Optimal spatial sampling design for the imputation of missing floor area data in the Southern California Association of Governments database

Xiaohong Che[1] and Richard Arnott[2]

[1]Department of Statistics, University of California, Riverside

[2] Department of Economics, University of California, Riverside

February 26, 2012

Contents

1 Introduction

Our overall aim in this part of the project is to estimate the aggregate floor area by land use and model zone for the Greater Los Angeles Metropolitan Area. The database was compiled by the Southern California Association of Governments from the assessment records provided by six counties (Imperial, Los Angeles, Orange, Riverside, San Bernardino, and Ventura), and contains records for some 5 million parcels. Each assessment record for a non-vacant parcel is supposed to contain the floor area of the structure(s) on the parcel. Unfortunately, this datum is recorded as either zero or blank for over one million parcels. The incidence of zeroes and blanks differs across land use and county. Collecting floor-area data on over one million parcels would be prohibitively expensive. Instead, floor area data will be collected for only a sample of parcels, with the floor area of the other parcels being imputed on the basis of the sample data.

The aim of this technical report is to propose a strategy for sampling “zero” and “blank” parcels with the object of imputing floor areas to all zero or blank parcels. Since some of the parcels are vacant, this entails imputing the status of “developed” or “vacant” to each of the “zero” or “blank” parcels, and then imputing a floor area to those parcels imputed as being developed. A sample datum is a satellite photograph, from which both development status and the length and width of buildings can be estimated, perhaps supplemented by a Google streetview photograph, from which the height of the buildings can be estimated. The strategy aims to maximize the accuracy of the imputed floor areas, subject to a time/budget constraint.

The UCSB team previously did some ground truth tracking of the floor areas of zero and blank parcels in Riverside and Orange Counties. A sample datum is a satellite photograph of the parcel, from which development status can be ascertained. They used a random sampling method, with the more limited objective of finding out which of the parcels are developed. On the basis of the sample, Gu and Arnott (2011) statistically imputed “developed” or “vacant” status to all the zero and blank parcels in the Los Angeles area (assuming that the statistical relationship estimated for Riverside County holds too in San Bernardino County, and that the statistical relationship estimated for Orange County holds too in Los Angeles and Ventura Counties).

2 Selective review of the literature

2.1 General principle of sampling design

Generally in statistics, one basic goal is to get the precise estimates for the parameters. Sampling design is not an exception either. Every sample design is associated with the cost of the sampling/survey and the precision (measured in terms of variance of the sampling parameters). The designs should be practical in the sense that it is possible to carry it through according to desired specifications. Out of all these designs, the one to be preferred is that which gives the highest precision for a given cost of the survey or the minimum cost for a specified level of precision. Statistically, high precision to a parameter means small variance to the parameter. So one major task in sampling design is to minimize the variance of the parameter under a given cost of survey/sampling (see, for example, Diggle and Lophaven (2006); Rao (1979); Smith (1988)).

The aim of a sample survey is to estimate the unknown population parameters like total/aggregate, ratio, median or mean based on a random sample drawn by some specified rules from the given population. In a sampling design, one pursues reduction in cost, greater speed, wider scope, higher accuracy, and the quantification of the uncertainty, i.e., the error. Therefore, sometimes in a sampling design, one tries to maximize the sample variance and minimize the sample size, etc., e.g. in the spatial sampling design (Kumar, 2009). These are some other principles in sampling designs.

2.2 Sampling design with missing data

Generally, a sampling scheme includes the following several aspects:

1. Identify the target area, variables in the design, and the accuracy required;

2. Consider the constraints from the following concerns: financial, logistical, operational etc.;
3. Set up the sampling design which includes determining number of samples etc.;
4. Decide method of taking samples including the following aspects: optimal time, size, devices etc.;
5. Write the protocols for data recording and fieldwork;
6. Propose the methods of statistical analysis.

For missing data problem in sampling design, the modification usually exists in the last step – methods of statistical analysis. In the statistical analysis of sampling design with missing data (the missing data appears in the sample that one gets from the sampling), one approach to handle such nonresponse is imputation of the missing data (Gao and Hui, 2000; Reiter et al., 2006; Rubin, 1987).

