Published in final edited form as:
*Atmos Environ*. 2009 February ; 43(5): 1153. doi:10.1016/j.atmosenv.2008.10.055.

# An Optimal Spatial Sampling Design for Intra-Urban Population Exposure Assessment

**Naresh Kumar**

## Abstract

This article offers an optimal spatial sampling design that captures maximum variance with the minimum sample size. The proposed sampling design addresses the weaknesses of the sampling design that Kanaroglou et al. (2005) used for identifying 100 sites for capturing population exposure to $NO_2$ in Toronto, Canada. Their sampling design suffers from a number of weaknesses and fails to capture the spatial variability in $NO_2$ effectively. The demand surface they used is spatially autocorrelated and weighted by the population size, which leads to the selection of redundant sites. The location-allocation model (LAM) available with the commercial software packages, which they used to identify their sample sites, is not designed to solve spatial sampling problems using spatially autocorrelated data. A computer application (written in C++) that utilizes spatial search algorithm was developed to implement the proposed sampling design. This design was implemented in three different urban environments - namely Cleveland, OH; Delhi, India; and Iowa City, IA - to identify optimal sample sites for monitoring airborne particulates.

### Keywords

optimal spatial sampling design; spatial-autocorrelation; intra-city exposure; and variance maximization

## 1. INTRODUCTION

There is an increasing interest in air pollution data at a high spatial resolution, because a few centrally located monitoring stations fail to capture spatial variability in air pollution within an urban area.1 Thus, the data from these stations alone can under- or overestimate exposure, which in turn could result in error and uncertainty in the quantification of the burden of diseases of air pollution. A spatial sampling design is required to capture spatial variability. In the article published in this journal, Kanaroglou et al.[2] suggested a location-allocation model (LAM) with the maximum attendance objective function for establishing a network of air pollution monitoring stations for capturing intra-urban population exposure to $NO_2$. The two goals of their paper were: (1) to develop a formal method of optimally locating a dense network of air pollution monitoring stations, and (2) to derive an exposure assessment model based on data monitored at these stations and related land use type, population, and biophysical environment.

The goals outlined in Kanaroglou et al. are extremely important for two reasons. First, air pollution monitoring for capturing spatial variability in exposure is expensive and time consuming. Therefore, it is critically important to optimize the locations of sample sites, so

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

that they capture and represent population exposure. Second, our understanding of the exposure in microenvironments is far from complete due to the limited spatial coverage of air pollution monitoring networks. While these goals are important, the methodology they proposed to identify an optimal network of monitoring stations suffers from a number of weaknesses and failed to meet the objectives outlined in their paper. The remainder of this article is organized into three sections. The first section presents an overview of LAM for operational research. The second section describes the main weaknesses of the sampling design Kanaroglou et al. used. The final section details the optimal spatial sampling design that captures maximum variance with the minimum sample size and addresses the weaknesses of Kanaroglou et al.'s sampling design.

## 2. LOCATION ALLOCATION MODEL – INTRODUCTION

### 2.1 Location-Allocation Model for Air Pollution Sampling Network

Operational research utilizes the location-allocation model (LAM) for various purposes, including service delivery, strategic planning and operational research. The implementation of LAM requires data on three different things – (a) a list of demand sites (i.e. all areas for which air pollution estimates are required), (b) potential candidate locations of service centers (i.e. a list of all logistically feasible locations where samplers can be deployed), and (c) a network that will connect both demand points and candidate locations. The third, however, is not essential if demand sites and candidate sites are not connected through a transport network, and Cartesian distance is used to establish the connectivity between these two. A theoretical framework and comparison of different LAMs are available elsewhere.[3]

## 3. Critique of Kanaroglou et al.'s Sampling Design

### 3.1 LAM for spatial sampling for air pollution monitoring

Kanaroglou et al. recognize the importance of intra-city variability in air pollution and utilize it as one of two criteria for determining the weights of candidate locations. The methodology they used for capturing spatial variability and weighting it by the population is problematic. First, weighting spatial variability (in the air pollution surface) by the population can result in the selection of redundant sites and degrade the efficiency of the sample design to capture population exposure. Weighting demand by population biases the site selection in favor of densely populated areas and does not necessarily capture spatial heterogeneity in air pollution. For example, selecting a site in an area of high spatial autocorrelation (in air pollution distribution) is adequate to capture population exposure in that area. And of course, one would like to restrict the selection of sample sites to residential areas only, if the goal of sampling is to assess population exposure.

