Model-based Sampling Selection under an Anisotropic Population

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

Abstract:

To select n sampling units out of N population units to predict the population quantity of interest in spatial statistics, an appropriate spatial sample design is required to recognize and account for the spatial auto-correlation in the spatial process; for example, a spatial systematic design traditionally would be used for better prediction results. This is, however, only effective under certain population covariance structures, such as an isotropic population model. For more general cases, the optimal sampling strategies can be used to select the optimal sample with which the mean-square error is minimized. Nevertheless, the practical interest of such optimal sampling strategies is seriously restricted by the intensive computational load and model assumption required to select the optimal sample. The object of this study is to construct spatial sampling designs under an anisotropic population to predict the population quantity of interest, such as population mean level, with lower prediction error. The performances of the proposed designs based on the relative efficiency of the proposed designs to simple random sampling will be illustrated with simulation study.

Keywords: Model-based Sampling; Optimal Sampling Strategy; Eigensystem; Gaussian Model; Log-Gaussian Model.

1 Introduction

Under a finite population setting, the population consists of N units labelled 1,2,...,N. From the model-based view point, the vector of the values of the population variable of interest, $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$, is considered as a realization of a random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$. Let s be the sample of n units selected from the population and $\mathbf{y}_s$, the vector of y values associated with s, be the vector of observed values. Then the data d is a collection of the sample s and the observed values $\mathbf{y}_s$, $d = (s, \mathbf{y}_s)$. The inference problems ni sampling can be categorized into two approaches, design-based and model-based approaches. In the rather more traditional design-based approach, $\mathbf{y}$ is considered as a constant vector and the inference is established based on the design probability only. In the model-based approach, on the other hand, $\mathbf{y}$ is considered 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. Notice that the design probability sometimes can be ignored depending on difference types of inference (Thompson and Seber 1996).

One major difference between the design-based and model-based is the existence of the optimal sampling strategy. Under the design-based approach, there is no so-called optimal sampling strategy which is always better than any other design among all possible populations(Godambe 1955, Thompson and Seber 1996). Intuitively it is not such a surprising result since an optimal sampling design-based strategy has to be better than any other design under any possible population, which is impossible because no population model is assumed. With an assumed population distribution, however, it is possible to established a model-based optimal sampling strategy. Furthermore, with an given or Bayesian population model and a fixed sampling size n, the optimal model-based sampling strategy is in general an n-stage adaptive one. That is, each sampling units should be selected based on what have been observed during the survey (Zacks 1969).

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. They utilized a modified annealing algorithm to search for the optimal sample under a fixed sample size n. 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. For the purpose of a more practical application, Chao (2003) described the extension of this two-stage optimal strategy to a Bayesian population model. In this

Support for this research was provided by the National Science Council, Taiwan, NSC 93-2118-M-006-007- .

extension, the optimal sample is selected by Markov chain Monte Carlo method as well as the modified annealing algorithm. Nevertheless, all these strategies share the common disadvantages, such as intensive computation required to determine the optimal sample and dependence of the optimal sampling selection on the assumed population model and predictor. Chao (2004) proposed two sampling designs that are based on the eigensystem of the population covariance to select sampling units. These two designs required less population information than the optimal strategies proposed in the past. Only the population covariance structure, but not the the exact population distribution, is assumed. Further, the selection is free of the predictor to be used. Also, the selection procedure of the sampling units is much easier to be implemented in practice. The prediction results are usually than what provided by the simple random sampling.

The covariance is a function of both the distance and direction in an anisotropic population, which is often seen in practice especially in a large area survey (Arbia and Lafratta 2002). Under such a population, the usual systematic or symmetric sampling locations which are widely used in a spatial sampling situation would become more difficult or even impossible depending on whether the type of anisotropic structure is geometric, which can be transformed back to an isotropic one with certain linear transformation of the coordinate system, or zonal anisotropic. In this research, we will examine the performances of these two designs under the anisotropic populations. Based on the result, we will evaluate the possibilities of further modification to improve the designs proposed in Chao (2004).

The intuition and algorithms of the proposed sampling designs will be briefly described in Section 2. The proposed methods are examined by simulation results in terms of the sampling locations and their relative efficiencies to Simple Random Sampling (SRS). Both of the geometric and zonal anisotropic population are considered in this research. For better visual evaluation, the sampling locations with a small population size selected by the proposed sampling methods are illustrated in Section 3. Section 4 presents the results of the relative efficiency to SRS with a larger population size. Results show that these two designs should be utilized depending in different population covariance structures. Applications and further research are discussed in Section 5.

2 Sampling Selection Methods

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

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 $\boldsymbol{\Sigma}$,

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

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 designs 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.

