A NONPARAMETRIC GEOSTATISTICAL METHOD FOR ESTIMATING SPECIES IMPORTANCE[1]

Andrew J. Lister, Rachel Riemann, and Michael Hoppus[2]

Abstract—Parametric statistical methods are not always appropriate for conducting spatial analyses of forest inventory data. Parametric geostatistical methods such as variography and kriging are essentially averaging procedures, and thus can be affected by extreme values. Furthermore, non normal distributions violate the assumptions of analyses in which test statistics are generated and compared to a theoretical distribution, such as analysis of variance or stepwise multiple linear regression. Here, we offer guidelines and an example of the use of the indicator approach for dealing with nonparametric data distributions, using data from a study conducted in northern Vermont and New Hampshire.

INTRODUCTION

Recently authors have used the USDA Forest Service's database of Forest Inventory and Analysis (FIA) plots to produce maps of species distributions (Iverson and others 1999, Moeur and Riemann Hershey 1999, Riemann Hershey and others 1997), pockets of high-value commercial trees (King 2000), and forest distribution (Zhu 1994). The methods used to produce these maps have varied from geostatistical simulation (Riemann Hershey and others 1997, King 2000) to advanced multivariate regression-based techniques (Zhu 1994, Moeur and Riemann Hershey 1999, Iverson and others 1999). Few of the techniques, however, have addressed the theoretical and practical problems associated with analyzing highly skewed distributions using parametric statistics. For example, positively skewed data distributions can affect semivariance calculations and kriging weights if the extreme values are located within patches of homogeneous patches of low values. Similarly, traditional statistical methods, such as multiple linear regression, determine the significance of a given model by calculating an F statistic and comparing it to a theoretical distribution. If the data from which the model was built are not normally distributed, erroneous inferences can be made.

Geostatistical techniques such as ordinary kriging and its variants do not inherently require normally distributed data; rather, they assume a multi-point Gaussian random function, described thoroughly in Isaaks and Srivastava (1989), Goovaerts (1997) and Myers (1994). The random function model, which can actually be thought of as a conceptual model, was formulated in part to account for the inherent uncertainty surrounding a set of spatially referenced observations. A random function, in effect, is a set of random variables for each location within a given spatial domain. A random variable is a variable whose values at any location are determined by some probabilistic mechanism. In other words, a reported estimate is drawn from a distribution of estimates that have some probability of occurring at the estimate's location.

Each data point can be conceived of as a random variable whose true value is known, and each estimate to be made is a linear combination of random variables (the known data).

The distribution of values making up a random variable can be described by a cumulative distribution function (CDF), or, as with class variables, by a probability density function (PDF). At any unknown location, a CDF constructed without any additional information regarding the form of the random variable would resemble that in figure 1A. In this situation, the best estimate of a variable at an unknown location would be the sample mean. However, in many earth science datasets, data are spatially dependent, and this knowledge can be used to update the CDF to one that might resemble that in figure 1B. In this instance, one might choose different percentiles of the CDF as an estimate to report, depending upon the goals of the study.

For example, assume there is a set of spatially referenced observations of the importance of species X within a study area. The importance of this species might be dependent upon variables such as soil chemistry, climatic factors, topographic relationships, or the presence or absence of other species. In general, these factors can be assumed to vary relatively smoothly across space. It can be inferred, thus, that areas with high levels of species X are surrounded by other areas of high levels of species X. In other words,

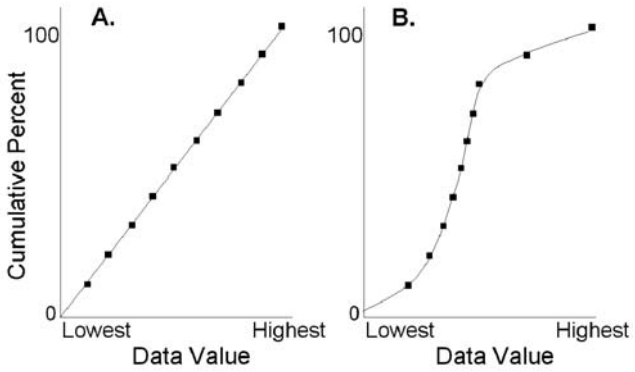

Figure 1—Example of two cumulative distribution functions defining the random variable at an unknown location: A—when no additional information is known about the values; B—when additional information, such as the form of the model of spatial dependence, is known. Points can be defined along B using indicator geostatistics.

