Mathematical Geology, Vol. 22, No. 4, 1990

Model-Free Estimation from Spatial Samples: A Reappraisal of Classical Sampling Theory[1]

J. J. de Gruijter[2] and C. J. F. ter Braak[2]

A commonly held view among geostatisticians is that classical sampling theory is inapplicable to spatial sampling because spatial data are dependent, whereas classical sampling theory requires them to be independent. By comparing the assumptions and use of classical sampling theory with those of geostatistical theory, we conclude that this view is both false and unfortunate. In particular, estimates of spatial means based on classical sampling designs require fewer assumptions for their validity, and are therefore more robust, than those based on a geostatistical model.

KEY WORDS: Fixed population, superpopulation, design-based inference, model-based inference, spatial dependence, p-unbiasedness, sampling strategy.

INTRODUCTION

"Classical techniques cannot be used because they are based on independence of sample data" (Yfantis et al., 1987).

These and similar statements have been made time after time in geostatistical literature (Russo and Bresler, 1981; Dahiya et al., 1985; Barnes, 1988). Our impression is that many geostatisticians now consider this as an elementary fact that need not be reiterated. It seems to be a geostatistical paradigm. But is it true? Our answer is: no. The answer has important practical implications for spatial sampling, as we will show.

Independence is indeed a key assumption in classical statistical inference (developed by R. A. Fisher and others). But no independence assumptions are made in classical *sampling* theory as developed by J. Neyman and others. This is explained in the sequel.

Modern textbooks on sampling (e.g., Cassel et al., 1977; Krishnaiah and Rao, 1988) distinguish two fundamentally different approaches that can be followed in spatial inference about population parameters. Following Särndal (1978) we refer to these as the design-based and the model-based approach.

[1]Manuscript received 8 May 1989; accepted 6 November 1989.

[2]Agricultural Mathematics Group, P.O. Box 100, 6700 AC Wageningen, The Netherlands.

0882-8121/90/0500-0407$06.00/1

Synonyms are, respectively, fixed population and superpopulation approach (see also Chaudhuri and Vos, 1988).

The design-based approach uses classical sampling theory. The central notion in classical sampling theory is the population, which is defined as the set of all units of interest. In geostatistics, the population is, for example, a region A or, more precisely, the set of all possible sampling locations in the region of interest. In the design-based approach, a *fixed* value is associated with each unit. Sampling consists of selecting a subset of units. Measuring the characteristic of interest reveals the value of each of them. In the extreme case of a sample covering the entire population and errorless measurement, there would be nothing more to know. Excluding measurement error, the only variation that plays a role in this approach is that resulting from the sampling process. Inference is therefore primarily based on the sampling design used: the design defines the probability of including any given set of units in the sample and thus enables valid inference procedures without any additional assumptions.

The model-based approach, when applied in a spatial context, boils down to the use of geostatistical theory, notably the part of it concerned with predicting spatial averages. This approach treats the value associated with any given location not as fixed but as *random*. The set of values associated with all possible locations in the region of interest is thus considered as just one realization of an underlying random process. At least some features of this process are assumed known and these assumptions are formalized in a geostatistical model. This model plays the role of what is called a superpopulation model in sampling theory. If, again in the extreme case, a sample would cover the entire region, one realization would be completely known, but uncertainty about the parameters in the model would remain. The models used in applications typically account for spatial structure, reflecting the fact that measurements at near locations often tend to yield more similar values than when taken farther apart.

In the model-based approach, locations need not be selected at random. They typically aren't. The only source of stochasticity is then the postulated underlying process. In this approach inference is therefore primarily based on the model formulated. The nature of the stochasticity involved in the model-based approach is thus fundamentally different from that in the design-based approach where, as we have seen, it originates from a physical sampling process. The latter is in our hands. The design-based approach thus requires fewer assumptions than the model-based approach. It is therefore advantageous with respect to robustness to use the design-based approach whenever possible.

The purpose of this article is to clarify the distinction between the two approaches. Failure in recognising the difference can easily lead to misunderstanding and fallacy. We discuss the matter in a rather informal way, using the terminology and some of the main concepts of Cassel et al. (1977), who developed a theoretical framework encompassing both approaches. Point-sampling in the plane is used as an illustration.

