Spatial Sampling for Demography and Health Survey

Naresh Kumar
Assistant Professor
Department of Geography
University of Iowa, Iowa City, IA 52242
Email: naresh-kumar@uiowa.edu

Spatial Sampling for Demography and Health Survey

Abstract: The recent advances in global position systems (GPS), geographic information systems (GIS) and remote sensing (RS) can be exploited for spatial sampling design for demographic and health survey. These technologies are, particularly, useful when a sampling frame is unavailable and/or location (of household) is important for data collection, such as population exposure to ambient air pollution can be greatly impacted by the location of residence. Building on these technologies, this article presents a methodology of spatial sampling adopted for the respiratory health and demographic survey conducted in Delhi and its environs from January-April 2004. The overall goal of the survey was to select households that adequately represented exposure to ambient air pollution. The proposed methodology involved constructing a sampling frame of residential areas and the simulation of weighted random points within the residential areas. The simulated locations were navigated with the aid of GPS to identify households at these locations and to acquire their consent to participate in the survey; a total of 1576 households at the 2000 simulated locations were found suitable and participated in the survey. The average ambient air pollution at the sample sites was not significantly different from the average air pollution observed in the study area, which demonstrates the robustness of the proposed sampling method.

Keywords: Spatially sampling, demographic and health survey, GIS and random point.

1. INTRODUCTION

The methods of spatial sampling have been in practice for a while, but their applications have been restricted to sample natural phenomena, such as plants, soil types and mineral deposition (Amblard-Gross et al., 2004; Dessard and Bar-Hen, 2005; Di Zio et al., 2004; Fleishman et al., 2005) and continuous phenomena, such as air pollution (Arbia and Lafratta, 2002). The application of spatial sampling to sample human population, however, is relatively new. With the increasing interests in the role of place and space in the theories of social-sciences, recording spatial contexts of demographic and health data and hence spatial sampling design is becoming increasingly important. Recent literature suggests that place and space matters in our day-to-day life (Goodchild and Janelle, 2004; Longley et al., 1999) and the local/neighborhood environment can greatly influence socio-economic, demographic, economic and health outcomes (Cerin et al., 2006; Gordon-Larsen et al., 2006; Liao et al., 2006; Schwartz, 1996). Given the increasing importance of understanding spatial contexts, this article demonstrates the use of GPS, GIS and RS for constructing a spatial sampling design for demographic and health surveys.

An effective spatial sampling design is required to capture spatial contexts. Borrowing from the conventional theories of sampling, a myriad of spatial sampling designs has been developed to sample natural and/or continuous phenomena (Amblard-Gross et al., 2004; Angulo et al., 2005; Di Zio et al., 2004; Stevens and Olsen, 2004; Thompson, 2002). Generally, these designs employ a type of systematic sampling where sampling domains are represented by regular polygons as strata and individual grids are selected randomly within the identified strata. In sampling human population, however, the goal is to sample households from a discrete geographic space with inhomogeneous population distribution. Therefore, the methods of spatial sampling adopted in natural sciences cannot be directly extrapolated to sampling human population.

Advances in GPS, GIS and RS technologies provide a unique opportunity for constructing an effective spatial sampling design. The proposed sampling design can be particularly valuable when conventional methods of sampling cannot be implemented because of the non-availability of a sampling frame. These new technologies can be exploited to identify residential areas, which can serve as an indirect measure of the sampling frame of households. This paper stems from our experiences in administering a respiratory health and demographic survey of 1576 households in Delhi Metropolitan from January-April 2004. The survey aimed at collecting data required to model respiratory health outcomes as a function of exposure to air pollution, particularly exposure to ambient air pollution and potential confounding (socio-economic and demographic) variables. Thus, capturing spatial variability in air pollution at the household level was at the heart of formulating the sampling strategy. An additional aspect was to assess the contribution of air pollution from different sources at the household location.

The article is organized into four sections. After a brief introduction in the first section above, the second section presents the background development of spatial-sampling methods. The third section describes the database used for constructing a sampling frame

of residential areas. The fourth section discusses the implementation strategy, which is followed by a discussion and conclusions in the final section.

