Mathematical Geology, Vol. 24, No. 7, 1992

Letter to the Editor

Design-Based Versus Model-Based Sampling Strategies: Comment on R. J. Barnes' "Bounding the Required Sample Size for Geologic Site Characterization"

Two fundamentally different sources of randomness exist on which design and inference in spatial sampling can be based: (a) variation that would occur on resampling the same spatial population with other sampling configurations generated by the same design, and (b) variation occurring on sampling other populations, hypothetically generated by the same spatial model, using the same sampling configuration. The former leads to the design-based approach, which uses classical sampling theory; the latter leads to the model-based approach and uses geostatistical theory. Failure to recognize these two sources of randomness causes misunderstanding about dependence of variables and the role of randomization in sampling, unwarranted narrowing down the choice of sampling strategies to those that are model-based, and abuse in simulation experiments. This is exemplified in Barnes' publication on the required sample size for geologic site characterization by nonparametric tolerance intervals. A basic design-based strategy like Simple Random Sampling is shown to require smaller sample sizes than the model-based strategy advocated by Barnes. In addition, Simple Random Sampling is completely robust against model errors and less complicated.

KEY WORDS: nonparametric tolerance interval, design-based sampling, model-based sampling, spatial dependence, sampling strategy.

INTRODUCTION

Barnes (1988) suggested a heuristic method to calculate the required sample size when the objective of sampling lies not in estimating a spatial average across a geologic site, but in establishing a nonparametric tolerance interval. Tolerance intervals are useful in site characterization "to minimize the chance of unknown and unexpected extremes" (Barnes, 1988). For independent, identically distributed random variables, the probability that a sample of size N covers the β percentile is given by:

$$\Pr(\text{maximum of } N \text{ samples} \geq \beta \text{ percentile}) = 1 - \beta^N \qquad (1)$$

This equation can be solved for the minimum sample size required to achieve a given acceptable probability of coverage P:

$$N_{\text{low}} = \log(1 - P)/\log(\beta) \qquad (2)$$

0882-8121/92/1000-0859$06.50/1 © 1992 International Association for Mathematical Geology

Barnes argued that (i) the classical formula (1) cannot be used for site characterization, because spatial data are correlated, and that (ii) the required sample size exceeds the size required with independent data, because spatial data usually show positive correlation.

We argue that the statements are not universally true. A counter example is as follows. If the site is sampled at completely random points, say $X_1 \ldots X_N$, then the observed values, say $z(X_1) \ldots z(X_N)$, are independent and identically distributed random variables (because $X_1 \ldots X_N$ are independent), regardless of the spatial variation of the property (De Gruijter and Ter Braak, 1990). It follows that the classical sample size formula for independent, identically distributed random variables can validly be applied. Thus, the validity of application hinges on the sampling design; the spatial structure of the property is immaterial. This example counters both statements. It is even possible to devise sampling designs that require less sampling points, instead of requiring more as Barnes (1988) lets us believe.

After summarizing Barnes' heuristic, we show that his experiment to test it is erratic in a way that illustrates the misunderstanding about the application of classical sampling theory to spatial samples as discussed by De Gruijter and Ter Braak (1990).

BARNES' HEURISTIC METHOD

To account for the effect of spatial correlation, Barnes defined the equivalent number of uncorrelated samples, N_{eq}, such that

$$\Pr(\text{maximum of } N \text{ correlated samples} \geq \beta \text{ percentile}) = 1 - \beta^{N_{\text{eq}}} \tag{3}$$

This yields the design requirement:

$$N_{\text{eq}} \geq N_{\text{low}} = \log(1 - P)/\log(\beta) \tag{4}$$

Barnes established an upper bound for N_{eq}:

$$N_{\text{eq}} \leq N \tag{6}$$

The numbering of equations follows Barnes (1988). Further, he proposed a heuristic method to estimate N_{eq} from the data. This method consists of two steps. First, an effective sample size is calculated according to

$$N_{\text{eff}} = \mathbf{1}'\mathbf{C}^{-1}\mathbf{1} \tag{9}$$

where $\mathbf{C}$ denotes the sample-to-sample correlation matrix and $\mathbf{1}$ denotes the N-vector of ones. Then the equivalent sample size is estimated via the empirical relationship

$$N_{\text{eq}} \approx N_{\text{eff}} \cdot \exp(1 - N_{\text{eff}}/N) \tag{11}$$

established from simulated values of N_{eff} and N_{eq}. This equation actually fits the data not nearly as well as the curve in Barnes' Fig. 1 suggests; curiously, the curve drawn does not represent Eq. (11).