[1] Paper presented at the Second Annual Forest Inventory and Analysis (FIA) Symposium, Salt Lake City, UT, October 17–18, 2000.

[2] Forester, Research Forester, and Research Forester, USDA Forest Service, Northeastern Research Station, 11 Campus Blvd., Newtown Square, PA 19008, respectively.

the random function exhibits spatial dependence, or autocorrelation. Autocorrelation, an index of similarity analogous to variance, can be calculated for points separated by distances placed in discrete distance classes. Once the relationship between autocorrelation and separation distance is modeled using variography, the variogram can be used in the estimation procedure (e.g., kriging) to update the CDF to create a conditional CDF (CCDF). Therefore, a given location's CDF is altered, or conditioned on the surrounding data, using the model of spatial dependence constructed for the random function. Under the multi-Gaussian assumption (i.e., that the random variables composing the random function are normally distributed), the mean of the random variable is the simple kriging estimate at that location, and the variance is the simple kriging variance (Isaaks and Srivastava 1989, Goovaerts 1997, Myers 1994).

Again, it is important to note that the non-normality of a distribution of samples does not necessarily imply a non-multiGaussian random function. An observed set of samples can be thought of as one realization of the random function; i.e., the samples could theoretically have had an infinite number of distributions. In geostatistics, the problem with non-normal distributions is that the modeling procedure (the variography) and the estimation procedure (the kriging) are essentially averaging techniques, and can be affected by small numbers of extreme values, a common phenomenon in earth science datasets. To resolve this, sample data with highly skewed distributions are normal score transformed (Deutsch and Journel 1998, Goovaerts 1997). In this procedure, the original CDF is mapped onto the standard normal CDF, giving the transformed distribution perfect symmetry, with a mean of zero and a standard deviation of one. The relationship between the two distributions is defined on a case-by-case basis so that after the estimation procedure is carried out, back transformation can be performed (Deutsch and Journel 1998).

The normal score transform is useful when performing conventional statistical estimation as well. For example, with stepwise multiple regression, variable inclusion and parameter estimates are determined by calculating test statistics that are compared to a reference distribution that was created under the assumption of normality. Large deviations from normality can lead to undesirable outcomes such as heteroscedasticity in regression residuals (Zar 1984), and thus should be carefully evaluated.

INDICATOR GEOSTATISTICS

Another estimation method used when working with highly skewed data distributions is indicator geostatistics. The goal of this paper is to present the general theory and methodology behind indicator kriging (IK), and to present an example of indicator kriging with varying local means (IKLVM). Both are nonparametric geostatistical approaches that avoid many of the abovementioned pitfalls. In IK and IKLVM, the CCDF is constructed by defining discrete points across the entire range of data values (fig. 1B), and then interpolating between these points to arrive at the completed CCDF for each point to be estimated. IK is a univariate approach, while IKLVM allows for the incorporation of ancillary covariates into the estimation procedure.

INDICATOR KRIGING

In IK, the first step in defining the CCDF values at a given location is to determine a series of threshold values or cutoffs from within the range of data values. In practice, deciles of the sample data distribution are chosen because the goal is to define the CCDF for the entire data range. Once these thresholds have been determined, each value is coded as a "1" or a "0", with "1" being assigned to values below that threshold, and "0" being assigned to values above the threshold. Thus in the example where deciles of the distribution are used as cutoffs, 10 datasets consisting of 1's and 0's will be created, one for each cutoff.

The second step in defining the CCDF values is to model the spatial autocorrelation for each of the coded datasets (e.g., deciles) using variography. In order to construct smoothly varying CCDF's, Goovaerts (1997) recommends using the same model or combination of basic models for all of the variograms. For example, the indicator variograms (correlograms) for each of seven percentiles of a dataset shown in figure 2 should be modeled using the same basic structure.

