SAMPLING AND MAPPING HETEROGENEOUS SURFACES BY OPTIMAL TILING

Ferenc Csillag
Institute for Land Information Management, University of Toronto,
Erindale College, Mississauga, ONT, L5L 1C6
tel: 416-828-3862; fax: 416-828-5273; e-mail: fcs@geomancer.erin.utoronto.ca

Miklós Kertész, Ágnes Kummert
Research Institute for Soil Science and Agricultural Chemistry,
Hungarian Academy of Sciences, H-1022 Budapest, Herman Ottó u. 15.

ABSTRACT

A novel approach has been developed, implemented and tested to approximate, or to map, large heterogeneous surfaces with predefined accuracy and complexity. The methodology is based on tiling, the hierarchical decomposition of regular grid tessellations. The quadtree-construction is guided by a measure of local homogeneity and the predefined number of leaves, or level of accuracy. A modified Kullback-divergence is applied for the characterization of goodness of approximation. The procedure is aimed to find the quadtree-represented map of limited number of leaves which is the most similar to a given image in terms of divergence. Various statistical and computational advantages are demonstrated within the framework of spatial data processing and error handling in geographic analysis. This methodology minimizes information loss under constraints on size, shape and distribution of varying size mapping units and the residual heterogeneity is distributed over the map as uniformly as possible. As such, it formally defines a cost-versus-quality function in the conceptual framework of the cartographic uncertainty relationship. Our straightforward decomposition algorithm is found to be superior to other combinations od sampling and interpolation for mapping strategies. It is advantageous in cases when the spatial structure of the phenomenon to be mapped is not known, and should be applied when ancillary information (e.g., remotely sensed data) is available. The approach is illustrated by the SHEEP (Spatial Heterogeneity Examination and Evaluation Program), an environmental soil mapping project based on high-resolution satellite imagery of a salt-affected rangeland in Hortobágy, NE-Hungary.

INTRODUCTION

The spatial structure of phenomena, its relationship to sampling, mapping, resolution and accuracy, has been the focus of a wide array of geographic research (Moellering and Tobler 1972; Goodchild 1992). Geographic information systems (GIS) are being increasingly used to utilize these findings related to sampling design to perform analyses on spatial databases and to evaluate their quality (Burrough 1986). These advances,

however, are rarely applied simultaneously in reported case studies, even though conceptually and methodologically they often have common foundations (for example, spatial covariation may be utilized in interpolation but not in sampling design and regionalization, or quality assessment).

We consider the conceptual framework for database development from sampling to the measurement of accuracy, and explicitly formalize the cost-versus-quality function. Sampling strategies for mapping are always based on some preliminary knowledge and a priori assumptions about the phenomenon to be mapped. They are based on expertise in the field of application or on mathematical statistics or on both. The "goodness," or efficiency of sampling, in general, is dependent on the relationship between the cost of sampling and the quality of the final map product. Regardless of the actual circumstances, the goal of sampling is to collect "representative" samples (Webster and Oliver 1990). We need a means of sampling that will ensure appropriate information that can predict characteristics at locations where no samples were taken. In terms of characterization, our analysis and discussion will be confined to the task of predicting the local (expected) value of a variable, which is a quite widely explored problem in mapping[1] (Ripley 1981).

Our assumptions about the circumstances of database development are as follows:

- a fixed budget is given for sampling, for which we seek maximum accuracy (or conversely, an accuracy threshold is given for which minimum sampling effort should be determined);
- the spatial structure of the phenomenon to be mapped is not homogeneous (i.e., it varies over a range of scales); therefore, no a priori partitions can be defined; and
- some ancillary data sets are available whose spatial pattern is in close correspondence with the phenomenon to be mapped.

A CONSTRAINED OPTIMAL APPROACH TO MAPPING A LATTICE

We provide here a method for constrained optimal approximation of lattices. The mapping of a two-dimensional heterogeneous surface is treated with the following constraints: (1) a data set (**A**) is available on a lattice, which will be sampled and approximated by a map (**M**); (2) both **A** and **M** are two-dimensional discrete distributions; (3) the location of each datum on **M** corresponds to the location of a datum or data on **A**; and (4) **M** consists of a finite number of patches that are homogeneous inside, and the value associated with a patch approximates the value(s) of **A** at corresponding locations. No assumptions are made about the exact nature of the spatial statistics of the surface to be mapped (e.g., stationarity).