COMMENT ON BARNES' CRITICAL EXPERIMENT

Barnes (1988) tested his method to determine coverage probabilities via a simulation experiment with random subsampling from existing data sets. He described his experiment as follows:

> Step (1) Select a data set of interest, a percentile of interest (β), and a subset size N (less than 10% of the original data set). (The data set must be large enough to enable accurate estimation of the variogram and the true β percentile of the underlying distribution.)
>
> Step (2) Randomly select a subset of size N from the original data set.
>
> Step (3) Using the "known" variogram, Eq. (9), Eq. (11), and Eq. (3) calculate and save p_i for the current subset, where
>
> $$p_i = 1 - \beta^{N_{\text{eq}}}$$
>
> (The p_is are heuristic assessments of the probability that the largest of the N samples is greater than the β percentile of the distribution.)
>
> Step (4) Determine the largest value included in the selected subset and compare it to the "known" β percentile. Let T_i equal 1 if the largest value is greater than the β percentile and 0 otherwise.
>
> Using the original data set, repeat Step (2) through Step (4) M times, counting the number of cases where the largest subset value is greater than the β percentile.

Barnes tested his method on seven data sets, with $\beta = 0.95$ and $M = 1000$. He argued that, if the method is correct, the sum of the T_is follows an approximately normal distribution with known mean and variance, and calculated from this the expected counts and Z scores as reproduced in column 5 and 6 of Table 1. Finally Barnes concluded: "The results (Table 1) demonstrate in all cases that the N_{eq} concepts prove satisfactory; that is, they appear to comprise a useful tool."

It cannot be denied that the expected counts are close to the observed counts. Our criticism is different. We claim that the observed counts are inconsistent with the supposed sampling design used. The crucial point is that subsets of size N are said to be randomly selected from the original data set (Step 2). With no other qualification this has to be interpreted as Simple Random Sampling with or without replacement (the difference is immaterial here because the sample size (N) is less than 10% of the original data set). However, under Simple Random Sampling with replacement the observed values, say $z(X_1), \ldots z(X_N)$, are independent and identically distributed random variables (because the locations $X_1, \ldots X_N$ are independent), regardless of the spatial variation of the property. This follows, for instance, from Theorem 6A in Parzen (1960, p.

Table 1. Barnes' Experiment Re-examined

				Barnes (1988)		Simple random sampling	
Case	Sample size	Subsample size	Actual count[a]	Exp. count	Z score	Exp. count	Z score[b]
A	250	25	546	547	−0.08	723	−12.5
B	203	20	467	474	−0.46	642	−11.5
C	248	24	514	535	−1.42	708	−13.5
D	305	30	605	600	+0.34	785	−13.9
E	154	15	415	401	+0.96	537	−7.7
F	186	18	465	444	+1.30	603	−8.9
G	109	10	267	286	−1.34	401	−8.7

[a] Number of times out of 1000 that the maximum in the subsample is greater than the 95-percentile of the sample.

[b] Z score = (actual − expected)/(standard deviation).

295): "Let the random variables Y_1 and Y_2 be obtained from the random variables X_1 and X_2 by some functional transformation, so that $Y_1 = g_1(X_1)$ and $Y_2 = g_2(X_2)$ for some Borel functions $g_1(.)$ and $g_2(.)$ of a real variable. Independence of the random variables X_1 and X_2 implies independence of the random variables Y_1 and Y_2." See also Theorem 2 in Ash (1970, p. 84).

It follows that the classical formula for independent, identically distributed random variables (Eq. 1) can validly be applied. The last two columns of Table 1 show the expected counts and Z scores on the basis of formula (1). Clearly, Barnes' observed counts are inconsistent with the counts expected under random sampling (all Z-scores are less than -8), hence the results of the experiment are inconsistent with the description of the experiment. Said simply, Barnes did not use Simple Random Sampling. We emphasize that if, for instance, a special purposive or cluster sampling technique was used this should have been mentioned explicitly and specified in detail, in view of the large effect this apparently has on the results and on the conclusion cited before.

COMMENT ON BARNES' PROPOSED SAMPLING STRATEGY

Barnes incorporated his method to estimate N_{eq} in a two-phase model-based sampling strategy, described as follows:

Step (1) Considering the risk economics of the question at hand, select an appropriate percentile β and probability of coverage P.

Step (2) Using Eqs. (6) and (4), calculate the lower bound on the number of samples required, N_{low}.