2. SPATIAL SAMPLING-BACKGROUND: The spatial sampling differs from non-spatial sampling methods, because a sample is selected based on geographic locations and/or their associated characterizes. Borrowing from the theory of conventional sampling, a myriad of spatial sampling methods have been developed and tested. The methods of spatial sampling have been exploited extensively to sample natural phenomena using a regular geometric pattern (Amblard-Gross et al., 2004; Dessard and Bar-Hen, 2005; Dumitrescu et al., 2006; Stevens and Olsen, 2004). The main idea behind spatial sampling is to determine an optimal sample size and select sampling locations/sites $\{s_1, \ldots s_n\}$ from which data $\mathbf{Z} = (Z(s_1),\ldots Z(s_2))$ can be used to estimate $g(Z(.))$ (Cressie, 1990) where g refers to some geographic area/extent.

Spatial sampling methods have been consistently used to collect data for natural resources with the main objective of predicting or estimating natural resources in the entire geographic extent (Stevens and Olsen, 2004). Unlike conventional methods of sampling, spatial sampling does not rely on a sampling frame of entities of interest, such as plants and animal species, soil and mineral resources. Because it is impractical to construct a sampling frame of natural resources. Therefore, the first step in spatial sampling is to construct a frame of a finite population of identifiable geographic units. Generally, this is achieved by overlaying a geometric pattern onto the geographic area of interest that, in essence, generates a sampling frame by partitioning the geographic extent into a finite number of identifiable units N. In the next step, one of the four classic methods of sampling – simple random, systematic random, stratified random, cluster random – or a combination of two or more of these four can be employed to select n sample sites from the set of N units (Cressie, 1990).

An overarching goal of a spatial sampling design is to predict/estimate the variable of interest with the least number of monitoring sites. Two methods are usually suggested to achieve this, namely space filling (Nychka and Saltzman, 1998) and optimal Bayesian design using inhomogeneous Markov chain simulation (Müller, 1998). Much of the focus on spatial sampling in recent years been on geometric structure for generating a random grid and the optimization with respect to sample size, variance maximization and spatial autocorrelation (Dessard and Bar-Hen, 2005; Dumitrescu et al., 2006; Getis et al., 2000; Salehi, 2004; Zhu and Stein, 2006; Zhu and Zhang, 2006). The most common regular grids are the equilateral triangular grid, the rectangular/square grid and the hexagonal grid (Cressie, 1990). Among the regular sampling plans, the equilateral triangular plan is the most efficient in terms of averages and maximum Kriging variances (McBratney and Webster, 1981).

The optimal sampling design seeks to capture maximum variability by the minimum possible sites, which involves an appropriate distance/spacing between sample sites so that spatial autocorrelation can be minimized/eliminated. One would discard autocorrelated sites for two reasons. First, avoiding spatial autocorrelation reduces redundancy. Second, the spatial correlation structure can have further implication for

choices regarding spatial-sampling design, estimation and prediction and observational method (Salehi, 2004). A model-based approach utilizing such correlation patterns has been particularly influential in the geographic survey of mineral and fossil fuel resources (Thompson, 2002). Different methods of modeling correlation structure in spatial data are discussed by Cressie (1990).

Recent literature on spatial sampling, especially in the context of natural sciences, draws attention to variability in 'inclusion probability' for identifying *n* sites (Arbia and Lafratta, 2002; Lafratta, 2006). For human and natural phenomena that experience an inhomogeneous distribution, it is reasonable to assume differential probability in their distribution. As a result, a good spatial sampling design recognizes the spatial inhomogeneities in the processes and accounts for them while choosing sample size *n* and sample locations $\{s_1, \ldots, s_n\} \subset R$ (Cressie, 1990). In a recent article, Stevens Jr. and Olsen (2004) suggest a spatially balanced approach to sampling natural resources by translating geographic space into an inclusion probability. In other words, larger geographic areas (longer in case of linear features) are more likely to be selected than smaller ones.