The present paper is wholly confined to inference about population parameters such as the spatial mean, and hence is not related to spatial prediction of values at individual locations or to contour-mapping. For instance, a spatial mean can be predicted in geostatistics by block-kriging with the entire region of interest as a single block.

A CLOSER LOOK AT THE DESIGN-BASED AND THE MODEL-BASED APPROACH

To clarify the difference between the design-based and the model-based approach we give a hypothetical example using planar point-sampling. Suppose an errorless measurement of a quantity z is taken at n points in a region A. The values obtained will be denoted by $z(x_i)$, where x_i is the vector of coordinates of the i^{th} sampling point. Either the value or the coordinates (or both) may be random variables. To distinguish between random and fixed components we use the convention to write random variables in uppercase and nonrandom in lowercase.

In the design-based approach, the n sample points are randomly selected according to a sampling design p, and the values are fixed. The variables involved can thus be written as $z(X_i)$ $(i = 1, \ldots, n)$. They represent random variables because the locations (X_i) are random. Whether $z(X_i)$ and $z(X_j)$ for $i \neq j$ are stochastically independent or not is completely determined by the sampling design and not by the spatial variation in A.

This simple fact seems to have been overlooked by some users of geostatistical techniques. Specifically, if measurements are taken at locations which are selected at random *and* also independently from each other, the corresponding stochastic variables are always mutually independent, not by model-assumption but by design of the observational process. In the example this could be verified empirically by repeated Simple Random Sampling with $n = 2$ (i.e., by selecting independently two points from the uniform probability density over A), thus producing a long series of pairs of values. If the series is long enough, the scatter diagram of the first value against the second will show no dependency, regardless of the spatial variation in A. Dependence between variables in the design-based approach need not be *disregarded*: it can be *avoided*.

Suppose that the quantity of interest is the spatial mean:

$$m(A) = \int_A z(x)\, dx \Big/ \int_A dx \tag{1}$$

The usual estimator of $m(A)$ in combination with Simple Random Sampling (srs) is the unweighted sample mean,

$$T_{um} = \frac{1}{n} \sum_{i=1}^{n} z\,(X_i) \tag{2}$$

The sampling strategy (srs, T_{um}) is said to be p-unbiased, since

$$E_P\,(T_{um}) = m(A) \tag{3}$$

where E_p denotes expectation over repeated sampling under the given design. Because of this p-unbiasedness, the p-MSE of T_{um} equals its p-variance (e.g., Cochran, 1977):

$$E_p\left[T_{um} - m(A)\right]^2 = V_p\,(T_{um}) = v(A)/n \tag{4}$$

where $v(A)$ denotes the variance between points in A, defined as:

$$v(A) = \int_A \left[z(x) - m(A)\right]^2 dx \Big/ \int_A dx \tag{5}$$

In the terminology of classical sampling theory, the set of all locations in region A is the population of interest. The locations are termed the units of this (infinite) population; the spatial mean $m(A)$ and the spatial variance $v(A)$ are parameters of this population.

To increase efficiency and/or to facilitate the fieldwork, other sampling designs may be called for. The formulae given here for Simple Random Sampling can be easily generalized to other classical sampling designs, such as Stratified Random Sampling (Cochran, 1977), which also enable model-free inference.

The model-based approach starts with a random function ξ that generates random values over A. Sampling at fixed points thus yields in our notation the random variables $Z(x_i)$ $(i = 1, \ldots, n)$. The uppercase for Z stresses that the values are random; the lower case for x_i that the locations are fixed. Generally, $Z(x_i)$ and $Z(x_j)$ for $i \neq j$ are stochastically dependent as determined by the random function ξ, which is partly specified by the assumed geostatistical model.

Suppose that ξ is second-order stationary in A. This means that $E_\xi Z(x)$ exists and does not depend on x, and that $\mathrm{Cov}[Z(x), Z(x + h)]$ exists and depends only on h (see Myers, 1989, for a discussion of this and other forms of stationarity used in geostatistics):

$$E_\xi\, Z(x) = \mu \tag{6a}$$

$$\begin{aligned}\mathrm{Cov}\left[Z(x), Z(x + h)\right] &= E_\xi\left[Z(x) \cdot Z(x + h)\right] - \mu^2 \\ &= C(h)\end{aligned} \tag{6b}$$

where E_ξ denotes expectation over realizations from ξ. Suppose further that the model mean μ is unknown and that the covariance function $C(h)$ is known, with model variance $C(0)$. In applications, the covariance function or its equivalent, the semivariogram function, usually has to be estimated from the data too, but this is immaterial in the present discussion.