Once these variograms have been created, IK is performed. In IK, each estimate is actually a weighted average of the sample data (1's and 0's) surrounding it, with the weights being derived from the variogram. An IK estimate can be interpreted as the probability of an outcome of 1, or, more specifically, the probability of the actual value at that location being below the threshold used to code the data. This process is repeated for each point to be estimated, and for each threshold dataset. In effect, in keeping with the above decile example, 10 continuous probability maps are created with the value of each pixel being the probability of falling below the threshold used to code the data. Thus, 10 pairs of x,y coordinates (cutoff value, probability of being less than that cutoff value) can be obtained for each location, and 10 points can be placed along the CCDF as in figure 1B. Interpolating between and extrapolating beyond these discrete points to fill in the CCDF should be undertaken with care; guidelines are given in Goovaerts (1997) and Deutsch and Journel (1998).

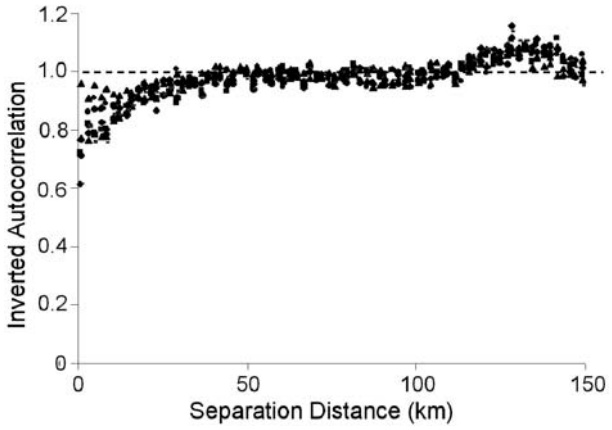

Figure 2—Examples of indicator variograms obtained from spruce-fir importance data from a study in northern New Hampshire and Vermont. See Lister and others (2000) for details. All of the variograms from different cutoff levels show some degree of spatial dependence.

The final step of the process is to choose a percentile of the CCDF to report as a value. Goovaerts (1997) gives a detailed discussion of different criteria that might be employed to make this decision. In essence, the decision will be based on some sort of an optimality criterion defined by the goals of the study. The following section will not only present an example of an indicator technique in practice, but also elucidate some of the steps one might take to develop an optimality criterion for the choice of the estimate to report.

INDICATOR KRIGING WITH LOCALLY VARYING MEAN—AN EXAMPLE

A study was conducted using data from 760 FIA plots in northern New Hampshire and Vermont (for details see Lister and others 2000). The relative importance (relative basal areas) of a combination of red spruce (*Picea rubens* Sarg.) and balsam fir (*Abies balsamea* L.) was determined on each plot, and IKLVM was applied. IKLVM is in principle identical to univariate IK, however the CCDF values are determined by a combination of logistic regression (Montgomery and Peck 1982) and simple kriging of the residuals of the logistic regression. The technique is implemented in a manner similar to that of simple kriging with varying local means, described in Metzger (1997), Majure and others (1996), and Hunner and others (1998). The general expression for the logistic regression estimate is

$$E(y \mid xi..xn) = \frac{\exp(b0 + b1*x1 + ... + bn*xn)}{1 + \exp(b0 + b1*x1 + ... + bn*xn)}$$

where E(y) can be interpreted as the probability of an outcome of "1" occurring (assuming that the data are coded as 0 and 1), b0…bn are the coefficients and x1…xn are the ancillary data layers.

The first step of the process was to determine the appropriate thresholds as described above. The frequency histogram of the data indicates a strong positive skewness (fig. 3), with 30 percent of the data values having 0 percent spruce-fir importance. Consequently, the cutoffs chosen for indicator coding were the 30th – 90th deciles of the original data's distribution, or values of 0, 3, 10, 17, 27, 40 and 60 percent spruce-fir importance. Once these seven coded data sets were constructed as described above, logistic regression was applied to determine, for each point to be estimated, the probability of being below or equal to one of the seven cutoff values (an outcome of "1").