The methods of spatial sampling adopted in natural sciences cannot be directly used for sampling human population for several reasons. First, most of these methods heavily rely on overlaying a regular geometric structure onto the geographic extent of an area to identify a finite number of *N* units and assume homogeneity in each unit. The use of a regular grid, however, does not match with the highly irregular shapes of residential areas. Therefore, a regular grid design is of little use for constructing a sampling frame of households, required for survey based social-science and/or health research.

Second, some households can spread across two adjacent grids, which can violate the assumption of independence in the distribution within each grid (cell). Third, overlaying a regular grid assigns equal probability to each unit, but the number of households can vary across grids. This necessitates the use of differential probability of inclusion in the sample design. Fourth, the major focus of these methods has been on collecting data for a single attribute. Sampling households, however, requires data collection for hundreds of attributes of both households and individuals living in them.

Because of the mentioned constraints, researchers have begun to adopt a more realistic approach to spatial sampling with varying probability of selection across discrete geographic spaces (Arbia and Lafratta, 2002; Cressie, 1990; Lafratta, 2006; Stevens and Olsen, 2004). The research by Lee et al. (2006) is particularly relevant for this study. They demonstrate the use of GIS to draw a sample of respondents in an urban environment, and their main goal was to understand how built environment facilitates/hinders walking and biking. Using the parcel data they constructed a sampling frame of built-up environment, and respondents were drawn randomly from the identified sampling frame.

Although the conceptual idea of constructing a sampling frame is somewhat similar to that adopted by Lee et al. (2006), the proposed sampling design is unique for several

reasons. First, exploiting RS and GIS technologies a sampling frame was constructed from scratch. Second, a customized application was designed (in GIS) that simulated the required number of sampling points in discrete geographic spaces with varying probability of inclusion. Third, it was plausible to integrate various types of data (including pollution sources) with each simulated location, because both the simulated points and other data were georeferenced. Finally, the interface between GIS and GPS allowed navigation and identification of households at the simulated locations and the finalization of household inventory.

Evaluation of sample design: The effectiveness of a sampling design can be evaluated by how precisely it can estimate population parameters, such as mean and covariance (Ackers et al., 2005; Lafratta, 2006). Testing a sample of random points, however, requires the test of complete spatial randomness (CSR). In a Cartesian space, it is relatively simple to evaluate the performance of randomness in points, in which distribution of points can be assumed as the realization of a Poisson process and the number of points within an area $n(a) \sim \text{Poisson}(\lambda)$. Testing randomness in point locations, however, can be challenging when a sample is derived from discrete geographic spaces with varying probability. Therefore, most studies recommend using Monte Carlo simulation with the assumption of an inhomogeneous Poisson distribution to construct a theoretical estimate (Cressie, 1990; Lewis and Shedler, 1979), in which $n(a) \sim \text{Poisson}(\lambda(s))$ and the λ varies by space. Although we constructed the theoretical estimates of k-statistics by simulating a set of 2000 points 99 times, as recommended for a 95% confidence interval (Schabenberger and Gotway, 2005), we also evaluated the performance of our sample by comparing the mean air pollution at the selected household locations with the average estimate of air pollution in the study area, because an overarching goal our sampling design was to draw households that represent exposure to ambient air pollution in the study area.

3. DATABASE

The data for this research come from a number of sources, including the National Remote Sensing Agency (NRSA) and Survey of India. The following data were used for constructing a sampling frame of residential areas and strata identification:

- Particulate matter in a range of 1 to 10μm in aerodynamic diameter recorded at 113 sites spread across Delhi and its neighboring areas from July-December 2003.
- Indian Remote Sensing (IRS) satellite imageries – Panchromatic and mutli-spectral (LISS) mode, 2002.
- Topographic maps of Delhi and neighboring states from Survey of India.
- Street map of Delhi from Eicher (EICHER, 2001).

4. Spatial Sampling Design: formulation, implementation and validation

This section describes the methodology adopted for sample selection, its implementation and evaluation. The details about these three are organized across six parts – (a) spatially

detailed air pollution data, (b) stratification by air pollution levels and their sources, (c) constructing a sampling frame of residential areas, (d) random point simulation with varying probability of inclusion, (e) household identification at the simulated locations and (f) the evaluation of the sample.

