applied to identify contributing sources in the metropolitan area of Mexico City, by taking samples of atmospheric aerosol in three sites and analysed with particle induced X-ray emission[2]. The principal component and cluster analysis were used in order to establish whether air-quality affinity areas existed for Birmingham, UK, on a spatio-temporal dataset comprising daily sulphur dioxide concentrations for a 571-day period and a 17-station network. The result revealed the existence of four air-quality affinity areas, which reflected the nature of land use in these areas[3]. In addition to considering only pollutant data, many studies have also incorporated meteorological data for cluster analysis. The use of cluster analysis for better elucidation of the dependence of air quality on meteorology has been reported[4]. The climatic classification of Saudi Arabia with an application of factor and cluster analysis by using Ward's method of clustering is studied[5].

One of the first requirements in the understanding of pollution is the state of pollution itself. This requires elaborate data over a long period of time covering different hours of the day and different seasons. In Kathmandu Valley, PM_{10} was monitored using ambient low volume air samplers (LVS) installed at six sampling sites within the valley area by the Ministry of Population and Environment, Government of Nepal. To obtain better spatial representation of the valley, air-monitoring stations were located in all directions representing all predominant urban areas associated with high, medium and low human activities (Figure 1).

Daily 24-h continuous air monitoring was carried out at each monitoring site. Air monitoring sites were characterized by urban traffic (Putalisadak), urban hospital (Patan Hospital), city core residential (Thamel), urban background (Bhaktapur, Tribhuvan University) and valley background location (Matsyagaon). A similar establishment

Figure 1. Map showing air-monitoring sites in Kathmandu Valley.

of air-monitoring sites and PM_{10} monitoring conducted between 1994 and 1999 from the network of air-sampling stations in Taiwan has been reported[6].

An LVS sampler, M/S Instrumatic Model 85-02 designed specifically for the Kathmandu Air Quality Monitoring Programme, capable of sampling fine particulates PM_{10} and $PM_{2.5}$ continuously for 24 h without pneumatic movement of filters was used. The particulate samples, collected and stored on daily basis, were shifted for standard gravimetric analysis on weekly basis from the monitoring sites.

The PM_{10} data available for three years, 2003 through 2005 in all the monitoring sites on a particular day were considered for analysis. To perform cluster analysis, only those data that were recorded at the same point of time in all monitoring stations were considered. The day/days during which PM_{10} data were missing in any of the sites were not considered.

In this study, Euclidean distance was chosen to measure the similarity between two sites and Ward's strategy was followed to unite two clusters. A feature of the Euclidean distance was that it is a weighted measurement; the higher the absolute value of the variable the higher will be its weight.

Ward's technique produced spherical clusters which were roughly of the same size. The aim was to join objects together into ever-increasing sizes of clusters using a measure of similarity of distance. This was a bottom-up approach, starting with n groups, each containing one object. Two of the cases were then combined into a single cluster. At the next stage, either a third case was added to the cluster containing two cases or two other cases were merged into a new cluster. This process continued until all cases belonged to one cluster. Once a cluster was formed it could not be split, but only be combined with other clusters. In addition to choosing the similarity or dissimilarity measure for comparing two observations, there was also the choice of what should be compared between groups that contain more than one observation. This is referred to as the linkage method. Ward's method joined the two groups that minimized the error sum of the squares (i.e. the within-cluster sum of squares). This procedure linked the pair of groups that produce the smallest variance in the merged group. So for each pair of groups, they were tentatively linked and the centroid determined. The average squared distance to the centroid (variance) was calculated and the pair that produced the smallest variance in the merged groups was linked. Due to the agglomerative nature of Ward's method, the cluster means changed as new cases were added.