In the model-based approach, inference about central tendency can be directed to at least two different target quantities: (a) the model mean μ of Eq. 6a, or (b) the spatial mean $M(A)$, defined as,

$$M(A) = \int_A Z(x)\, dx \Big/ \int_A dx \tag{7}$$

The model mean is fixed and can only be *estimated*. The spatial mean, however, being the mean of a realization from ξ, is in this approach a random variable and can be *predicted*.

The quality criterion often used in model-based inference is the ξ-MSE. Applied to any estimator $\hat{\mu}$ of the model mean, this is defined as

$$\xi\text{-MSE}(\hat{\mu}) = E_\xi\,(\hat{\mu} - \mu)^2 \tag{8}$$

Minimizing this criterion while confining $\hat{\mu}$ to linear combinations of the sample data leads to the well-known ξ-BLU estimator $\tilde{\mu} = \lambda_\mu' Z_S$ with

$$\lambda_\mu' = (1' C_S^{-1} 1)^{-1}\, 1'\, C_S^{-1} \tag{9}$$

where λ_μ, Z_S, and 1, respectively, denote the vector of optimal weights, the vector of Z-values at the sample points, and the vector of one's (all n-dimensional if the sample size is n), and C_S denotes the corresponding $n \times n$ matrix of covariances between the sample points. The ξ-variance of $\tilde{\mu}$ is:

$$V_\xi\,(\tilde{\mu}) = (1' C_S^{-1} 1)^{-1} \tag{10}$$

Similarly, the ξ-BLU predictor of $M(A)$ is another weighted mean, $T_{wm} = \lambda_Z' Z_S$, now with weights

$$\lambda_Z = \lambda_\mu + (C_S^{-1} - \lambda_\mu\, 1' C_S^{-1})\, \bar{C}_{S,A} \tag{11}$$

where $\bar{C}_{S,A}$ denotes the n-dimensional vector of mean covariances between each sample point and all points in A. T_{wm} has variance

$$V_\xi\,(T_{wm}) = \bar{C}_{A,A} + \lambda_Z' C_S \lambda_Z - 2\lambda_Z' \bar{C}_{S,A} \tag{12}$$

where $\bar{C}_{A,A}$ denotes the mean covariance between all pairs of points in A. Both $\tilde{\mu}$ and T_{wm} are ξ-unbiased because, for any given set of sample points,

$$E_\xi\,(\tilde{\mu} - \mu) = E_\xi\,[T_{wm} - M(A)] = 0 \tag{13}$$

As demonstrated by Corsten (1989), explicit expressions as above for the BLUE, BLUP, and their variances can be obtained as Generalized Least Squares solutions of the corresponding regression problems. This avoids the use of Lagrange multipliers. The same expressions can also be obtained by substitution of the Lagrange multipliers in the usual formulae (in, e.g., Matheron, 1960, Secs. 3–4; Journel and Huijbregts, 1978, Ch. V).

If not only the weights, given the locations, but also the locations themselves are to be optimized with respect to the ξ-MSE criterion, this typically leads to purposive sampling rather than some form of random sampling. The optimal design (p_ξ) for estimating μ will differ in general from the optimal design (p_M) for predicting $M(A)$. It is important to note here that the strategy (p_M, T_{wm}) is ξ-unbiased but not p-unbiased, whereas, e.g., (srs, T_{um}) and (srs, T_{wm}) are both ξ-unbiased and p-unbiased for $M(A)$.

The example illustrates points that are relevant to the choice of a sampling strategy.

1. Model-based inference may be directed at two different types of quantities (in our example, model means and spatial means) which differ in interpretation, optimal design, and optimal estimator/predictor, and which have different variances (Eqs. 10 and 12). Model quantities may be of interest in process-oriented studies, whereas spatial means are more relevant to inventory studies. It is important to be entirely clear about what type of quantity the study is aimed at.