4.1 Air Pollution in Delhi and its environs: An overarching goal of the proposed sampling was to recruit households such that the ambient air pollution at the selected household locations truly represented exposure to ambient air pollution in the study area. A secondary goal was to identify potential sources of air pollution around the sample sites. Although there are various air pollutants in the environment, we particularly focused on airborne particles of different sizes ≤2.5μm and ≤10μm in aerodynamic diameter, $PM_{2.5}$ and PM_{10}, respectively, because these two have been recognized as standard measures of air quality worldwide (WHO, 2000). The terms air quality and air pollution will be referring to $PM_{2.5}$ and/or PM_{10} in the remaining parts of this article.

Given the limited spatial-temporal coverage of air pollution data for the study area, a field campaign was conducted and air pollution was monitored from July 23 to December 3, 2003 at 113 sites in Delhi and its surroundings. For selecting these sample sites, a spatially dispersed sampling design was adopted, in which sample sites were identified using a two-step process. In the first, a rectangular grid was overlaid onto the entire study area, which ensured full coverage of the area. In the second step, a random location was simulated within each cell (of size 1x1.5km), and the simulated locations were transferred to a Garmin Global Positioning System (GPS) in order to navigate them and examine their suitability. Some sites, which were inaccessible, were discarded and re-simulated, resulting in a final sample of 113 suitable sites (Figure 1). At each site air was sampled at two different times between 7:30AM and 10:00PM every third day. Each sample involved four readings – two each in mass and count modes; each reading was based on two minutes of sampling in the mass mode and one minute of sampling in (particles) count mode.

The Aerocet 531, a real time photometric sampler, from Met One Instruments, Inc., was used to collect air pollution data (Met One Inc, 2003). It is an automatic instrument that can estimate particulate mass (PM) in a range of ≤1, ≤2, ≤5, ≤7 and ≤10μm in aerodynamic diameters in mass mode, and PM ≤ 0.5 and PM ≤ 10μm in count mode. The instrument uses laser technology and uses a right angle scattering method at 0.78μm, which is different from gravimetric measurements. The source light travels at a right angle to the collection system and detector, and the instrument uses the information from the scattered particles to calculate a mass per unit volume. A mean particle diameter is calculated for each of the 5 different sizes. This mean particle diameter is used to calculate a volume (cubic meters), which is then multiplied by the number of particles and then a generic density (μgm^{-3}) that is a conglomeration of typical aerosols. The resulting mass is divided by the volume of air sampled for a mass per unit volume measurement (μgm^{-3}).

This instrument also recorded relative humidity (RH) and temperature with every sample. The main flaw of the instrument is that the mass values can be easily inflated with the

increase in RH, especially when it is > 40% (Thomas and Gebhart, 1994). A standard relationship between photometric and gravimetric measurements as discussed by Ramachandran et al. (2003) can be used to calibrate the data for relative humidity. Data from this instrument were compared against that from gravimetric samplers in order to evaluate the robustness of photometric samplers (Kumar, 2006).

4.2 Stratification by air pollution levels and their source identification

On an average we had more than 65 samples at each of the 113 sites, which represent sampling at different times of a day and different days of a week. Each sample included two readings (four minutes of sampling). Average estimate of $PM_{2.5}$ and PM_{10} across six months were computed for all 113 sites. These estimates were used to interpolate spatially detailed air pollution surfaces in ArcGIS 9.x (ESRI, 2005) with the aid of Kriging method, which estimates air pollution at given location as an inverse function of distance weighted by spatial autocorrelation among the sample sites (Cressie, 1990; Isaaks and Srivastava, 1989). The averages of $PM_{2.5}$ and PM_{10} in the study area were $35.99 \pm 2.25 \mu gm^{-3}$ and $194.15 \pm 22.57 \mu gm^{-3}$ (95% CI), respectively. The average values of both fine and coarse particles were several folds greater than the recommended standards. Figure 2a and 2b shows substantial spatial variability in ambient air pollution in Delhi and its surroundings. Using the PM_{10} surface, the study area was partitioned into three strata – less than 150[1], 150 to 250 and ≥250m (Figure 3).