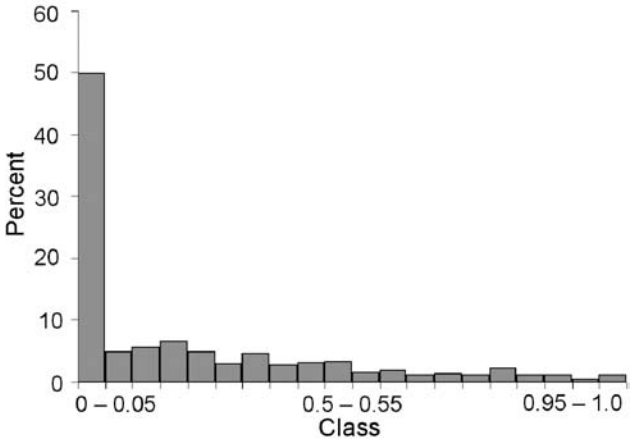

Figure 3—Frequency histogram of the spruce-fir importance data used in the example. The distribution exhibits a strong positive skew, with 50 percent of the data falling in the first class.

In order to build the logistic regression model, exhaustively sampled data layers (i.e., secondary data that were collocated with FIA plots and at all points to be estimated) were chosen based on a combination of user judgment, exploratory data analysis, and stepwise logistic regression. The variables included in the final regression model were Landsat band 4, digital elevation model (DEM) -derived slope and the square root of elevation, and the square root of latitude. One logistic regression model was built for each of the seven cutoffs. All regressors contributed significantly to the model for each cutoff at the 0.01 level, with the majority being significant at the 0.0001 level.

The value of each pixel in the maps in figure 4 was predicted using the logistic regression model for that cutoff. Each value represents the probability that a pixel falls below the cutoff used to code the data. For example, the upper left map represents the probability of a pixel's value being lower than or equal to the original data's 30th percentile, which is 0 percent spruce-fir importance. The highest probabilities (lighter pixels) of having 0 percent spruce-fir importance are seen in areas such as valleys or in clearly nonforest areas, where the logistic regression procedure yielded low values. As the cutoff values increase, the amount of area which probably falls below that cutoff's level of spruce-fir increases until finally, at the 90th percentile, spruce-fir forest has a relatively high probability of occurring at importance levels of at most 60 percent everywhere except forested ridge tops far from roaded areas (shown as darker areas on the map). In these areas, the forests have a high chance of having greater than 60 percent spruce-fir importance.

The next step of the IKLVM procedure was to calculate the regression residuals from each model by subtracting the probabilities from the regression output from the coded data. These residuals are then assessed for spatial dependence using variography. In our example, the inverted correlograms (hereafter referred to as variograms) of the residuals indicate that spatial dependence does exist in the regression residuals (fig. 5). None of the variograms exhibited substantial anisotropy, i.e. the model of spatial continuity did not change with direction.

The next to final step was to use simple kriging to estimate for each map the error at every point in the study area, based on the variograms of the logistic regression residuals (fig. 5). These error (residual) maps were then combined with the regression-based maps with simple addition to arrive at updated maps of IKLVM probability estimates. It is these updated estimates that were used to complete the CCDF.

To reiterate, we created IKLVM maps for each of our cutoff values. If we were to stack these maps one on top of the other and randomly sample any pixel of the stack, we would obtain a set of x, y coordinates that could be used to locate

