J. ENVIRONMENTAL SYSTEMS, Vol. 29(3) 245-267, 2002-2003

# AIR QUALITY MONITORING NETWORK DESIGN USING INFORMATION THEORY

**VINEET KUMAR JAIN**
*University of Cornell, Ithaca, New York and*
*Indian Institute of Technology Kanpur, India*

**MUKESH SHARMA**
*Indian Institute of Technology Kanpur, India*

**ABSTRACT**

Most of the techniques currently available in literature for designing an Air Quality Monitoring Network (AQMN) are complex in nature and are suited to one or more specific objective(s) and particular conditions. The present study proposes a simple and generalized method for designing an optimum AQMN based on entropy concepts, which are central to the information theory. This study considers the AQMN as an environmental information system. The AQMN provides the information about the random events (pollution levels) occurring in the area of interest. Information observed at one station can be inferred partially from observations at other stations. This concept is used to form a network that conveys the maximum possible information about the environment of the area for a given number of stations. The optimum size of the network is determined when addition or a new station does not add significant information to the existing network. AQMN design based on multiple pollutant leads to different optimum AQMNs. A combined AQMN based on equal weightage to each pollutant is suggested. It is observed that design based on discrete random variables becomes computationally very intensive in large networks. As a possible solution, AQMN is designed based on continuous variables and a comparison is done with the discrete variables based design. This methodology is applied to the existing network of nine stations in Delhi being operated under the Indian National Ambient Air Quality Monitoring (NAAQM) Program.

## INTRODUCTION

The basic goal of air quality monitoring is the protection of human health and welfare. This broad goal has produced broad objectives and thus an air quality monitoring program can have one or more of the following objectives:

1. Assess impact of selected sources(s);
2. Assess violation of ambient standards in a region (all sources combined);
3. Plan control strategies;
4. Evaluate risk to environment (human/soil/vegetation/monument); and
5. Activate episode controls; and others (e.g., model validation, land use planning etc.).

Although it is tempting to design a system that can serve multiple objectives, in practice, it appears that only a few combinations of objectives are realizable for a given network. For example, it is not possible to use a network designed to monitor long-term trends of air pollution levels to investigate a specific complaint.

Significant work has been done in designing the network for different sets of objectives, (e.g., maximum detection of violation, maximum coverage of area). Lee et al. use two criteria 1) "detection criterion" to maximize the probability of detecting violations of an environmental standard and 2) "protection criterion" to maximize the people living in the grid squares with monitoring stations [1]. Nakamori et al. [2] and Nakamori and Sawaragi [3] divided the monitoring domain into sub-areas that minimized the sum of concentration variances within sub-areas based on the spatial distribution pattern of long-term average. A monitoring station was put in each sub-area. Modak and Lohani assigned station locations, one by one, to the grid points where "spatial average" was maximum and the overlap with the coverage of the other stations as the minimum based on the spatial correlation field [4]. They also extended their method to network design for multiple pollutants [5, 6]. Handscombe and Elsom [7] used spatial correlation analysis to define delimiting areas containing two or more stations with highly correlated data. One station in the area was preserved as a "reference station," while the others were eliminated as redundant ones. Katoch et al. eliminated those stations where the data could be explained with the required accuracy stations where the data could be explained with the required accuracy by a regression equation using the data from four or five stations [8].

These studies have their own advantages. These works are mainly based on conventional statistical methods, which lead to adoption of very narrow defined statistical objectives and suited to some particular conditions. Many methods used the linear correlation between stations to find the optimum locations. Some methods use the inverse of the variance as the information criterion for selection of stations but these methods consider the performance of the station as an individual. Many of these methods are quite complex. This might be the reason why most of the current design practices of a monitoring network are heavily based upon

experience and judgment with few analytical guides. The guidelines available (e.g., [9, 10]) provide quantitative advice about members of monitors but not their geographic locations.

Developments in information theory can assist in addressing some of the above limitations of current approaches. Its abstract formulations are applicable to any probabilistic statistical system of observations. The information theory provides a general criterion, i.e., maximization of information, which is more appropriate to the multifaceted and often unpredictable role which is ultimately played by the air quality data. Information Theory has recently found its uses in the hydrological processes [11-14]. In information theory, a station is ranked based on its performance within the system not as an individual.

## CONCEPTS OF INFORMATION THEORY

Information theory has its mathematical roots in concepts or disorder or entropy in thermodynamics and statistical mechanics. An extensive literature devoted to studies of the relation between the notions and mathematical form of entropy and information exists [15, 16].

### Entropy Concepts and Communication Network

In this section, the concepts of entropy as developed by Shannon [17] are presented in the context of AQMN design.

#### *Definition of Information*

Let $x$ be some event, which occurs with probability $p[x]$. If one is told that event $x$ has occurred then it implies that the amount of information obtained is given by

$$1[x] = -\log p\,[x] \tag{1}$$

Where, $1[x]$ is amount of information received in units nats.

If $p_1, p_2, \ldots, .p_n$ are the probabilities for "$n$" different states of random variable $X$ values having possible outcomes $x_i$; $i = 1, 2, \ldots, n$, then expectation of information (termed as entropy) is given by

$$H(X) = -\sum_i p[x_i]\log p[x_i] \tag{2}$$

Consider now another random variable $Y$ having possible outcomes $y_j$, $j = 1, 2, \ldots, m$. If $Y$ is related to $X$, the mutual information between them can be expressed as

$$T(X; Y) = H(X) + H(Y) - H(X, Y) \tag{3}$$

H($X$) and H($Y$) are individual entropy of $X$, $Y$ respectively. H($X$, $Y$) is the joint entropy of $X$ and $Y$. It can be expressed mathematically as

$$H(X, Y) = -\sum_i \sum_j p[x_i, y_j] \log p[x_i, y_j] \quad (4)$$

H($X$, $Y$) measures the uncertainty associated with outcome pairs $x$, $y$. It is dependent upon the association between the two variables as well as their dispersion. T($X$; $Y$) is mutual information which is due to dependency between $X$ and $Y$ and it is, in fact, repetition of uncertainty between the $X$ and $Y$. This is the basis of entropy based design of air quality monitoring system. $X$ and $Y$ may be considered as the data sets of air quality obtained at two locations in area. Knowing the data set at one station one can predict to some extent the information on air quality at other station.

