International Council for the Exploration of the Sea

CM 1993/D:31
Refs B,G,H,J,K,N
Statistics Committee

SURVEY STRATEGIES FOR STRUCTURED POPULATIONS PT II: PRECISION OF VARIANCE ESTIMATORS

by

E J Simmonds and R J Fryer

SOAFD Marine Laboratory
PO Box 101, Victoria Road
Aberdeen, AB9 8DB
Scotland, UK

SUMMARY

For surfaces with local positive correlation more precise estimates of the surface mean can often be obtained using stratified random or systematic sampling rather than simple random sampling. The increase in precision depends on the relationship between the spatial correlation, the sampling intensity and the region to be sampled. However, the final choice of strategy also depends on the objectives of a survey. The decision may depend not only on the need to estimate the mean but also on the need to estimate the precision of the mean. This paper provides insight into the comparative performance of some methods for estimating variance and the impact of different sampling strategies on this process.

Simulated surfaces were generated with a range of statistical properties similar to those observed on annual acoustic surveys of North Sea herring. These surfaces contained elements of local positive correlation, a random process and a non stationary component. The proportions of these were varied to study the effects of different situations.

Eight different sampling strategies were investigated, all with a sampling intensity similar to the annual herring surveys, varying from 40 transects randomly located in one stratum to 40 transects with systematically spaced and with a centred starting point. The bias and the true error variance of the sample mean associated with each sampling strategy were estimate. For the surfaces generated, all the sampling strategies were approximately unbiased. The highest error variance was obtained from the simple random sample; the lowest error variances were obtained from the systematic strategies.

From each sampling strategy, four estimators of the variance of the sample mean were investigated. These were based on a) the sample variance; b) the pooled within strata variance; c&d) two geostatistical estimation variances, using spherical and exponential models with nugget. The simulations were used to estimate the mean, median and 90% intervals for each variance estimator for each sampling strategy.

The variance estimator based on the pooled within strata variance is unbiased for strategies from simple random to two transects per strata. For more structured surveys it is biased upwards. The two geostatistical variance estimators closely followed the true error variance for all strategies. In these simulations, 90% intervals of the pooled within strata and geostatistical variance estimators are narrowest at two transects per strata, and 90% intervals of the abundance estimator are narrowest for systematic strategies.

INTRODUCTION

The interaction between survey precision and survey strategy was examined by simulation by Simmonds and Fryer (1992). They concluded, that for surfaces with local positive correlation more precise estimates of the surface mean can often be obtained using stratified random or systematic sampling rather than simple random sampling. The increase in precision depends on the relationship between the spatial correlation, the sampling intensity and the region to be sampled.

Examples of the effects of different strategies were given for herring populations in the Orkney Shetland area of the North Sea. These indicated decreases in error variance of between 20 and 80% as the strategy was changed from simple random to systematic with the same sampling intensity. It was pointed out that the final choice of strategy depends on the objectives of the survey. The decision may depend not only on the need to estimate the mean but also on the need to estimate the precision of the mean. This paper considers the latter problem by comparing four variance estimators over a range of sampling strategies.

METHODS

Examination of Data from Herring Surveys

The statistical properties of three acoustic surveys were investigated in one dimension by examining the transect abundance. The amplitude distribution of transect values was estimated by normalising the abundances in each year and combining the transect values over years (Fig. 1). Variograms were calculated from each survey to determine the spatial statistical properties of the transect abundances. Two of these variograms are shown in Simmonds and Fryer (1992). All the surveys show a small positive correlation with a range of about two to three transects, a nugget effect of between zero and half the sill, and a non-stationary component illustrated by a continually increasing variogram. A number of plausible variogram functions describe these data. There are some indications of trend, since the southern half of the area contains lower abundance than the northern part for all three surveys. However, it is not possible to conclusively distinguish between trend and large scale spatial correlation from the data so both possibilities are considered.

Simulation Procedures

To investigate all these possible characteristics a range of surfaces with differing statistical properties were simulated. The surfaces contained elements of local positive correlation, random components and non stationary components. The local spatial correlation was generated using an auto-regressive function chosen to give spatial autocorrelation similar to the herring surveys. The random component was derived from

a random number generator. The non-stationary component was derived in three ways; a simple random walk, a linear trend and a cosine trend function from $-\pi/4$ to $5\pi/4$. As the relative proportions of these components is difficult to establish from the survey data, the proportions were varied to see how sensitive the conclusions were to a wide range of situations. Finally the surface amplitude values were modified so that the amplitude distribution was similar to the distribution observed on the three surveys. For each set of conditions 1,000 surfaces of 400 locations were generated. In all cases the amplitude distribution and the range of spatial autocorrelation were fixed but the proportions and type of non-stationary component varied. The details of the surface generation method are included as Appendix A.

To examine the relationship between variance estimation and sampling strategy we considered eight survey strategies and four variance estimators. The sampling intensity of 40 transects was similar to the coverage used on the surveys being studied. The sample strategies were:

1.	40 Transects Randomly Located in 1 Stratum	(40/1)
2.	20 Transects Randomly Located in 2 Strata	(20/2)
3.	10 Transects Randomly Located in 4 Strata	(10/4)
4.	5 Transects Randomly Located in 8 Strata	(5/8)
5.	2 Transects Randomly Located in 20 Strata	(2/20)
6.	1 Transects Randomly Located in 40 Strata	(1/40)
7.	40 Transects with Systematic Spacing and a Random Start	(Sys-Rand)
8.	40 Transects with Systematic Spacing and Centred	(Sys-Cent)

The strata boundaries were located systematically with equal spacing throughout the area.

Let the survey area be denoted by X and have a size $|X|$. For each realised (simulated) surface let $S(x)$ be the surface value at x. An example of a surface and systematic samples is shown in Figure 2.