FROM MULTIPLE RESOLUTION TO VARYING RESOLUTION

Spatial pattern can play a significant role in the characterization of patches from sampling through interpolation to regionalization.

[1] For a review of sampling, resolution and accuracy see Csillag et al. (1993).

Understanding spatial pattern generally aims at the design of a sampling scheme, which is reasonable under certain statistical assumptions and models (Kashyap and Chellapa 1983, Kabos 1991), and for the application of "best" parameters defined by those models in interpolation and representation (Tobler 1979; Jeon and Landgrebe 1992). Advancements in computing and storage performance in GIS, paralleled with the apparent contradiction between finding the best resolution for regular samples and identifying (a priori) patches (partitions) for samples (Webster and Oliver 1990), has increased the popularity of mapping with multiple representation (Dutton 1984). Beside some storage- and processing-related technical issues, the identification and representation of mapping units (or area-classes; see Mark and Csillag 1989) have not been adequately addressed. Several soil or vegetation maps, whose patches often contain inclusions, can serve as simple illustrations: when a partition is optimal for the patches, it misses the inclusions, and when it is optimized for the inclusions, it becomes redundant for the patches (Csillag et al. 1992). Furthermore, it is prohibitive, because of size, to examine all possible partitions for optimization.

In the construction of databases, the hierarchy of resolutions in multiple representations based on uniform grids, a pyramid, offers the possibility of creating a GIS, which adjusts the level of detail to particular queries. Furthermore, it has become feasible to create varying resolution representations, of which quadtrees have become most well known and widely applied (Samet 1990). Beside storage and processing efficiency, the advantage of varying resolution representations is to ensure uniform distribution of accuracy (quality) over the entire data set. This would require criteria for creating quadtrees from pyramids.

We have developed and implemented a method (Kertész et al. 1993) to create a quadtree as a constrained optimal approximation of the lowest (most detailed) level of a pyramid. A quadtree can also be thought of as a spatial classification (partition on tiles) with the advantage that not only does each location belong to one and only one leaf, but the location, size and arrangement of all possible leaves is known a priori. Our method starts from the highest (least detailed) level (i.e., approximating the entire lattice by one value, its mean) and proceeds by straightforward decomposition. Because of the strong constraint on the shape, size, arrangement and potential number of patches on a map represented by a quadtree as a function of the number of levels (Samet 1990), with the application of an appropriate measure, the accuracy of all potential maps can be compared -- hence the process can be optimized.

DIVERGENCE MEASURES TO CHARACTERIZE DIFFERENCES IN SPATIAL PATTERN

To quantify the dissimilarity between the lattice and its (potentially varying resolution) map we apply a measure that (1) directly compares the lattice and the map and returns a scalar, (2) has a value independent of the size of the lattice, (3) is nonparametric, (4) is independent of the scale of data values (i.e., invariant to multiplication by a constant), and (5) provides the opportunity for additive application due to element-by-element computation (Figure 1).