Entropy terms defined above can be extended for more than two variables, for example, joint entropy of a random vector is defined as

$$H(X_1, X_2, ....., X_v) = -\sum_i \sum_j \ldots\ldots \sum_q p(x_i, x_j, ..... x_q) \log p(x_i, x_j, ..... x_q) \quad (5)$$

in which $x_i$, $x_j$, ..., and $x_q$, represent discrete outcomes of $X_1$, $X_2$,....., and $X_v$, respectively.

The above mathematical description of entropy and mutual information is presented here so that the reader can get insight into these principles. Shannon [17] did the seminal work in this field.

### *Measure of Entropy for Continuous Variables*

The entropy of the continuous variable $X$ knowing the length ofthe class interval ($\Delta x$) can be defined as

$$H(X; \Delta x) \cong \int_{-\infty}^{\infty} f(x) \log f(x) dx - \log(\Delta x) \quad (6)$$

The above expression is dependent on the unit chosen for $X$. A possible solution to make entropy independent of the unit chosen suggested by Chapman [13] considers the proportional class intervals instead of fixed length intervals, which is equivalent to taking the logarithm of $X$ and dividing into small equal parts. The entropy of a log normal variable ($Z = \log X$) for proportional class interval ($\Delta x/x$) is given by:

$$H(X; \Delta x/x) = \frac{1}{2} \log(2\pi \quad e\sigma_z^2) - \log(\Delta x/x) \quad (7)$$

This above expression is independent of the unit chosen. Similar to discrete variables, entropy terms can be defined for continuous random vectors, for example entropy for a normal random vector can be defined as

$$f(\mathbf{X}_n) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{X}_n - \mathbf{X}_{on})^T \Sigma^{-1} (\mathbf{X}_n - \mathbf{X}_{on})\right] \quad (8)$$

where $\mathbf{X}_n = (X_1, X_2, ..., X_n)^T$ is the normal random vector of $n$ variables

$$H = \frac{1}{2}\sum_{i=1}^{n} \log\ \sigma_i^2\ \ \pi 2e + \log|R| \tag{9}$$

As the correlation matrix R is positive definite

$$|R| \le \prod_{1}^{n} \rho_{ij} = 1 \tag{10}$$

The value of the joint entropy H can be varied by changes in standard deviations of the variables or by changing the correlation matrix of the variables. It is obvious from the above expression that entropy (uncertainty) will be more of an uncorrelated set than that of the correlated set.

## FRAMEWORK FOR AIR QUALITY MONITORING NETWORK DESIGN

This section develops methodology for the optimum design of the network using the concepts of the information theory discussed earlier in this article.

The AQMN is considered as an air quality information system. Since it is not practical to put stations everywhere in the area of interest, the system needs to be optimized to transmit the maximum information of the area for a given number of stations (termed as permanent station locations). It is known that the information observed at one site can be inferred partially from observations at other sites. The measurement obtained at the "*n*" permanent station locations will be used to provide an estimate of events which have occurred at the remaining locations using some unbiased estimators ($F(\ )$). Assume "*m*" locations in the area form a dense network (termed as temporary network) enough to reduce all the uncertainty in the system, i.e., temporary network provides all information of the area. If the role of the permanent network is viewed as providing information that ultimately reduces uncertainty concerning events at the "*m*" locations, then design criterion is to maximize the reduction in uncertainty. This is equivalent to maximizing the information transmitted between events at the "*m*" locations and the measurements provided by the "*n*" number of stations of permanent network. The design problem considered here is the selection of the "*n*" (which can be 1,2,3..... or *m*) air quality monitoring stations location subset "*i*" ($S_i^n$) out of tie total ${}^mC_n$ location sets so as to maximize the joint entropy H ($S_i^n$) for that specific "*n*" or, equivalently, information transmission between the temporary network and permanent network formed by a given selection of "*n.*" Information transmitted by a permanent network is equivalent to the joint entropy of the permanent network as given by Caselton and Hussain [11] is:

$$T(S;\ S_i^n\ F(S_i^n)) = H(S_i^n) \tag{11}$$

In short, the optimization problem can be stated as

$$\text{Max H } (S_i^n) \text{ to find critical } S_i^n$$

Where, i = 1, 2, ..., ${}^mC_n$

## STUDY AREA AND DATA COLLECTION

### Study Area

The applicability of "Entropy" concept for AQMN design is investigated for city of Delhi (which includes New Delhi). Delhi is rapidly becoming nucleus of trade, commerce, and industry in the northern region of India. It has been found that it is one of the most polluted cities of the world and a significant part of the pollution in the city is due to vehicular pollution [18]. Major sources of pollution in Delhi are vehicles, power plants, industries and domestic fuel uses. An estimated quantity of about 1600 metric tons of pollutants is emitted in the atmosphere every day in Delhi.

Consequently, the development of systematic monitoring networks is very important for the city of Delhi. In India, "Central Pollution Control Board" (CPCB), an Environmental Protection Agency (EPA), has established a national network of air quality monitoring stations. The pollutants measured are sulfur dioxide ($SO_2$), nitrogen dioxide ($NO_2$) and Suspended Particulate Matter (SPM).

### Data Collection

NAAQM network in Delhi comprised stations at nine locations (Figure 1). The present status of air quality indicates alarmingly high value of SPM at all locations. Even in the residential areas the air quality in respect of SPM exceeds by a factor of two when compared to the standard (i.e., 200 $\mu g/m^3$). Although the levels of $SO_2$ and $NO_2$ have remained within the standards, the trends show that levels are rising. For the purpose of this work, monthly average data of the parameters (SPM, $SO_2$, and $NO_2$) were collected for all nine locations for five years (i.e., 1991-95) from the published reports of the CPCB (1990-95) [19]. Table 1 presents the basic statistics (Correlation, Matrix, Mean, Standard Deviation, and Coefficient of Variance (COV)) of air pollution data from nine locations.