The true mean value of the surface is then:

$$E = \frac{1}{|X|} \int_X S(x)\,dx \qquad (1)$$

For each survey let:

- x_{ij} be the location of the j^{th} transect in the i^{th} strata, regarding the systematic surveys as 1 transect in each of 40 strata.
- s_{ij} be the surface value at x_{ij} (ie $s_{ij} = S(x_{ij})$)
- J be the number of transects in each strata and I the number of strata.
- N be the total number of transects (IJ), which for these simulations was 40.

The sample mean in the i^{th} strata is:

$$\bar{s}_i \frac{1}{J} \sum_{j=1}^{J} S_{ij} \tag{2}$$

The overall sample mean is:

$$\bar{s} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} s_{ij} \tag{3}$$

The mean abundance is estimated by the overall sample mean:

$$\hat{E} = \bar{s}$$

The four different variance estimators are:

1. "Sample variance":

$$Var_s(\hat{E}) = \frac{1}{N(N-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} (s_{ij}-\bar{s})^2 \tag{4}$$

2. "Pooled within strata variance":

$$Var_p(\hat{E}) = \frac{1}{N(N-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} (s_{ij}-\bar{s}_j)^2 \tag{5}$$

For those surveys with only one transect per strata, adjacent strata are combined in pairs.

3. Geostatistical estimation variance using a spherical model with nugget, $Var_{gs}(\hat{E})$, fitted by an iterated least squares procedure. The expression for the variance is given in Matheron (1971):

$$Var_{gs}(\hat{E}) = 2\gamma(X,X_{ij})-\gamma(X,X)-\gamma(x_{ij},x_{kl}) \tag{6}$$

where:

$$\gamma(X,X) = \frac{1}{|X|^2}\iint_X \gamma(x,x')\,dx\,dx'$$
$$\gamma(X,x_{ij}) = \frac{1}{Xn}\sum_{ij}\int_X \gamma(X,x_{ij})\,dx \qquad (7)$$
$$\gamma(x_{ij},x_{kl}) = \frac{1}{n^2}\sum_{ij}\sum_{kl}\gamma(x_{ij},x_{kl})$$

where γ is the variogram:

$$\gamma(d) = a + b\,(1.5\,d/R - 0.5\,(d/R)^3) \quad \ldots\ldots\ d<R \qquad (8)$$
$$= a + b \quad \ldots\ldots\ d\geq R$$

In equation (9) a is the nugget, a+b the sill and R the range.

4. Geostatistical estimation variance using an exponential model with nugget, $Var_{ge}(\dot{E})$, fitted by an iterated least squares fitting procedure. The variance is estimated using equations 7 and 8 above and with γ the variogram :-

$$\gamma(d) = a + b\,e^{(1-d/R)}$$

where a is the nugget, a+b the sill and R the range.

The iterated least squares procedure for both geostatistical estimators used the experimental variogram derived from pooled samples in 39 distance bins and weighted by the number of samples per bin. The experimental variogram is given by:

$$\gamma(d) = \frac{1}{2N(d)}\sum_{ijkl}^{N(d)}[s_{ij}-s_{kl}]^2 \qquad (9)$$
$$d = |x_{ij}-x_{kl}|$$

The distance d assigned to an interval was defined as the mean distance between sample pairs within the interval. The parameter estimates were indistinguishable from the fit to the cloud of 780 sample pairs obtained from 40 data values (not put into bins) and computational much faster.

The statistical properties of the abundance estimate $\hat{E}$ were investigated by comparing the simulated estimates say $\hat{E}_m$ with the simulated abundances E_m. Negligible bias was found in these simulations for any of the sampling strategies. The true error variance $Var_t(\hat{E})$ is given by:

$$Var_t(\hat{E}) = \frac{1}{999} \sum_{m=1}^{1,000} (\hat{E}_m - E_m)^2 \qquad (10)$$

The variogram of the complete simulated surfaces was also computed to check on the statistical properties of the generated surfaces.

The mean, median and 90% intervals of each variance estimator were estimated for each sampling strategy. The lower 90% limit, the median and upper 90% limit were obtained by sorting the 1,000 simulated variance estimates and selecting those in locations 50, 500 and 950 respectively.

RESULTS

The variograms for two sets of simulations are shown in Figures 3 and 4. Each figure shows 10 variograms from surfaces with different proportions of nugget effect, stationary positive correlation and non-stationary components. These variograms have statistical properties which cover the range of models derived from data collected on North Sea herring surveys. An example of the cloud of point pairs, the experimental variogram and the fitted models, from one realisation is shown in Figure 5.

Precision of Abundance

The estimates of error variance $Var_t(\hat{E})$, for the three types of non-stationary component are similar (Fig. 6) The simple random sample, on the left, has the highest error variance. The error variance decreases monotonically to the systematic centred strategy on the right of the graph.

Distribution of Variance Estimators

The four variance estimators also show similar results for each method of surface generation. The results for the cosine trend and the linear trend are summarised in Figures 7 and 8.

The mean of the sample variance estimator $Var_s(\hat{E})$ is almost constant, independent of strategy, with a slight increase from random to systematic. The 90% interval is widest for the simple random strategy narrowing slightly with increasing order in the survey. The sample variance estimator is positively biased for all strategies except the simple random strategy.

The pooled variance estimator $Var_p(\hat{E})$ is unbiased from the simple random survey to two transects per strata. For strategies with only one transect per strata, the estimator is positively biased although less biased than $Var_s(\hat{E})$. The 90% interval narrows as the strategy becomes more ordered until the number of transects per strata reduces to two.

Both geostatistical estimators give similar results. The means of these estimators are close to the true error variance over the full range of strategies. The 90% interval narrows from a high for the simple random survey and reaches a minimum at two transects per strata, and widens slightly for strategies with one transect per strata. The exponential model performs better than the spherical model for the systematic strategies. This is because this model more correctly matches the choice of an autoregressive model for the local positive autocorrelation. It is not possible to determine which model is more appropriate for the herring surveys and the differences in the two models indicate the uncertainty due to the choice of model. Generally the geostatistical estimators give a wider spread of estimates than those from the pooled variance estimator. However, these two estimators involve the use of an iterated least squares fitting procedure to arbitrarily chosen models, without any non-stationary components. Alternative fitting procedures or theoretical variograms might improve on these results. For example Cressie (1991) suggests a fourth root transform of squared differences yields a more robust estimate of the variogram; however, this may not be applicable to data surfaces with a non-stationary component.

