ECOSCIENCE 9 (2) : 152-161 (2002)

Sampling and geostatistics for spatial data[1]

Jay VER HOEF, Alaska Department of Fish and Game, 1300 College Road, Fairbanks, Alaska 99701, U.S.A.,
e-mail: jay_ver_hoef@fishgame.state.ak.us

Abstract: The goals of classical statistical sampling (*e.g.* estimation of population means using simple random sampling, stratified random sampling, etc.) and geostatistics (*e.g.* estimation of population means using block kriging) can be identical. For example, both can be used to estimate the average value, or total amount, of a variable of interest in some area. The most fundamental difference between classical sampling and geostatistics is that classical sampling relies on design-based inference while geostatistics relies on model-based inference. These differences are illustrated with examples. Classical sampling usually considers sampling for finite populations, but in the spatial context, it is easily adapted to infinite populations. Geostatistics has only considered infinite populations, but methods for finite populations have been developed recently. To compare classical sampling to geostatistics for both infinite and finite populations, I consider the following data sets: 1) a fabricated fixed spatial pattern from an infinite population of a spatially-continuous variable; 2) a single, fixed, real data set from a finite population on a grid of spatial locations; and 3) simulated random patterns from an autocorrelated model from a finite population on a grid of spatial locations. For each data set, I select samples randomly. Then I use classical sampling estimators and geostatistical estimators of the mean values. Results show that both methods provide unbiased estimates and have variances and confidence intervals that are valid, but in general the geostatistical methods are more efficient, having estimates closer to the true values.
Keywords: block kriging, finite populations, model-based inference, simulations.

Résumé : Les buts poursuivis par l'échantillonnage statistique classique (*e.g.* estimation des moyennes de population en utilisant l'échantillonage aléatoire simple, l'échantillonage aléatoire stratifié, etc.) et la géostatistique (*e.g.* estimation des moyennes de population utilisant le krigeage ordinaire d'un bloc) peuvent être identiques. Par exemple, les deux approches peuvent être utilisées pour estimer la valeur moyenne, ou la quantité totale d'une variable pour un secteur donné. Il existe néanmoins une différence importante entre ces deux méthodes. L'échantillonnage classique repose sur l'inférence basée sur la théorie de l'échantillonnage alors que la géostatistique classique repose sur l'inférence basée sur un modèle. Cette différence est illustrée à l'aide d'exemples. En général, l'échantillonnage classique est approprié pour des populations de taille finie. Toutefois, dans un contexte spatial, il peut être facilement adapté pour l'étude de populations de taille infinie. Jusqu'à tout récemment, la géostatistique analysait uniquement des populations de taille infinie. Il est maintenant possible de les utiliser pour l'étude des populations de taille finie. Afin de comparer l'efficacité de l'échantillonnage classique et de la géostatistique, trois ensembles de données des populations de taille finie et infinie ont été employés : 1) un patron spatial fixe, fabriqué à partir d'une population de taille infinie d'une variable continue au niveau spatial, 2) un ensemble de données réelles issu d'une population de taille finie examinée grâce à une grille de localisations spatiales et 3) des patrons aléatoires simulés à l'aide d'un modèle autocorrélé d'une population de taille finie, elle aussi examinée grâce à une grille de localisations spatiales. J'ai choisi, de façon aléatoire, des échantillons pour chaque ensemble de données. J'ai ensuite utilisé les estimateurs de l'échantillonnage classique et des géostatistiques pour calculer les valeurs moyennes. Les deux méthodes ont permis d'obtenir des estimations correctes dont les variances et les intervalles de confiance sont valides. Toutefois, les méthodes géostatistiques sont en général plus efficaces que celles de l'échantillonnage classique. En effet, elles produisent des estimations plus proches des valeurs réelles.
Mots-clés : krigeage ordinaire d'un bloc, populations de taille finie, inférence basée sur un modèle, simulations.