### Entropy Based AQMN Design for Delhi

The present network of nine stations represents the temporary network and objective is to choose the permanent network (a subset of temporary network) that transmits maximum information transmission (Equation 11). The problem can be solved either by considering the variable as discrete or continuous; in this research both cases are considered.

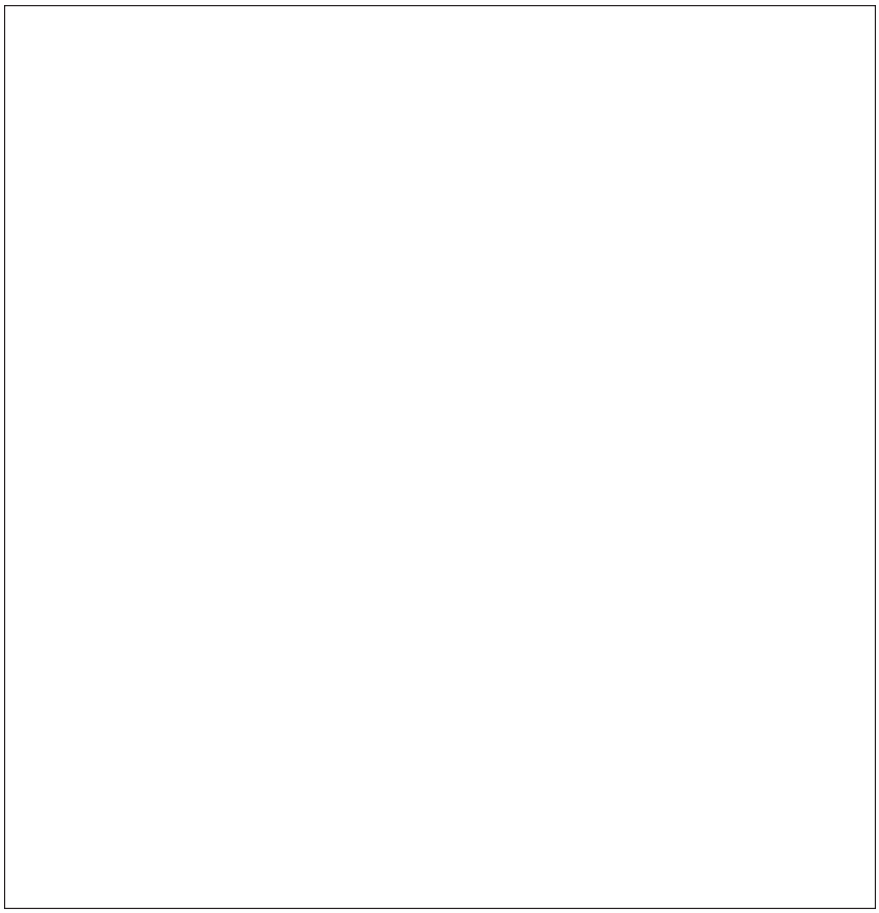

Figure 1. Running locations of air quality monitoring stations in Delhi.
Note: ● Locations of ambient air quality monitoring stations.

### Air Quality—Discrete Variable

A program in language "C" was written to compute the optimum information transmission. The intervals taken in the program for discretizing the $SO_2$, $NO_2$, and SPM data are given in Table 2. The basis for taking this interval is to obtain a reasonable distribution of data in each interval.

## RESULTS AND DISCUSSION

The computed values of H ($S_i^n$) (i.e., information transmission by various networks of given size) for each of the nine possible choices of single station location are given in Table 3 for each pollutant. Table 4 presents the optimal

Table 1. Basic Statistics of Pollutants Concentrations

| Station No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | $SO_2$ (Correlation matrix) | | | | |
| 1 | 1 | 0.2538 | 0.4305 | –0.0570 | 0.4606 | 0.3459 | –0.2203 | –0.1892 | –0.2195 |
| 2 | 0.2538 | 1 | 0.5930 | –0.0455 | 0.4469 | 0.2690 | 0.0542 | 0.0037 | 0.0047 |
| 3 | 0.4305 | 0.5930 | 1 | –0.2180 | 0.6652 | 0.6536 | –0.1589 | –0.1896 | –0.2790 |
| 4 | –0.0570 | –0.0455 | –0.2180 | 1 | –0.1017 | –0.0142 | 0.0635 | 0.0759 | –0.2246 |
| 5 | 0.4606 | 0.4469 | 0.6652 | –0.1017 | 1 | 0.5976 | –0.1760 | –0.2219 | –0.2187 |
| 6 | 0.3459 | 0.2690 | 0.6536 | –0.0142 | 0.5976 | 1 | –0.1649 | –0.3521 | –0.4450 |
| 7 | –0.2202 | 0.0542 | –0.1589 | 0.0635 | –0.1760 | –0.1649 | 1 | 0.4456 | 0.6732 |
| 8 | –0.1892 | 0.0037 | –0.1896 | 0.0759 | –0.2219 | –0.3521 | 0.4456 | 1 | 0.5561 |
| 9 | –0.2195 | 0.0047 | –0.2790 | –0.2246 | –0.2187 | –0.4450 | 0.6732 | 0.5561 | 1 |
| | | | | | $SO_2$ (Mean, standard deviation, COV) | | | | |
| Mean | 13.7302 | 15.7660 | 20.2849 | 20.0509 | 14.2189 | 11.9849 | 19.3717 | 27.4434 | 31.3834 |
| Std-dev | 6.2041 | 6.0801 | 10.3058 | 8.6919 | 4.8753 | 4.0876 | 14.1124 | 15.9782 | 22.7165 |
| COV | 0.4519 | 0.3856 | 0.5080 | 0.4335 | 0.3429 | 0.3411 | 0.7285 | 0.5822 | 0.7238 |

