Multivariate Analysis Techniques on Model-based Sampling

Chao, Chang-Tai[1] and Lin, Feng-Min[1]
Department of Statistics, National Cheng-Kung University, Taiwan[1]

Abstract

Previous work regarding lowering prediction error in model-based sampling strategies has different shortcomings. For example, optimal strategies often require intensive computation, also depend on the population model and the predictor. The spatial systematic design cannot work properly under an anisotropic population, or population with non-homogeneous variance. Intuitively, one would like to select units that account for as much population variability as possible. Ideally, it can be done by carefully examining the population covariance matrix, in which the information of the population variability is contained. The object of this study is to construct a model-based sampling design which makes use of various multivariate analysis techniques to explore the information contained in the population covariance matrix. The properties of this designs will be discussed and compared.

KEY WORDS: Model-Based Sampling; Optimal Sampling Strategy; Multivariate Analysis; Principal Component; Cluster Analysis.

1. Introduction

The inference problems in sampling can be categorized into two approaches, the design-based approach and model-based approach. In the design-based approach,the vector of the values of the population variable of interest, $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$, is considered as a constant vector and the inference is established based on the design probability only. On contrast to the design-based approach, the model-based approach consider $\mathbf{y}$ as a realization of a random vector $\mathbf{Y}$ with density function $f(\mathbf{y}; \boldsymbol{\theta})$, and the inference is based on the population stochastic model as well as the design probability. One major difference between these two approaches is the existence of an optimal sampling strategy. Under the model-based approach, Basu (1969) has showed an optimal sampling strategy is available and the optimal selection of sampling units depends on the observed values of the population variable of interest obtained during the survey. Furthermore, Zacks (1969) described a theoretical optimal sampling strategy under a given or Bayesian population model and a fixed sampling size n. This optimal strategy is an n stage adaptive one, that is, the sampling units are selected in a way such that the selection of the next unit should depend on the observed values obtained in all the previous stages as well as the population model.

Such an n stage adaptive strategy is in fact very complicated and computational consuming. Sacks and Schiller (1988) proposed an optimal conventional sampling strategy under a given population model, however, the selection of sampling units by this conventional strategy does not take the observed value into account. For making use of the observed values obtained during the survey, Chao and Thompson (2001) proposed a two-stage optimal adaptive strategy under a given population model to further improve the optimal conventional strategy proposed by Sacks and Schiller (1988) and compromise with the optimal n stage strategy. Chao (2003) also described the extension of this two-stage optimal strategy to a Bayesian population model.

The optimal strategies that were proposed by different authors in the past have certain common disadvantages, despite the fact that they can often locate the optimal sampling units. First of all, the computational load required can be extremely intensive to determine the optimal sample especially for a large population size and/or a complicated population model. These optimal strategies also assume an exact population model such as a given or prior population distribution, but such information might not be available in practice. In addition, the optimal sample varies from case to case in terms of different population distributions and prediction inferences. All these disadvantages restrict practical applications of these optimal strategies.

The purpose of this research is to construct a flexible sampling strategy for a better prediction result. The idea to select sampling units that can provide lower prediction mean-square is rather straightforward. Intuitively, units that have high variance themselves, strong covariance with other unselected units, also lower within correlation are preferred. With a careful evaluation of population covariance matrix, it is possible to select such units for a better prediction purpose. Multivariate analysis plays an important role in evaluating the covariance matrix of a random vector. Chao (2004) proposed two sampling selection methods motivated by one of the well-known multivariate analysis techniques, Principal Component Analysis, which objective is to summarize most of the variability using the principal components with the highest variances. Hence these methods could selected an appropriate sample to account for as much possible population variability.

Actually other multivariate analysis techniques are also of potential possibility to select sampling units for a better prediction result. In this research, we firstly utilized another well-known multivariate technique, Cluster Analysis, to divide the units into several groups such that the units within the same group are as similar as possible.

After the population has been partitioned, we select the with-cluster sample with the design proposed by Chao (2004) which could identify units that account for more population variability. The algorithm of proposed design is described in Section 2. The proposed method is examined by simulation results in terms of the sampling locations and the empirical relative efficiencies to Simple Random Sampling Without Replacement (SRSWOR). For better visual evaluation, the sampling locations with a small population size selected by the proposed sampling methods are illustrated in Section 3. Simulation study shows that they can always provide more precise prediction results than SRSWOR. Some of the simulation results are presented in Section 4. Comments on the current findings and the future research are addressed in Section 5.