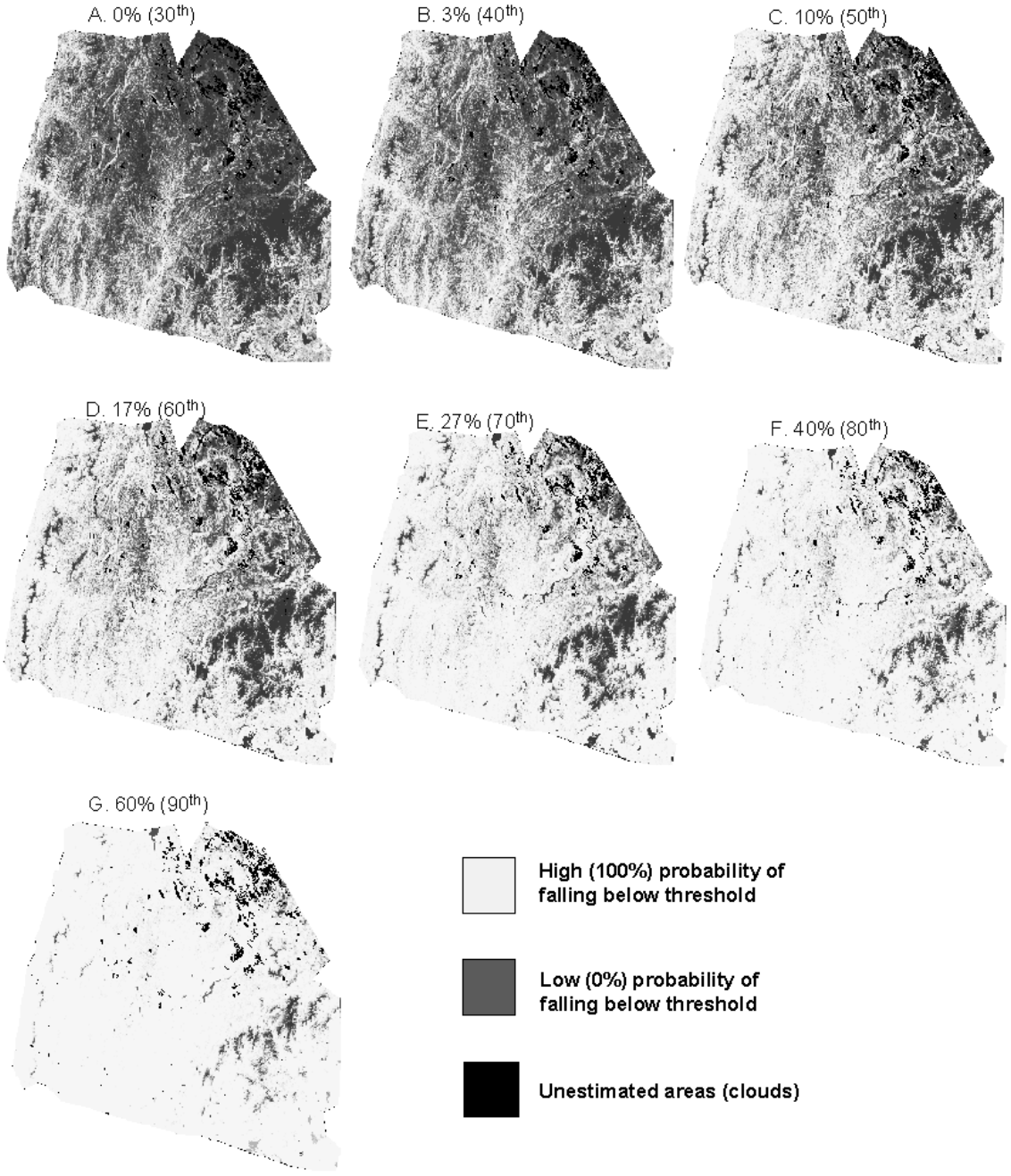

Figure 4—Logistic regression maps constructed from data coded as 1 or 0, based on whether they fall below the indicated threshold percentage of spruce-fir importance. Light pixels have higher probabilities of falling below that cutoff than dark pixels. The effects of topography are readily apparent in the southeast part of the study area.

discrete points on a CCDF similar to that in figure 1B. It is important to remember that each pixel in the map has its own CCDF. In our example, we chose to fill in the CCDF by implementing linear interpolation between points in the center of the distribution and hyperbolic interpolation in the tails. This choice was made based on examination of the resulting CCDF's and assessment of their plausibility, as well as on the suggestions of Deutsch and Journel (1998) and Goovaerts (1997).