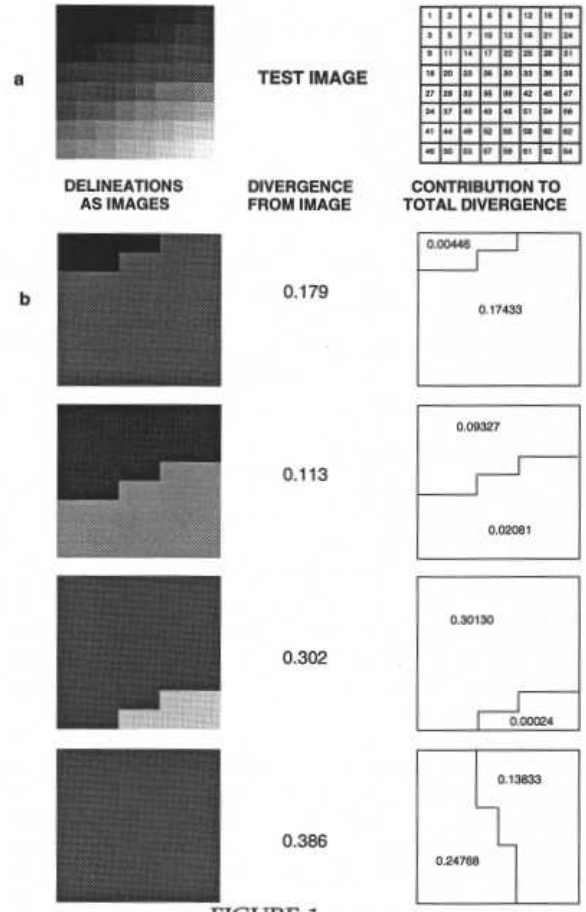

FIGURE 1.

Test data set (a) with four possible delineations with their corresponding total divergence and the contribution of the patches (b).

We have chosen Kullback-divergence (Csiszár 1975),

$$D_{Kullback}(p \mid q) = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \log_2(p_{ij}/q_{ij}) \qquad (1)$$

where

$$\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = \sum_{i=1}^{I} \sum_{j=1}^{J} q_{ij} = 1, \qquad p_{ij} \geq 0,\ q_{ij} \geq 0$$

and

p and q are discrete spatial distributions of I by J elements;

because its usage does not require any limitations concerning the nature of

the distributions, and described its general properties and its relationship to Shannon's information formula, χ^2 elsewhere (Kertész et al. 1993). Its real advantage is that for any delineated patch on a map, one can compute the contribution of that particular patch to the total divergence:

$$D(patch_{grid} \mid patch_{map}) = \sum_{i}^{patch} \sum_{j} [grid_{ij}/SUM] \; log_2(grid_{ij}/map_{ij}) \qquad (2)$$

where

$grid_{ij}$ is the value for i,jth cell of the grid
map_{ij} the value for i,jth cell of the map
I and J are the side lengths of the grid to be mapped and the map in corresponding cells, respectively.

DECOMPOSITION PROCEDURE AND CRITERIA

Our method utilizes the advantage of varying resolution adjusted to spatial variability because it provides a rule for selecting the necessary spatial detail to represent each mapping unit with similar (internal) heterogeneity. In other words, it leads to a measure of local variability not weighted by area. Therefore, at very heterogeneous locations it will lead to smaller units (i.e., higher levels of the quadtree), whereas at relatively homogeneous locations it will decompose the grid into larger tiles (i.e., lower levels of the quadtree).

The decomposition algorithm can be characterized by two features: the cutting rule and the stopping rule. In this paper we use "maximum local divergence" as the cutting rule, and "total number of mapping units" as the stopping rule. Further options will be discussed in the final section. The first rule means that the quadtree leaf is cut into four quadrants whose local divergence is the maximum among all existing leaves, while the second rule stops the decomposition at a threshold predefined by the number of leaves.

The approach embedded in this method avoids the major conceptual problems of spatial pattern analysis and measurement of accuracy tied to the existence and knowledge of the spatial covariance function. In several, primarily environmental, mapping tasks the uncertainty in the delineation of mapping (and sampling) units is intolerably high, especially when local factors control the landscape. Kertész et al. (1993), for example, characterized salt-affected, semiarid rangelands exhibiting scale-invariant patterns; certain (mostly transitional) patches occurred over several hectares as well as over a few square centimeters. The measurement of accuracy in databases constructed for such areas will heavily depend on the validity of statistical assumptions (Goodchild and Gopal 1989; Csillag 1991). Our method not only does not require strong assumptions about the spatial statistics of the lattice but (1) optimizes local accuracy of sampling and mapping units under a global threshold, (2) explicitly links accuracy to the number of mapping units (and vice versa) and (3) when those strong assumptions are valid, leads to identical results.

MAPPING WITH ANCILLARY DATA: SAMPLING QUADTREES

We demonstrate the characteristics of the method described in the previous section from a database development perspective (i.e., as if one had to actually sample a surface and approximate it on a lattice), with several illustrative examples. For comparison, besides sampling based on quadtrees, equal number of samples will be taken by random and regular (square) design. Reconstruction will be performed by Thiessen-polygonization, inverse squared Euclidean distance, and kriging to test the robustness of the sampling method over various levels of sophistication in interpolation.