Each of the three strata was cross classified by two main sources of air pollution – proximity to roads and industrial clusters. The road network and industrial clusters data were digitized from topographic sheets and Eicher street maps (EICHER, 2001). It is evident from the literature that the concentration of air pollution, especially that of coarse particles, decrease exponentially with increase in distance from roads (Violante et al., 2006). Therefore, short distances from road can greatly impact exposure to air pollution from traffic from roads. The proximity to major roads was partitioned into four categories – ≤250, 250-500, 500-1000 and ≥ 1000m and proximity to minor roads into two categories only ≤ 250 and > 250m, which were equated with the last two categories of proximity to the major road because of the less frequency of vehicles on the minor roads. In total, we have four categories of proximity to roads (Figure 4).

Air pollution from industries can have impact up to greater distances, industries are one of major sources of fine particles in the study area (Kumar, 2006), which can stay aloft longer distances. Therefore, categories of proximity to industrial clusters were spanned over longer distances. A total of five categories were identified - namely ≤ 1, 1-2, 2-3, 3-4 and ≤ 4km, resulting in five strata (Figure 5).

4.3 Constructing a sampling frame of residential areas: In the year 2000, the Census of India prepared a household list for the 2001 census enumeration. This list can serve as a household sampling frame, but in the absence of household location identifiers, the list

[1] Overall PM_{10} concentration in the study area was substantially higher than the EPA standards in the US that are (a) a 24-hour standard = $150 \mu g/m^3$, and (b) an annual 24-hour standard = $50 \mu g/m^3$. (EPA, 2005. http://www.epa.gov/ttn/oarpg/naaqsfin/pmfact.html)

was not adequate to capture spatial variability in air pollution levels and proximity to air pollution sources. Therefore, a sampling frame of residential areas was constructed for sampling households.

With the aid of satellite RS and GIS technologies different types of land use and land cover were identified. A strategy of stepwise exclusion was adopted to generate the frame of residential areas. In the first step, vegetation canopy cover (R_v) was extracted using the normalized difference vegetation index (NDVI) derived from the 2000 multi-spectral (LISS-III) Indian Remote Sensing (IRS-1D) satellite imagery (23.5m spatial resolution) and subtracted from the entire study area (R), which resulted in non-vegetated areas $|R-R_v|$. The LISS-III multi-spectral imageries were resampled at a higher spatial resolution in order to blend them with the panchromatic imagery (6.85m spatial resolution) for constructing signatures of different land use types, and a method of supervised classification that relied on these signatures was used to extract water bodies (R_w), roads (R_{rd}), barren land (R_b) which were then subtracted from $|R-R_v|$ to extract built-up areas (R_{bl}).

Built up area (R_{bl}) consists of different land-use types, including residential and industrial areas and roads. Extracting residential area alone from satellite remote imageries is a challenging task, due to the fact that there are only insignificant differences between the spectral signature of industrial and residential areas. Thus, the maps of industrial clusters, official building and 50m buffer of streets/roads (digitized from Eicher map of Delhi) were merged and subtracted from R_{bl} resulting in residential areas R_d (Figure 6). Subsequently, we could partition each stratum into the components R_d and R-R_d. If an entire region, R, is partitioned into k strata then $\bigcup_{i=1}^{k} R_d$ can serve as the sampling frame for implementing the suggested methodology. A combination of the three strata of air pollution, four strata of proximity to roads and five strata of industrial clusters resulted in a total of 60 categories of residential areas (Figure 7).

4.3 Weight assignment: A good spatial sampling design recognizes the spatial inhomogeneities in the process and accounts for them when choosing a set of sample locations (Cressie 1993). Since human population is not distributed uniformly across a given area, it is necessary that a differential weighting scheme is used in simulating random household locations. Generally, in a stratified random sampling the sample is selected in proportion to strata size, and for sampling human population it is reasonable to weight the sample by the size of population of each stratum. In the proposed sampling design, however, the sample was weighted by the area of each stratum, because the goal of our survey was to assess exposure to air pollution and its effect on respiratory health. Therefore, weighting sample by area under different air pollution strata was more appropriate than the number of individuals in the strata. Moreover, population data were not available to match the spatial resolution of air pollution strata.