Once we created our percentile maps and arrived at a CCDF for each pixel, the final step in the IKLVM approach was to choose a percentile of the distribution to report as the final estimate. Goovaerts (1997) and Deutsch and Journel (1998) discuss criteria that can be used to make this decision. In general, they suggest that the user establish an "optimality criterion", or set of conditions that a "good" estimate must satisfy, and then use this criterion to make the choice. Figure 6 shows scatterplots for both the model fit (A-G) and a set of

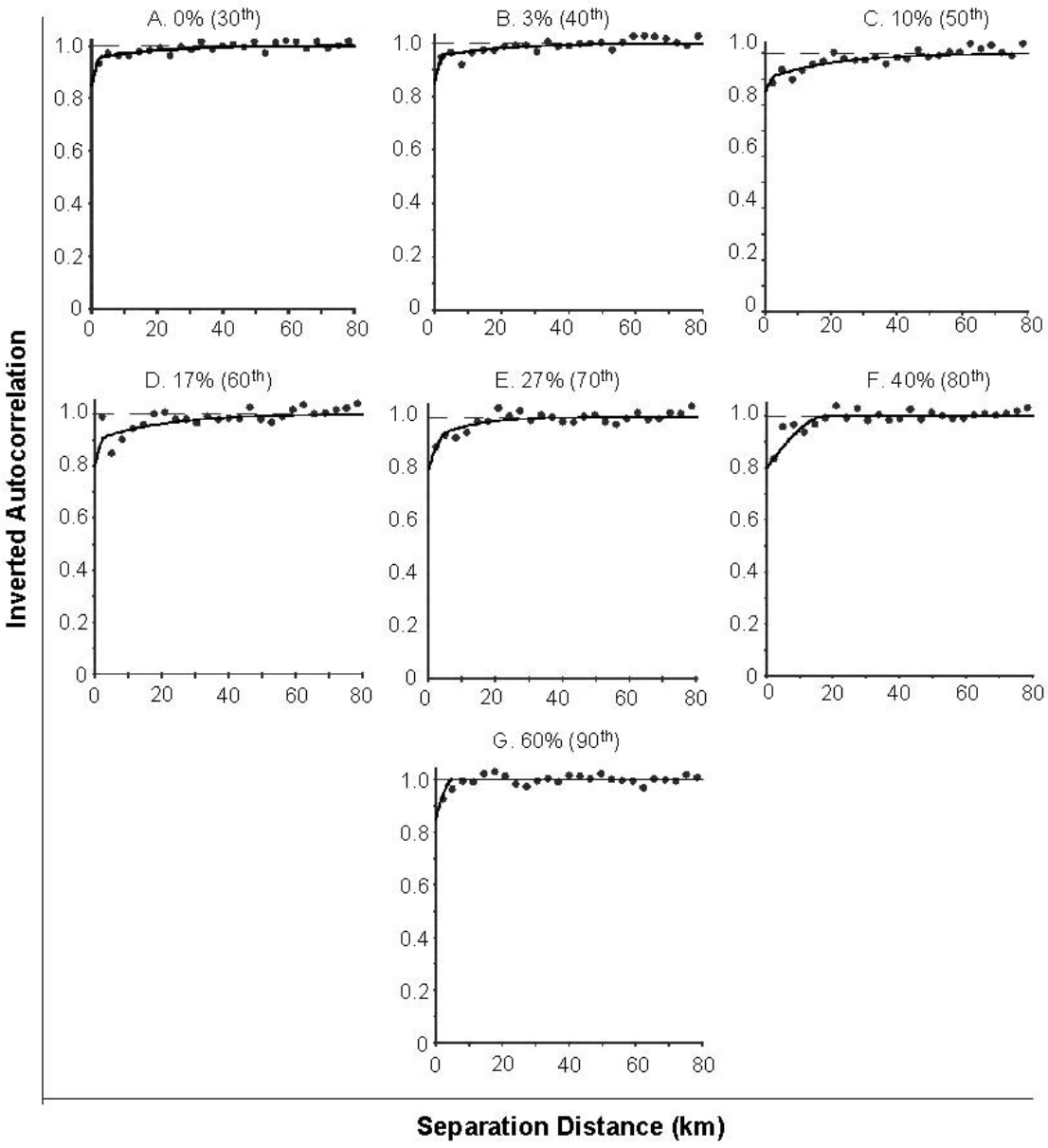

Figure 5—Variograms of residuals from the logistic regression procedure for the indicated cutoff. All residual variograms exhibit spatial dependence.

validation data's fit (H-N) for a range of deciles of the CCDF (the 20th through the 80th) that probably encompasses the final estimate. The dashed line passing through the cloud of points is the 45-degree line, along which all points would fall if the model perfectly predicted the sample points. The dark line is a least squares best fitting line describing the relationship between the actual value (x axis) and the estimate (y axis). The closer the agreement between the least squares best fitting line and the 45-degree line, the more accurate the model is, on average.