Introduction

For the most part, the geostatistical method of kriging is considered a method of spatial interpolation (Robertson, 1987), primarily used for making maps. However, the goals of classical statistical sampling and geostatistics can be identical. For example, both can be used to estimate the average value, or total amount, of a variable of interest in some area. In fact, it was this goal for mining that led Matheron (1963) and others (Journel & Huijbregts, 1978) to develop kriging (for a historical review, see Cressie, 1990).

Classical statistical methods of sampling can also be used to estimate the average, or total amount, of a variable. The word sample can be confusing, so I will give definitions here. The physical unit that is measured or observed will be called the sample unit. All possible sample units will be called the population. All sample units that are observed, taken collectively, will be called the sample. The statistical theory that relates to randomly choosing sample units and making estimates of the population from the sample will be called "classical sampling." This is consistent with most statistical texts on the subject (e.g., Cochran, 1977, or Thompson, 1992).

First consider the case of a spatially continuous population. An example would be the estimation of the total volume of snow in a study area. Conceptually, our sample units will be points. There are an infinite number of points for the spatially continuous population, so the population is infinite. From classical sampling, we would obtain our estimate by selecting n sample units (points) at random, mesuring the depth of snow at the n sites, taking their average, and multiplying by the area of the field. Classical sampling theory also allows us to obtain the variance of this estimate. We can accomplish the same task by using geostatistics. Geostatistics began as a way to estimate the amount of gold

[1]Rec. 2001-06-20; acc. 2002-01-03.

in an area, which is similar to estimating the amount of snow in a study area. This is known as block kriging.

Next, consider the case of a spatially discrete population. Suppose that a small study area has been partitioned into a finite set of samples of, say, N plots that are 1 m by 1 m, and we wish to estimate the average biomass in the study area. We randomly select n of N plots, clip and weigh the n samples, and then use the mean of the samples to estimate the mean value of the study area. Surprisingly, there has been no geostatistical counterpart to this until recently (Ver Hoef, 2001). The statistical estimation methods and types of data can be classified into a simple table (Table I).

TABLE I. Classification of methods based on type of population and type of statistical theory.

	Statistical theory	
Population	Classical sampling	Geostatistics
Infinite (spatially continuous)	Classical sampling methods	Block kriging
Finite (spatially discrete)	Classical sampling methods	Finite population Block kriging

Some questions that ecologists might ask are: Which of these methods should I use? What are the differences between the methods? What are the assumptions of each method? Which method is more powerful? This paper attempts to answer some of these questions. The objectives of this paper are to compare classical sampling methods with block kriging geostatistical methods through simulation and example. Some of this is review, but I will also introduce finite population block kriging and compare it to classical sampling methods for finite populations.

Design-based versus model-based statistics

One of the questions that I asked above was: What are the assumptions of each methods? The most fundamental difference between classical sampling and geostatistics is the underlying assumption about what is random and what is fixed. This is best illustrated with an example in one dimension. On the left side of figure 1, a fixed pattern is generated from the function

$$z(x) = \alpha_{s1}\sin(\beta_{s1}x) + \alpha_{s2}\sin(\beta_{s2}x) + \alpha_{c1}\cos(\beta_{c1}x) + \alpha_{c2}\cos(\beta_{c2}x) + \alpha_e(\exp(x)-1)$$

where $\alpha_{s1}=1$, $\alpha_{s2}=8$, $\alpha_{c1}=3$, $\alpha_{c2}=6$, $\alpha_e=10$, $\beta_{s1}=2\pi$, $\beta_{s2}=22\pi$, $\beta_{c1}=8\pi$ and $\beta_{c2}=58\pi$. For this pattern, 10 samples were drawn at random. This was done three times. Notice in figure 1, on the left, that the spatial pattern is fixed – it does not change – but the samples do. Statistical inference based on random samples, as given on the left of figure 1, is called design-based inference. Classical methods of statistical sampling as given, for example, by Cochran (1977) and Thompson (1992), are examples of design-based inference. For design-based inference, we obtain estimators from the way in which the sample was taken, using, for example, Horvitz-Thompson (1952) estimation.