2.3 Spatial sampling design (with presence of spatial autocorrelation)

Spatial sampling design is one special design in sampling studies. Spatial sampling involves determining a limited number of locations in geographic space for faithfully measuring phenomena etc. Basic spatial sampling schemes include random, clustered and systematic which are the classic probability sampling methods. These basic schemes can be applied at multiple levels in a designated spatial hierarchy. It is also possible to exploit ancillary data, for example, using property values as a guide in a spatial sampling scheme to measure educational attainment and income. Spatial models such as autocorrelation statistics, regression and interpolation could also dictate sample design. Kumar (2009) presented two principles related to spatial sampling design – maximizing variance and minimizing sample size. Our project involves spatial autocorrelation, therefore we will describe how the spatial sampling design can be modified by the presence of spatial autocorrelation with an example from Kumar (2009).

The design of Kumar (2009) is a model based spatial sampling design. Model based design (Hansen et al., 1983) consists of an evaluated quantity or an objective function (e.g. sample variance) under an assumed population model, and a sampling plan that each sample is selected with the probability to optimize the quantity or the objective function (e.g. maximizing the sample variance).

In Kumar's method, preliminary estimates of spatial variance need to be calculated to identify sample locations. The local semivariance [1] is calculated as below

$$\hat{r}_i = \frac{1}{k} \sum_{j=1}^{k} \underset{i \neq j, d_{ij} \leq h}{(z_i - z_j)^2} \tag{1}$$

where k is the number of neighbors around ith candidate within distance range h, z_i is the variable of interest at the location i, it can be air pollution concentration, or whatever we are interested in, d_{ij} is the distance from location i to location j, and it is bounded by some distance range h when one calculates the local semivariance.

After the preliminary estimation of variance, Kumar (2009) tries to determine the sample size which requires the estimates of variance σ_z^2. Without considering the spatial autocorrelation, the variance σ_z^2 may be overestimated. After Kumar (2009) takes the spatial autocorrelation into consideration, the formula of σ_z^2 is revised as:

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

where $\forall_{ij} = 0$ if $d_{ij} < h$. Therefore, we can see that the modification exists in the formulas during the procedures of the spatial sampling designs when one considers spatial autocorrelations. Kumar (2009) further proposed an optimal network method to identify the optimal locations as the final step in their spatial sampling design where the spatial autocorrelation also influenced the design. We will not state the details here.

[1] Semivariance is a widely used concept in spatial statistics. The total semivariance over the entire study (without the control for spatial autocorrelation) is defined as $r = \sum_{k=1}^{K} \sum_{l \neq k} (z_k - z_l)^2$. Let $\sigma^2 = \sum_{k=1}^{K} (z_k - \bar{z})^2/(K-1)$ denote the sample variance. It can be shown that the average semivariance and the sample variance are about equal, i.e., maximizing semivariance of Z by selecting n sample locations, we also maximize sample variance σ^2.

2.4 Statistical packages for sample design procedures

There have been numerous packages/softwares to implement sampling design procedures and post analysis, such as SAS, R, SPSS, STATA Minitab. Lohr (1999) gave a short review on the sampling packages/softwares. The website www.hcp.med.harvard.edu/statistics/survey-soft/ provides an up-to-date overview of these programs.

In statistics, the major packages/softwares used in sampling designs are SAS and R. SAS mainly provides five procedures to do the sampling design – PROC SURVEYSELECT, PROC SURVEYMEANS, PROC SURVEYREG, PROC SURVEYLOGISTIC, and PROC SURVEYFREQ. In R, there are two major packages for sampling design – "survey" and "sampling". Both these two packages have some build-in functions to do the survey/sampling designs. For example, one can use "balancedcluster" to select a balanced cluster sample with the package "sampling". With the package "Survey", one can use function "twophase" to design a two-phase sampling.

3 Two proposed methods

3.1 Descriptions of the two methods

3.1.1 Two-step adaptive spatial sampling design

We propose two sampling design here. The first one is a two-step sampling design. The sampling frame is all the parcels with zero/blank floor area which we are interested in. Below is a detailed description of this design.

In the first step, we will determine the sample size n with buget constraint B by $n \times t \leq B$, where t is the time for sampling one parcel. The second step is an adaptive spatial sampling design which is revised from Cox Jr (1999). This method does not require any statistical model for the spatial distribution, instead it constructs an increasingly nonparametric approximation to it as sampling proceeds. So, it is not a model-based sampling design. The procedures of the second step are listed as below:

1. Sample n^* previously unsampled locations in the area of concern. The method to sample

the n^* samples may be a proportionate stratified method.The whole database are stratified with respect to the land use type.Based on the total number of parcels in each stratum, the sample size from each stratum can be calculated in the proportionate stratified sampling design, i.e., after we determine the sample size n^*(n^* is not so big a number, and it is definitely less than the total sample size n), for a stratum with population size N_h, the sample size n_h^* within the stratum will be calculated by $n_h^* = \frac{N_h}{N} \times n^*$. And the method to determine the specific sampling parcels can be a simple random sampling (SRS).