It is apparent that for low percentiles, the model dramatically underestimates (fig. 6). For larger percentiles, however, the situation is reversed with values being over predicted. Intermediate values tend to be predicted the best near the middle of the distribution. For some applications, the user might be very concerned about correctly estimating values close to zero, for example, when trying to accurately locate areas with small amounts of some rare but valuable tree species. Before investing in field reconnaissance, a user might want to be as certain as possible that the species occurs at a location, so he or she might choose a percentile where the values are underestimated, for example, the 20th or the 30th percentiles. Similarly, a user might want to find all areas where there are large amounts of a species of interest, in which case the user might choose a percentile

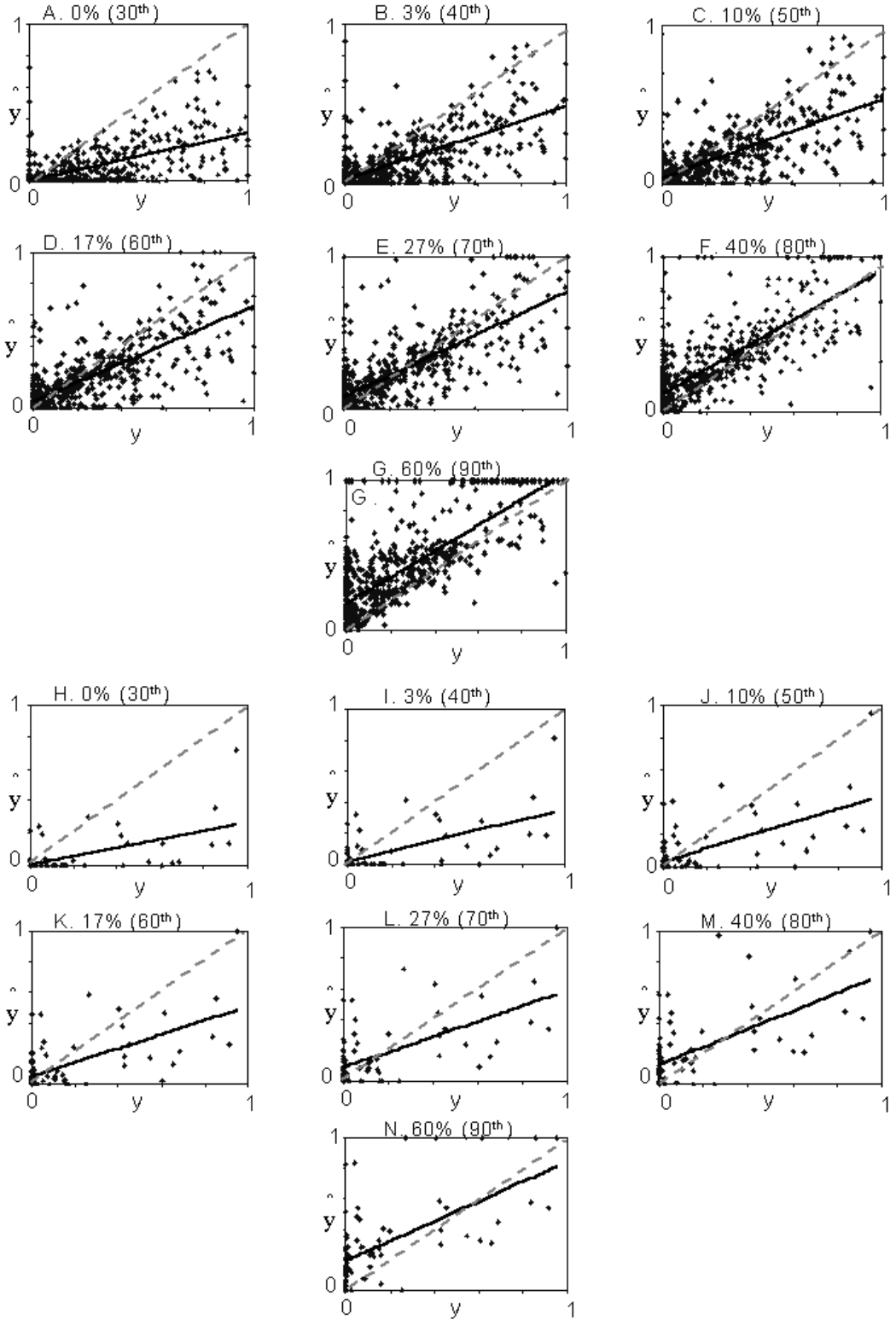

Figure 6—Scatterplots of the model fit of estimates from the 20th to 80th deciles (A-G, respectively), and for the 70 validation data (H-M, respectively). The actual data value is on the x axis, and the predicted value is on the y axis. The gray dashed line is the 45-degree line (perfect agreement), and the black solid line is the least squares best fitting line through the points.

such as the 80th or 90th. The percentile of the CCDF to report as an estimate is therefore chosen based on the relative impact of overestimation or underestimation.

Another criterion might be avoidance of heteroscedasticity of the residuals. The percentile values surrounding the median (figure 6D, E, F and K, L, and M) appear to exhibit roughly equal variance for the entire range of estimates, with the points approximately following the 45-degree line. In our example, we might choose the 70th percentile based on this criterion.