This study presents the result of PM_{10} measurement conducted at six sampling sites within the Kathmandu Valley area during 2003–05. The objective was to identify air-monitoring sites with similar pollutant behaviour characteristics by means of cluster analysis and to explore pollutant sources in such clusters. From Figure 2, it can

Table 1. Inter-station correlation

Station	Matsyagaon	Tribhuvan University	Bhaktapur	Thamel	Patan Hospital	Putalisadak
Matsyagaon	1					
Tribhuvan University	0.891	1				
Bhaktapur	0.823	0.923	1			
Thamel	0.751	0.876	0.890	1		
Patan Hospital	0.683	0.783	0.777	0.781	1	
Putalisadak	0.652	0.773	0.793	0.846	0.785	1

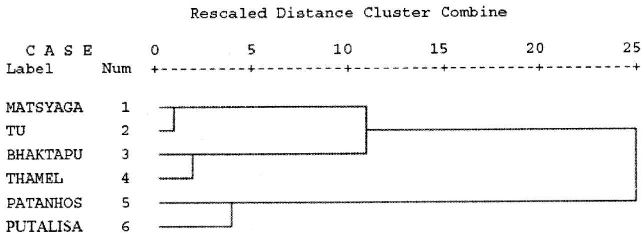

Figure 2. Dendrogram using Ward's method for the study period.

be observed that two distinct clusters based on average 24 h PM_{10} concentration were observed during the study period 2003–05. Matsyagaon and Tribhuvan University (TU) together with Bhaktapur and Thamel representing the valley and urban background areas were merged to form one cluster, whereas Patan Hospital and Putalisadak representing city core area associated with high commercial and vehicular activities formed another cluster. The fact that the groups were the same for PM_{10} pollutants suggested that the concentration was dependent on emission from local sources of area and land-use pattern in question, and not influenced by the existence of meteorological or topographical conditions of the area.

Mann–Whitney test indicated significant difference ($P < 0.05$) in mean PM_{10} concentration between two distinct clusters.

Pearson coefficients of correlation for PM_{10} concentration, as calculated between the monitoring sites during the study period, are displayed in Table 1. All correlations were found significant ($P < 0.01$).

Although the concentration levels exhibited significant spatial variability, the strength of the Pearson's coefficient correlation between Bhaktapur and TU was highest (0.923) compared to other air-monitoring stations. This could be because both these stations are located outside the ring road area, referred to as urban background areas in the Kathmandu Valley air-monitoring network. This finding suggested that continuing air monitoring at both stations was not essential. Instead either one of the two sites could be retained. This will not hamper the quality of representation of air-monitoring stations. Alternatively, another site, preferably in the northeastern region of Kathmandu Valley, beyond Putalisadak area, in the outskirts could be located. It has been reported that the since winds blow in the valley in a southwesterly direction[7], such a positioning of the new air-monitoring station would be more meaningful.

Performing cluster analysis of PM_{10} data for different seasons observed in six air-monitoring sites of Kathmandu Valley, two distinct clusters were observed irrespective of the season. Figure 3 *a* depicts the nature of cluster formation with respect to the pre-monsoon season in Kathmandu Valley. It can be observed that TU and Matsyagaon grouped up in one distinct cluster and the remaining monitoring stations of the network grouped in another cluster. Cluster analysis indicated that the chief pollutant sources prevalent in these two air-monitoring stations had some commonality. It is relevant to indicate that Matysagaon and TU are at relatively elevated places having higher altitudes compared to other air-monitoring sites of the network. Secondly, the nature of land-use pattern, topographical characteristics of the area and local emission sources predominantly influence the PM_{10} levels here than meteorological parameters.

The picture with respect to the nature of the existing pollutant sources as brought out by cluster analysis in the monsoon season with respect to Patan Hospital and Putalisadak air-monitoring sites is characteristic. These two sites emerged as a one distinct cluster and the remaining four monitoring sites in network formed into another cluster (Figure 3 *b*). The reason for this was the tendency of areas associated with parameters like anthropogenic activities getting concentrated over a limited area of the city core. Normally monsoon rain exerts a cleansing effect on ambient dust particulate levels at least during and soon after the spell of rainfall. However, in places where vehicular traffic and associated commercial nature of the area predominate, the extent of the cleansing effect is nullified considerably. The level of PM_{10} therefore remained consistently higher even through the area was experiencing rainfall. This fact was observed from the set of data on PM_{10} obtained from Putalisadak and Patan Hospital areas. The fact that the nature of the monitoring sites in the groups was similar for PM_{10} pollutant levels, suggests that the concentration was mostly governed by the land-use pattern of the area, and not by meteorological or topographical characteristics. The distinction is marked as confirmed by Mann–Whitney test of significance ($P < 0.05$) with respect to PM_{10}. The test revealed that significant

