Model-Based Optimal Selection of Sampling Units with Computational Ease

Chao, Chang-Tai
Department of Statistics,National Cheng Kung University,Taiwan 70101
ctchao@email.stat.ncku.edu.tw

Key Words: Model-based Sampling, Optimal Sampling Strategy,Eigensystem, Gaussian Spatial Model, Log-Gaussian Spatial Model.

1 Introduction

For the prediction problem in model-based survey sampling, an *optimal sampling strategy* consists of an unbiased predictor $\hat{T}(d)$ and a sampling design that can select the sample s to minimized the conditional mean-square prediction error given s (Thompson and Seber 1996). Sacks and Schiller (1988) proposed a modified annealing algorithm to select the optimal *conventional* sampling sites under a given population model when the best unbiased estimator is used. Since the optimal sampling strategy is actually one with a n-stage adaptive sampling design (Zacks 1969), Chao and Thompson (2001) proposed an simplified two-stage optimal adaptive sampling strategy when the values of the parameters of the population model are given. Chao (2003) also proposed a two-stage adaptive optimal sampling strategy under a Bayesian population model with a given prior. Although the optimality of the optimal sampling strategies have been illustrated in both simulation studies and a real data set (e.g Chao and Thompson 2001, Chao 2003), the computation of the optimization algorithms used are usually very computationally intensive. Also these optimal strategies require exact population model. Therefore, the practical usage of these strategies is restricted (Chao 2003).

In practice, a simple sampling strategy that can provide more efficient predictor with an easy sampling procedure, less population assumption and affordable computation load is desired. In this research, a convenient method to select sampling units with a fixed sample size under a given population covariance structure, but not the exact population distribution is proposed. The sampling units are selected based on the eigensystem of the population covariance matrix. The intuition and algorithm of the proposed sampling design are introduced in Section 2. Instead of the full knowledge of the population model, only the covariance structure is required to select the sampling units by a simple algorithm. In Section 3, the performance of the proposed method is examined in terms its relative efficiency to Simple Random Sampling (SRS). Results show that the proposed sampling scheme is usually better than SRS under a moderate correlated population. The proposed sampling procedure can also be extended to the optimal adaptive two-stage sampling strategy with slight modification. Application and further research are discussed in Section 4

Support for this research was provided by the National Science Council, grant NSC 90-2118-M-006-007-.

2 Method

To select sampling units that can give lower mean-square prediction error, the units that have better prediction ability to other unselected units or higher variance themselves are preferred. In other words, one would like to select the units that account for as much total population variability as possible. Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be the ordered eigenvalues of $\mathbf{\Sigma}$,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N, \tag{1}$$

and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N$ be the associated normalized eigenvectors. Then the original N-dimensional coordinate system can be rotated into a new N-dimensional orthogonal coordinate system, in which the N axes are the linear combinations of the original variables, such that the coefficients of the i_{th} linear combination, denoted as X_i, $i = 1, \ldots, N$, are the components of the i_{th} eigenvectors $\mathbf{e}_i$. That is

$$X_i = \mathbf{e}_i'\mathbf{Y} = e_{i1}Y_1 + e_{i2}Y_2 + \ldots + e_{iN}Y_N,$$

where e_{ij} is the j_{th} component in the i_{th} eigenvectors. X_i is also known as the i_{th} principal component in Principal Component Analysis (PCA). The original covariance structure can then be explained by X_i's. The variability in $\mathbf{Y}$ is extracted into the variances of uncorrelated random variables, X_i's, and (e.g. Anderson 1984)

$$\sum_{i=1}^{N} \mathrm{Var}(X_i) = \sum_{i=1}^{N} \mathrm{Var}(Y_i).$$

In addition, the variance of X_i is

$$\mathrm{Var}(X_i) = \lambda_i, \forall i = 1, \ldots, N$$

Hence, if one would like to select the units that can account for more variability in $\mathbf{Y}$, then the unit that is associated with component having a large absolute value in the leading eigenvectors are reasonable candidates. Based on this intuition, we propose the following sampling design to select

$$s = \{i_1, i_2, \ldots, i_n\}, i_j \in \{1, 2, \ldots, N\}, i_j \neq i_{j'}, \forall j \neq j'$$

with a fixed sample size n.

Both the sign and magnitude of e_{ij} provide information regarding the role corresponding unit j plays in X_i. In the following proposed design, the sampling units are selected depending not only the magnitude, but also the sign of their corresponding components in the leading eigenvectors. The sampling units are selected by the following algorithm:

$n = 1$ **:** $s = \{j\}$, $|e_{1j}| = \max_i |e_{1i}|$.

$n > 1$ **: Step 1:** Let $s' = \{j_1, j_2, \ldots, j_m\}$, $m < N$, where

$$|e_{1j_1}| \geq |e_{1j_2}| \geq \ldots \geq |e_{1j_m}| \geq \ldots \geq |e_{1j_N}|$$

and m is an integer that indicates the number of units in s'. m can be appropriately specified before the survey according to the population size N.