Second, they used a linear regression model to generate a variability surface of ($NO_2$) based on the relationship between $NO_2$ observed at the existing 16 monitoring stations and the land-use characteristics around these sites. The variables they used, such as lengths of expressways within 0–50m and 50–200m of the monitoring stations, were significantly autocorrelated, and the value of this variable gradually declines as the distance from the expressway increases. As evident in Table 1 (*in their paper, p.2405*) the regression coefficients of expressway lengths within 0–50m and 50–200m were -592 and 253, respectively. The first was negatively correlated with the $NO_2$ and the second was positively correlated with $NO_2$. Since both were highly autocorrelated, both canceled out each-other's influence and resulted in the biased estimation of regression results. Given the spatial dependence structure in the variables they used in the regression, the $NO_2$ surface they predicted (at 5m spatial resolution) had very high spatial autocorrelation. Without controlling for spatial autocorrelation, the sum of the demand (and also the total variance) is significantly exaggerated, and results in the selection of redundant sites. LAM, available in the commercial software packages such as ArcInfo, do not account for spatial autocorrelation, as these were not designed for solving the spatial sampling problems.

Third, while the formalization of sampling design is important, formalization of the sample size is even more important, because air pollution monitoring is expensive and time-consuming. Therefore, it is critically important to determine the sample size such that it adequately captures the representative estimates of population exposure. The classical sampling theory suggests that the sample size ($n$) for computing sample mean with 95% confidence interval can be defined as $n = (1.96*\sigma/e)^2$, where σ is the standard deviation and $e$ is the acceptable margin of error. The computation of variance ($\sigma^2$) based on spatially autocorrelated data can result in over estimation of variance and hence oversampling. Kanaroglou et. al. determined $n \sim 100$ in an ad-hoc fashion and did not provide any justification. Therefore, it is not clear it is difficult to evaluate the statistical robustness of the exposure estimation based on the air pollution data monitored the identified 100 sample sites.

### 3.2 Allocation modeling

Kanaroglou et al. state that ArfInfo software (ESRI[3, 4]) "offers two options: the P-Median Problem or P-MP and the Attendance Maximizing Problem" (p. 2404) – to implement LAM. In fact, this software offers four different location criteria for implementing LAM, namely mindistance, mindistpower, maxattend and maxcover.[4] Among these four the distance minimization and exponential distance minimization belong to P-Median group. They described P-Median and Attendance Maximization in their paper and made a case for employing the latter to identify 100 sites. The locational criterion they chose maximizes the attendance at the service centers, assuming that attendance declines with distance in a linear fashion. This criterion, however, does not account for spatial autocorrelation and could chose spatially clustered locations if population size is relatively large at the adjacent location.

The goal of an optimal network for air pollution monitoring should be to capture the best representation of air pollution exposure with the available sample size (or the minimum sample size) rather than optimizing geographic access or attendance at the monitoring stations. Given the demand surface of spatial-variability was weighted by the population, LAM with the attendance maximization must have biased sample sites in the favor of areas with relatively higher population concentration and resulted in clustering the sample sites. As stated earlier, to optimize the location of sample sites, the spatial autocorrelation must be minimized and location candidates should be restricted to the residential areas only if the goal of sampling is to compute population exposure.