| $NO_2$ (Correlation matrix) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.0586 | 0.5962 | 0.1967 | 0.2262 | 0.3825 | 0.2524 | 0.1255 | 0.3664 |
| 2 | 0.0586 | 1 | –0.2100 | 0.2504 | –0.1719 | 0.0077 | –0.0166 | –0.0167 | –0.0021 |
| 3 | 0.5962 | –0.2100 | 1 | 0.0553 | 0.3812 | 0.3579 | 0.3509 | 0.2975 | 0.4427 |
| 4 | 0.1967 | 0.2504 | 0.0553 | 1 | 0.0593 | 0.0834 | –0.1861 | –0.2126 | –0.2111 |
| 5 | 0.2262 | –0.1719 | 0.3812 | 0.0593 | 1 | 0.6316 | 0.1324 | -0.0838 | 0.2483 |
| 6 | 0.3825 | 0.0077 | 0.3579 | 0.0834 | 0.6316 | 1 | 0.1401 | 0.1243 | 0.2240 |
| 7 | 0.2524 | –0.0165 | 0.3509 | –0.1861 | 0.1324 | 0.1401 | 1 | 0.8415 | 0.8638 |
| 8 | 0.1255 | –0.0167 | 0.2975 | –0.2126 | 0.0838 | 0.1243 | 0.8415 | 1 | 0.7653 |
| 9 | 0.3664 | –0.0021 | 0.4427 | –0.2111 | 0.2483 | 0.2240 | 0.8638 | 0.7653 | 1 |
| $NO_2$ (Mean, standard deviation, COV) | | | | | | | | | |
| Mean | 28 | 30.3264 | 33.4189 | 29.0396 | 32.5347 | 24.9774 | 40.7132 | 41.3377 | 64.4792 |
| Std-dev | 9.1255 | 6.8465 | 9.3310 | 7.9405 | 8.0705 | 5.0500 | 21.2377 | 22.8543 | 41.4108 |
| COV | 0.3259 | 0.2258 | 0.2792 | 0.2734 | 0.2481 | 0.2022 | 0.5216 | 0.5529 | 0.6422 |

(Table 1 continues on next page)

Table 1. (Cont'd.)

| Station No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| SPM (Correlation matrix) | | | | | | | | | |
| 1 | 1 | 0.6125 | 0.2291 | 0.4474 | 0.5234 | 0.4444 | 0.5542 | 0.1394 | 0.3299 |
| 2 | 0.6125 | 1 | 0.2789 | 0.6372 | 0.6090 | 0.5724 | 0.5540 | 0.2095 | 0.2786 |
| 3 | 0.2291 | 0.2789 | 1 | 0.1809 | 0.0672 | 0.2931 | 0.2240 | –0.0003 | –0.0243 |
| 4 | 0.4474 | 0.6372 | 0.1809 | 1 | 0.6383 | 0.5304 | 0.4340 | 0.2915 | 0.2740 |
| 5 | 0.5234 | 0.6090 | 0.0672 | 0.6383 | 1 | 0.5331 | 0.7305 | 0.3002 | 0.4376 |
| 6 | 0.4444 | 0.5724 | 0.2931 | 0.5304 | 0.5304 | 1 | 0.3347 | 0.0520 | –0.0089 |
| 7 | 0.5542 | 0.5540 | 0.2240 | 0.4340 | 0.7305 | 0.3347 | 1 | 0.3959 | 0.6324 |
| 8 | 0.1394 | 0.2095 | –0.0003 | 0.2915 | 0.3002 | 0.0520 | 0.3959 | 1 | 0.4389 |
| 9 | 0.3299 | 0.2786 | –0.0243 | 0.2740 | 0.4376 | –0.0089 | 0.6324 | 0.4389 | 1 |
| SPM (Mean, standard deviation, COV) | | | | | | | | | |
| Mean | 336.0980 | 326.9020 | 415.3992 | 355.2549 | 375.6078 | 335.3137 | 354.0588 | 440.0196 | 494.54 |
| Std-dev | 89.3795 | 107.2512 | 120.2052 | 95.5169 | 95.4024 | 93.1287 | 184.1715 | 241.5528 | 202.6098 |
| COV | 0.2659 | 0.3281 | 0.2894 | 0.2689 | 0.2540 | 0.2777 | 0.5202 | 0.5490 | 0.4097 |

Table 2. Intervals for Discretisation of Air Quality Data

| Pollutant | Intervals | | | | | |
|---|---|---|---|---|---|---|
| $SO_2$ ($\mu g/m^3$) | 0–20 | 20–40 | 40–60 | 60–80 | 80–100 | 100–200 |
| $NO_2$ ($\mu g/m^3$) | 0–20 | 20–40 | 40–60 | 60–80 | 80–100 | 100–200 |
| SPM ($\mu g/m^3$) | 0–200 | 200–400 | 400–600 | 600–800 | 800–1000 | 1000–1200 |

Table 3. Information Transmission by Permanent Network of Single Stations (Discrete Variables)

| Station location no. | Information transmission, in nats | | |
|---|---|---|---|
| | $SO_2$ | $NO_2$ | SPM |
| 1 | 0.218 | 0.777 | 0.714 |
| 2 | 0.484 | 0.425 | 0.989 |
| 3 | 0.663 | 0.809 | 0.947 |
| 4 | 0.826 | 0.615 | 0.779 |
| 5 | 0.353 | 0.529 | 0.820 |
| 6 | 0.094 | 0.424 | 0.724 |
| 7 | 0.884 | 1.378 | 1.252 |
| 8 | 1.122 | 1.399 | 1.596 |
| 9 | 1.447 | 1.665 | 1.321 |

network for three pollutants depending upon the size of the network. It may be noted from Table 4 that optimal network is a function of pollutant type.

