Claude Bellehumeur and Pierre Legendre

Aggregation of Sampling Units: An Analytical Solution to Predict Variance

Geographical variables generally show spatially structured patterns corresponding to intrinsic characteristics of the environment. The size of the sampling unit has a critical effect on our perception of phenomena and is closely related to the variance and correlation structure of the data. Geostatistical theory uses analytical relationships for change of support (change of sampling unit size), allowing prediction of the variance and autocorrelation structure that would *be observed if a survey was conducted using different sampling unit sizes.*

To check the geostatistical predictions, we use a test case about tree density in the tropical rain forest of the Pasoh Reserve, Malaysia. This data set contains exhaustive information about individual tree locations, so it allows us to simulate and compare various sampling designs. The original data set was reorganized to compute tree densities for 5 × 5-, 10 × 10-, *and* 20 × 20-*meter quadrat sizes. Based upon the* 5 × 5-*meter data set, the spatial structure is modeled using a nugget effect (white noise) plus an exponential model. The change of support relationships, using within-quadrat variances inferred from the variogram model, predict the spatial autocorrelation structure and new variances corresponding to* 10 × 10-*meter and* 20 × 20-*meter quadrats. The theoretical and empirical results agreed closely, whereas neglecting the autocorrelation structure would have led to largely underestimating the variance. As the quadrat size increases, the range of autocorrelation increases, while the variance and the proportion of noise in the data decrease.*

INTRODUCTION

Geography and many other scientific disciplines use data arranged into areal sampling units (that is, surfaces). The size of the sampling units or the level of

The large-scale forest plot at the Pasoh Forest Reserve is an ongoing project of the Malaysian Government, directed by the Forest Research Institute of Malaysia through its Director General, Datuk Dr. Salleh Nor, under the leadership of N. Manokaran, Peter S. Ashton, and Stephen P. Hubbell. Supplemental funds from the following sources are gratefully acknowledged: National Science Foundation (USA) BSR Grant No. INT-84-12201 to Harvard University through Peter S. Ashton and Stephen P. Hubbell; Conservation, Food and Health Foundation, Inc. (USA); United Nations, through the Man and the Biosphere programme, UNESCO-MAB grants nos. 217.651.5, 217.652.5, 243.027.6, 213.164.4, and also UNESCO-ROSTSEA grant no. 243.170.6; NSERC grant No. OGP0007738 to P. Legendre.

Claude Bellehumeur is a research associate in environmental studies at Universite de Sherbrooke. Pierre Legendre is professor of biology at Universite de Montreal.

Geographical Analysis, Vol. 29, No. 3 (July 1997) © 1997 Ohio State University Press
Submitted 11/27/95. Revised version accepted 11/8/96.

aggregation of these units is an important component of the scale of an investigation and may critically influence our perception of phenomena spread out in space. Changing sampling unit size induces changes in the statistical parameters estimated for a population. This problem is very well known in geography and was studied among others by Openshaw (1977, 1984), Dudley (1991), and Cressie (1991, pp. 284–89).

Classical statistical and geostatistical theories give analytical solutions to predict the change in variance due to different sampling unit sizes. Classical statistical theory works well to predict these changes when the hypothesis of independence of the sampling units is valid (no spatial autocorrelation in the data). This, however, is rarely the case in geographical sciences. Most variables relating to spatial environments present spatial structures such as gradients, patches, trends, etc. These structures can exist at many scales and correspond to intrinsic features of the environment.

This paper presents a simple analytical method, already known in geostatistics, to perform change of support operations enabling the prediction of the statistical parameters and features of the spatial autocorrelation structure resulting from the aggregation of sampling units (Journel and Huijbregts 1978; Cressie 1991). The method considers the within-unit variance which is calculated from a variogram model. We will study a rain forest plot of Malaysia as a test case, using the variable "tree density" for different quadrat sizes. A tract of mapped forest (1 kilometer long and 0.5 kilometer wide), located at 102°18′ W and 2°55′ N, was established in the Pasoh Reserve, Malaysia, to monitor long-term changes (Kochummen, LaFrankie, and Monokaran 1991). The survey enumerated all trees and positioned each one by geographic coordinates. We reorganized the data into nonoverlapping quadrat units and calculated tree densities (number of trees per square meter in a quadrat) corresponding to 5 x 5-, 10 x 10-, and 20 x 20-meter quadrats.