### 3.3 Variance surface

Using the regression coefficients from the linear regression model they estimated initial air pollution surface at 5m spatial resolution, and then computed spatial variability at the centriod of each 5m cell. Their equation (1) *on page 2400* adopted from Cressie[5] is misleading for two different reasons. First, in the original equation of the average semivariogram ($\hat{\gamma}_h$) the denominator ($k$, i.e. all pairs of points within a distance range $h$) is multiplied by 2, because $\hat{\gamma}_h$ includes *all pairs of points* within a distance range $h$, and each pair of points is included twice. Second, dividing the numerator by 2 does not make any sense to compute the local estimate of variance of $NO_2$ (at given pixel) from all its neighbors within the specified $h$ distance (≤300m in their case). If the goal was to compute average variance, then denominator must include just the number of neighbors around each $i^{th}$ candidate site within $h$ distance.

Another important thing for generating variance surface is to define $h$ empirically and determine whether $h$ should vary locally or remain constant throughout the study area.

Kanaroglou et al., however, do not shed any light on why they chose *h* ~ 300m, and if *h* should have varied from the central parts of the city to suburban areas or for individual candidate site.

## 4. POTENTIAL SOLUTIONS

### 4.1 Calculating local estimates of spatial variance

Preliminary estimates of spatial variance are needed to identify sample sites for air pollution monitoring. By adopting directly from the formula of global average semivariance, the local semivariance can be calculated as

$$\widehat{\gamma}_i = \frac{1}{k}\sum_{j=1}^{k}(z_i - z_j)^2_{i \neq j, d_{ij} \to h} \tag{1}$$

Where $z_i$ is the value of air pollution at $i^{th}$ candidate/demand location; $z_j$ is the air pollution concentration at $j^{th}$ neighbor within distance range l of *h; k* is the number of neighbors around $i^{th}$ candidate within distance range *h*; and $\hat{\gamma}_i$ is the average semi-variance at $i^{th}$, and indicates how different air pollution concentration is at this location with respect to its neighbors. A low value of $\hat{\gamma}_i$ means air pollution concentration at $i^{th}$ site is quite similar to that at its neighboring locations with distance *h*, and monitoring air pollution there could capture representative estimates of air pollution in its neighboring areas.

The value of $\hat{\gamma}_i$ can largely be influenced by the selection of *h*. Therefore, it is critically important to define *h* empirically. A semivariogram can help us determine *h*, i.e. the extent within which spatial autocorrelation is statistically significant. The global semi-variogram, however, does not provide insight into whether the geographic extent of spatial autocorrelation is constant or varies regionally or locally. Therefore, another important thing to consider is to develop regional and local semivariograms and vary $h_i$ by sites $i$ = {1,…N} or by regions. Building on this concept, equation (1) can be rewritten as

$$\widehat{\gamma}_i = \frac{1}{k}\sum_{j=1}^{k}(z_i - z_j)^2_{i \neq j, d_{ij} \to h_i} \tag{2}$$

The way $\hat{\gamma}_i$ is calculated (using a moving window) enhances the intensity of spatial autocorrelation in the variance surface. Thus, air pollution ($z_i$) observed at the $i^{th}$ location should suffice as the weight for a potential candidate for the site selection.

### 4.2 Spatial autocorrelation corrected variance estimation

Once the air pollution surface is calculated the sample size needs to be determined, which requires the estimates of variance $\sigma_z{}^2$ and error tolerance. Since most air pollution and environmental data observe significant spatial autocorrelation, the classical way of computing variance could overestimate $\sigma_z{}^2$. An alternative strategy would be to compute variance controlling for spatial autocorrelation as given below

$$\sigma^2 = \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{k}\forall_{ij}}\sum_{i=1}^{N}\sum_{j=1}^{k}(z_i - z_j)^2\forall_{ij} \tag{3}$$

where $\forall\ _{ij}$= 0 if $d_{ij} \leq h$

### 4.3 Optimal network

Optimal location identification can be considered as a two step process. In the first, identify all potential candidates $i$={1,…,$N$}. If the sampling goal is to estimate population exposure, residential areas would be an ideal choice to define candidate locations.[6] In the second, define the objective function to identify the sample sites $i$={1,…,$n$} ∈ {1,…,N}. Since the goal of the proposed sampling design is to capture maximum variance in $z$ by the set of $n$ sample sites, the objective function can be written as

$$\max |z|=\frac{1}{n}\sum_{\substack{i=1 \\ z_i\neq z_j, d_{ij}>h}}^{n}(z_i-\bar{z})^2 \tag{4}$$