## Discussion

The single stations providing the optimum information transmission are location number 9 for $SO_2$, and $NO_2$, and number 8 for SPM respectively (Table 3). These stations have highest COV (Table 1) for the respective pollutants and individually provide maximum information and thus are selected at first place. The optimum combination for two-stations is provided by locations 8 and 9 in case of SPM and $SO_2$ and 7 and 9 in case of $NO_2$ (Table 4). It may be noted that when more than two stations are to be selected, it is just not the individual variance/entropy, it is the maximum joint entropy (of all combinations) which decides the selection of new station(s). The basis of obtaining the optimum

Table 4. Optimal Networks of Different Sizes for $SO_2$, $NO_2$, and SPM (Discrete Variables)

| No. of stations, ($n$) | Optimal station location numbers | | | | | |
|---|---|---|---|---|---|---|
| | $SO_2$ network | $H(Si^n)$ | $NO_2$ network | $H(Si^n)$ | SPM network | $H(Si^n)$ |
| 1 | 9 | 1.447 | 9 | 1.665 | 8 | 1.596 |
| 2 | 8, 9 | 2.211 | 7, 9 | 2.563 | 9, 8 | 2.566 |
| 3 | 4, 8, 9 | 2.776 | 7, 9, 8 | 3.108 | 3, 9, 8 | 3.203 |
| 4 | 7, 4, 8, 9 | 3.124 | 1, 7, 9, 8 | 3.416 | 3, 7, 9, 8 | 3.491 |
| 5 | 3, 7, 4, 8, 9 | 3.343 | 3, 1, 7, 9, 8 | 3.663 | 2, 3, 7, 9, 8 | 3.723 |
| 6 | 2, 3, 7, 4, 8, 9 | 3.477 | 5, 3, 1, 7, 9, 8 | 3.803 | 1, 2, 3, 7, 9, 8 | 3.801 |
| 7 | 5, 2, 3, 7, 4, 8, 9 | 3.565 | 2, 5, 3, 1, 7, 9, 8 | 3.866 | 6, 1, 2, 3, 7, 9, 8 | 3.846 |
| 8 | 1, 5, 2, 3, 7, 4, 8, 9 | 3.591 | 4, 2, 5, 3, 1, 7, 9, 8 | 3.892 | 4, 6, 1, 2, 3, 7, 9, 8 | 3.846 |
| 9 | 6, 1, 5, 2, 3, 7, 4, 8, 9 | 3.591 | 6, 4, 2, 5, 3, 1, 7, 9, 8 | 3.918 | 5, 6, 1, 2, 3, 7, 9, 8 | 3.846 |

locations for a given size of the network is as follows: 1) for the given size (say 5 stations), first all possible combinations of stations must be determined (total combinations for five stations = 120); 2) for each combination, estimate the information transmission; and 3) select that combination which transfers the maximum information. For example, Figure 2 presents various combinations of five stations on x-axis and information transmitted by these combinations on y-axis. It is apparent from Figure 2 that for SPM, the combination number 91 transmits the information and thus the corresponding five locations for combination number 91 will make the optimum network (combination 91 corresponds to station number 2, 3, 7, 8, 9 in the present case). Similarly, for each pollutant and for each subset of combinations of various sizes of the network, the transmitted information was obtained and the optimum locations thus obtained are given in Table 4. It may be noted that the optimum locations are not exactly the same for all pollutants. Results also show that stations 8 (Netaji Nagar) and 9 (Town Hall) contribute relatively more information compared to other stations for all pollutants. From Table 4, it is clear that station 6 (Siri Fort) provides least information for $SO_2$ and $NO_2$ network and station 5 (Janakpuri) for SPM network.

Figure 2. Information transmission vs. AQMNs of size $n = 5$.

It is worth mentioning that selecting stations based on only individual entropy will not lead to optimum networks as there is overlap of information between them. For example, for a network of three stations for SPM, 3, 8, and 9 are selected based on their individual entropies while network of stations 4, 8, and 9 are selected based on the joint entropy. Linear correlation is not a good measure of association between two variables. It assumes relation between variables is linear and it is ignorant of the individual distributions. Joint entropy is a dependent upon the association among the variables and their individual dispersions. The above example can be explained intuitively using the correlation matrix for SPM. From Table 1, it is clear that station 7 has relatively good correlation with stations 8 (0.40) and 9 (0.63) while station 3 has very poor correlation with stations 8 (–0.00) and 9 (–0.02).

The optimal information transmission for different sizes of network can be normalized by computing the ratio of the uncertainty resolved by the optimal network of given size to the total uncertainty in the region (explained by all temporary stations being in operation). This is equal to the ratio of information transmission by the optimal network to the maximum information that can be transmitted.

So far, station locations have been prioritized, but it is still not clear as to with how many stations one can achieve the required information. To address this issue, in Figure 3 a set of curves indicated by B shows the plot between

Figure 3. Normalized information transmission by optimal AQMNs ($SO_2$/$NO_2$/SPM).

number of stations and cumulative normalized optimal information provided by the number of stations for present case study of Delhi. The curves A and C represent two extreme conditions; A: all stations in the network are completely dependent on each other and C: all stations in the network are completely independent of each other.

## Optimal Size of AQMN

In Figure 3, under the extreme condition shown by curve A, it is clear that a single station constitutes the most appropriate network and that marginal information gain for additional stations is zero. At the other extreme, curve C indicates the same marginal information gain for all station additions. In other words, for situation C it could be argued that all "*m*" stations are equally desirable and therefore, "*m*" station network is justified. However, an actual result typified by curve B, falls between these extreme conditions and provides a less clear indication of the appropriate size of a permanent network. A criterion such as a minimum marginal information gain from a new station to be added can be used to resolve this question of optimal network size.

It is clear from Figure 4 that as the size of the network increases the marginal incremental increase of information decreases. For $SO_2$ network, there is no addition of information after adding 6th station to the network (Figure 4). For

Figure 4. Marginal increase in information transmission with increase in size of AQMNs ($SO_2$/$NO_2$/SPM).

SPM network there is no further addition of information after 7th station (Figure 4) and for $NO_2$ network, there is no marginal increase of information after 8th station (Figure 4).