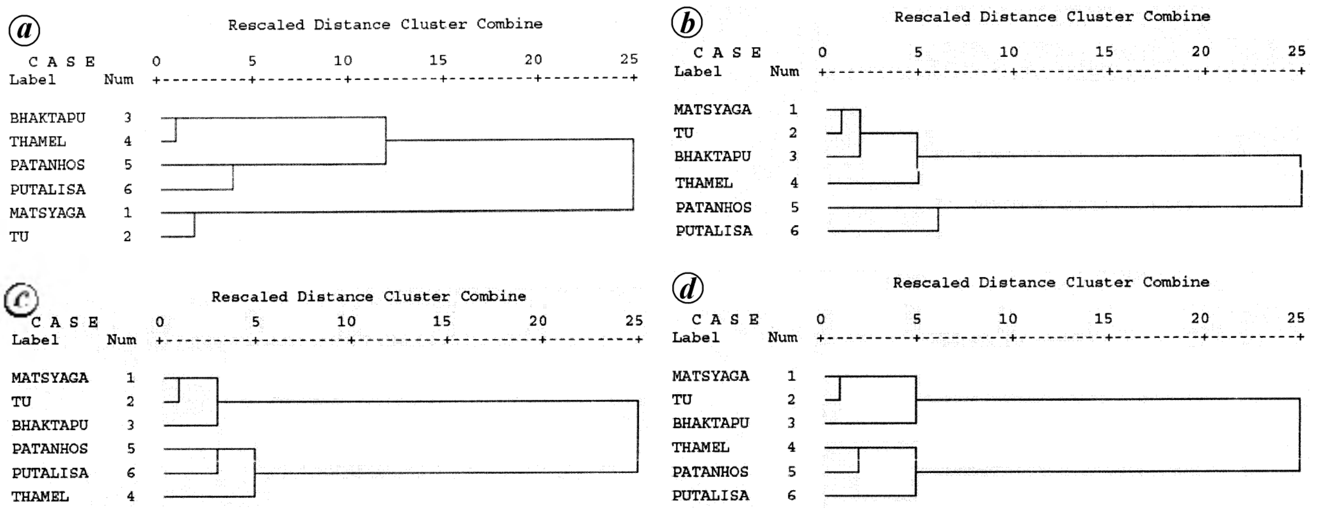

Figure 3. Dendrogram using Ward's method for the study period: ***a***, Pre-monsoon season; ***b***, Monsoon season; ***c***, Post-monsoon season, and ***d***, Winter season.

difference in the mean PM_{10} was observed between the two clusters during both pre-monsoon and monsoon seasons.

During the post-monsoon and winter seasons, air-monitoring sites representing valley background and urban background merged to form one cluster, whereas urban areas of Kathmandu Valley formed another cluster (Figure 3 *c* and *d*). For the post-monsoon season Matsyagaon and TU as well as Patan Hospital and Putalisadak formed the initial clusters. Later, Bhaktapur was incorporated into the first cluster along with Matsyagaon and TU; Thamel entered into the second cluster along with Patan Hospital and Putalisadak. During winter, Matsyagaon and TU as well as Thamel and Patan Hospital formed the initial clusters. Later, Bhaktapur and Putalisadak were merged respectively in the two clusters. Significant difference in the means with respect to PM_{10} was observed using the Mann–Whitney test between clusters classified as urban valley background and core city area during the post-monsoon and winter seasons.