METHODS

Changing Sampling Unit Size

Several scientific disciplines dealing with data spread out in space have observed that the sampling unit size influences the estimates of the statistical parameters of a population. Chou (1991), Openshaw (1984), and Clark and Avery (1976), among others, pointed out that census data are frequently aggregated over geographical areas, and the aggregation level influences the statistical parameters of a distribution. The simultaneous change of the sampling unit size and the variance of plant density have been extensively used, in plant community analysis, to measure empirically the occurrence of spatial patterns at several scales (among others, Greig-Smith 1952, Ludwig and Goodall 1978). Levin (1992) presents empirical results concerning the description of ecosystems, which show the complexity of the relationships between the variance and the sampling unit size when the spatial pattern displays spatial autocorrelation structures. In the field of remote sensing, Marceau, Howarth, and Gratton (1994) resampled remote sensing images to different spatial resolutions and deduced empirical relationships between variances and spatial resolutions. These studies, however, construct specific empirical relationships and do not provide a general framework to make predictions of statistical and spatial structure parameters.

Classical statistical relationships attempt to predict the change in variance due to different sizes of sampling units. If one neglects spatial correlation, a classical relationship suggests that the variance of aggregated samples should decrease linearly with the number of sampling units in an aggregated sample:

$$\mathrm{Var}(V_{\mathrm{Agg}}|A) = \mathrm{Var}(v|A)/N \tag{1}$$

where $\mathrm{Var}(V_{\mathrm{Agg}}|A)$ is the variance of the aggregated samples V_{Agg} in area A, $\mathrm{Var}(v|A)$ is the original variance of the sampling units v in the same area, and N is the number of aggregated sampling units. An aggregated sample is formed by combining several sampling units into a single sample. This relationship is only valid for homogeneous areas where sampling units are independent of each other. It is necessary to consider the spatial structure of phenomena in order to correctly predict the effect of aggregation on the statistical parameters of a distribution.

Spatial Structure

Several techniques have been developed for the description of spatial patterns of populations (Cliff and Ord 1981; Haining 1990). The variogram is a tool to characterize the spatial variability of a variable distributed across a geographic area. The traditional estimator of the variogram is defined as (Journel and Huijbregts 1978, pp. 26–40; Cressie 1991, p. 40):

$$\gamma^*(\mathbf{h}) = (2N(\mathbf{h}))^{-1}\sum[z(\mathbf{x}) - z(\mathbf{x}+\mathbf{h})]^2 \tag{2}$$

where $z(\mathbf{x})$ and $z(\mathbf{x}+\mathbf{h})$ are measurements of a given variable at locations $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$, separated by the vector of directional distance $\mathbf{h}$, and $N(\mathbf{h})$ is the number of pairs of samples considered in the given distance class. Generally, the variogram tends to level off at a *sill* equal to the variance of the variable. The distance at which this occurs is referred to as the *range*. The discontinuity at the origin (nonzero intercept) is called the *nugget effect*. It is a random component corresponding to local variations occurring at scales smaller than the sampling interval, such as fine-scale spatial variability and measurement error (Cressie 1991, pp. 59–61).

Change of Support Operations

Problems of change of support have received a lot of attention in the geostatistical literature because ore reserve estimation requires estimation of the grade of large blocks, based upon small drill core data (Journel and Huijbregts 1978, pp. 61–94; Lantuejoul 1988; Isaaks and Srivastava 1989, chap. 19).