Step (3) Using all available information, locate and collect N_{low} samples. This will be called Phase I sampling.

Step (4) Using the N_{low} samples collected during Phase I, estimate the variogram for the site.

Step (5) Solicit Phase II candidate sampling plans in an ordinary manner but, using the estimated variogram Eqs. (9) and (11), select only from the plans which satisfy Eq. (4).

This strategy is inferior to the basic design-based strategy of Simple Random Sampling combined with Eq. (1), for the following reasons.

1. Barnes' strategy would normally require a considerably larger sample size for the same coverage probability. Under Simple Random Sampling not more than $\log(1 - P)/\log(\beta)$ samples are needed (Eq. 2), where P denotes the required probability of coverage. This is the lower bound of the number required by Barnes' strategy.

2. Barnes' strategy is approximative only and liable to impairment by model errors, whereas Simple Random Sampling with Eq. (1) is exact and independent of any model.

3. Barnes' strategy is much more complicated.

In the class of design-based strategies it is possible to reduce the sample size as required with Simple Random Sampling even further by employing stratification, grid sampling, and other variance-reduction techniques. Sedransk and Smith (1988) discuss this for the related problem of quantile estimation. Classical papers on the sample size problem for distribution-free tolerance limits are those of Wilks (1941), Scheffé and Tukey (1944), and Murphy (1948). Sample size for other types of tolerance limits has been discussed by Faulkenberry and Weeks (1968), Miller (1989), and Odeh et al. (1989).

CONCLUDING REMARKS

It is worth noting that interpretation of the coverage probability in design-based strategies differs from that in model-based strategies. The former is the probability that the largest sample value exceeds a given population percentile under repeated sampling according to given design. The latter is the probability of exceeding the percentile in a set of fixed sample points, for a random realization of the postulated spatial model. We feel that in the context of site characterization the design-type of coverage probability may be more valuable than the model type. In addition, is seems appropriate to have statistical control over selection of sample locations by a suitable form of randomization. Without this precaution even unconscious personal preferences may lead to significantly biased results, as has been repeatedly demonstrated (Yates, 1935).

ACKNOWLEDGMENT

We thank Dr. A. A. M. Jansen and Dr. M. J. W. Jansen (Agricultural Mathematics Group) for their valuable comments on previous versions of this paper.

REFERENCES

Ash, R. B., 1970, Basic Probability Theory: Wiley, New York, 337 p.

Barnes, R. J., 1988, Bounding the Required Sample Size for Geologic Site Characterization: Math. Geol., v. 20, p. 477–490.

De Gruijter, J. J., and Ter Braak, C. J. F., 1990, Model-Free Estimation from Spatial Samples: A Reappraisal of Classical Sampling Theory: Math. Geol., v. 22, p. 407–415.

Faulkenberry, G. D., and Weeks, D. L., 1968, Sample Size Determination for Tolerance Limits: Technometrics, v. 10, p. 343–348.

Miller, R. W., 1989, Parametric Empirical Bayes Tolerance Intervals: Technometrics, v. 31, p. 449–459.

Murphy, R. B., 1948, Non-Parametric Tolerance Limits: Annals of Mathematical Statistics, v. 19, p. 581–588.

Odeh, R. E., Chou, Y. M., and Owen, D. B., 1989, Sample-Size Determination for Two-Sided β-Expectation Tolerance Intervals for a Normal Distribution: Technometrics, v. 31, p. 461–468.

Parzen, E., 1960, Modern Probability Theory and Its Applications: Wiley, New York, 464 p.

Scheffé, H., and Tukey, J. W., 1944, A Formula for Sample Size for Population Tolerance Limits: Ann. Math. Stat., v. 15, p. 217.

Sedransk, J., and Smith, P. J., 1988, Inference for Finite Population Quantiles, in P. R. Krishnaiah and C. R. Rao (Eds.), Sampling. Handbook of Statistics, Vol. 6: North-Holland, Amsterdam, p. 267–289.

Wilks, S. S., 1941, Determination of Sample Sizes for Setting Tolerance Limits: Ann. Math. Stat., v. 12, p. 91–96.

Yates, F., 1935, Some Examples of Biased Sampling: Ann. Eugen., v. 6, p. 202–213.

J. J. de Gruijter
The Winand Staring Centre for Integrated Land, Soil and Water Research
P.O. Box 125
6700 AC Wageningen
The Netherlands

C. J. F. ter Braak
Agricultural Mathematics Group
P.O. Box 100
6700 AC Wageningen
The Netherlands