2. Sampling Design

Let $\mathbf{Y}$ be the population random vectors with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)'$, and covariance matrix

$$\mathrm{Var}(\mathbf{Y}) = \boldsymbol{\Sigma} = \{\sigma_{ij}\}_{i,j=1,\ldots,N},$$

where

$$\sigma_{ij} = \begin{cases} \mathrm{Var}(Y_i) & \text{if } i = j \\ \mathrm{Cov}(Y_i, Y_j) & \text{if } i \neq j. \end{cases}$$

The objective is to select n sampling units out of the N population units to predict the population quantity of interest $T(\mathbf{Y})$ with some unbiased predictor $\hat{T}(d)$. In particular, we consider the prediction of population total $T(\mathbf{Y}) = \sum_{i=1}^{N} Y_i$, and the best unbiased predictor, $\hat{T} = E[T|d]$ in this research.

One of the basic principles of sampling is to select sampling units that are as less similar as possible. The essence of Cluster Analysis is to partition the population in the way that the units within a cluster are as similar as possible. Hence, a sample selected by some design within each cluster can be representative of the population as a whole. Then, within each cluster we select the units which could account for more variability with the design proposed by Chao (2004).

The proposed sampling design utilizing Cluster Analysis and Principal Component Analysis is a straightforward approach.

1. Partition the population into $g \leq n$ clusters, denoted as $\mathbf{u}_1, \ldots, \mathbf{u}_g$, $\mathbf{u}_i \cap \mathbf{u}_j = \emptyset$ and $\cup \mathbf{u}_i = \mathbf{u}$.

2. Select n_i sampling units, denoted as s_i, within cluster $\mathbf{u}_i$ with the design proposed in Chao (2004), such that $n_i \propto N_i$ and rounded to the nearest integer. If the rounded $n_i = 0$, then $n_i = 1$.

3. The final sample of size n is the collection of s_i.

The only population information required in the proposed sampling method is the population correlation matrix $\mathbf{R}$ has to be given. The population is clustered by treating $\mathbf{D} = \mathbf{1}_{N\times N} - \mathbf{R}$ as the distance matrix, where $\mathbf{1}_{N\times N}$ is a $N \times N$ matrix with all elements equal to 1. In addition, the algorithm used in Cluster Analysis to divide the population into several disjoint groups is the $K - means$ method (e.g. Johnson and Wichern 1998).

3. Sampling Locations

The sampling locations selected by sampling designs proposed in Section 2 are illustrated under two different possible population locations in this section. The population random vector $\mathbf{Y}$ is assumed to follow a multivariate normal distribution

$$\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

where

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

In this section, the population and sample sizes are set to be $N = 25$ and $n = 6$, respectively. A Gaussian-shaped spatial covariance function (Cressie 1993) is used to generate $\boldsymbol{\Sigma}$, $\sigma_{ij} = \sigma^2 \exp(-||\mathbf{h}||^2/a^2)$, where $\mathbf{h}$ is the Euclidean distance between unit i and j. The parameter a determines the strength of covariance in the study region. The larger a is, the stronger the covariance between population units is, and vice versa.

First we consider the possible population units are the cross points of a 5×5 rectangular grid. Figure 1 illustrates the sampling locations selected by the design described in Section 2 with $n = 5$ and the number of cluster g is 3. The correlation matrix $\mathbf{R}$ is calculated by $\boldsymbol{\Sigma}$. In Figure 1 it is clear the design provides different results because the clustering algorithm gives different kinds of partitions. It seems the design can successively select the units spread evenly on the rectangular grid. Figure 2 illustrates the sampling locations when the population units are randomly distributed. In this situation the results are more stable.

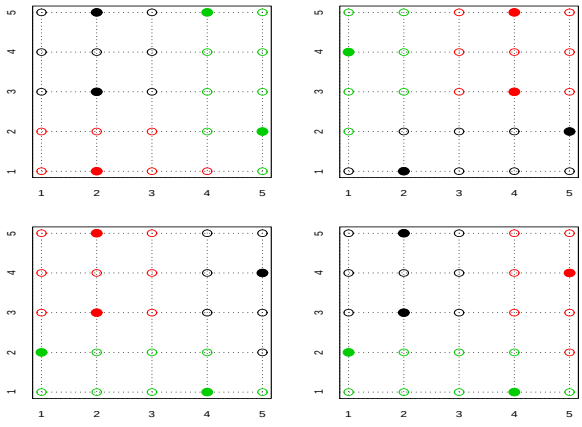


Figure 1: The possible population locations and sampling units selected by the design proposed under regularly distributed population locations

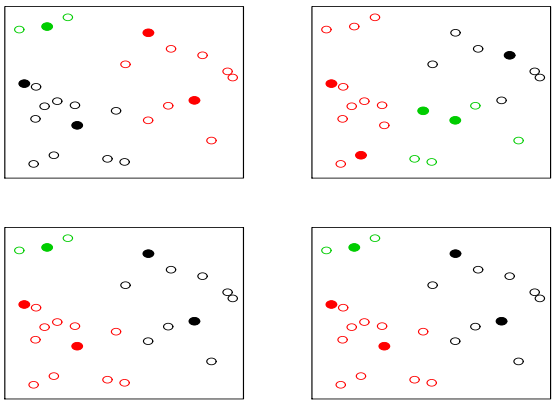


Figure 2: The possible population locations and sampling units selected by the design proposed under randomly distributed population locations