$\vdots$

Step k **:** Let $s_{tmp} = \{l_1, l_2\}$, where l_1 and l_2 satisfy

1. l_1, l_2 have not been selected into s.
2. $$|e_{kl_1}| = \max_i |e_{ki}|$$
3. $$|e_{kl_2}| = \max_{\substack{j \\ e_{kj} \cdot e_{kl_1} < 0}} |e_{kj}|$$

Units l_1 and l_2 will be added into s by

$$\begin{cases} i_{2(k-1)} = l_1, i_{2k-1} = l_2, \text{ if } n \geq 2k-1 \\ i_{2(k-1)} = l_1, \text{ if } n = 2(k-1) \end{cases}$$

$\vdots$

Repeat step k till $n = 2k-1$ or $n = 2(k-1)$

Final Adjustment : Let $s_{-i_1} = \{i_2, \ldots, i_n\}$, and $i_1 = j_p$, $j_p \in s'$ such that j_p satisfies

$$mcor(j_p, s_{-i_1}) = \min_{\substack{j_k \in s' \\ j_k \notin s_{-i_1}}} mcor(j_k, s_{-i_1})$$

where $mcor(j_k, s_{-i_1})$ is the multiple correlation coefficient between unit j_k and the set s_{-i_1},

3 Simulation Study

In the spatial Gaussian model, the population random vector $\mathbf{Y}$ is assumed to follow a multivariate normal distribution

$$\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2}$$

where

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)', \ \boldsymbol{\Sigma} = \{\sigma_{ij}\}, \ i, j = 1, \ldots, N.$$

In this article, a Gaussian-shaped spatial covariance function (Cressie 1993) is used to generate $\boldsymbol{\Sigma}$, that is

$$\sigma_{ij} = \sigma^2 \exp(-||\mathbf{h}||^2/c^2) \tag{3}$$

where $\mathbf{h}$ is the Euclidean distance between sites i and j. The parameter c determines the strength of covariance in the study region. The larger c is, the stronger the covariance between population units is, and vice versa. In the following simulation, parameter values $c = 3.5$ $\mu_i = 0, ^{\forall}$ i and $\sigma^2 = 1$ are used. The parameter c is set to be 3.5 for a moderate correlated spatial population. The population size N is chosen to be 81.

The population quantity of interest is the population total.

$$T(\mathbf{Y}) = \mathbf{1}_N'\mathbf{Y} = \sum_{i=1}^{N} Y_i,$$

where $\mathbf{1}_N$ is a vector of length N in which all elements are 1. The Best Linear Unbiased Predictor (BLUP) for the population total,

$$\hat{T}_1 = \mathbf{1}_n'\mathbf{w}_s + \mathbf{1}_{N-n}'[\boldsymbol{\nu}_{\bar{s}} + \boldsymbol{\Lambda}_{\bar{s}s}\boldsymbol{\Lambda}_{ss}^{-1}(\mathbf{w}_s - \boldsymbol{\nu}_s)], \quad (4)$$

(cf. Bolfarine and Zacks 1992 p.25) is used. In Equation 4, $\bar{s}$ is an index set containing the labels of all the unselected units, $\mathbf{w}_s$ is the vector of observed values, $\boldsymbol{\nu}_s$ and $\boldsymbol{\nu}_{\bar{s}}$ consist of the mean values associated with s and $\bar{s}$. $\boldsymbol{\Lambda}_{\bar{s}s}$ is the covariance matrix between $\mathbf{W}_{\bar{s}}$ and $\mathbf{W}_s$ and $\boldsymbol{\Lambda}_{ss}$ is the covariance matrix of $\mathbf{W}_s$. Note that the Best Unbiased Predictor (BUP) and BLUP are equivalent under the Gaussian model.

The relative efficiency of a design to SRS is defined as the ratio of the mean-square prediction error obtained with SRS to that obtained with the design, so that a value greater than 1 indicates the proposed design is more efficient. In this article, mean-square prediction error was estimated with simulation by producing K realizations of the model and design and calculating

$$\mathbf{E}(T - \hat{T})^2 = \frac{1}{K}\sum_{j=1}^{K}(T_j - \hat{T}_j)^2,$$

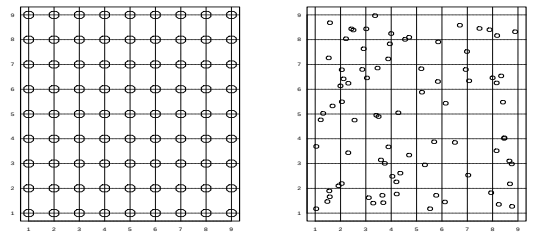
where T_j and $\hat{T}_j$ are the true and predicted population total of the j_{th} realization. For each case, $K = 15,000$ realizations are simulated for each case.