Now consider the right side of figure 1. Here, the samples are fixed at locations x = 0.03, 0.07, 0.10, 0.13, 0.16, 0.20, 0.30, 0.55, and 0.87. For each of the three panels on the right, the sample locations do not change. Instead, the pattern is random, changing from panel to panel. The random pattern was generated from a first order autoregressive process,

$$z(x_i) = \rho z(x_{i-1}) + \varepsilon(x_i)$$

for x_i=0.0, 0.001, 0.002,...,1.0, where ρ = 0.95 and $\varepsilon(x_i)$ is an independent, normally distributed random variable with mean 0 and standard deviation 2.5. To start the process, we set $z(x_0) = 0$. Statistical inference based on a random mechanism governing the way that data are generated, as given on the right of figure 1, is called model-based inference. Kriging is an example of model-based inference. For model-based inference, we obtain estimators from the assumptions that we make about the model that generated the data. Further discussion of design versus model based inference is provided by Särndal (1978), de Gruijter and Ter Braak (1990) and Brus and de Gruijter (1993).

Finite and infinite populations in a spatial context

Texts on classical sampling (Cochran, 1977; Thompson, 1992) typically consider sampling for finite populations. However, we may be estimating quantities that are spatially continuous. For spatially continuous data, if we use sample units that are points, then the population is infinite; see for example Cordy (1993) and Stevens (1997). We rarely have sample units that are true points. For example, when investigating pollution the sample unit might be a cubic cm of air at 100 locations throughout a state. Because the sample unit has some volume (a cubic centimeter), there are a finite number of sample units, but for a whole state, that number is very large. It is often impossible to enumerate millions of sample units and then chose a sample randomly from among them. In this case, we consider the population to be infinite. Other ecological populations that could be considered spatially continuous are biomass, soil moisture, etc. The populations can be made finite if we use sample units that have areas that are large enough relative to the study area to make it reasonable to label all possible sample units and thus choose randomly from among them.

Taking a simple random sample from a finite population is easy. We make a list of all the N samples and choose n at random. This is usually done without replacement. To take a simple random sample for a spatially continuous variable, an infinite population, we randomly choose an x-coordinate and then a y-coordinate from a uniform distribution over the area of interest and repeat this process n times.

Inference for infinite populations

Let us consider the case of a spatially continuous, or infinite population, for a fixed pattern. First, we use classical sampling ideas. As an example, again consider the depth of snow in a small study area. Mathematically, the total volume of snow in the study area is the integral of the snow depth over the whole field (equation [1] in the Appendix 1); let us denote this as τ. The average snow depth is the integral divided by the area (equation [2] in the appendix); let us denote this as α. If we take a random sample uniformly over the field, then the estimate of the average snow depth is the sample mean, $\hat{\alpha}_{RS} = \bar{z}$, and the estimate of the total snow volume is $\hat{\tau}_{RS} = |A|\,\bar{z}$, where $|A|$ denotes the area of the field. The sample variance is calculated as an average sum