Air pollution $z_i$ observed at the $i^{th}$ sample site must be significantly different from that observed at its neighboring sample site ($z_j$). The distance ($d_{ij}$) between a sample site ($z_i$) and its neighbors ($z_j$) must be > $h$ so that the spatial autocorrelation $\rho_z \sim 0$ in z across the set of $n$ sample sites is zero. Sample size is critically important to define the amount of variance $\sigma^2{}_z$ captured by the set of $n$ sample sites, and the $n$ can be conditioned on $m$, as

$$\frac{\sum_{i=1}^{n}(z_i-\bar{z})^2}{\sum_{i=1}^{N}(z_i-\bar{z})^2\forall} \geq m \tag{5}$$

where 0 < $m$ < 1 and determines the extent of variance to be captured by set of n sample sites, and ∀= 0 if $d_{ij} \leq h$. The decision $m$ will be governed by the availably of resources and the required precision in the exposure estimation. A value of 1 will require monitoring air pollution at all candidate locations avoiding for spatial autocorrelation, meaning $d_{ij} > h$. The proposed methodology logs the total variance captured by each site selection incrementally and hence can provide information about the fraction of total variance likely to be captured by air pollution monitored at the proposed optimal sample sites.

## 5. SUMMARY

I have read the work of this group and I am impressed by the way they have brought spatial aspects of air pollution monitoring and its associated health effects to the forefront of epidemiological studies. While the spatial variability in air pollution and the methods of capturing spatial variability are critically important, one should not overlook the limitations of the existing tools that are not particularly suited for developing an optimal network of air pollution monitoring, such as LAM used by Kanaroglou et al.[2] It is important to conceptualize these models and evaluate them carefully as to whether they meet the identified goals of capturing spatial variability (in air pollution) and assessing population exposure. They have introduced LAM to develop an optimal network of air pollution monitoring stations. Their methodology, however, fails to achieve the goals outlined in their paper, and suffers from a number of weaknesses, including redundancy in site selection, placing sites in uninhabited areas, wrong formulation of local spatial variance and misuse of regression. Reading this paper can provide insight into an optimal spatial sampling design that computes the minimum sample size, captures the maximum variance (in the air pollutant in question) and addresses the weakness that Kanaroglou et al.'s sampling design suffers from. The proposed optimal spatial sampling design was implemented and tested successfully for monitoring air pollution in three very different urban environments – Cleveland, OH, Delhi, India and Iowa City, IA.

## References

1. Ott D, Kumar N, Thomas P. Passive sampling to capture spatial variability in $PM_{10-2.5}$. Atmospheric Environment 2008;42:746–56.

2. Kanaroglou PS, Jerrett M, Morrison J, Beckerman B, Arain MA, Gilbert NL, Brook JR. Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. Atmospheric Environment 2005;39(13):2399–409.

3. Kumar N. Changing Geographic Access to and Locational Efficiency of Health Services in Two Indian Districts between 1981 and 1996. Social Science and Medicine 2004;58(10):2045–67. [PubMed: 15020019]

4. ESRI. ArcGIS, Version 9.1, Redlands. CA: Environmental Systems Research Institute; 2005.

5. Cressie, N. Statistics for Spatial Data. New York: Wiley; 1993.

6. Kumar N. Spatial Sampling Design for Survey Based Research: Some Experiences from Respiratory Health and Demographic Survey of Households in Delhi, India. Population Research and Policy Review. 2007