The iterated least squares fitting procedure is also discussed in Cressie (1991). The method used here is a weighted least squares procedure which compares favourably with other methods examined in Cressie. A more detailed examination of different variogram estimators and fitting procedures is required before the most appropriate method can be selected for this type of situation.

To examine the relative performance of the different estimators scatter plots of variance estimates derived from the pooled within strata variance and both geostatistical estimators obtained from the same realisations are shown in Figures 9 and 10. The two geostatistical estimators are very similar, suggesting that the choice of model is not critical in these circumstances. Figure 10 also shows that the geostatistical estimators have a longer tail than the pooled variance for the simple random survey.

To summarise the results Figure 11 shows the mean 90% intervals for the abundance estimates, and for the pooled variance estimator and the two geostatistical estimators, over all the simulations. The pooled estimator is only shown for strategies for which it is unbiased, 40 to two transects per strata inclusive. The "sample variance" estimator has been omitted from this graph as it is biased for all strategies except the simple random strategy and this is included as the first point for the pooled variance estimator.

Choice of Survey Strategy

These results provide information on the precision of both estimates of abundance and estimates of precision and how these change with survey strategy. One method of utilising this is to use a decision surface to show which survey strategy is optimum, given user defined weights for the allocation of effort to improve either the precision of the abundance estimate or the precision of the variance estimate. The absolute levels of abundance and variance and their precision are very different. A function is required that expresses the relative change in precision for abundance and for variance with changing strategy. One such function can be provided by the normalised 90% interval ($I_{90s}/\bar{I}_{90}$) for both abundance and variance. There are a number of methods of deriving approximately unbiased variance estimates from stratified random strategies with two or more transects per strata. The pooled variance estimator, which has the narrowest 90% interval, has been selected as the best method for these strategies. For strategies with one transect per

strata the geostatistical estimators provided the best estimate of variance and have been selected. The best strategy for any chosen survey objectives, or weight regime, thus corresponds to the maximum of:

$$\left[\frac{1}{W_a\frac{I_{90as}}{\bar{I}_{90a}} + W_v\frac{I_{90vs}}{\bar{I}_{90v}}}\right] \quad \text{where} \quad W_a + W_v = 1$$

where:

I_{90as} is the 90% interval for abundance using strategy s
I_{90a} is the mean 90% interval for abundance for all strategies
I_{90vs} is the 90% interval for variance using strategy s
I_{90v} is the mean 90% interval for variance for all strategies

The decision surface based on our simulations is shown in Figure 12. Thus, for example, if the requirement is to allocate 80% of effort to estimate abundance and 20% to estimating variance then a systematic survey is the optimal strategy. If between 57 and 100% of effort is allocated for estimating abundance, 0 to 43% for variance, then the best strategy is a systematic survey with variance estimated using a geostatistical model. Conversely if between 0 and 57% of effort is allocated for estimating abundance, 43 and 100% for variance, the best strategy is two random transects per strata using pooled variance estimation.

CONCLUSIONS

If a stock distribution can be represented by a statistical distribution within the range examined in this paper, and the mean abundance is estimated by the arithmetic mean of the data values, the best strategies, selected from those examined, are; systematic strategies, when the main aim is abundance estimates, and two transects per strata when the aim is primarily to estimate the precision of the abundance estimate. The statistical models examined included varying proportions of a random process, local positive correlation, non stationary component given by a linear or a low frequency cosine trend, or a simple random walk. In all the simulations the amplitude distribution and local positive autocorrelation were constrained to be similar to those observed on the North Sea herring surveys. The conclusions do not depend on the choice of non-stationary model or on the exact combination of the different components.

REFERENCES

Cressie, N.A.C. 1991. Statistics for Spatial Data. John Wiley and Sons, New York, 900pp.

Matheron G. 1971. The theory of regionalised variables and its application. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau, Fasc. 5, Ecole Nat. Sup. des Mines de Paris.

Simmonds, E.J. and Fryer, R.J. 1992. A Simulation Study of Survey Strategies for Structured Populations. CM 1992/D:24.

APPENDIX A

Surface Generation Method

The aim of the surface generation was to provide amplitude and spatial distributions similar to those observed from the survey data. For this purpose, the following expression was used to derive 400 surface values S_i at points x_i within a surface area X. (For the autoregressive functions 1,000 values were generated and the last 400 used):

$$S_i = T\left(\sqrt{a}p_i + \sqrt{b}q_i + \sqrt{c}r_i + M\right) \quad (11)$$

where: the p_i is independent and normally distributed with mean 0 and variance 1

$$p_i = N(0,1) \quad (12)$$

the q_i is an autoregressive series given by:

$$q_i = \alpha q_{i-1} + N(0,d) \quad (13)$$

in which α is fixed to give a range of autocorrelation of 2.5 transect spacings and d is chosen so that the realised variance of q_i within the surface is 1

r_i is a non stationary component generated by one of 3 different methods.

a) By a simple random walk:

$$r_i = r_{i-1} + N(0,d) \quad (14)$$

in which d is chosen so that the realised variance of r_i within the surface is 1

b) By a linear trend

$$r_i = f + gx_i \quad (15)$$

where f and g are chosen so that r_i has a mean of 0 and a variance of 1

c) By a cosine trend

$$r_i = h\,Cos\left(\frac{3x_i}{2X} - \frac{\pi}{4}\right) \qquad (16)$$

where h is chosen to give a variance of 1

The coefficients a, b and c in equation 14 are varied to combine these components in ten different proportions.