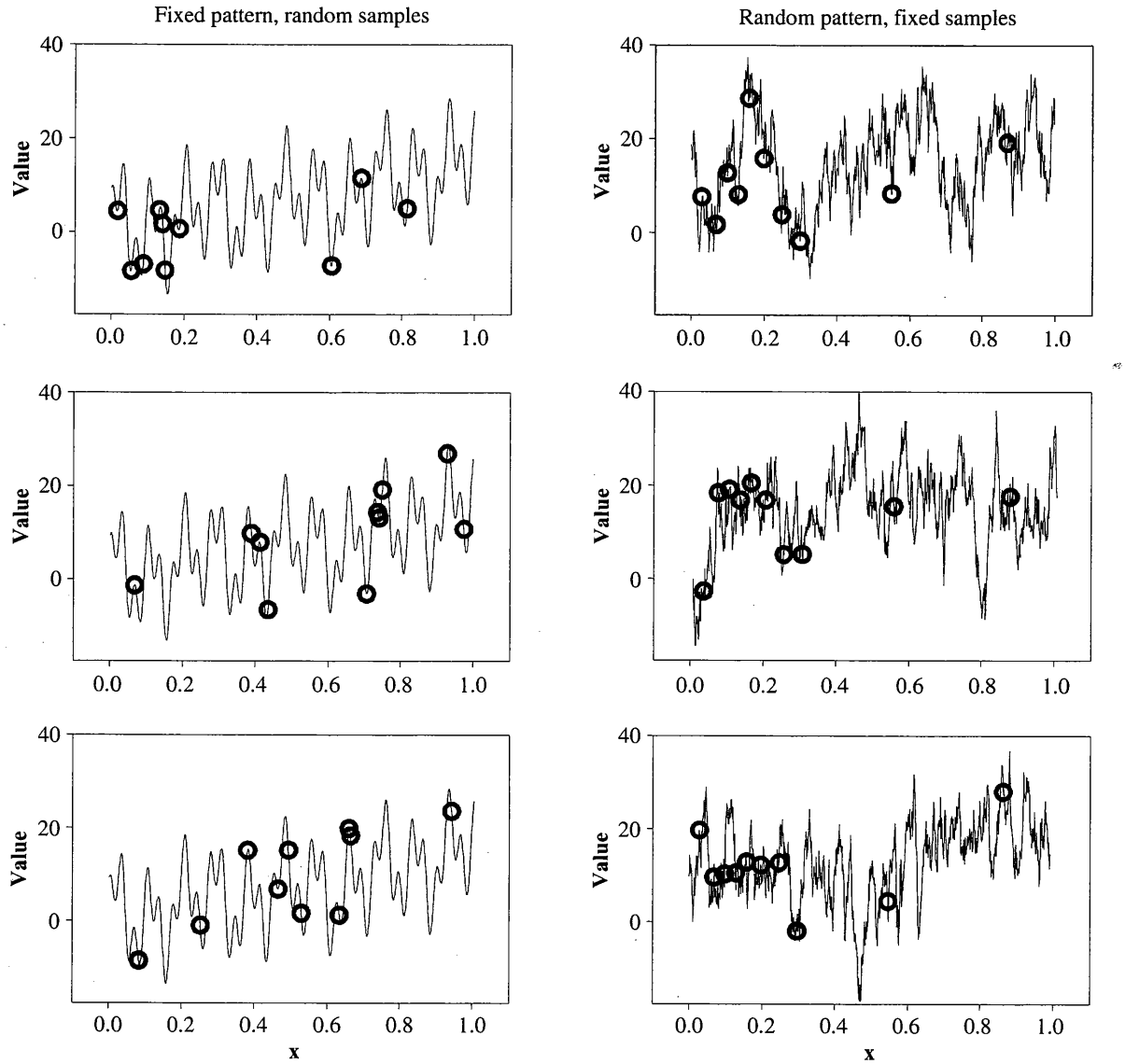

FIGURE 1. A comparison of design-based inference and model-based inference. On the left are 3 panels where the spatial pattern remains fixed and the samples are from randomly chosen locations. On the right are 3 panels where the spatial pattern is from an autocorrelated random process and the samples are from fixed spatial locations. The circles show the locations of the samples.

of squares, often denoted as S^2 (equation [4] in the Appendix). Then the estimated variance of $\hat{\alpha}_{RS}$ is S^2 / n, and the estimated variance of $\hat{\tau}_{RS}$ is $|A|^2 S^2 / n$.

Next, let us consider block kriging for predicting α and τ, where now we are assuming that the pattern is the result of a random process, and the samples unit locations are fixed. Let us denote the block kriging predictor of α as $\hat{\alpha}_{BK}$; the formula is given by equation [5] in the Appendix. The block kriging predictor for τ is $\hat{\tau}_{BK}=|A|\,\hat{\alpha}_{BK}$. For $\hat{\alpha}_{BK}$ we have an estimate of the prediction variance, denoted $\text{var}(\hat{\alpha}_{BK})$ (equation [6] in the Appendix). The estimated prediction variance for τ is $\text{var}(\hat{\tau}_{RS})=|A|^2\,\text{var}(\hat{\alpha}_{BK})$. More details on these formulas are given in the Appendix.