The additivity property of variances in nested designs implies (Isaaks and Srivastava 1989, pp. 476–80):

$$\mathrm{Var}(v|A) = \mathrm{Var}(v|V) + \mathrm{Var}(V|A) \tag{3}$$

where $\mathrm{Var}(v|A)$ is the variance of a small sampling unit v in area A, $\mathrm{Var}(V|A)$ is the variance of a large sampling unit V in area A, and $\mathrm{Var}(v|V)$ is the variance of a small sampling unit v in the large sampling unit V. This relationship shows that the variance of sampling units v in a certain area A can be expressed as a sum of within and between sampling unit variances. Journel and Huijbregts (1978, pp. 66–67) show that the variance $\mathrm{Var}(v|V)$ is related to the variogram:

$$\mathrm{Var}(v|V) = \bar{\gamma}(V,V) - \bar{\gamma}(v,v) \tag{4}$$

where $\bar{\gamma}(V,V)$ is the average point variogram value calculated over all possible distance vectors $\mathbf{h}$ contained in V, and similarly for $\bar{\gamma}(v,v)$. Equations (3) and (4) allow one to calculate the variance corresponding to a new sampling unit size V, if we know the autocorrelation structure for a point.

When v is used to compute an empirical variogram, a regularized form of variogram is estimated. We then deduce a point model $\gamma(\mathbf{h})$ (that is, $v = 0$) from a regularized model $\gamma_v(\mathbf{h})$. If $\gamma_v(\infty) = C_{1v}$, which is the sill value or the variance component of the spatial structure for v, and if v is smaller than the range, then,

$$C_{1v} = C_{1\bullet} - \bar{\gamma}(v, v) \tag{5}$$

where $C_{1\bullet}$ is the sill value for a point support. This relationship leads to

$$C_{1\bullet} = C_{1v}/(1 - F) \tag{6}$$

where F is equal to $\bar{\gamma}_1(v, v)$, representing the mean variogram value for a point variogram model with a sill equal to 1. Then, F can be computed from only the knowledge of the type of model and its range (Journel and Huijbregts 1978, p. 109). This correction only concerns the spatially structured part of the variance. The variance component ascribed to random variation and modeled by a nugget effect (C_0) follows the classical relationship [equation (1)].

The range of the spatial structure is affected by the size of the sampling units. The range of a spatial component C_{1v}, estimated from a support of size $l \times l = v$, is $a_{1\bullet} + l$, where $a_{1\bullet}$ is the practical range that would be measured if the support was a point (Journel and Huijbregts 1978, p. 84).

In practice, if the data are defined for a support v, we deduce first an approximate point model $\gamma(\mathbf{h})$ which is coherent with the empirical variogram $\gamma_v(\mathbf{h})$. Obtaining the point support variogram from a regularized variogram is, strictly speaking, impossible as it requires knowledge of the point scale structure, which is not available. However, the main features (sill and range) of the new model $\gamma_{v'}(\mathbf{h})$ can be deduced from the model $\gamma_v(\mathbf{h})$. The following rules provide acceptable approximations to deduce the point variogram and the regularized variograms $\gamma_{v'}(\mathbf{h})$ corresponding to new supports v':

1. For the spatially structured part of $\gamma_v(\mathbf{h})$, the point variogram is approximated by a variogram of the same type with a practical range of $a_{1\bullet} = a_{1v} - l$ and a sill $C_{1\bullet} = C_{1v}/(1 - F)$.
2. For the variogram of the new support v', the above defined point variogram is used, assuming that $\gamma_{v'}(\mathbf{h})$ is of the same type with a range $a_{1\bullet} + l'$ and a sill $C_{1v'} = C_{1\bullet} - \bar{\gamma}(v', v')$.
3. The nugget effect component, corresponding to v', is computed as $C_{0v'} = C_{0v} \bullet v/v'$, where C_{0v} is the nugget effect corresponding to support v. This random component is added to the spatially structured model $\gamma_{v'}(\mathbf{h})$ defined in step 2.

RESULTS

Summary Statistics

Summary statistics of tree density values for quadrats of 5 × 5, 10 × 10, and 20 × 20 meters, show that as size increases, extreme values disappear because they are diluted and combined into larger quadrats (Table 1). The mean remains constant but the variance decreases. The empirical results show an important departure from classical relationship predictions (Figure 1). Considering the empirical counts in 5 × 5-meter quadrats as our base for calculations (variance = 0.0610), we would expect variances of 0.0153 and 0.00381 for quadrats of 10 × 10 and 20 × 20 meters, respectively. These results are much smaller than the empirical variances of 0.0275 and 0.0161.