2. Place on a search list the k parcels with the zero/blank sample floor area observed so far. The k parcels comprise the search list.

3. Measure the floor area from each of its previously unsampled immediate neighbors in the sampling frame (Neighbors can be the parcels which are closest to a parcel from the east, west, south and north directions). Remove each parcel from the list when it has been measured.

4. Place onto the search list any parcel with zero/blank floor area from its immediate neighbors which are defined and sampled in step 3.

5. Continue steps 3 and 4 until the search list is empty and the number of sampled parcels reaches predetermined total sample size n.

3.1.2 Bayesian spatial sampling design

The second sampling design is a Bayesian sampling design. The procedures are as follow:

1. Assume that the spatial distribution of floor area follows a parametric model, say $(x, y; q)$, where q is a vector of parameters. Assume a prior probability density function for q, say, $f_q(q)$.

2. Obtain any set of observed sample values for floor area at different locations. Here sample size is n^*. And the method to specify the n^* parcels can be a proportionate stratified sampling design as described in the first sub-step of the second step in the two-step adaptive

sampling design. After this, update beliefs about q by conditioning on the observed values using Bayes' rule. (This requires knowing or assuming a likelihood function relating q to the observed values.)

3. Choose the terminal decision, a^*, to maximize the expected value of a utility function $u(c, b)$, where the expectation is taken with respect to the posterior (conditioned on observed data) probability distribution for q.

4. Recursively choose the sampling locations and stopping rule via backward dynamic programming to maximize expected utility.

5. Stop when the total number of sampled parcels reaches the total sample size n obtained in the first step of the method in section 3.1.1.

Here the utility function is a bounded function, used in economic studies. Some examples of the utility functions are $u(x) = 1 - exp(-x/r)$; $u(x, y) = ax + by, \text{where} \quad a, b > 0$; $u(x, y) = x^a y^b, \text{where} \quad a, b > 0$ etc. A utility function has some properties such as monotonicity, convexity.

3.2 Pros and cons of each

Here two sampling designs are proposed. One is a nonparametric adaptive procedure. It is not a model-based sampling design. The other is a Bayesian one which is a model-based approach. It has been shown that model-based approach is the best choice (Brus and De Gruijter, 1997) when

1. We want to map the target structure.

2. Sample size large enough for calibrating a model of variation.

3. Strong autocorrelation exists from which we may profit in mapping, etc..

Our problem has the above characters. So, model-based sampling design is quite appropriate here. Adaptive method is not a model-based method and is less mathematical than the Bayesian method. But it is easier to implement. Bayesian method is more scientific.

3.3 Recommendation

The two-phase adaptive sampling design is advocated here since it is easier to understand and implement.

3.4 Imputation of the missing value

Having undertaken both phases of the adaptive spatial sampling design, one needs to proceed with statistical inference and imputation. We have been asked to provide standard deviation of our estimates, if possible. The contribution to statistical imputation mainly goes to Robin (1987). Schafer (1997) proposed an imputation scheme as follow:

Denote the observed-data posterior as $P(\theta|Y_{obs})$, here, Y_{obs} are the observed data. Denote $P(\theta|Y_{obs}, Y_{mis})$ as the complete-data posterior, here Y_{mis} are the missing data.

1. Given a current guess $\theta^{(t)}$ of the parameter, first draw a value of the missing data from the conditional predictive distribution of Y_{mis},

$$Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta(t)) \tag{3}$$

2. Conditioning on $Y_{mis}^{(t+1)}$, draw a new value of θ from its complete-data posterior,

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t+1)}) \tag{4}$$

3. Repeating (3)-(4) from a starting value $\theta^{(0)}$ yields a stochastic sequence $\{(\theta^{(t)}, Y_{mis}^{(t)}) : t = 1, 2, ...\}$ whose stationary distribution is $P(\theta|Y_{mis}|Y_{obs})$, and the subsequences $\{\theta^{(t)} : t = 1, 2, ...\}$ and $\{Y_{mis}^{(t)} := 1, 2, ...\}$ have $P(\theta|Y_{obs})$ and $P(Y_{mis}|Y_{obs})$ as their misrespective stationary distribution.