To decide the optimal numbers, two criteria are suggested. These include: (i) increase in marginal information by subsequent addition of every new station and (ii) percentage of total cumulative uncertainty explained by subsequent addition of every new station. For criterion (i), the constant slope (0.11) in curve C (of Figure 3) is obtained under the conditions where all nine stations might be regarded as being equally justified. The slope of curve C is, therefore, adopted as the minimum acceptable marginal information gain per station and applied to curve B, and then a network of three stations appears optimal for all pollutants as shown in Table 5.

For criterion (ii), it is decided that network must explain at least 90% of the total uncertainty, and then a network of four stations as shown in Table 5 appears optimal for all pollutants. The concepts of both the criteria are justified but the point of cut-off in the criteria (e.g., 0.11 increases in normalized marginal information and 90% explanation in uncertainty) are arbitrary. If greater financial resources are available, the cut-off point can be relaxed and more stations can be added. However, it does not really indicate number of stations should be less than number of stations presently located in Delhi. But it shows existing stations are not the optimal ones and the other locations should be explored to form a better AQMN. Thus it is a good idea to collect data for a large number of points in the area using direct measurement and/or air quality modeling to find the potential locations. It is quite possible that some other locations in the study area form a bigger optimal network and conveying more information about the area.

## Combined AQMN Design Considering Three Pollutants

As evident from the above discussion, the results suggest different optimal networks for each pollutant. But this is not a practical solution for developing countries like India as one cannot monitor different pollutants at different locations due to limited resources. Thus constrained optimal network where all pollutants are measured simultaneously at every location is desirable. If $SO_2$,

Table 5. Optimal Networks Based on Criterion (I) and Criterion (II) (Discrete Variables)

| | Station numbers for | | |
|---|---|---|---|
| Criterion number | $SO_2$ network | $NO_2$ network | SPM network |
| Crirterion (I) | 4, 8, 9 | 7, 8, 9 | 3, 8, 9 |
| Criterion (II) | 4, 7, 8, 9 | 1, 7, 8, 9 | 3, 7, 8, 9 |

$NO_2$, and SPM concentrations are considered independent, then basis for optimal design is proposed. Optimal location set can be found by optimizing the total information transmissions from all pollutants, i.e., sum of the information transmissions by $SO_2$, $NO_2$, and SPM network. The results of this proposed method are given in Table 6.

## Computational Environment

A combinatorial and sorting algorithm were developed to find the joint entropy for different combinations and to optimize the information transmission. The program takes the input as data file of pollutant concentration, the total number of stations in the permanent network (minimum number of stations taken) and intervals taken for discretisation of the concentration data set. Program was run on a Pentium-II machine. It was noted that time taken to run the program increases exponentially with increase in the selection of the minimum number of stations (i.e., the permanent network). It is due to the exponential nature of the problem, which is explained below with three stations in the network.

For a three-station network, there are 84 (*COMB*) ways of choosing three stations out of total nine stations. For each combination, program has to calculate joint probability 125 ($PERM = N \times N \times N$) times. See Table 7 for permutation and combination and computed time required as number of station increases.

From Table 7, it is clear that problem is of exponential nature. It is so costly to compute as the $n$ (size of the permanent network) of $N$ (number of intervals) increases. It is of the $N^k$ type problem, which is feasible for some limiting values

Table 6. Constrained* Optimal AQMNs of Different Sizes Measuring All Pollutant Simultaneously (Discrete Variables)

| Number of stations, $n$ | Optimal information transmission, in nats | $H(Si^n)/H(S)$ | Optimal station location number |
|---|---|---|---|
| 1 | 4.433 | 0.390 | 9 |
| 2 | 7.296 | 0.643 | 8, 9 |
| 3 | 8.8320 | 0.778 | 7, 8, 9 |
| 4 | 9.9360 | 0.875 | 3, 5, 8, 9 |
| 5 | 10.557 | 0.9297 | 7, 3, 5, 8, 9 |
| 6 | 10.889 | 0.9589 | 6, 4, 3, 5, 8, 9 |
| 7 | 11.117 | 0.979 | 1, 2, 3, 4, 6, 7, 8 |
| 8 | 11.284 | 0.994 | 5, 1, 2, 3, 4, 6, 8, 9 |
| 9 | 11.355 | 1 | 7, 1, 2, 3, 4, 5, 6, 8, 9 |

*All three pollutants must be measured at selected stations.

Table 7. Exponential Nature of Computational Problem

| "n" | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| COMB | 1 | 36 | 84 | 126 | 126 | 84 | 36 | 9 | 1 |
| PERM | 5 | 25 | 125 | 625 | 3125 | 15625 | 78125 | 390625 | 1953125 |
| Time | <1 sec | <1 sec | <1 sec | <1 sec | 0.07 min | 0.36 min | 1.8 min | 9 min | 45 min |

PERM = N × N × N . . . . .; N is the number of intervals taken in data.

of N (it is number of intervals here) and k (it is the number of minimum stations). Thus, as the number or the stations will increase, time of computation of joint entropy will increase exponentially. And after some value of *n,* it will become impossible to compute entropy in real time (see last column of Table 7).

## Addressing the $N^k$ Problem

Determination of entropy directly is computational intensive and thus feasible for limiting value of "*n.*" By knowing the statistical distribution followed by the data, it is possible to compute the joint entropy by considering the variable as a continuous variable. Entropy determination of a continuous variable is computationally less intensive. However, it has some drawbacks like it can be negative and it is dependent on the unit of the variable chosen. Therefore, sometimes it becomes difficult to interpret the results: for example, the value of the joint entropy will be at variance if units chosen are microgram/cubic meter or parts per million (for the same air quality level). However, there are ways to overcome these shortcomings as presented earlier (see section Definition of Entropy for Continuous Variable).