4.4 Random point simulation: Simulating a random point in a continuous Cartesian space requires generating a pair of pseudo random numbers and then plotting each

against x-axis and y-axis (generally, both range from 0-1). Simulating geographically weighted random points in a discrete geographic space is somewhat different, and is implemented in three steps: first, simulating a pair of pseudo random numbers and then projecting them into the coordinate system of the background layer (the identified strata in our context); second, identifying the container of the simulated location using the projected X-Y coordinates; and finally, placing or rejecting a point location after checking the proportion of random points allocated to the identified container (Figure 8).

The allocation of random points in strata can be complicated when the strata are not continuous and spread across many polygons. In such cases, the number of points assigned to a container (polygons of residential areas in our case) may not be an integer value and many smaller strata can be assigned a small fraction (such as 0.2, 0.6). Theoretically, the allocation of a point to a polygon cannot be counted as a fractional value; it needs to be counted as one. Therefore, the allocation of integer part and the decimal part should be handled separately. In the proposed methods, the integer part was assigned and simulated first and then decimal part was handled. For the decimal parts a threshold value is chosen based on the number of units with fractional values across the candidate containers, and points are assigned randomly across the candidate containers that qualify the threshold value. The simulation stops when the number of simulated points reaches the required sample size even though some candidate polygons can be left without any sample point. A customized application was designed in ArcGIS 9.x to implement the proposed sampling method (Figure 9). The application requires a background layer (or shape) with the weight attribute, and if no weight is assigned area is assumed as the default weight.

4.5 Surveying households – experiences: The first step in administering the survey was to identify households at the simulated locations and acquire their consents to participate in the survey. Household locations were identified with the aid of street maps (Figure 5) and global positioning systems (GPS). The simulated point locations were transferred onto the street map (Figure 10), which served both to verify that all simulated points were in the residential areas, and to identify the neighborhood (locations) of the simulated points for navigation purposes. The simulated locations were also loaded onto GPS units with point identifiers. To navigate to a location, first we traveled to the neighborhood in which the point was located, and then navigated to the point location with the aid of the GPS, which has a spatial resolution of less than 5m. Once the household location was identified, consent was acquired for the final survey. While navigating the simulated locations, two main problems were encountered. First, 11% of the simulated locations were placed midway between two or more households. Second, about 2% of points were placed onto multi-story household complexes. For both problems, we prepared a list of all probable candidate households at the spot, and one was picked up randomly using a lottery system. From a list of 2000 simulated random points, only about 1576 turned out to be viable for the household survey (Figure 11), and the rest either did not consent to participate in the survey or were placed at inaccessible locations. The survey was administered in the identified households from January to April 2004.

4.6 Effectiveness of the selected ample: To evaluate the effectiveness of our sample, the average estimate of air pollution at the household location was compared with that reported in the study area.

$PM_{2.5}$ and PM_{10} were interpolated at the sampled (household) locations using Kriging with optimal parameters that resulted in the least difference in the observed and predicted values. The mean of $PM_{2.5}$ and PM_{10} were estimated as 35.25±0.21 μgm^{-3} and 195.54±2.04μgm^{-3}, respectively. The mean value of $PM_{2.5}$ and PM_{10} at the selected household locations were not significantly different from that observed in the study area (for $PM_{2.5}$ $|\bar{x} - \mu| = 0.250$ $p \neq t = 0.61$ and for PM_{10} $|\bar{x} - \mu| = 1.38$ ($p \neq t = 0.61$)), which leads us to conclude that the ambient air pollution at the selected household locations adequately represents the exposure to ambient air pollution at the household locations in the study area.