All procedures are carried out on a data set, with spatial characteristics illustrated in Figure 2, taken from a satellite image used for database design and development in SHEEP (Spatial Heterogeneity Examination and Evaluation Program), an environmental rangeland degradation mapping project in East Hungary (Tóth et al. 1991; Kertész et al. 1993).

Selecting optimal samples by the proposed decomposition method requires some information in the form of a lattice about the surface to be sampled and mapped. At first, the entire lattice is approximated by its (global) mean (the root of the quadtree), and its Kullback-divergence is measured from the actual distribution. Then, the lattice is approximated by four quadrants (level 1 on the quadtree), and for each leaf their contribution to the total Kullback-divergence is computed according to Equation (2). If the threshold in the number of samples or in total Kullback-divergence has not been met, decomposition is continued by cutting the leaf with the highest contribution (i.e., the most heterogeneous one). Thus, at each step, the number of (potential) sampling units increases by three, and the decomposition proceeds least intensively over homogeneous areas.

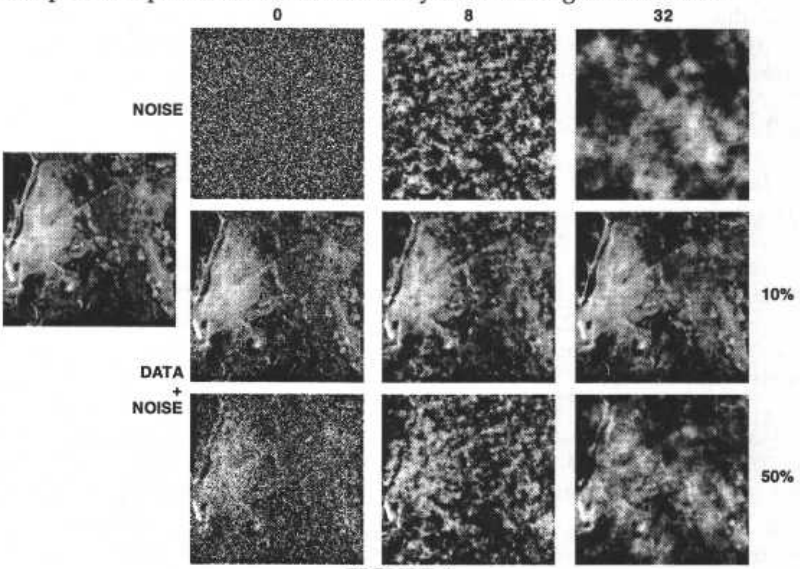

FIGURE 2.
The original data set (left), some of the noise fields (top) and their combinations (bottom).

To compare the efficiency of the different sampling designs we reconstructed the 128-by-128 data sets and measured their Kullback-divergence from the original data. For realistic illustration we set the number of samples to 256. Three interpolations (quadtree leaves, Thiessen-polygonization, and inverse squared Euclidean distance) are summarized for the sampling quadtrees (256 leaves), and three interpolations (kriging, Thiessen-polygonization, and inverse squared Euclidean distance) are shown for the regular and random sampling. The numerical results are summarized in Table 1.

Table 1.
Accuracy of approximations (D_K*10^4 to the original data based on 256 samples) by sampling design (Q=quadtree, G=regular grid, R=random) and interpolation (QTR=sampling quadtree, THI=Thiessen-polygonization, DI2=inverse square distance, KRI=kriging).

INTERPOLATION	QTR	THI	DI2	KRI
SAMPLING				
Q	8.937	16.864	11.094	---
G	---	22.864	17.550	17.646
R	---	25.545	19.824	18.468

The sampling quadtrees lead to approximately half, or less, the Kullback-divergence than any other sampling and interpolation method, because they not only are more sensitive to local variability but also are obtained by using all data from the approximated distribution.[2] In addition, samples selected by this method carry over so much information about the distribution of variance in the data set that they systematically result in significantly smaller Kullback-divergence than the other sampling methods over all interpolation techniques. Furthermore, sampling based on quadtrees and interpolation by inverse squared Euclidean distance is, by far, the superior method among those tested.