## Air Quality Monitoring Network Design for Continuous Variable Goodness of Fit Test

The Kolomogorov-Smirnov test was carried out to find the goodness of fit of the assumed distribution [20]. An investigation of two types of distribution, multivariate normal, and log normal, showed that there is no significant difference between their goodness of fit; both distributions qualified the goodness of fit test for 95% cases of data. However, in this research log normal distribution is assumed as earlier research shows that air quality data are log normally distributed [21-23].

## Results and Discussion

As discussed earlier (Measure of Entropy for Continuous Variables), determination of joint entropy is dependent on the unit chosen. Therefore to overcome

this problem, joint entropy is determined for proportional class interval, which is independent of the unit chosen using the log normal distribution (Equation 7).

The results obtained for optimal information transmission in the case where air quality parameters have been assumed as continuous variables are shown in Table 8.

## DISCUSSION

The single station providing the optimum information transmission is location 9 for $SO_2$, $NO_2$, and SPM as shown in Table 8. The optimum combination for two-stations is provided by locations 3 and 9 in case of $SO_2$, 5 and 9 in case of $NO_2$, and 8 and 9 in case of SPM as shown in Table 8. It is clear from Table 8 that station 6 (Siri Fort) provides least information for $SO_2$ and $NO_2$ network and station 1 (Nizammudin) for SPM network; thus these stations are least desirable.

### Optimal Size of Network

Based on criterion (i) (as explained earlier) for optimal size, number of stations comes out to be "4" for all the three pollutants as shown in Table 9 while on the basis of criterion (ii) the number of stations come out to be five for $SO_2$ and $NO_2$ and 6 for SPM.

### Combined AQMN Design Considering Three Pollutants

Using the same approach as discussed in combined design for discrete case is adopted in the combined design for continuous case (Table 10).

### Comparison of Results for Discrete and Continuous Variable Case

The joint entropies have come out to be somewhat larger in the case of continuous variable. The optimal network is not exactly matching for all values of "n" (specified number of stations) as found in the discrete case. This can be attributed to the assumed distribution that is not exactly giving the distributions followed by of the data. However, differences in final network designs are not significant. Table 10 shows the comparison between the continuous and discrete combined network design.

## CONCLUSION

The study has developed a generalized method based on entropy concepts for the design of an air quality monitoring network (AQMN). In the study, an AQMN is viewed as a communication channel which transmits information concerning the region served by the network. The information transmission capabilities of an

Table 8. Optimal Networks of Different Sizes for (n = 1 to 9) for Three Pollutants

| No. of stations, (*n*) | Optimal station location numbers and information transmission $H(Si^n)$ | | | | | |
|---|---|---|---|---|---|---|
| | $SO_2$ network | $H(Si^n)$ | $NO_2$ network | $H(Si^n)$ | SPM network | $H(Si^n)$ |
| 1 | 9 | 1.948 | 9 | 2.218 | 9 | 1.727 |
| 2 | 3, 9 | 3.081 | 5, 9 | 3.349 | 8, 9 | 3.123 |
| 3 | 8, 3, 9 | 4.008 | 5, 8, 9 | 4.355 | 7, 8, 9 | 4.158 |
| 4 | 7, 8, 3, 9 | 4.719 | 7, 5, 8, 9 | 5.120 | 3, 7, 8, 9 | 5.085 |
| 5 | 4, 7, 8, 3, 9 | 5.243 | 1, 7, 5, 8, 9 | 5.811 | 4, 3, 7, 8, 9 | 5.785 |
| 6 | 2, 4, 7, 8, 3, 9 | 5.440 | 4, 1, 7, 5, 8, 9 | 6.223 | 2, 4, 3, 7, 8, 9 | 6.269 |
| 7 | 1, 2, 4, 7, 8, 3, 9 | 5.559 | 3, 4, 1, 7, 5, 8, 9 | 6.603 | 5, 2, 4, 3, 7, 8, 9 | 6.660 |
| 8 | 5, 1, 2, 4, 7, 8, 3, 9 | 5.559 | 2, 3, 4, 1, 7, 5, 8, 9 | 6.857 | 6, 5, 2, 4, 3, 7, 8, 9 | 6.994 |
| 9 | 6, 5, 1, 2, 4, 7, 8, 3, 9 | 5.559 | 6, 2, 3, 4, 1, 7, 5, 8, 9 | 6.827 | 1, 6, 5, 2, 4, 3, 7, 8, 9 | 7.125 |

Table 9. Optimal Networks Based on Criterion (I) and Criterion (II)

| | Station numbers for | | |
|---|---|---|---|
| Criterion number | $SO_2$ network | $NO_2$ network | SPM network |
| Crirterion (I) | 3, 7, 8, 9 | 5, 7, 8, 9 | 3, 7, 8, 9 |
| Criterion (II) | 3, 4, 7, 8, 9 | 1, 5, 7, 8, 9 | 2, 3, 4, 7, 8, 9 |

Table 10. Comparison between Discrete and Continuous Constrained* Optimal Networks

| | Constrained Optimal Networks | | | |
|---|---|---|---|---|
| Number of stations, *n* | Optimal station location number (discrete variables) | $H(Si^n)/H(S)$ | Optimal station location number (continuous variables) | $H(Si^n)/H(S)$ |
| 1 | 9 | 0.390 | 9 | 0.304 |
| 2 | 8, 9 | 0.643 | 8, 9 | 0.487 |
| 3 | 7, 8, 9 | 0.778 | 7, 8, 9 | 0.631 |
| 4 | 3, 5, 8, 9 | 0.875 | 3, 7, 8, 9 | 0.762 |
| 5 | 7, 3, 5, 8, 9 | 0.9297 | 4, 3, 7, 8, 9 | 0.851 |
| 6 | 6, 4, 3, 5, 8, 9 | 0.958 | 2, 4 3, 7, 8, 9 | 0.905 |
| 7 | 1, 2, 3, 4, 6, 7, 8 | 0.979 | 5, 2, 4, 3, 7, 8, 9 | 0.956 |
| 8 | 5, 1, 2, 3, 4, 6, 8, 9 | 0.994 | 6, 5, 2, 4, 3, 7, 8, 9 | 1.00 |
| 9 | 7, 1, 2, 3, 4, 5, 6, 8, 9 | 1 | 1, 6, 5, 2, 4, 3, 7, 8, 9 | 1.00 |

*All three pollutants must be measured at selected stations.