5. DISCUSSION AND CONCLUSION

Sample sites for collecting spatial data must be chosen with care, for locational-choices can affect the quality of the results from statistical analysis (Müller 1998). Most conventional methods adopted for sampling natural phenomena, such as plant species, mineral ore or fossil fuel, rely on selecting *n* units from a set of *N* derived by partitioning the entire geographic extent into *N* shapes of some regular geometric structure, such as hexagon, rectangle or square. In other words, a regular grid of the chosen geometry is overlaid onto the study area to identify *N* units. Assuming contiguity in natural phenomena, say soil or air pollution or air pressure, the use of a regular grid seems reasonable, but most human activities and their geographic distribution are neither contiguous nor distributed uniformly. As a result, regular spatial sampling methods cannot be extended to sample human population spread across discrete geographic spaces with varying density.

Advances in GIS and RS technology have simplified the collection and analysis of spatial data (Longley et al., 1999). These technologies are equipped with tools that enable us to integrate data from different sources and different geographic scales. This article has demonstrated the application of these technologies for constructing a sampling frame of discontinuous residential areas and the selection of households with references to air pollution levels and proximity to the sources of air pollution. Using a customized application designed in ArcGIS (9.x) (ESRI, 2005) random points weighted by the residential area were simulated. Finally, these points were translated into household locations with the aid of street maps and GPS technology.

The suggested methodology has several advantages. First, it can be employed to develop a sample even when a sampling frame is not available, particularly in developing countries where it is almost impossible to acquire a relevant sampling frame. Second, the methodology can play a vital role in collecting sample data for research that incorporates place/space as an important context of social-behavioral processes. Telephone interviews that rely on random digit dialing do not necessarily provide a reliable locational context due to widespread usage of mobile/cell phones in recent years. The use of the proposed

methodology, however, can ensure a reliable locational context. Third, the availability of spatial context(s) of household locations can provide an opportunity for the analysis and modeling of spatial-dependency and causality in social, behavioral and health outcomes; e.g., ambient air pollution at household location or in the neighborhood can serve as a proxy of exposure and can be associated with respiratory health.

Sampling of households for survey based research has begun to draw attention in recent years (Lee et al., 2006) and its usage is likely to increase in the near future with the increasing importance of place and space in social science and public health research, particularly given the usefulness of understanding – (a) how spatial variations in exposure to environmental contaminants affect health outcomes, and (b) how local/neighborhood contexts shapes the spatial patterns of social, economic, demographic and behavioral outcomes.

References

Ackers, M. et al., 2005. Gender, age, and ethnicity in HIV vaccine-related research and clinical trials - Report from a WHO-UNAIDS consultation, 26-28 August 2004. Lausanne, Switzerland. Aids, 19(17): W7-W28.

Amblard-Gross, G., Maul, A., Ferard, J.F., Carrot, F. and Ayrault, S., 2004. Spatial variability of sampling: Grid size impact on atmospheric metals and trace elements deposition mapping with mosses. Journal of Atmospheric Chemistry, 49(1-3): 39-52.

Angulo, J.M., Ruiz-Medina, M.D., Alonso, F.J. and Bueso, M.C., 2005. Generalized approaches to spatial sampling design. Environmetrics, 16(5): 523-534.

Arbia, G. and Lafratta, G., 2002. Anisotropic spatial sampling designs for urban pollution. Journal of the Royal Statistical Society Series C-Applied Statistics, 51: 223-234.

Cerin, E., Saelens, B.E., Sallis, J.F. and Frank, L.D., 2006. Neighborhood Environment Walkability Scale: validity and development of a short form. Med Sci Sports Exerc, 38(9): 1682-91.

Cressie, N., 1990. The Origins of Kriging. Mathematical Geology, 2(3): 239-52.

Dessard, H. and Bar-Hen, A., 2005. Experimental design for spatial sampling applied to the study of tropical forest regeneration. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere, 35(5): 1149-1155.

Di Zio, S., Fontanella, L. and Ippoliti, L., 2004. Optimal spatial sampling schemes for environmental surveys. Environmental and Ecological Statistics, 11(4): 397-414.

Dumitrescu, A., Granados, A., Wallace, J. and Watson, S., 2006. Demand-driven evidence network in Europe. Bulletin of the World Health Organization, 84(1): 2-2.

EICHER, 2001. Delhi: City Map. Eicher Goodearth Ltd., New Delhi.