M is 35. This provides a surface with all the required properties except that the amplitude distribution is normal. The function T, found experimentally, transforms the amplitude distribution to one similar to the observed distribution:

$$\begin{aligned} T(z) &= z \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{ for } z > 40 \\ &= 0.00321\,z^2 + 0.4542\,z + 16.66 \ldots\ldots \text{ for } z \le 40 \end{aligned} \qquad (17)$$

The resulting surfaces have properties similar to those of North Sea herring: a mean of 45, a sample variance of 1,200 a range of autocorrelation of 2.5 transect spacings and an amplitude distribution similar to Figure 1. The proportions of spatial variation are illustrated in Figures 3 and 4 for the cosine and linear trend. The match between the observed and simulated amplitude distributions can be seen in Figure 1 where the bars show the observed distribution and the solid line shows the simulated distribution.

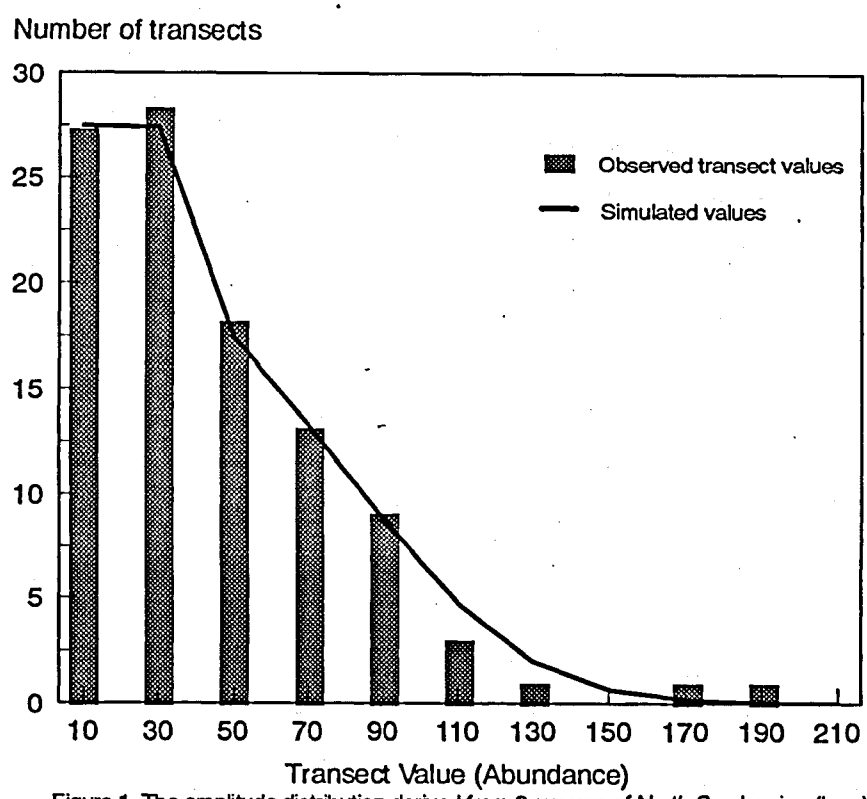

Figure 1. The amplitude distribution derived from 3 surveys of North Sea herring (bars), and the simulated values .

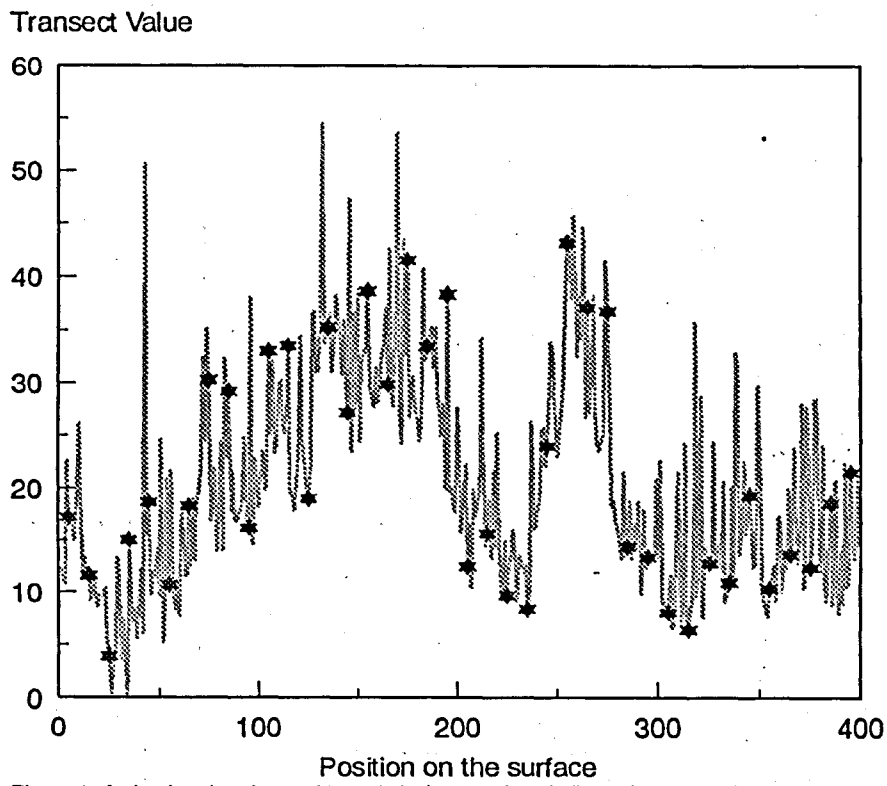

Figure 2. A simulated surface with statistical properties similar to North Sea herring showing sample locations and values from a systematic sampling strategy.