Cluster analysis also confirmed that sites associated with heavy traffic and densely populated areas exhibited similar characteristics during post-monsoon and winter seasons of the year, a fact repeatedly observed based on PM_{10} levels during the study period of three consecutive years. Thus cluster analysis has confirmed that it is a reliable method to identify the sources of air pollution. The analysis could facilitate city-planners as well as air-quality managers to take the necessary steps towards control measures.

The purpose of this study was to understand the spatial patterns of air pollution in Kathmandu Valley over a short period using cluster analysis based on the criteria of PM_{10} pollutants. The dataset was subjected to cluster analysis to identify the nature of sources of air pollutants in the valley.

Cluster analysis indicated that annual seasons had a definitive role in the formation of clusters, justifying observation in variation of PM_{10} concentration in the respective areas monitored. The analysis confirmed the influence of seasonal effects, and their extent and limitations on the observed PM_{10} variations.

The monsoon season was expected to reduce ambient dust concentration. While this is true in some areas, the findings of the present study and cluster analysis indicated that rainfall had minimum effect in the reduction of PM_{10} concentration in city core areas. This suggests that particulate levels were dependent more on commercial activities, level of urbanization, distance of air-monitoring sites with respect to city core areas and topo-geographical effects rather than rainfall during monsoon season.

Formation of cluster pattern is the same in Kathmandu Valley with respect to winter and post-monsoon seasons. Formation of the two clusters during pre-monsoon and monsoon was different, again indicating inter-grouping of air-monitoring stations with similar characteristics.

The study shows that the cluster analysis method is useful in identifying sources of air pollutants. This could help in air-quality management in urban areas. The study has also provided a methodology for researchers, practitioners and regulatory authorities to investigate spatial patterns of air pollution in the Kathmandu Valley.

1. Lin, N. H., Lee, H. M. and Chang, M. B., Evaluation of the characteristics of acid precipitation in Taipei, Taiwan using cluster analysis. *Water, Air Soil Pollut.*, 1999, **113**, 241–260.
2. Miranda, J. *et al.*, PIXE analysis of atmospheric aerosols from three sites in Mexico City. *X-Ray Spectrom.*, 2005, **34**, 315–319.
3. MaGregor, G. R., Identification of air quality affinity areas in Birmingham, UK. *Appl. Geogr.*, 1996, **16**, 109–122.
4. Dorling, S. R., Davies, T. D. and Pierce, C. E., Cluster analysis: A technique for estimating the synoptic meteorological controls on air and precipitation chemistry – Method and applications. *Atmos. Environ. Ser. A*, 1992, **26**, 2575–2581.

5. Ahmed, B. Y. M., Climatic classification of Saudi Arabia: An application of factor – cluster analysis. *GeoJournal*, 1997, **41**, 69–84.
6. Kuang, L. Y., Spatial and seasonal variation of PM_{10} mass concentrations in Taiwan. *Atmos. Environ.*, 2002, **36**, 3403–3411.
7. Kitada, T. and Regmi, R. P., Dynamics of air pollution transport in late wintertime over Kathmandu Valley, Nepal: As revealed with numerical simulation. *J. Appl. Meteorol.*, 2003, **42**, 1770–1798.

ACKNOWLEDGEMENTS. We thank the Heads of the Department of Natural Sciences and Department of Environmental Sciences and Engineering, and administrative authorities of Kathmandu University, Nepal. Help rendered by Mr Chiranjibi Gautam, Urban Environment and AQM Adviser, Environmental Sector Programme Support, Ministry of Environment and Population, Government of Nepal is acknowledged. Kothmandu University provided financial support for this study.