SAMPLING AND RECONSTRUCTION: NOISY DATA

In general, at best, one can assume to have a data set that corresponds to the phenomenon to be mapped but contains a certain amount of noise. This available data set may be remotely sensed data, as in the illustrative example, or it can be any existing data set (on a lattice) in a database. In practice, these are exactly the sources of sampling design and database development. Therefore, it is important to examine the proportion and spatial structures of noise and its effects on the accuracy of sampling and reconstruction.

We generated noise fields with five different levels of spatial autocorrelation (correlation distance or range = 0, 4, 8, 16, 32 cell units)[3] and mixed them with 10%, and 50% weights to the original data sets (Figure 2),

[2] These results refer to the ideal situation of when the data set to be approximated in our database is entirely well known. The Kullback-divergences of the Q_QTR (sampling quadtree with mean values assigned to the leaves) method can be interpreted, therefore, as measures of the cost of data compression.

[3] Gerard Heuvelink kindly made his software available (Heuvelink 1982).

preserving the original mean and variance. The Kullback-divergences between the noisy and original data sets are summarized in Table 2. It helps to "scale" Kullback-divergence to certain amounts of noise, which reveals that while there is a strong relationship between the amount of noise and Kullback-divergence, it does not change significantly with its correlation length.

Table 2.
Kullback-divergences between the noisy and the original data sets (D_K*10^4)

	CORRELATION LENGTH				
	0	4	8	16	32
NOISE-LEVEL					
10%	4.782	4.802	4.856	5.128	4.912
50%	26.234	26.200	27.296	30.537	28.004

We examine the effects of noise by designing the sampling quadtree on the noisy data, and approximate the original one. Numerical results indicate that the longer range noise is added (i.e., the smoother the ancillary data set becomes), the better the approximation is to the ancillary data, and the accuracy of reconstructing the original data decreases. At 10% random noise, the 256 samples taken from the noisy ancillary data set approximate the original data almost as well as if samples were taken from the original ($9.532*10^{-4}$ versus $8.937*10^{-4}$ for tiling reconstruction).

Reconstructions based on sampling quadtrees with 256 samples using inverse squared Euclidean distance interpolation (providing the best results among the methods tested) outperform reconstructions based on random and regular square sampling using the same interpolation, regardless of the amount and spatial structure of noise (Table 3). The stronger the pattern, the more sampling quadtrees provide advantage (Figure 3). At low (10%) noise levels, consistently over all noise structures, reconstructions based on sampling quadtrees lead to approximately one-third less Kullback-divergence than reconstructions based on other sampling methods. All accuracies slightly increase when approximating original data.

Table 3.
Kullback-divergences (D_K*10^4) between reconstructions, the original data set and noisy data sets; 256 sampling locations determined on noisy data by sampling quadtree (Q), regular square grid (G) and random (R); samples taken from noisy data and inverse squared Euclidean distance interpolation.

	CORRELATION LENGTH				
	0	4	8	16	32
10%					
(sampled vs. sampled)					
Q	16.781	16.225	15.653	12.096	11.363
G	22.704	21.909	19.573	17.402	16.321
R	23.786	23.651	21.095	19.308	19.092
(sampled vs. original)					
Q	12.937	13.321	14.532	14.224	14.723

G	19.071	19.511	20.037	20.248	21.395
R	20.889	21.992	21.240	22.631	22.416
50%					
(sampled vs. sampled)					
Q	39.874	34.879	27.003	15.906	9.383
G	40.994	36.858	28.183	17.048	12.071
R	39.921	37.998	30.225	19.777	16.476
(sampled vs. original)					
Q	25.756	26.478	31.245	32.198	34.615
G	26.779	28.618	32.015	34.518	37.932
R	27.742	31.464	31.392	38.259	38.713