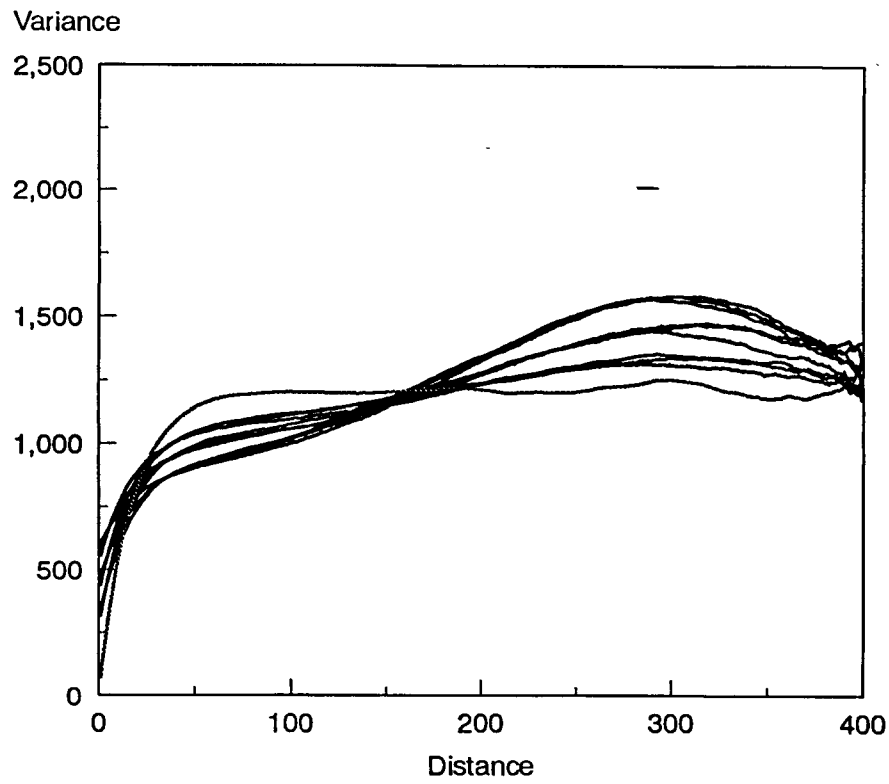

Figure 3. The mean variogram for ten simulations using cosine trend function with varying proportions of nugget and positive local correlation.

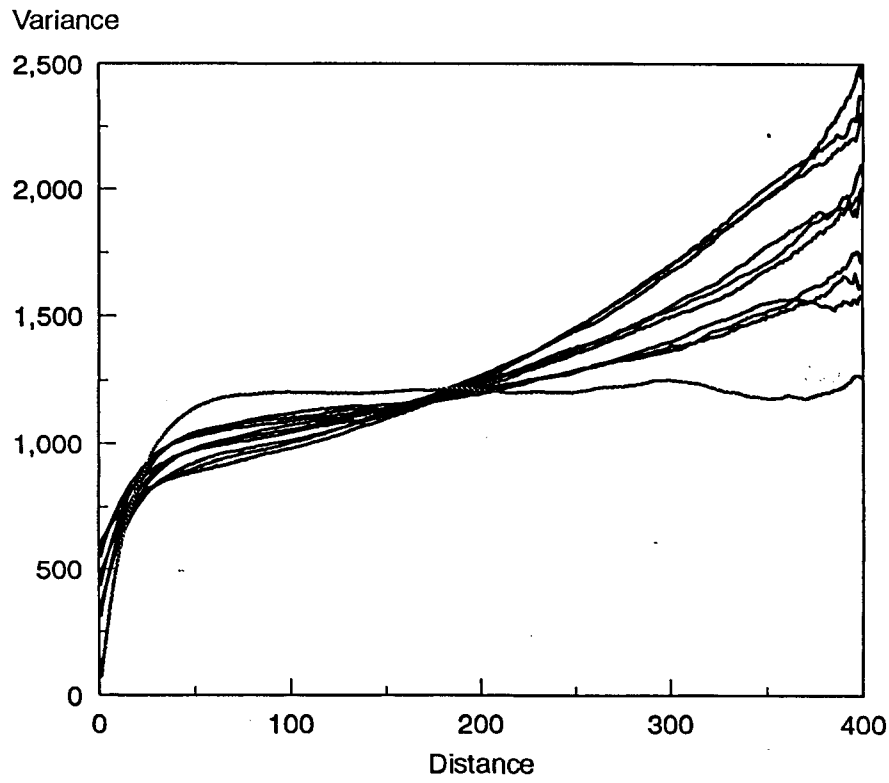

Figure 4. The mean variogram for ten simulations using linear trend function with varying proportions of nugget and positive local correlation.