Received 8 September 2006; revised accepted 22 May 2007

Comparative assessment of REMAP and ISSR marker assays for genetic polymorphism studies in *Magnaporthe grisea*

Sonia Chadha* and T. Gopalakrishna

Nuclear Agriculture and Biotechnology Division,
Bhabha Atomic Research Centre, Trombay, Mumbai 400 085, India

Retrotransposon microsatellite polymorphism (REMAP) and inter simple sequence repeat (ISSR) are simple PCR-based assays targeting retrotransposon and/or microsatellite regions (variable regions) of the genome. In fungi, ISSR has been applied for genetic variation studies, whereas REMAP was recently shown as a DNA fingerprinting assay. However, these techniques have not been compared for their utility in genetic polymorphism studies. In this study, comparative evaluation of ISSR and REMAP assays for percentage polymorphism, polymorphic information content, marker index, genetic similarity values and cost per polymorphic markers generated demonstrated REMAP as a marker of choice for genetic variability studies in *Magnaporthe grisea*. The results suggest a faster rate of evolution of these elements in *M. grisea*, thereby leading to its diversification. The growing number of reports on the presence of dispersed families of LTR-retrotransposon in fungi suggests wider application of REMAP assay in fungi.

Keywords: Genetic variation, inter simple sequence repeat, *Magnaporthe grisea*, retrotransposon microsatellite polymorphism.

MAGNAPORTHE grisea, a causative agent of rice blast disease, exhibits a high degree of genetic variability[1]. In *M. grisea* several repetitive sequences are reported[2], whose genetic rearrangement leads to race and pathogenic variations[3]. Long terminal repeat (LTR)-retrotransposon *MAGGY*, is a functional element present in *M. grisea*[4]. Nakayashiki *et al.*[5] demonstrated transpositional activity of *MAGGY* in various fungi and suggested it as a candidate for a gene-tagging tool in filamentous fungi, especially in ascomycete.

Microsatellite or simple sequence repeat (SSR) regions consisting of tandemly repeated short DNA sequences (1–6 base pairs) are abundant in the eukaryotic genome[6]. Although analyses of SSRs provide a codominant, highly reproducible and genetically informative marker system, development of SSRs is time-consuming and laborious[7]. Inter simple sequence repeat (ISSR), a variant of the SSR approach amplifies the DNA sequence between two SSR loci[8]. ISSR primers are based on di-, tri-, tetra- or penta-nucleotide repeats with 5′ or 3′ anchored base(s). ISSR technique has been employed for the characterization of genetic variations within fungi[9]. Another PCR-based DNA fingerprinting assay REMAP (REtrotransposon Microsatellite Amplified Polymorphism) detects retrotransposons inserted near the SSRs[10]. The assay involves use of an ISSR primer in combination with a primer based on the LTR region of the retrotransposon to generate multilocus profiles. There are various reports of REMAP in plants[11,12]. In an earlier report, we used the REMAP technique based on *MAGGY* to generate multilocus DNA profile[13]. The simplicity, rapidity, reproducibility and cost-effectiveness of REMAP and ISSR make these assays an ideal tool to study population genetics, genome stability and evolutionary aspects in fungi. However, before application of any technique on a larger scale, various factors have to be evaluated to decide the suitability of the technique[14]. In this study, REMAP and ISSR assays were compared for their utility in generating polymorphic DNA profiles, genetic variation studies in *M. grisea* isolates and to understand the degree of congruency between the two marker systems.

M. grisea isolates used in the study (Table 1) were subcultured on potato dextrose agar (PDA, Hi Media, India)

Table 1. Indian *Magnaporthe grisea* isolates used in the study

Sl. no.	Isolate name	State
1	LUN 8.3	Maharashtra
2	KVL 7.3	Maharashtra
3	Karjat CV11	Maharashtra
4	Karjat 184	Maharashtra
5	KN 1.6.1	Kerala
6	KN 4.6.3	Kerala
7	KN 1.4.4.5	Kerala
8	KN 1.2.1	Kerala
9	Warangal	Andhra Pradesh
10	Maruteru	Andhra Pradesh
11	Pondicherry	Puducherry
12	Nawagam P203	Gujarat
13	Nawagam SV/CV102	Gujarat
14	Titabar	Assam
15	Chiplima	Orissa

*For correspondence. (e-mail: soniachaddha@yahoo.com)