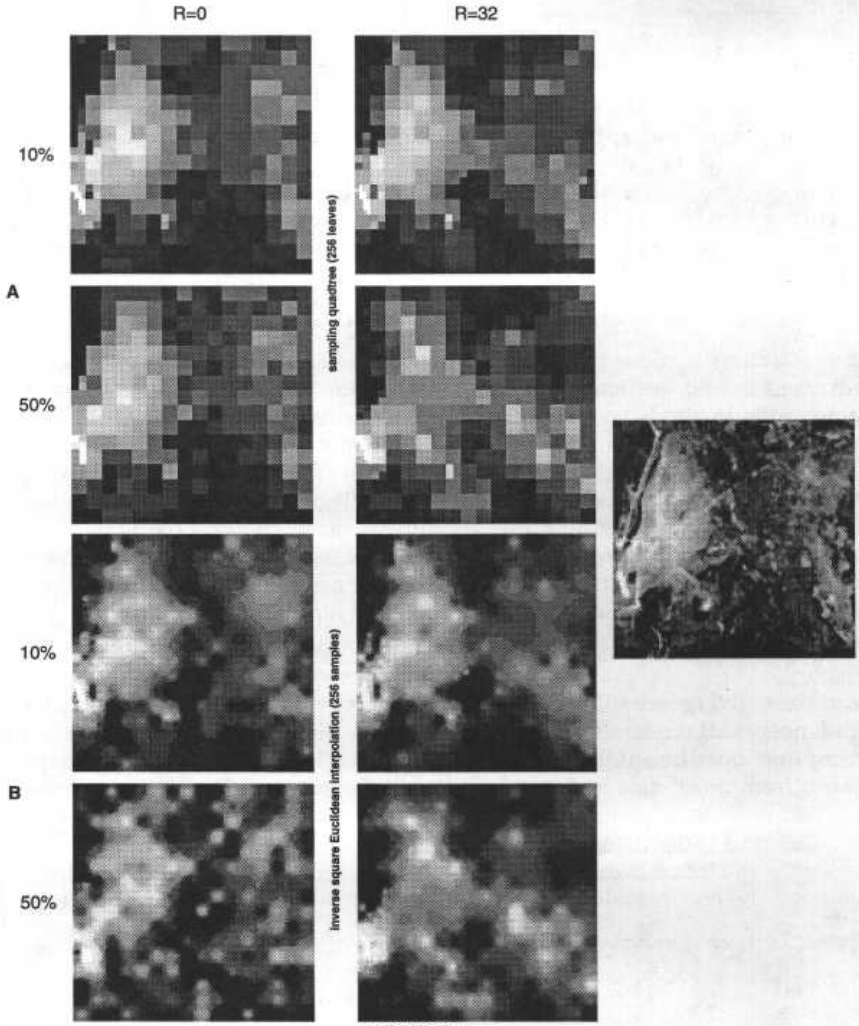

FIGURE 3.

Reconstructions based on 256 samples under various amounts and spatial structure of noise: sampling quadtree (A), inverse square distance (B).

At high (50%) noise levels the differences are more related to the spatial structure of noise. Whereas the Kullback-divergence of 50% noisy data from the original is about six times that of 10% noisy data (Table 2), the Kullback-divergences of reconstructions increase by a factor of only three. Reconstruction by inverse squared Euclidean distance interpolation based on 256 samples, selected by sampling quadtrees from the noisy ancillary data, approximate the original data set comparably than 256 samples selected by other sampling methods from the original data set (e.g., D_K increases from $12.937*10^{-4}$ to $14.723*10^{-4}$ as the correlation length of noise increases; these values are below the ones obtained for regular square sampling [G, $17.550*10^{-4}$] or random sampling [R, $19.824*10^{-4}$] of the original data and significantly below the D_K values between reconstructions and the noisy data set).

CONCLUDING REMARKS

The approximation or mapping procedure described above is a part of a larger project aimed to develop optimal resolution mapping for heterogeneous landscapes, and salt-affected grasslands in particular. The optimization is carried out by locally adjusting (changing) the spatial resolution of the map so that it conveys maximum information for the user with given (predefined) number of patches. Hence, it is a sampling effort constrained optimization.

The sampling quadtrees are computed controlling accuracy by Kullback-divergence, an information theoretical measure to characterize spatial pattern. Several modifications of the current algorithm (with the cutting rule tied to maximum contribution to total Kullback-divergence and straightforward decomposition) are under investigation, as well as are extensions to other, more flexible, hierarchical data structures and links to efficient computations of spatial statistical characteristics based on tile-size distributions (Kummert et al. 1992).