Chao (2004) proposed two design, denoted as I and II which make use of the information provided by the eigensystem of the population covariance structure. For detailed procedures and insights of these designs, please refer to Chao (2004). In short, design I makes use of the magnitudes/absolute values of the components in the eigenvectors, and design II also takes the sign of the components into account.

3 Sampling Locations

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

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

where

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

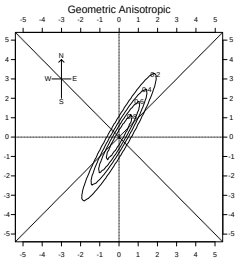
In this article. The covariance matrices are generated by two different anisotropic models, geometric anisotropic and zonal anisotropic, are considered. The population and sample sizes are set to be $N = 25$ and $n = 5$, respectively.

3.1 Geometric Anisotropic

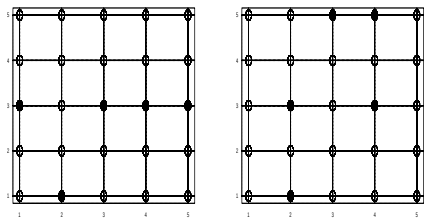
In the spatial geometric anisotropic population, the covariance matrix can be transformed to an isotropic one by a linear transformation. Equation 3 is the geometric covariance function used in this section (Eriksson and Siska 200).

$$\begin{aligned}\sigma_{ij} &= \sigma^2 \exp(-||\mathbf{h}||^2/c^2)\mathrm{Cov}(Y_i, Y_j) \\ &= \sigma^2 \exp\left[h^2 \frac{\cos^2(\phi-\theta) + \lambda^2 \sin(\phi-\theta)^2}{a_{max}}\right]\end{aligned} \tag{3}$$

where h is the Euclidean distance between site i and j, θ is the angle of rotation of the coordinate system, a_{max} is the range on the direction of θ, ϕ is the separation angle, and $\lambda > 1$ is the ratio of anisotropy of the ellipse. Equation 3 indicates that the larger h is or the closer ϕ to θ, the stronger the covariance



(a) Covariance



(b) 1 (c) 2

Figure 1: The covariance structure given in Equation 3 with $\sigma^2 = 1$, $\lambda = 7$, $\theta = \pi/3$, and $a_{max} = 3$. and the sampling locations selected by design I and II.

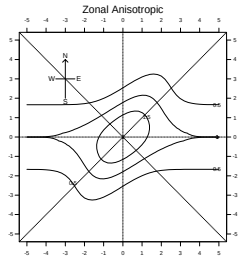
between population units is, and vice versa. First we consider that the possible sampling sites (also the population units) are the cross points of a 5×5 rectangular grid. Figure 1 illustrates the covariance structure and the sampling locations selected by the designs described in Section 2 with $n = 5$. It is clear that design II seems be able to select more plausible sample.

3.2 Zonal Anisotropic

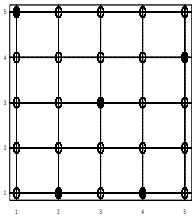
In contrast to the geometric anisotropic model, the zonal anisotropic covariance matrix cannot be transformed to an isotropic one by a linear transformation. Equation 4, which is essentially a sum of two different geometric covariances, is the zonal covariance function used in this section (Eriksson and Siska 200).

$$\begin{aligned}\mathrm{Cov}(Y_i, Y_j) = \sigma_1^2 \exp&\left[h^2 \frac{\cos^2(\phi-\theta_1) + \lambda_1^2 \sin(\phi-\theta_1)^2}{a_{1,max}}\right] \\ + \sigma_2^2 \exp&\left[h^2 \frac{\cos^2(\phi-\theta_2) + \lambda_2^2 \sin(\phi-\theta_2)^2}{a_{2,max}}\right]\end{aligned} \tag{4}$$

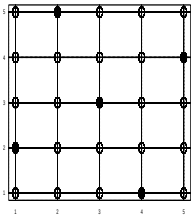
Figure 2 shows the covariance structure and the sampling locations selected by the designs described in Section 2 with $n = 5$. Again, design II seems be able to select more plausible sample.



(a) Covariance



(b) 1



(c) 2

Figure 2: The covariance structure given in Equation 4 with $\sigma_1^2 = \sigma_2^2 = 1$, $\lambda_1 = 3$, $\lambda_2 = \infty$, $\theta_1 = \pi/3$, $\theta_2 = 0$, $a_{1,max} = 4$, and $a_{2,max} = \infty$. and the sampling locations selected by design I and II.

