

Sampling and geostatistics for spatial data¹

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'échantillonnage aléatoire simple, l'échantillonnage 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, measuring 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

¹Rec. 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.

Population	Statistical theory	
	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 $x_i=0.0, 0.001, 0.002, \dots, 1.0$, where $\rho = 0.95$ 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