TABLE 1
Summary Statistics for Tree Density (number of trees per square meter) for Each Quadrat Size

	Quadrat size (m^2) 5 × 5	 10 × 10	 20 × 20
n	20,000	5,000	1,250
Mean	0.668	0.663	0.665
Variance	0.0610	0.0275	0.0161
Minimum	0.0	0.0	0.33
Maximum	1.72	1.60	1.23

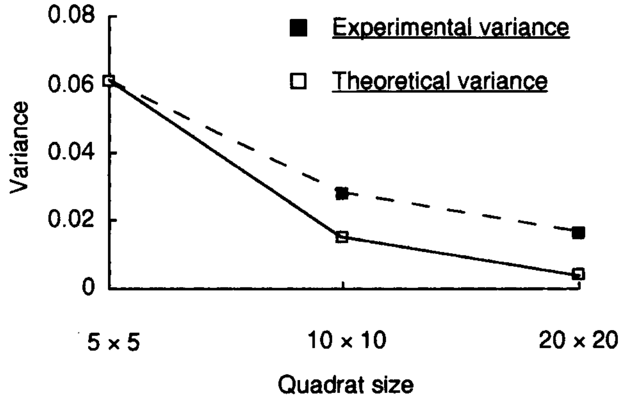

FIG. 1. Relationship between the Variance and Quadrat Size for the Empirical Results and the Results Expected from the Classical Relationship (equation 1)

Spatial Structure of Tree Density

Empirical variograms of tree density corresponding to the 5 × 5-, 10 × 10-, and 20 × 20-meter sampling units, for the north-south and east-west directions, show well-defined sills (Figure 2). The underlying process is considered to be isotropic, that is, $\gamma(\mathbf{h})$ does not depend on the direction of $\mathbf{h}$. Exponential models with nugget effect provided good adjustment to the empirical variograms:

$$\gamma(\mathbf{h}) = C_0 + C_1(1 - \exp(-\mathbf{h}/k)) \qquad (7)$$

where C_0 is the nugget effect, C_1 is the variability due to the structure in the exponential model, and k is a shape parameter (Table 2). The ratio of the nugget effect to the sill, called the relative nugget effect, can be used to evaluate sampling error and fine-scale spatial effects. The exponential model reaches its sill $(C_0 + C_1)$ asymptotically. The practical range of an exponential model is defined as $a = 3k$, the distance at which the variogram is 95 percent of C_1.

As the quadrat size increases, sill values decrease and ranges increase. The most important effect is the decrease in relative nugget effect. For the 5 × 5-meter quadrats, the proportion of random variation is very high (75 percent), and the process does not seem very strongly spatially structured. On the other hand, for 20 × 20-meter quadrats, the process displays an important spatially structured component accounting for 83 percent of the spatial variance.