Estimating spatial autocorrelation

Before proceeding with block kriging, we need models for spatial autocorrelation and ways to estimate the parameters of the spatial models. A common assumption for spatial data is that they come from a stationarity model. Let $Z(\mathbf{s})$ be a random variable at location $\mathbf{s}$, where $\mathbf{s}$ is a vector of the x- and y-coordinates. The assumption that all random variables have a common mean is called mean stationarity; $E[Z(\mathbf{s})]=\mu$. An additional assumption is that the autocovariance, given as

$$C(\mathbf{h}) = \text{cov}(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})),$$

depends only on the spatial relationship between variables, not their exact locations. Together mean stationarity and $C(\mathbf{h})$ form a second-order stationarity assumption. It is also possible to model autocorrelation using variograms, but in this paper I will use autocovariances. The autocovariance given by $C(\mathbf{h})$ needs a particular parametric form. In this paper I use an isotropic exponential model,

$$C_e(\|\mathbf{h}\|) = \theta_n I(\|\mathbf{h}\| = 0) + \theta_s \exp(-\|\mathbf{h}\| / \theta_r),$$

where $\|\mathbf{h}\|$ denotes Euclidean distance and $I(f)$ denotes the indicator function, which is equal to 1 if f is true, otherwise it is 0. This is an isotropic model because the autocovari-

ance depends only on the distance between two locations and not on the directional orientation between them. The model depends on three parameters: θ_n, which is often called the nugget effect; θ_s, which is often called the partial sill; and θ_r, which is a range parmeter. The exponential autocovariance function is shown in figure 2.

If the parameters of the autocovariance function are known, then we can use the block kriging equations directly. However, the autocovariance function is rarely known, so it must be estimated from the data. In this paper, I will use restricted maximum likelihood (REML). REML was developed by Patterson and Thompson (1971, 1974), and used in the spatial context by Kitanidis (1983). Zimmerman (1989) gives computational details. For a general discussion of REML for spatial data, see Cressie (1993, *p.* 91), and for ecological applications, see Ver Hoef and Cressie (2001) and Ver Hoef *et al.* (2001). For the case of independent data (no autocorrelation) with a single common variance parameter δ^2 for all variables, the REML estimate of δ^2 is given by S^2, which is equation [4] in the appendix and the same as for classical sampling. Note that this has less bias than full maximum likelihood, where δ^2 is estimated by

$$\hat{\delta}^2_{MLE} = \sum_{i=1}^{n}(z(\mathbf{s}_i) - \bar{z})^2 / n.$$

For the spatial case, REML also has less bias than full maximum likelihood (Mardia & Marshall, 1984). Some authors claim that REML requires strictly Gaussian data (*e.g.*, Chiles & Delphiner, 1999, p. 110), but this is not true. It is true that REML was developed for Gaussian data, but both Heyde (1994) and Cressie and Lahiri (1996) show that REML estimates solve unbiased estimating equations, so they work much more generally than for Gaussian data. This should be evident from the example given above. For independent data with a common variance δ^2, the REML estimate of δ^2 is given by S^2, but S^2 is well known to be unbiased for δ^2 for more general situations than Gaussian data.

There are some advantages to using REML. The main advantage is that it is more automatic than using a least squares method. The main alternative to REML is weighted least squares (Cressie, 1985). Weighted least squares generally requires that you to bin the lags (distances) between locations, so you must select the number of bins and size of the bins. In this paper, I make use of simulations, so REML is desirable because it does not require binning the lags.