4. Simulation Study

The performances of the design proposed is evaluated in this section by the Empirical Relative Efficiency (ERE) to SRSWOR. The ERE of a design to SRSWOR is defined as the ratio of the mean-square prediction error obtained with SRSWOR to that obtained with the design, so that a value greater than 1 indicates the proposed design is more efficient. In this study the mean-square prediction error is estimated with simulation by producing K realizations of the model and design and calculating

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

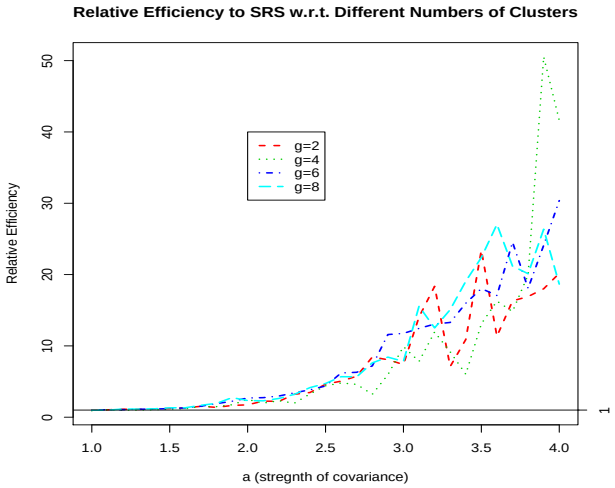
where T_j and $\hat{T}_j$ are the true and predicted population total of the j_{th} realization. For each case, $K = 15,000$ realizations are simulated for each case.

The studied cases are essentially the same as those in Section3, only with a larger study region and population size. The population size used in this section is $N = 81$. The population quantity of interest is the population total.

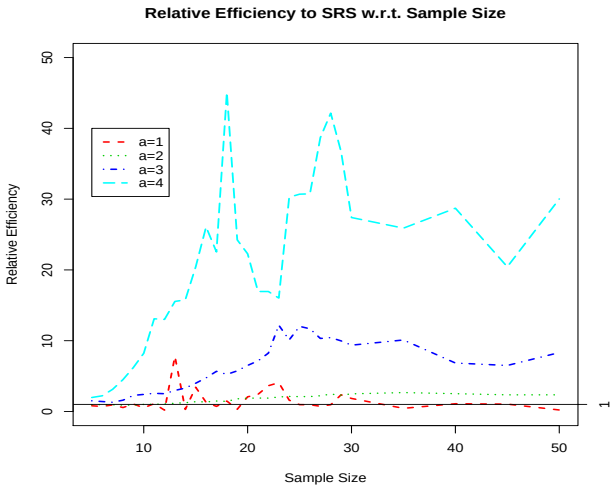
$$T(\mathbf{Y}) = \mathbf{1}_N'\mathbf{Y} = \sum_{i=1}^{N} Y_i,$$

where $\mathbf{1}_N$ is a vector of length N in which all elements are 1. The predictor used is the Best Linear Unbiased Predictor (BLUP). (Simulation results regarding other spatial population model and predictor to be used are not discussed in this article.)

In this simulation, value of parameter a ranges from 1.0 to 4.0, and values of population parameters $\mu_i = 0, \forall\, i$ and $\sigma^2 = 1$ are used. We consider two possible population units, the cross points of a 9×9 rectangular grid and the randomly distributed population locations. The ERE of the proposed design to SRSWOR under the two different cases plotted in Figure 3 and 4, respectively.



(a) The ERE with respect to a from 1 to 4 under $g = 2$, 4, 6, 8, and $n = 30$

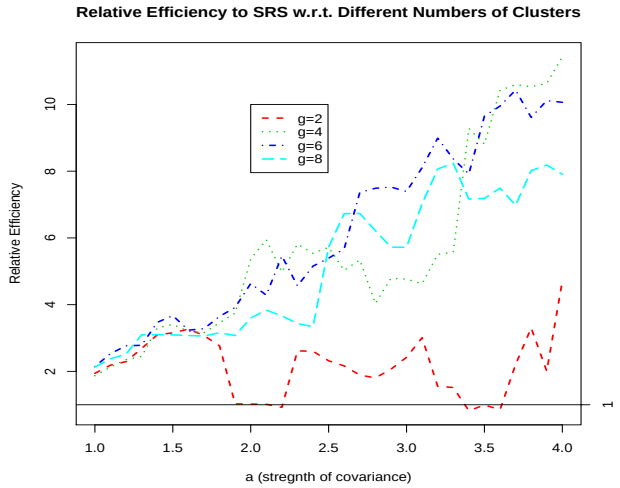


(b) The ERE with respect to the sample size from 1 to 50 under $a = 1$, 2, 3, 4, and $g = 5$