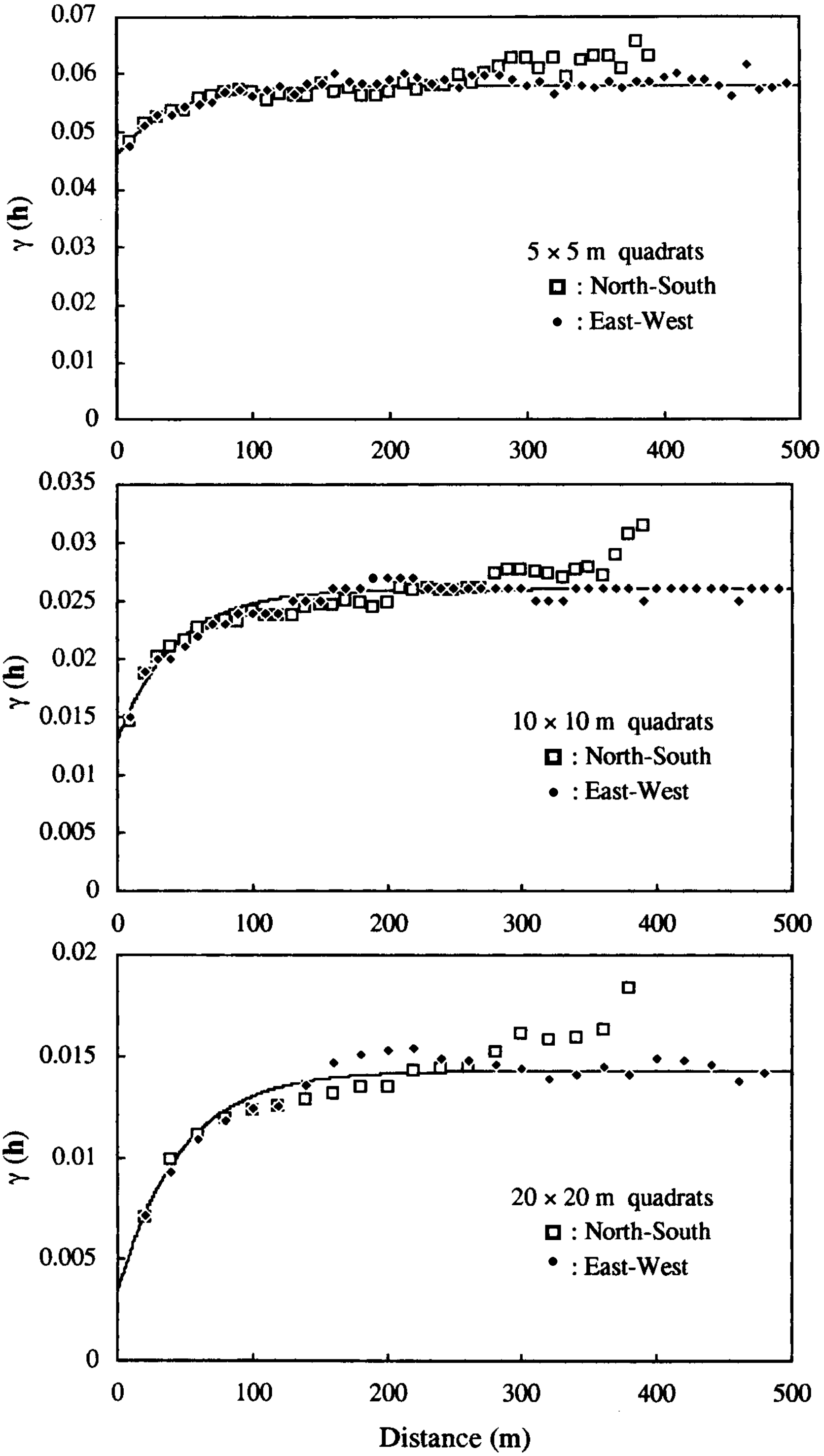

FIG. 2. Directional Variograms of the Tree Density Variable for 5 × 5-, 10 × 10-, and 20 × 20-Meter Quadrats

TABLE 2
Parameters of Exponential Variogram Models for Each Quadrat Size

Quadrat size (m^2)	C_0	C_1	Sill	a	$Rel\,C_0$
5 × 5	0.0446	0.0151	0.0597	104	0.75
10 × 10	0.0113	0.0139	0.0253	110	0.45
20 × 20	0.00245	0.0118	0.0145	129	0.17

C_0 is the nugget effect, C_1 is the variance component associated to the structured spatial scale, the sill is $C_0 + C_1$, a is the practical range, and $Rel\,C_0$ is the relative nugget effect $(C_0/(C_0 + C_1))$.