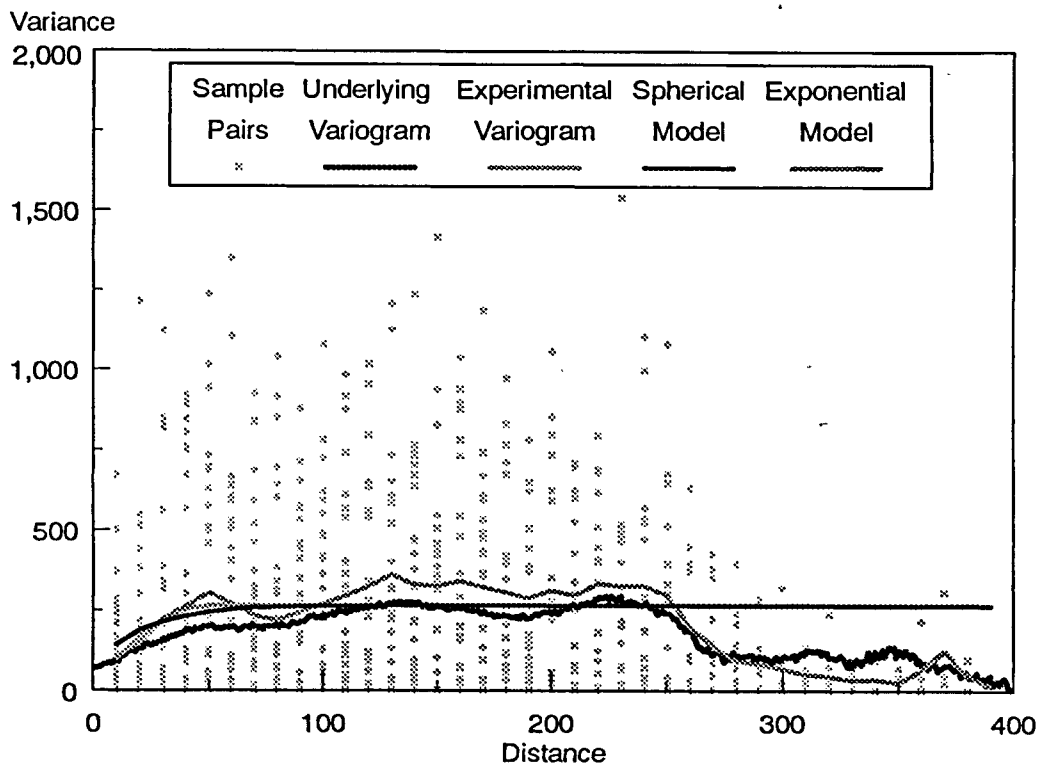

Figure 5. The experimental variogram compared with variance values computed from sample pairs, the variogram of the underlying surface and models fitted by least squares.

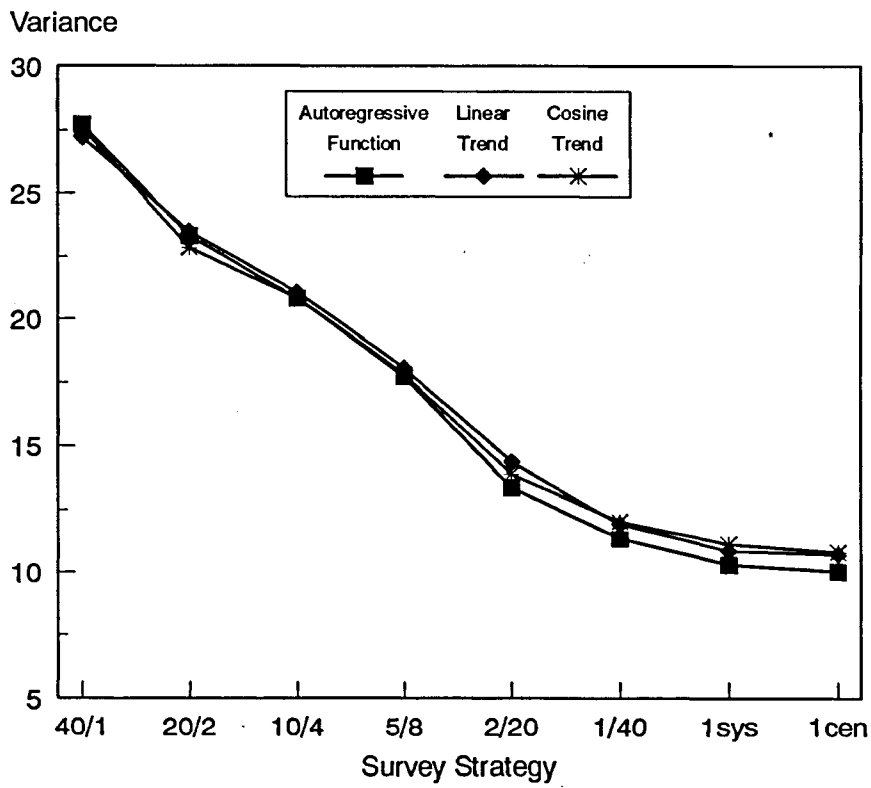

Figure 6. The error variance calculated from sample means for three different methods generating non-stationary components.

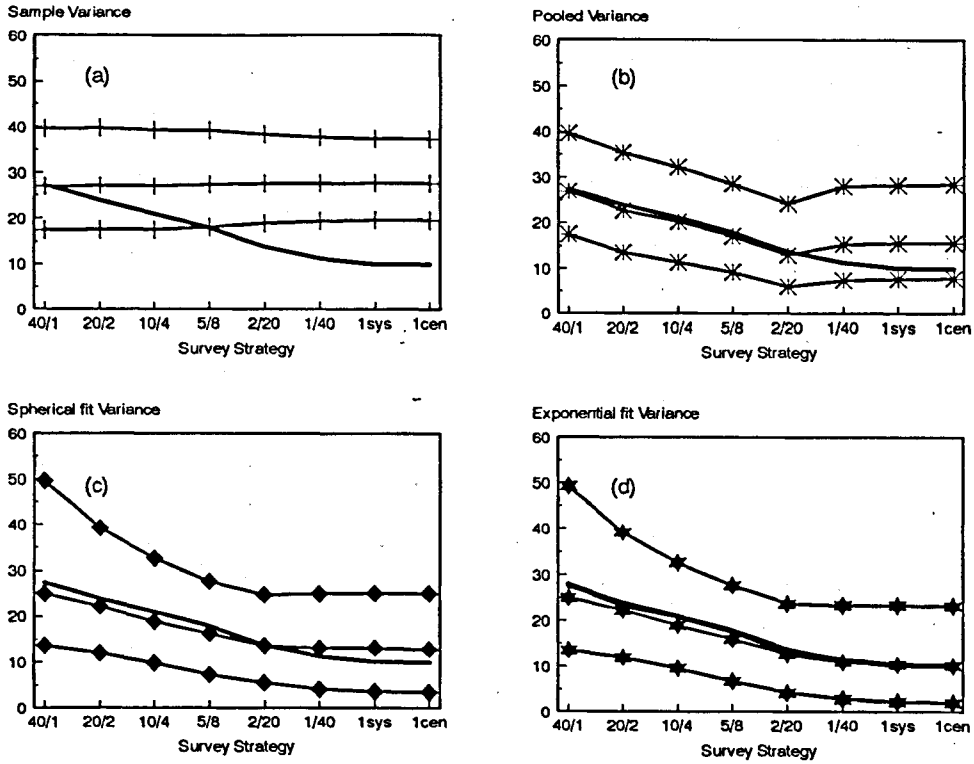

Figure 7. The true error variance (thick line) and the median and 90% intervals for 4 variance estimators (a) sample , (b) pooled, (c) spherical model and (d) exponential model variances for cosine trend.