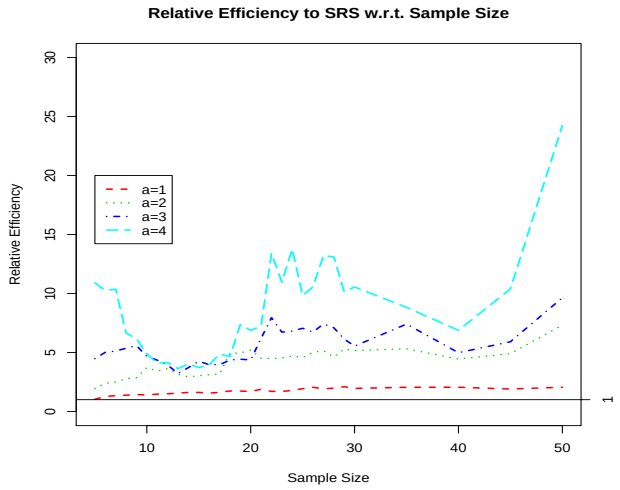
Figure 3: Empirical Relative Efficiency to SRSWOR on a region with regular distributed population locations.

Figure 3(a) illustrates the ERE of the proposed sampling design to SRSWOR for $g = 2$, 4, 6, and 8 under a from 1 to 4 when sample size n is equal to 30. The values of ERE greater than one indicating that the proposed design is more efficient than SRSWOR in terms of providing lower mean-square error. It can be seen that the values of ERE are usually greater than one and get better as a increases since the proposed design can take advantage of the stronger covariance. No matter what value of g is chosen, the performance does not make much difference. Figure 3(b) is the ERE of the proposed design to SRSWOR with respect to sample size from 1 to 50 under a = 1, 2, 3, and 4 when the number of clusters is five. In this case the proposed design performs better with larger sample size but the results are not stable.

The simulation conditions are the same as in which the



(a) The ERE with respect to a from 1 to 4 under $g = 2$, 4, 6, 8, and $n = 30$



(b) The ERE with respect to the sample size from 1 to 50 under $a = 1$, 2, 3, 4, and $g = 5$

Figure 4: Empirical Relative Efficiency to SRSWOR on a region with random distributed population locations.

population units are regularly distributed at the study region as in Figure 3(a) and (b) and the ERE are plotted in Figure 4(a) and (b). The proposed design performs clearly better than SRSWOR under $g = 4$, 6, and 8 regardless of a selected. However the performances are not very stable when the population is divided into two clusters. Note that the instability might result from which the number of clusters is not fitting; hence it is also important to choose an appropriate number of clusters in the proposed design. In addition, the performances of the proposed design are always better than SRSWOR regardless of sample size.

5. Discussions and Future Research

The proposed designs usually can provide more efficient prediction results than SRSWOR. Although the result is not optimal, the computation required is much less and easier than the optimal sampling strategies proposed the past. In addition, they do not depend on the exact population distribution and the predictor to be used. The population correlation matrix is the only population information required in the design. In fact, only the covariance pattern has meaningful impact on the sampling selection but not the exact values of the entries in $\mathbf{R}$. The correlation matrix can be replaced by the empirical correlogram in practice, and such information is often available in a spatial sampling situation. Hence, the proposed designs are more flexible and robust than the theoretical optimal designs. We still need to examine the performance of the proposed design when the population variance are not homogeneous and when the population is an anisotropic one. The performance of the proposed design is not vary stable in the simulation study and hence further modification of the design is certainly necessary and worthy for the future research. Related work and results are expected to be proposed in the near future.

REFERENCES

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, New York.

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

Chao, C.T. (2003), "Markov Chain Monte Carlo on Adaptive Sampling Selections," *Environmental and Ecological Statistics*, **10**, 129-151.

Chao, C.T., (2004), "Selection of Sampling Units under a Correlated Population Based on the Eigensystem of the Population Covariance Matrix ," *Environmetrics*, **15**, 757-775.

Chao, C.T., Thompson, S.K. (2001), "Optimal Adaptive Selection of Sampling Sites," *Environmetrics*, **12**, 517-538.

Cressie, N., (1993), *Statistics for Spatial Data*, Wiley, New York.

Mardia, K.V, Kent, J. T., Bibby, J.M., (1979), *Multivariate Analysis*, Academic Press Inc., London.

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

Thompson, S.K., Seber, G.A.F. (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.