TABLE 3
Parameters of the Change of Support Transformation

	Quadrat size (m^2) 5 × 5	10 × 10	20 × 20
$\bar{\gamma}(v, v)$	0.00123	0.00234	0.00429
F	0.0755	0.143	0.263
C_0(emp)	0.0446	0.0113	0.00245
C_0(inf)	0.0446	0.0112	0.00279
C_1(emp)	0.0151	0.0139	0.0118
C_1(inf)	0.0151	0.0140	0.0120
Variance	0.0610	0.0275	0.0161
Var($v\|A$)	0.0597	0.0252	0.0148

$\gamma(v, v)$ is the average point variogram value calculated for a quadrat of size v.
F is the mean variogram value for the point variogram model with a sill equal to 1.
C_0(emp) is the empirical nugget effect.
C_0(inf) is the nugget effect inferred from the empirical value of the 5 × 5 quadrat size.
C_1(emp) is the empirical structured variance component.
C_1(inf) is the structured variance component inferred from the theoretical model.
Var($v|A$) is the variance of a unit v in the study area A, given by the analytical relationship.

Empirical Verification of the Change of Support Relationships

The previous sections have shown empirically that changing the sampling unit size modifies the variance, as well as the spatial autocorrelation structure of data. We will now check whether the analytical solution allows the prediction of our empirically obtained results. For the 5 × 5-meter quadrat size, the practical range is 3 × 34.67 meters = 104 meters (Table 2). Given that for an exponential model, the practical range equals $3k$, then the parameter $a_{1\bullet}$ of a point model is equal to $(3k - l)/3 = 33.0$. Estimating the point sill value of the structured component requires the evaluation of the within-quadrat variance $\bar{\gamma}(v, v)$. The mean value $\bar{\gamma}(v, v)$ and the parameter F can be calculated numerically from function $\gamma(\mathbf{h})$ by discretizing sampling unit v into a finite number of points or by generating random lags within v (stochastic integration), and calculating the average variogram values for lags contained in v (Table 3).

Using the mean variogram values for the 5 × 5-meter quadrat size (Table 3), the point sill value of the structured component is given by formula (6) as

$$C_{1\bullet} = C_{1v}/(1 - F)$$

$$C_{1\bullet} = 0.0151/(1 - 0.0755)$$

$$C_{1\bullet} = 0.0163.$$

The theoretical point support variogram is an exponential model:

$$\gamma(\mathbf{h}) = 0.0163(1 - \exp(-\mathbf{h}/33)). \tag{8}$$

Such a point variogram model could be deduced for any other quadrat size, as long as the quadrats are not too large relative to the range of the point variogram. From this theoretical point model, it is possible to calculate the variance of any given sampling unit size in the whole area and to find an appropriate variogram model describing the spatial structure features for various quadrat sizes.

The variogram models shown in Figure 2 have two components each: a random and a spatially structured component. The change in the random component due to a change of support follows the classical relationship [equation (1)]. The random component for the 5 × 5-meter quadrats is 0.0446. Therefore, for 10 × 10-meter quadrats, the random component should be 0.0112 (0.0446/4), and for 20 × 20-meter quadrats, 0.00279 (0.0446/16) (Table 3). The effect of a change of support operation on the spatially structured component of variance is given by equation (5). For 10 × 10-meter quadrats; $\bar{\gamma}(v, v) = 0.00234$ (Table 3), the structured variance component for the 10 × 10-meter quadrat size is

$$C_{1(10\times 10)} = C_{1\bullet} - \bar{\gamma}(10, 10)$$

$$C_{1(10\times 10)} = 0.0163 - 0.00234$$

$$C_{1(10\times 10)} = 0.0140.$$

This analytical solution gives a variance of 0.0252 for 10 × 10-meter quadrats ($C_{0(10\times 10)} + C_{1(10\times 10)}$). The empirical value is 0.0275, while the classical approach would have given 0.0153 (0.0610/4). The analytical solution above is closer to the empirical value than the classical relationship (Table 3). The slight underestimation may be due to a long-range spatial structure in the north-south direction which is not modeled, considering the size of this structure compared with the size of the study area.

DISCUSSION

Changing the size of sampling units induces changes in the variance and in the spatial autocorrelation structure of the data. Geostatistical theory considers the autocorrelation structure to perform change of support operations, using the within-support variance inferred from the variogram model. The method allows the prediction of the statistical parameters and the features of the spatial structure which would be observed for aggregated sampling units.