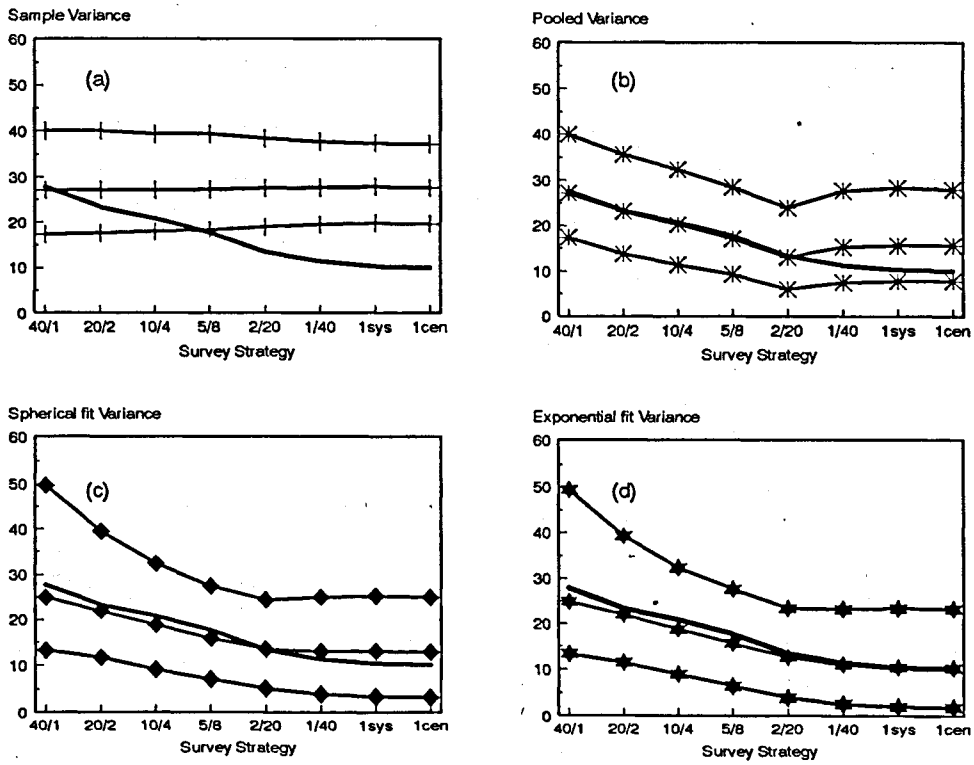

Figure 8. The true error variance (thick line) and the median and 90% intervals for 4 variance estimators (a) sample , (b) pooled, (c) spherical model and (d) exponential model variances for linear trend.

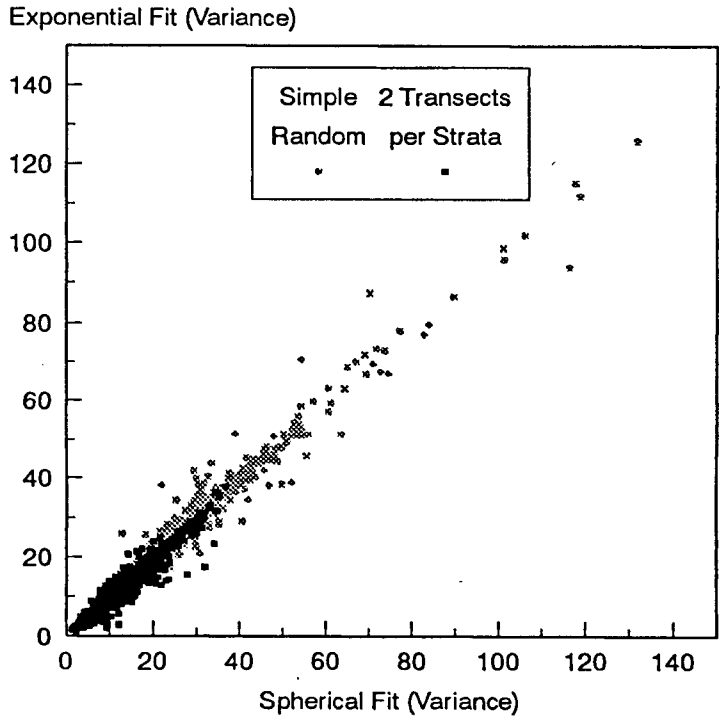

Figure 9. Scatter plot of variance estimates by two geostatistical models for simple random strategy and two transects per strata.

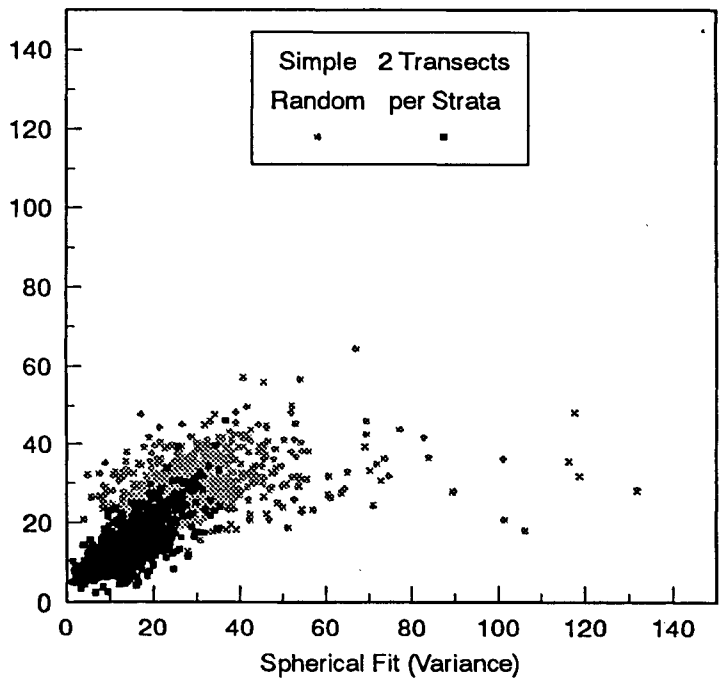

Figure 10. Scatter plot of variance estimates by geostatistical model and pooled variance for simple random strategy and two transects per strata.