Two sampling situations are considered: the regularly and randomly distributed possible sampling locations. In Figure 1(a), $N = 81$ possible sampling locations are regularly located at the cross points of a 9×9 rectangular grid (case 1). On the other hand, the 81 locations in Figure 1(b) are generated by a bivariate uniform distribution. Let (A_i, B_i) be the coordinates of site i, then

$$A_i \overset{i.i.d}{\sim} Unif(1, xlim)$$

$$B_i \overset{i.i.d.}{\sim} Unif(1, ylim)$$

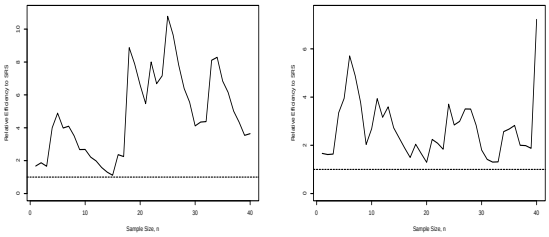
and A_i is independent from B_i, $\forall i$. In this case, $xlim$ and $ylim$ are both selected to be 9.



(a) Case1 (b) Case2

Figure 1: Regularly and randomly distributed population locations.

The relative efficiencies of the proposed design to SRS with respect to the sample sizes from 1 to 40 under two different cases are plotted in Figures 2(a) and (b). It is clear that the performance of the proposed design is in general better than SRS in both cases. The relative efficiency to SRS can be as high as 10 in case 1 and 6 in case 2. The reason is that, symmetrically and evenly placing the sampling sites is advantageous with the regularly distributed population sites as in case 1. By taking the sign of e_{ij} into consideration, the proposed design implicitly arranges the sampling units symmetrically. In addition, the final adjustment of selecting i_1 from s' is helpful to arrange s more evenly.



(a) Case 1 (b) Case 2

Figure 2: Relative efficiencies to SRS

Though the relative efficiency is often greater than 1, the proposed design does not perform as well when the possible sites are randomly distributed. A possible reason is that the proposed evaluates the magnitude and sign of e_{ij} separately. In each step of the proposed design, the component e_{kl_2} has the largest absolute value among all e_{kj}, $\forall j = 1, \ldots, N$ and $e_{kj} \cdot e_{kl_1} < 0$, under the condition that l_2 has not been selected. However, the magnitude of e_{kl_2} might not be large enough for l_2 to be a "good" sampling unit,

especially when the population units are distributed randomly. It is possible to improve the proposed with further modification. The possible approaches for further improvement is briefly discussed in Section 4.

4 Discussion

With a population covariance function that is not a monotonically decreasing function of the distance, it is not appropriate to locate the sampling sites in the study region systematically and evenly. One would need an alternative sampling design to select sampling units for lower prediction error. As illustrated in Section 3, the proposed design can usually provide better selection of sampling units than SRS for predicting the population quantity of interest under a moderately correlated population. Unlike the earlier results of model-based optimal sampling strategies, the proposed design does not depend on the exact population distribution nor the predictor used for the population quantity of interest. No intensive computation is required to select the sampling sites, and the computation load does not increase too much with the population size. Furthermore, the sampling procedure is easy and straightforward. Hence, the proposed designs should be of more practical interest than the earlier optimal sampling strategies.

The two-stage optimal adaptive sampling strategy proposed by Chao and Thompson (2001) also needs intensive computation to determine the sampling sites. Another future research possibility is the extension of the proposed design to a two-stage adaptive sampling strategy - the observed values will be considered in the sampling design. A similar selection procedure as in the proposed design should be able to locate the second-stage sampling sites for lower prediction error with much less computation load.

Although the proposed sampling design can often select better sampling units than SRS, their performances are not stable. Appropriate modification is required for further improvement. The goal is to improve its performance as close to the optimal sampling strategy as possible. One possible approach is to evaluate the magnitude and sign of e_{ij} with a more delicate procedure. For another possibility, some of the techniques used in PCA to select the number of principal components which can effectively summarize the covariance matrix might be helpful (e.g Rencher 1995 pp.434-437). For example, a screen plot is a simple procedure widely used in PCA to determine the number of principal components that account for most of the population variability, therefore, another possible modified design is to select the sampling sites only based on the eigenvectors corresponding to the selected "important" principal component.

5 References

Anderson TW. 1984. An Introduction to Multivariate Statistical Analysis, 2nd Ed., John Wiley & Sons Inc.

Bolfarine H. ,Zacks S. 1992. *Prediction Theory for Finite Population.* Springer Verlag, New York.

Chao CT ,Thompson SK. 2001 Optimal Adaptive Selection of Sampling Sites, *Environmetrics*, 12: 517-538.

Chao CT. 2003. Markov Chain Monte Carlo on adaptive sampling selections. *Environmental and Ecological Statistics*, **10**, 129-151.

Cressie NAC. 1993. Statistics for Spatial Data. Wiley, New York. revised version.

Rancher AC. 1995 *Methods of Multivariate Analysis.* John Wiley and Sons, Inc..

Sacks J, Schiller S. 1988. Spatial design. In Gupta, S.S. and Beregr, J.O., editors, *Statistical Decision Theory and Related Topics IV*, volumn 2. pp.385-395. Springer, New York.

Thompson SK, Seber GAF. 1996. *Adaptive Sampling.* Wiley, New York.

Zacks S. 1969 Bayes sequential design of fixed size samples from finite population, *Journal of American Statistical Association*, **64**, 1342-69