A change of support operation involves the following steps: (i) A variogram model $\gamma_v(\mathbf{h})$ is derived from the empirical data, corresponding to a regularized form of variogram for a given sampling unit size v. (ii) A point model $\gamma(\mathbf{h})$ is deduced from the regularized model, using equation (6): $C_{1\bullet} = C_{1v}/(1 - F)$. The variance component ascribed to random variation follows equation (1). (iii) Once the point model $\gamma(\mathbf{h})$ and its parameters have been found, another expression $\gamma_{v'}(\mathbf{h})$ can be derived for another sampling unit size v'. Knowledge of the point model allows one to calculate the mean variogram values $\bar{\gamma}(v, v)$ for any sampling unit size.

The geostatistical predictions were verified using a data set about tree density in the tropical rain forest of the Pasoh Reserve. The comparison confirms that the results computed from the change of support relationships agree closely with empirical results. We have shown four key results:

(1) As the size of the sampling units increases, the variance decreases while the mean remains constant.

(2) With an increase in quadrat size, the range of autocorrelation increases, while the variance and the proportion of noise in the data decrease.

(3) The reduction in variance for the aggregation of spatially autocorrelated sampling units is less important than for aggregated independent sampling units. Consider a spatially autocorrelated process. Within an aggregated sampling unit, the original sampling units are more similar than if they were the result of a process corresponding to white noise only. As a consequence, the within-unit variance is smaller than the variance expected by classical statistical theory; on the other hand, the among-unit variance is larger.

(4) From an empirical variogram, we can deduce a theoretical point model that enables the estimation of appropriate models corresponding to various support sizes.

LITERATURE CITED

Chou, Y. H. (1991). "Map Resolution and Spatial Autocorrelation." *Geographical Analysis* 23, 228–46.

Clark, W. A. V., and K. L. Avery (1976). "The Effects of Data Aggregation in Statistical Analysis." *Geographical Analysis* 8, 428–38.

Cliff, A. D., and J. K. Ord (1981). *Spatial Processes, Models, and Applications.* London: Pion.

Cressie, N. A. C. (1991). *Statistics for Spatial Data.* New York: John Wiley and Sons.

Dudley, G. (1991). "Scale, Aggregation, and the Modifiable Areal Unit Problem." *Operational Geographer* 9, 28–33.

Greig-Smith, P. (1952). "The Use of Random and Contiguous Quadrats in the Study of Structure in Plant Communities." *Annals of Botany, New Series* 16, 293–316.

Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge: Cambridge University Press.

Isaaks, E., and R. M. Srivastava (1989). *An Introduction to Applied Geostatistics.* New York: Oxford University Press.

Journel, A. G., and C. J. Huijbregts (1978). *Mining Geostatistics.* London: Academic Press.

Kochummen, K. M., J. V. LaFrankie, and N. Manokaran (1991). "Floristic Composition of Pasoh Forest Reserve, a Lowland Rain Forest in Peninsular Malaysia." *Journal of Tropical Forest Science* 3, 1–13.

Lantuejoul, C. (1988). "On the Importance of Choosing a Change of Support Model for Global Reserves Estimation." *Mathematical Geology* 20, 1001–19.

Levin, S. (1992). "The Problem of Pattern and Scale in Ecology." *Ecology* 73, 1943–67.

Ludwig, J. A., and D. W. Goodall (1978). "A Comparison of Paired- with Blocked-Quadrat Variance Methods for the Analysis of Spatial Pattern." *Vegetatio* 38, 49–59.

Marceau, D. J., P. J. Howarth, and D. J. Gratton (1994). "Remote Sensing and the Measurement of Geographical Entities in a Forested Environment. 1 The Scale and Spatial Aggregation Problem." *Remote Sensing of Environment* 49, 93–104.

Openshaw, S. (1977). "A Geographical Solution to Scale and Aggregation Problems in Region Building, Partitioning, and Spatial Modeling." *Transactions of the Institute of British Geographers New Series* 2, 359–472.

——— (1984). *The Modifiable Areal Unit Problem.* CATMOG (Concepts and Techniques in Modern Geography) Norwich, England: Geo Books 38.