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

Fleishman, E., Mac Nally, R. and Murphy, D.D., 2005. Relationships among non-native plants, diversity of plants and butterflies, and adequacy of spatial sampling. Biological Journal of the Linnean Society, 85(2): 157-166.
Getis, A. et al., 2000. Geographic Information Science and Crime Analysis. URISA Journal, 12(2): 7-14.
Goodchild, M.F. and Janelle, D.G. (Editors), 2004. Spatially Integrated Social Science. Oxford University Press, Oxford.
Gordon-Larsen, P., Nelson, M.C., Page, P. and Popkin, B.M., 2006. Inequality in the built environment underlies key health disparities in physical activity and obesity. Pediatrics, 117(2): 417-24.
Isaaks, E.H. and Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York.
Kumar, N., 2006. Air Quality Regulation and Spatial Dimension of Air Pollution in Delhi and Its Surroundings. Economic and Political Weekly.
Lafratta, G., 2006. Efficiency evaluation of MEV spatial sampling strategies: a scenario analysis. Computational Statistics & Data Analysis, 50(3): 878-890.
Lee, C., Moudon, A.V. and Courbois, J.Y.P., 2006. Built environment and behavior: Spatial sampling using parcel data. Annals of Epidemiology, 16(5): 387-394.
Lewis, P.A.W. and Shedler, G.S., 1979. Simulating non-homogeneous Poisson processes by thinning. Naval Research Logistics Quarterly, 26: 403-13.
Liao, D.P. et al., 2006. GIS approaches for the estimation of residential-level ambient PM concentrations. Environmental Health Perspectives, 114(9): 1374-1380.
Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W., 1999. Geographical Information Systems. John Wiley & Sons, New York.
McBratney, A.B. and Webster, R., 1981. Detection of ride and furrow pattern by spectral analysis of crop yield. International Statistics Review, 49: 45-52.
Met One Inc, 2003. AEROCET 531: Operation Manual, Grants Pass: OR.
Müller, W.G., 1998. Collecting Spatial Data: Optimum Design of Experiments for Random Fields. Physica-Verlag, New York.
Nychka, D. and Saltzman, N., 1998. Design of air-Quality monitoring Networks. In: D. Nychka, W. Piegorsch and L. Cox (Editors), Case Studies in Environmental Statistics. Springer-Verlag, New York.
Ramachandran, G., Adgate, J.L., Pratt, G.C. and Sexton, K., 2003. Characterizing Indoor and Outdoor 15 Minute Average PM2.5 Concentrations in Urban Neighborhoods. Aerosol Science and Technology, 37: 33-45.
Salehi, M., 2004. Optimal sampling design under a spatial correlation model. Journal of Statistical Planning and Inference, 118(1-2): 9-18.
Schabenberger, O. and Gotway, C.A., 2005. Statistical Methods for Spatial DAta Analysis. Texts in Statistical Science. Chapman & Hall/CRC.
Schwartz, J., 1996. Air pollution and hospital admissions for respiratory disease. Epidemiology, 7(1): 20-8.
Stevens, D.L. and Olsen, A.R., 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association, 99(465): 262-278.
Thomas, A. and Gebhart, J., 1994. Correlations between gravimetry and light-scattering photometry for atmospheric aerosols. Atmospheric Environment, 28(5): 935–938.

Thompson, S.K., 2002. Sampling: Wiley Series in Probability and Statistics. John Wiley and Sons, London.
Violante, F.S. et al., 2006. Urban atmospheric pollution: personal exposure versus fixed monitoring station measurements. Chemosphere, 64(10): 1722-9.
WHO, 2000. Guidelines for air quality, World Health Organization, Geneva.
Zhu, Z.Y. and Stein, M.L., 2006. Spatial sampling design for prediction with estimated parameters. Journal of Agricultural Biological and Environmental Statistics, 11(1): 24-44.
Zhu, Z.Y. and Zhang, H., 2006. Spatial sampling design under the infill asymptotic framework. Environmetrics, 17(4): 323-337.

Manuscripts with figures can be downloaded from the link below.

http://jh302-nk-01.iowa.uiowa.edu/papers/NK_PrPr_March_2007WithFigures.pdf