4 Relative Efficiency to SRS

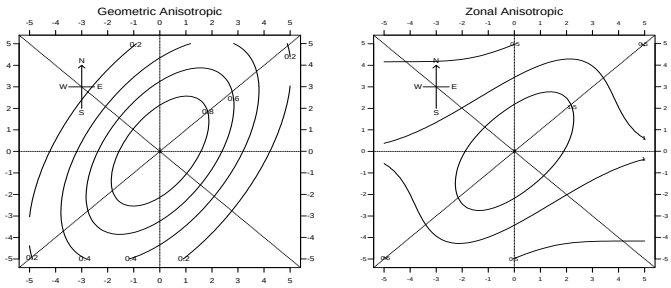
The performances of the proposed designs are examined under some moderate correlated populations based on the relative efficiencies of designs I and II to SRS. The cases studied are essentially the same as those in Section3, only with a larger study region and population size. The population size used 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 Best Linear Unbiased Predictor (BLUP) for the population total (cf. Bolfarine and Zacks 1992 p.25).

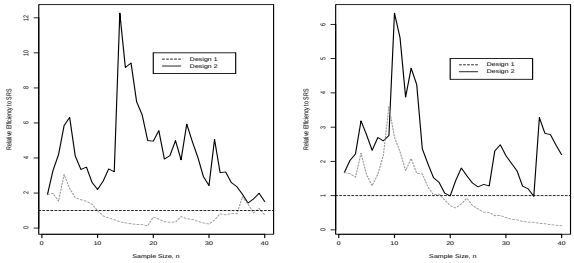
The relative efficiency of a design to SRS is defined as the ratio of the mean-square prediction error obtained with SRS to that obtained with the design, so that a value greater than 1 indicates the proposed design is more efficient. In this article, mean-square prediction error was estimated with simulation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

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



(a) Geometric Anisotropic (b) Zonal Anisotropic

Figure 3: Geometric and Zonal Anisotropic Covariance structure



(a) Geometric Anisotropic (b) Zonal Anisotropic

Figure 4: Relative Efficiencies of Design 1 and 2 to SRS on a region with sampling sites on the cross points of a 9×9 rectangular grid.

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. Figures 3 describes the covariance structures under geometric and zonal anisotropic model used in this section.

Shown in Figures 4 are the relative efficiencies of design I and II to SRS. Design II is always better than SRS, but the performance of design I is not as good and it is often even worse than SRS.

In fact, design I is known to perform better with non-homogeneous population variances and randomly distributed possible sampling locations (Chao 2004). We also examine the performance of design I and II under such a situation. Figure 5 is the possible sampling locations as well as the variances. Figure 6 is the relative efficiencies of these two designs to SRS, and it indicates that both are better than SRS. In addition, design I is always better than

design II as expected.

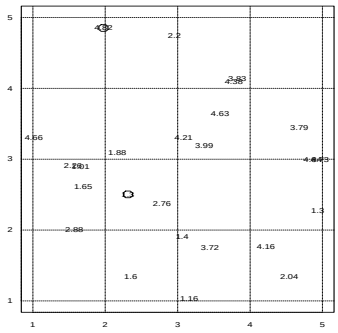


Figure 5: Randomly distributed possible sampling sites with nonhomogeneous variance.

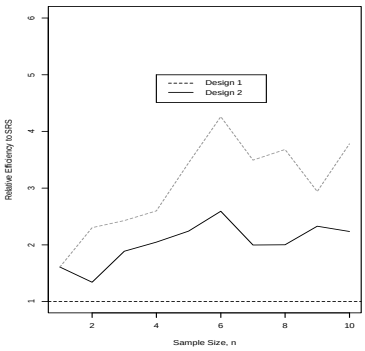


Figure 6: Relative Efficiencies of Design 1 and 2 to SRS on a region with random distributed sampling sites and nonhomogeneous population variance.

5 Discussions and Future Research

Design I and II should be used under different sampling situations. With homogeneous population variance, design II is always better than SRS and should be suggested. On the other hand, both can be used instead of SRS when when the population variance is nonhomogeneous. Notice that with randomly distributed possible sampling locations and/or nonhomogeneous population variance, the usual systematic/symmetric spatial sampling designs are not able to select "good" sample.

The performance of these two designs are not stable, and sometimes can be worse then SRS, especially when the population units are regularly located in the study region with homogeneous population variance. Further modification of these two 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 TW. (1984), *An Introduction to Multivariate Statistical Analysis*, 2nd Ed., John Wiley & Sons Inc.

Arbia G ,Lafratta G. (2002), "Anisotropic spatial sampling designs for urban pollution", *Applied Statistics* **51**:223-234

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.

Eriksson, M and Siska, P.P. (2000), "Understanding anisotropic computation", *Mathehatical Geology*, Vol.32 No.6, pp.683-700.

Godambe VP. (1955), "A unified theory of sampling from finite population", *Journal of the Royal Statistical Society* **B17**:269-278

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