AQMN have been assessed using the information theory. The proposed approach provides opportunity to find the priorities for air quality monitoring stations existing in an AQMN. The approach also provides the basis for optimal size of the network; specifically, answering the question: Does further addition of a new station in the network result in any significant information gain?

For large sizes of network, design considering the air concentrations as discrete variable is computationally difficult due to $N^k$ type of problem which can't be solved in real time. This issue has been addressed by taking the variables as continuous. It has been shown that two methods (discrete and continuous approaches) produce almost similar AQMNs.

The developed method is applied to the existing AQMN in Delhi. It has been shown that the generalized approach has successfully addressed the following issues for AQMN design of Delhi: 1) priority locations for sampling, and 2) optimal size of network.

Further research can be done to design an air quality monitoring network that not only considers the information content of the data but also the use-specific applications (e.g., trend analysis, standard compliance, etc.). In other words, design of the network strikes the balance between the mathematical measure of information content and practical user-oriented concerns that are relevant for AQMN. In this study, a combined network is designed based on giving equal weights to the environmental information provided by each pollutant. Since it might be possible that some pollutants are more important based on their health effects, different weights can be considered in future studies.

## REFERENCES

1. T. D. Lee, R. J. Graves, and L. F. McGinnis, A Procedure for Air Monitoring Instrumentation Location, *Management Science, 24,* pp. 1451-1461, 1978.
2. Y. Nakamori, S. Ikeda, and Y. Sawaragi, Design of Air Pollutant Monitoring System by Spatial Sample Stratification, *Atmospheric Environment, 13,* pp. 97-103, 1979.
3. Y. Nakamori and Y. Sawaragi, Interactive Design of Urban Level Air Quality-Monitoring Network, *Atmospheric Environment, 18,* pp. 793-799, 1984.
4. P. M. Modak and B. N. Lohani, Optimization of Ambient Air Quality Monitoring Networks (Part I), *Environment Monitoring and Assessment, 5,* pp. 1-19, 1985.
5. P. M. Modak and B. N. Lohani, Optimization of Ambient Air Quality Monitoring Networks (Part II), *Environment Monitoring and Assessment, 5,* pp. 21-38, 1985.
6. P. M. Modak and B. N. Lohani, Optimization of Ambient Air Quality Monitoring Networks (Part III), *Environment Monitoring and Assessment, 5,* pp. 39-53, 1985.
7. C. M. Handscombe and D. M. Elsom, Rationalization of the National Survey of Air Pollution Monitoring Networks Using Spatial Correlation Analysis—A Case Study of the Greater London Area, *Atmospheric Environment, 16,* pp. 1061-1070, 1982.
8. H. Katoch, S. Nagasawa, A. Ootaki, and K. Shiozawa, Study on Representativeness of Air Pollution Station by Statistical Model (in Japanese), *Journal of Japan Society of Air Pollution, 20,* pp. 384-393, 1985.
9. World Health Organization, *Air Monitoring Programme Design for Urban and Industrial Areas,* WHO Offset Publication No. 33, Geneva, 1977.
10. EPA, *Guidelines for Air Quality Maintenance Planning and Analysis, Volume II: Air Quality Monitoring and Data Analysis,* Report No. EPA-450/4-74-012, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 1974.
11. W. F. Caselton and T. Husain, Hydrologic Networks: Information Transmission, *Journal of Water Resources Planning and Management Division,* ASCE, 106 (WR2), pp. 503-520, 1980.
12. N. Harmanicioglu and V. Yevjevich, Transfer of Hydrologic Information Among River Points, *Journal of Hydrology, 91,* pp. 103-118, 1987.
13. T. C. Chapman, Entropy as a Measure of Hydrologic Data Uncertainty and Model Performance, *Journal of Hydrology, 85,* pp. 111-126, 1986.

14. V. P. Singh, Hydrologic Modeling Using Entropy, *Journal of the Institute of Engineering, Civil Engineering Division, 80* Part CV2, pp. 55-60, 1989.
15. S. Kullback, *Information Theory and Statistics,* Wiley, New York, 1959.
16. L. Brillouin, *Science and Information Theory,* Academic Press, New York, 1962.
17. C. E. Shannon, Mathematical Theory of Communications, I and II, *Bell System Technical Journal, 27,* pp. 379-423, 1948.
18. WHO-UNEP, *Urban Air Pollution in Mega Cities of the World,* Blackwell, Oxford, 1992.
19. CPCB, *National Ambient Air Quality Statistics of India,* NAAQM Series: NAAQM/1-8/1990-95, 1990-95.
20. I. Miller, J. E. Freund, and R. A. Johnson, Probability and Statistics for Engineers (4th Edition), Prentice-Hall, USA, 1990.
21. F. A. Glifford, Jr., *The Lognormal Distribution of Air Pollution Concentrations,* Air Resources Atmospheric Turbulence and Diffusion Laboratory, ESSA, Oak Ridge, Tennessee (pre print, 3p.), 1969.
22. J. B. Knox and R. K. Pollack, *An Investigation of the Frequency Distribution of Surface Air Pollution Concentrations,* proceedings of the Symposium on Statistical Aspects of Air Pollution Data, U.S. EPA, Research Triangle Park, North Carolina, No. EPA-650/4-74-038, pp. 9-1 to 9-17, 1974.
23. Y. Kalpasanov and G. Kurchatova, A Study of the Statistical Distribution of Chemical Pollutants in Air, *Journal of Air Pollution Control Association, 26*:10, pp. 981-985, 1976.

Direct reprint requests to:

Mukesh Sharma
Department of Civil Engineering
Indian Institute of Technology Kanpur
Kanpur 208016, India
e-mail: mukesh@iitk.ac.in