Comparison of random sampling and block kriging for infinite spatial populations

Similar to the fixed pattern seen in figure 1 in one dimension, I created a fixed pattern in two dimensions with

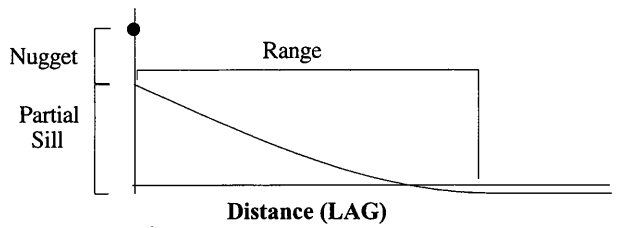

FIGURE 2. A generic autocovariance function, showing the nugget, sill and partial sill, and range.

$$z(x,y) = \alpha_{xs1}\sin(\beta_{xs1}x) + \alpha_{xs2}\sin(\beta_{xs2}x) + \alpha_{xc1}\cos(\beta_{xc1}x) + \alpha_{xc2}\cos(\beta_{xc2}x) + \alpha_{xe}(\exp(x)-1) + \alpha_{ys1}\sin(\beta_{ys1}y) + \alpha_{ys2}\sin(\beta_{ys2}y) + \alpha_{yc1}\cos(\beta_{yc1}y) + \alpha_{yc2}\cos(\beta_{yc2}y) + \alpha_{yq}y^2,$$

where $\alpha_{xs1}=1$, $\alpha_{xs2}=8$, $\alpha_{xc1}=3$, $\alpha_{xc2}=6$, $\alpha_{xe}=10$, $\beta_{xs1}=2\pi$, $\beta_{xs2}=22\pi$, $\beta_{xc1}=8\pi$, $\beta_{xc2}=58\pi$ and $\alpha_{ys1}=2$, $\alpha_{ys2}=7$, $\alpha_{yc1}=4$, $\alpha_{yc2}=5$, $\alpha_{yq}=-30(e-2)$, $\beta_{ys1}=4\pi$, $\beta_{ys2}=36\pi$, $\beta_{yc1}=6\pi$, and $\beta_{yc2}=66\pi$. Notice that for the region bounded by $0 \le x \le 1$ and $0 \le y \le 1$, from equations [1] and [2] in the Appendix, the true values are $\tau=0$ and $\alpha=0$. From this fixed surface, I took random samples of size 100. To take a single random sample I chose the x- and y-coordinates randomly from a uniform distribution. Figure 3 shows the fixed continuous surface along with the sample unit locations for 2 different random samples. For a sample of size 100, I computed the classical sampling estimate of the mean $\hat{\alpha}_{RS}$. I estimated the variance of $\hat{\alpha}_{RS}$ with $\hat{\text{var}}(\hat{\alpha}_{RS})$. Similarly, I computed the block kriging estimate $\hat{\alpha}_{BK}$ given by [5] in the Appendix and its variance estimate $\hat{\text{var}}(\hat{\alpha}_{BK})$ given by [6] in the Appendix, where I used REML to estimate the parameters of the exponential covariance model $C_e(\|\mathbf{h}\|)$. I repeated this for 1,000 random samples of size 100.

From the 1,000 estimates of both $\hat{\alpha}_{RS}$ and $\hat{\alpha}_{BK}$, and their estimated variances, I computed the following validation statistics:

1) bias, as the average of $\hat{\alpha}_m - \alpha$, where m is *RS* or *BK*;

2) root mean squared error (RMSE), as the square-root of the average of $(\hat{\alpha}_m - \alpha)^2$;

3) root average estimated variance (RAEV), as the square-root of the average of $\hat{\text{var}}(\hat{\alpha}_m)$; and

4) 80% confidence interval coverage, as the proportion of times that the confidence interval,

$$\hat{\alpha}_m \pm 1.28\sqrt{\hat{\text{var}}(\hat{\alpha}_m)}$$

contained the true value .

The results of the simulation are given in table II. Notice that there is no evidence of bias for either method because the bias value is very small compared to RMSE. Also notice that for this fixed pattern, block kriging has a smaller RMSE than classical sampling, indicating that the block kriging estimate is, on average, closer to the true value than the classical sampling estimate. If the estimated variances are valid, then the RAEV should be close to RMSE, which, from table II, appears to be true for both methods. Finally, the 80% confidence interval should contain the true value 80% of the time, and from table II it appears that both methods have valid confidence intervals.