2. Stochastic dependence or independence is not a property of any population or, in our case, of any region. It can be a property of a set of variables and is either induced by a sampling design or implied by a model. Confusion and misunderstanding about independence may have reached undesirable proliferation in geostatistical literature. Misleading statements in this respect are, for instance:

> The conventional statistical approach to describe variability of soil hydraulic properties treats the observations of a given property as being statistically independent regardless of their spatial position. (Russo and Bresler, 1981)

> In Part I (Dahiya et al., 1984a), spatial variability of some nutrient constituents, viz. NO_3, K, Mg, and organic C, of a loess soil field was evaluated by applying classical statistical analysis (i.e., probability density function, mean, and variance). An implicit assumption in that analysis is that the observations of a given soil property are independent of one another regardless of their location in the field. (Dahiya et al., 1985)

> The classic development of nonparametric tolerance intervals begins with an assumption of independent, identically distributed random variables. This is unrealistic for the geologic environment—in general, geologic site characterization data are not independent. (Barnes, 1988)

The truth is that the design-based approach with its classical sampling theory cannot be dismissed for invalidity of model assumptions because it hasn't any, but the model-based approach may.

3. The selection of a sampling strategy is primarily the selection of a quality criterion (Särndal, 1978). Such a criterion may relate to different sources of stochasticity: a sampling design or a model of the spatial variation or both. (Expressions like "minimum variance" and "unbiasedness" may therefore be ambiguous when used without further qualification.) For example, if the ξ-MSE

criterion is adopted, a model-based strategy will be chosen. On the other hand, if p-unbiasedness is judged necessary, a design-based strategy with some form of random sampling will be the solution.

We doubt whether users of results from geostatistical methods are generally aware of the fact that ξ-unbiasedness is only an "internal" guarantee inasmuch as it relates to a model, not to the particular region investigated. It therefore covers by no means the consequences of biased sampling as might arise, for instance, from selecting locations with expectedly high (or low) values or with best accessability, or from interference of a regular sampling pattern with cyclic variation in the region.

Such sampling will bias the results of any subsequent inference to an unknown degree. This is of special concern in the context of quality control and legislation with regard to natural resources.

CONCLUSION

By ruling out classical sampling theory for its supposed invalid assumption of independence, the choice of strategies in spatial sampling is improperly narrowed to those that are model-based. The design-based and model-based approaches clearly have their own pros and cons. Design-based strategies are inapplicable if for some reason probability sampling is impracticable. The model-based approach is inapplicable if reliable identification of a model is prevented by lack of data.

If both approaches are practicable, the primary choice to be made is that of the quality criterion to judge alternative strategies. In practical terms, one type of criterion relates to what would happen if sampling is repeated in the same region, but with other sampling configurations. This type leads to design-based strategies. The other type of criterion relates to repeated sampling with the same configuration, but in other regions. This yields model-based strategies. Geostatistical models may form a natural basis for inference in situations where part of the region is inaccessible, or measuring has been censored or impaired by systematic error (see Laslett and Sandland, 1989). Furthermore, if data are available from the vicinity of the region these can be used via the model-based approach, as in block-kriging. This is especially relevant if the region is sparsely sampled.

The introduction of randomness in a sampling design may lower its efficiency compared to the optimal model-based strategy when the model is correct and the true covariance function is known. However, this loss of efficiency can often be reduced by choosing sensible restrictions on the randomization of the sample locations (e.g. by stratification of the region). Much of classical sampling theory is devoted to this. The remaining loss might then well be a worthwhile premium for robustness against model errors and for p-unbiasedness.

Moreover, model-based strategies depend on *estimated* covariances or variograms in practice and therefore will not be optimal either.

In soil science the design-based approach to estimate population parameters has been succesfully practicised for many decades. Webster (1977) discussed the usefulness of various sampling designs, including nested sampling, in pedology. For a study on a specific design with random transects see De Gruijter and Marsman (1985).