We evaluated the performance of the proposed method under various levels and different spatial structures of noise and compared the results with other (regular square and random) sampling. Reconstructions of two-dimensional distributions on a regular lattice based on sampling quadtrees outperform other sampling designs. The more heterogeneous the surface to be mapped and the fewer (realistically limited number of) samples taken, the more benefit can be gained.

This study provided the foundations for the sampling and mapping procedures of the SHEEP project in northeast Hungary. Further studies of the efficiency and robustness of this and related methods (e.g., by pruning the quadtree and/or formalizing the correspondence among mapped variables) will be evaluated and tested in several other test sites.

ACKNOWLEDGMENT

This research was supported under Grant No. DHR-5600-G-00-1055-00, Program in Science and Technology Cooperation, Office of the Science Advisor, U.S. Agency for International Development.

REFERENCES

BURROUGH, P. A. 1986. Principles of Geographical Information Systems. (Oxford: Clarendon Press).

CSILLAG, F. 1991. Resolution revisited. Proceedings of AutoCarto-10. (Bethesda: American Society of Photogrammetry and Remote Sensing/ American Congress on Surveying and Mapping), pp. 15-29.

CSILLAG, F., KERTÉSZ, M., AND KUMMERT, Á. 1992. Resolution, accuracy and attributes: Approaches for environmental geographical information systems. Computers, Environment and Urban Systems **16**: 289-297.

CSISZÁR, I. 1975. I-divergence geometry of probability distributions. Annals of Probability. **3**: 146-158.

DUTTON, G. 1984. Geodesic modeling of planetary relief. Cartographica **21**: 188-207.

GOODCHILD, M. F. 1992. Geographical information science. Int. J. Geographical Information Systems **6**: 31-46.

GOODCHILD, M.F. and GOPAL, S. 1989. Accuracy of spatial databases. (London: Taylor & Francis).

HEUVELINK, G. 1992. An iterative method for multi-dimensional simulation with nearest neighbour models. In P. A. Dowd and J. J. Royer (Eds.), 2nd CODATA Conference on Geomathematics and Geostatistics, **31**: 51-57, (Nancy: Science de la Terre).

JEON, B. AND LANDGREBE, D.A. 1992. Classification with spatio-temporal interpixel class dependency contexts. IEEE T. on Geoscience and Remote Sensing **30**: 663-672.

KABOS, S. 1991. Spatial statistics. (Budapest: Social Science Information Center, in Hungarian).

KASHYAP, R.L. AND CHELLAPA, R. 1983. Estimation of choice of neighbors in spatial-interaction models of images. IEEE T. on Information Theory. **IT-29**: 60-72.

KERTÉSZ, M., CSILLAG, F. AND KUMMERT, Á. 1993. Mapping heterogeneous images by optimal tiling. (manuscript) submitted to the Int. J. Remote Sensing.

KUMMERT, Á., KERTÉSZ, M., CSILLAG, F. AND KABOS, S. 1992. Dirty quadtrees: pruning and related regular decompositions for maps with predefined accuracy. Technical Report, (Budapest: Research Institute for Soil Science), p. 71.

MARK, D. M. AND CSILLAG, F. 1989. The nature of boundaries on 'area-class' maps. Cartographica **26**: 65-79.

MOELLERING, H. AND TOBLER, W. 1972. Geographical variances. Geographical Analysis **4**: 34-50.

RIPLEY, B. D. 1981. Spatial statistics. (New York: J. Wiley & Sons).

SAMET, H. 1990. Applications of spatial data structures. (Reading: Addison-Wesley).

TOBLER, W. R. 1979. Lattice tuning. Geographical Analysis **11**:36-44.

TÓTH, T., CSILLAG, F., BIEHL, L. L., AND MICHÉLI, E. 1991. Characterization of semivegetated salt-affected soils by means of field remote sensing. Remote Sensing of Environment. **37**: 167-180.

WEBSTER, R. AND OLIVER, M. 1990. Statistical methods in soil and land resource survey. (Oxford: Oxford University Press).