Inference for finite populations

Now consider the finite population case. Suppose that for some spatial area A, there are N total sample units. The population total is the sum of the variable of interest over all N units; call it t (equation [7] in the Appendix). Let us denote the population mean as a ($a = t/N$, equation [8] in the Appendix). For a simple random sample without replacement, the estimators of the mean and the total are $\hat{a}_{RS} = \bar{z}$ and $\hat{t}_{RS} = N\bar{z}$, respectively, where $\bar{z}$ is the sample

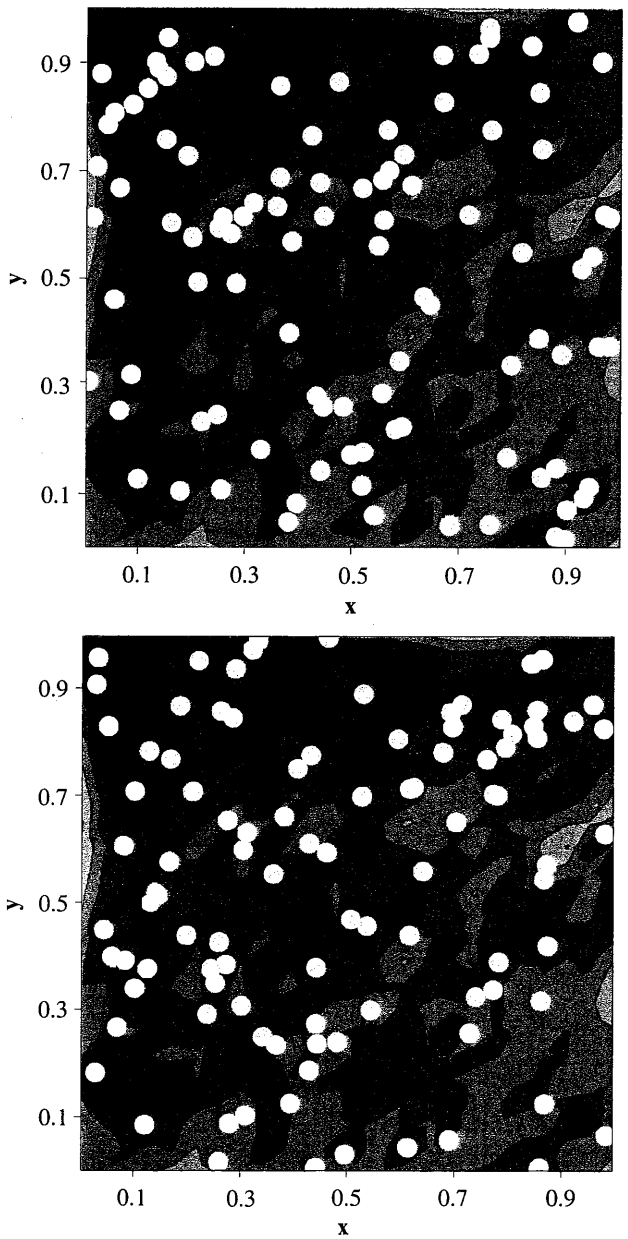

FIGURE 3. A fixed continuous spatial pattern. The darker areas are lower values, and the lighter areas are higher values. The white circles show the locations of random samples. The top figure shows the fixed surface with one random sample, and the bottom shows the fixed surface with another random sample.

TABLE II. Comparison of random sampling and block kriging. One thousand random samples were generated from a fixed continuous spatial pattern. Sample sizes were 100. For each simulation an isotropic exponential covariance model was estimated from the sample data using REML for FPBK.