Finally, it should be noted that, apart from serving directly as a basis for inference and design, spatial models are frequently employed by sampling theorists to evaluate design-based strategies. This hybrid approach uses quality criteria of the ξp-type, based on expectations over repeated sampling under p and realizations from ξ. See Cochran (1946), Quenouille (1949), Das (1950), and Matérn (1960) for early examples. Diggle and Ter Braak (1982) worked out specific cases in ecology. We believe that the practice of survey sampling could greatly benefit from more comparative studies on the qualities of different design-based and model-based strategies, under various models and model deviations.

ACKNOWLEDGMENT

We thank Ir. A. A. M. Jansen, Dr. M. J. W. Jansen (Agricultural Mathematics Group) and Dr. D. E. Myers (University of Arizona) for their valuable comments on previous versions of this paper.

REFERENCES

Barnes, R. J., 1988, Bounding the required sample size for geologic site characterization: Math. Geol., v. 20, p. 477–490.

Cassel, C-M., Särndal, C-E., and Wretman, J. H., 1977, Foundations of inference in survey sampling: Wiley, New York, 192 p.

Chaudhuri, A., and Vos, J. W. E., 1988, Unified theory and strategies of survey sampling: North-Holland, Amsterdam, 414 p.

Cochran, W. G., 1946, Relative accuracy of systematic and stratified random samples for a certain class of populations: Annals of Mathematical Statistics, v. 17, p. 164–177.

Cochran, W. G., 1977, Sampling techniques: Wiley, New York, 428 p.

Corsten, L. C. A., 1989, Interpolation and optimal linear prediction: Statistica Neerlandica, v. 43, p. 69–84.

Dahiya, I. S., Anlauf, R., Kersebaum, K. C., and Richter, J., 1985, Spatial variability of some nutrient constituents of an Alfisol from loess. II. Geostatistical analysis: Z. Pflanzenernaehr. Bodenk., v. 148, p. 268–277.

Das, A. C., 1950, Two-dimensional systematic sampling and the associated stratified and random sampling: Sankhya. v. 10, p. 95–108.

De Gruijter, J. J., and Marsman, B., 1985, Transect sampling for reliable information on mapping units, *in* D. R. Nielsen and J. Bouma (Eds.), Soil spatial variability: Pudoc, Wageningen, p. 150–165.

Diggle, P. J., and ter Braak, C. J. F., 1982, Point sampling of binary mosaics in ecology, *in* B. Ranneby (Ed.), Statistics in theory and practice—Essays in honour of Bertil Matérn: Swedish Univ. of Agricultural Sciences, Section of Forest Biometry, S-901 83 Umea, Sweden, p. 107–122.

Journel, A. G., and Huijbregts, Ch. J., 1978, Mining geostatistics: Academic Press, London, 600 p.

Krishnaiah, P. R., and Rao, C. R., 1988, Sampling—Handbook of statistics: v. 6., North-Holland, Amsterdam, 594 p.

Laslet, G. M., and Sandland, R. L., 1989, Precision and accuracy of kriging estimators with inter-laboratory trial information, *in* M. Armstrong (Ed.), Geostatistics, Vol. 2: Kluwer, Dordrecht, p. 797–808.

Matérn, B., 1960, Spatial variation: Medd. Statens Skogsforskningsinst., v. 49, p. 1–144.

Matheron, G., 1971, The theory of regionalized variables and its applications: Ecole Nationale Supérieure des Mines de Paris, 211 p.

Myers, D. E., 1989, To be or not to be . . . stationary? That is the question: Math. Geol., v. 21, p. 347–362.

Quenouille, M. H., 1949, Problems in plane sampling: Annals of Mathematical Statistics, v. 20, p. 355–375.

Russo, D., and E. Bresler, 1981, Soil hydraulic properties as stochastic processes: I. An analysis of field spatial variability: Soil Sci. Soc. Am. J., v. 45, p. 682–687.

Särndal, C-E., 1978, Design-based and model-based inference in survey sampling: Scand. J. Statist., v. 5, p. 27–52.

Webster, R., 1977, Quantitative and numerical methods in soil classification and survey: Clarendon Press, Oxford, 269 p.

Yfantis, E. A., Flatman, G. T., and Behar, J. V., 1987, Efficiency of kriging estimates for square, triangular, and hexagonal grids: Math. Geol., v. 19, p. 183–205.