The above is a typical statistical imputation procedure. In that book, they provided an example for this method in which the formulas for mean and variance were presented. And I am presenting the method as follow:

Suppose $Y = (y_1, ..., y_n)$ is an iid sample from a normal distribution with mean μ and variance ψ. Further suppose that only the first n_1 elements of Y are observed and the remaining $n_0 = n - n_1$ are missing. Use the iterating procedure, and after some derivation, we can get the expectation of the tth iteration for the mean μ as follow:

$$E(\mu^{(t)}) = \bar{y}_{obs} \tag{5}$$

and the variance of the mean:

$$V(\mu^{(t)}) = n_1^{-1}\psi \tag{6}$$

Further estimator for ψ can be obtained using some data augmentation method Schafer (1997). Therefore we can estimate the variance and stardard deviation for μ.

This is a parametric method which has the distributional assumption on the responses which might not be that feasible. But it gives a general idea on the statistical imputation.Without the assumption of the distribution on the responses, there exists a non-parametric method (Ning and Cheng 2010) as below:

In Ning and Cheng's (2010) method, they denoted the data as $(X_i, Y_i, \delta_i), i = 1, 2, ..., n$. The covariates X_i are observed, and $\delta_i = 1$ if Y_i is observed, otherwise $\delta_i = 0$. A nearest neighbor (NN) estimator is defined as:

$$\mu_{NN} = \frac{1}{n} \sum_{i=1} n\{\delta_i Y_i + (1-\delta_i) m_K(X_i)\} \tag{7}$$

Here $m_K(X_i) = \frac{1}{K}\sum_{j=1} KY_{i(j)}$, and $\{(X_{i(j)}, Y_{i(j)}) : \delta_{i(j)} = 1, j = 1, ..., K\}$ is a set of K observed data pairs, and $X_{i(j)}$ denote the jth nearest neighbor to X_i among all the covariates $X's$ corresponding to those $Y_k^{'}s$ with $\delta_k = 1$.

An asymptotic variance from Ning and Cheng (2010) is as follow:

$$\sigma^2(\mu_{NN}) = Var(Y) + (1 + \frac{1}{K})E[\sigma^2(X)(1 - P(X))] + E[\frac{\sigma^2(X)(1-P(X))^2}{P(X)}] \tag{8}$$

where $P(X) = P(\delta = 1|X, Y) = P(\delta = 1|X)$. This estimator for variance can be used to claculate the statistics such as statndard deviation in our project.

References

Brus, D. and De Gruijter, J. (1997). Random sampling or geostatistical modelling? choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, 80(1-2):1–44.

Cox Jr, L. (1999). Adaptive spatial sampling of contaminated soil. *Risk Analysis*, 19(6):1059–1069.

Diggle, P. and Lophaven, S. (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics*, 33(1):53–64.

Gao, S. and Hui, S. (2000). Estimating the incidence of dementia from two-phase sampling with non-ignorable missing data. *Statistics in medicine*, 19(11-12):1545–1554.

Goodchild, M. F., Li, W., Schild, A., and Royal, N. (2010). Scag parcel database validation report on accuracy of total floor space per parcel by ground truth trekking. Technical report.

Gu, Y. and Arnott, R. (2011). Floor area data adjustment for the parcel database of greater los angeles region. Technical report.

Hansen, M., Madow, W., and Tepping, B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, pages 776–793.

Kumar, N. (2009). An optimal spatial sampling design for intra-urban population exposure assessment. *Atmospheric Environment*, 43(5):1153–1155.

Lohr, S. (1999). *Sampling: design and analysis*. Thomson.

Ning, J. and Cheng, P. (2010). A comparison study of nonparametric imputation methods. *Statistics and Computing*, pages 1–13.

Rao, J. (1979). Optimization in the design of samples surveys. *Optimizing Methods in Statistics*, pages 419–434.

Reiter, J., Raghunathan, T., and Kinney, S. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32(2):143–149.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* John Willey & Sons.

Schafer, J. (1997). *Analysis of incomplete multivariate data*, volume 72. Chapman & Hall/CRC.

Smith, P. (1988). Survey design optimization for estimating the exploitable biomass of a fishery accounting for non-sampling errors. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 37(3):370–384.