Validation statistics	*SRS*[1]	*BK*[2]
Bias	0.002	-0.020
RMSE[3]	1.28	1.02
RAEV[4]	1.29	1.00
80%CI[5]	0.813	0.806

[1] Simple Random Sampling
[2] Block Kriging
[3] Root Mean Squared Errors
[4] Root Average Estimated Variance
[5] 80% Confidence Interval Coverage

mean. As for infinite populations, the sample variance is S^2 (equation [4] in the Appendix). Then the estimated variance of $\hat{a}_{RS}$ is $\text{var}(\hat{a}_{RS}) = (S^2/n)(1\text{-}n/N)$, and the estimated variance of $\hat{t}_{RS}$ is $\text{var}(\hat{t}_{RS}) = N^2(S^2/n)(1\text{-}n/N)$. Notice that the main difference in the variances between the finite case and the infinite case is due to the finite population correction factor, $(1\text{-}n/N)$, which goes to 0 as n goes to N. That is, if you observe every sample in a finite population, the variance of your estimate is 0 – you know the value exactly.

Next, consider a finite version of block kriging for estimating a and t, where now we are assuming that the pattern is the result of a random process, and the samples are fixed. An important but subtle point is that when estimating a, it is the average value of the actual pattern, not the mean over many simulations of the random process. Going back to the panels on the right side of figure 1, if we simulated spatial patterns an infinite number of times, we would find that the average value of a is 0, but for any one of the patterns a will not be 0. We want to estimate a for an actual, realized pattern. Let the vector $\mathbf{z}$ contain all sample units, with the data arranged so that $\mathbf{z} = (\mathbf{z}'_s, \mathbf{z}'_u)'$, where the subscript s indicates those sample units that are sampled and the subscript u indicates those sample units that are not sampled. The vector $\mathbf{b}$ contains weights for the quantity that we wish to estimate. For example, if $\mathbf{b} = (1,1,...,1)'$, then $t = \mathbf{b}'\mathbf{z}$. If $\mathbf{b} = (1/N, 1/N,...,1/N)'$ then $a = \mathbf{b}'\mathbf{z}$. In addition, we can consider small area estimates where the vector $\mathbf{b}$ contains mostly zeros, but with ones or other weights in positions that indicate weighting for particular samples. For an example of small area estimation, see Ver Hoef (2001). Then the finite population block kriging (FPBK) estimates of a and t are denoted $\hat{a}_{\text{FPBK}}$ and $\hat{t}_{FPBK}$, and their formulas are given by equation [9] in the Appendix. The prediction variances are denoted as $\text{var}(\hat{a}_{\text{FPBK}})$ and $\text{var}(\hat{t}_{FPBK})$ and the formulas are given by equation [11] in the Appendix. More details of these formulas are given by Ver Hoef (2001).

Comparison of classical sampling and FPBK for fixed populations

To compare classical sampling to FPBK, a set of species diversity values from a grid of 200 plots that measured 70 cm × 70 cm are shown in figure 4. These data are the number of different vascular plant species, and they come from glades in the Ozarks of southeastern Missouri (Ver Hoef, Reiter & Glenn-Lewin, 1993). Glades are grassy openings, usually caused by shallow, draughty soils in a predominantly forested landscape (Kucera & Martin, 1957). From this fixed pattern (Figure 4), I took random samples without replacement of size $n = 100$. For each random sample, I computed the classical sampling estimate $\hat{\alpha}_{RS} = \bar{z}$. I estimated the variance of $\hat{\alpha}_{RS}$ with $\text{var}(\hat{\alpha}_{RS}) = (S^2/n)(1\text{-}n/N)$. Similarly, I computed the FPBK estimate $\hat{\alpha}_{\text{FPBK}}$ given by [9] in the Appendix and its variance estimate $\text{var}(\hat{\alpha}_{\text{FPBK}})$ given by [11] in the Appendix. I used REML to estimate the parameters of the exponential covariance model $C_e(\|\mathbf{h}\|)$. I repeated this for 1,000 random samples of size $n = 100$. From the 1,000 estimates of both $\hat{a}_{RS}$ and $\hat{a}_{\text{FPBK}}$, and their estimated variances, I computed the same statistics as given for table II: bias, RMSE, RAEV, and the 80% confidence interval coverage.