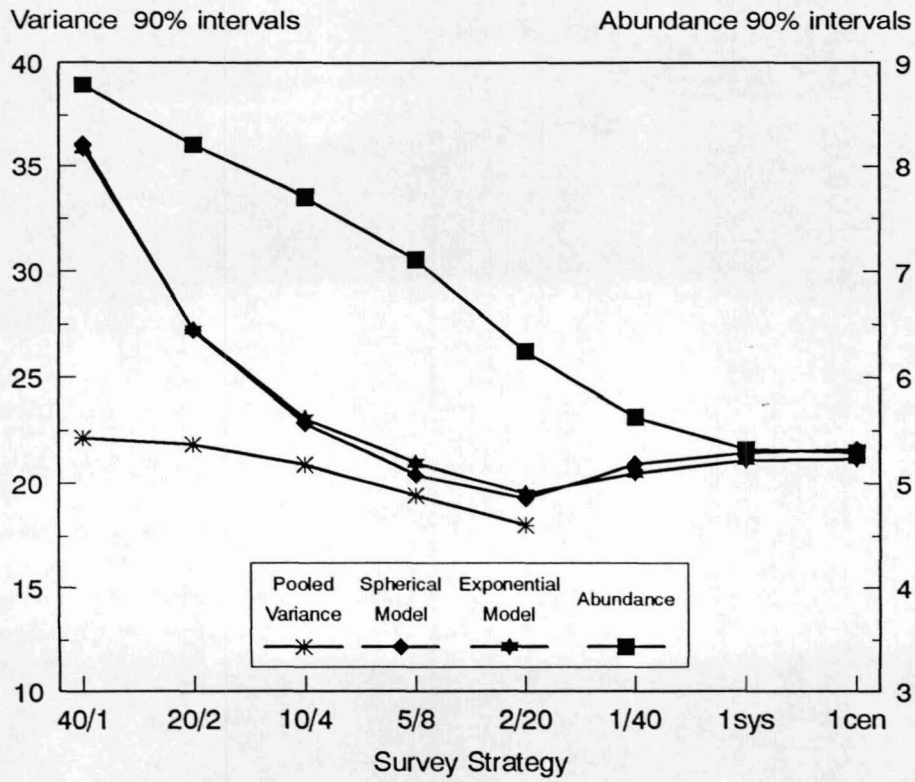

Figure 11. 90% confidence intervals for abundance estimation and variance estimation using pooled variance, exponential and spherical model estimators.

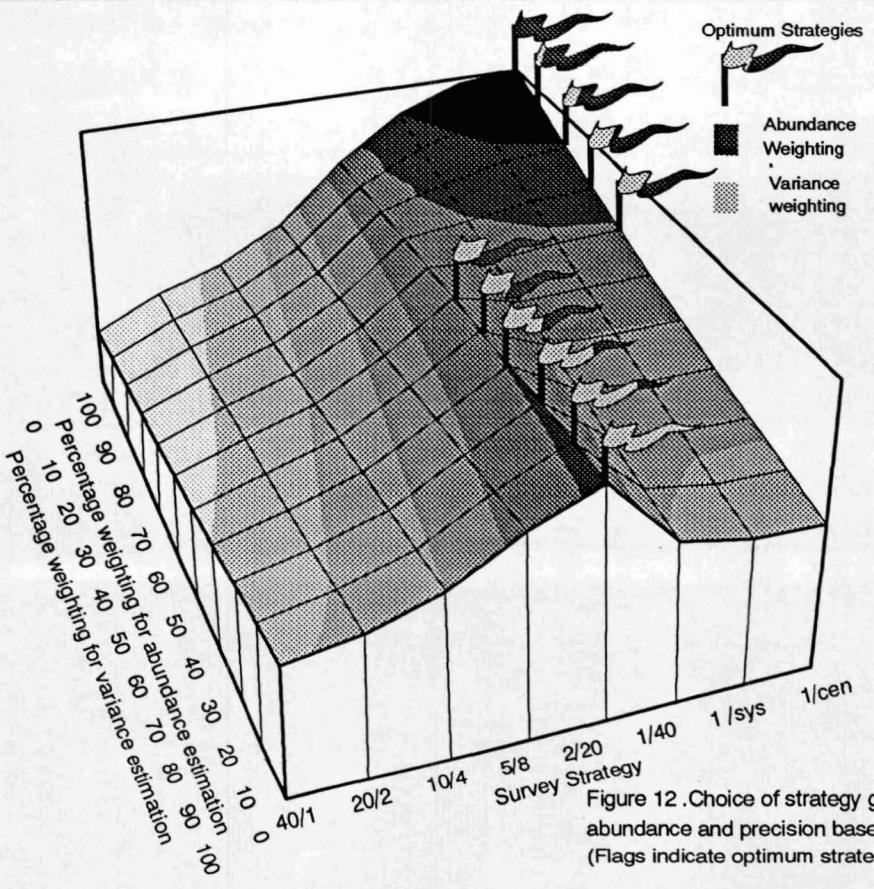

Figure 12 .Choice of strategy given weighting for abundance and precision based on survey objectives (Flags indicate optimum strategies)