Another criterion that can be applied is the ability of a given percentile to produce estimates with a distribution that resembles that of the original data, either for certain areas of the distribution, or for the distribution as a whole. Figure 7 shows a histogram of estimates from each percentile compared to that of the original data. Using this criterion, the estimates from the 40th percentile (the "x" symbol) of the CCDF's most closely agree with the original data in the first class, which encompasses the lowest importance values for spruce fir (< 0.05, or 5 percent importance) and has the largest class occupancy.

Table 1 shows the results of a quantitative method of comparing the estimate histograms with that of the original data. One might seek to minimize the squared difference between percentages of estimates in each class for the different techniques. In addition, one might want to weight these differences by the magnitude of class occupancy because differences in very populous classes might be more important than differences in less populous classes. For example, for the first class in table 1, the value of 301.991 was arrived at by squaring the difference between the

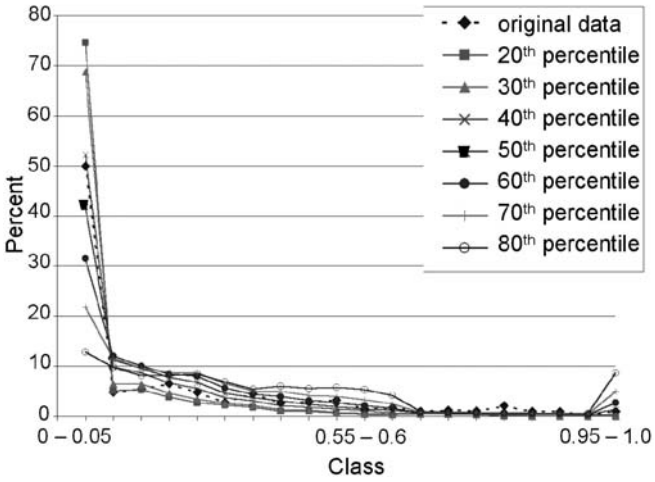

Figure 7—Histograms of the original data (dashed line) and the estimates from each of seven percentiles (solid lines; see legend for details). The 40th percentile histogram has the closest agreement with the original data's histogram.

Table 1—Quantitative assessment of the differences between the histograms in figure 7. For each percentile, the weighted average squared difference between the percentile histogram and the original data's histogram was calculated for each class. The 40th percentile has the minimum weighted average difference

Class	Percent of original data	Squared difference						
		20th	30th	40th	50th	60th	70th	80th
1	50.000	301.991	176.375	2.450	33.863	170.107	398.375	692.871
2	4.816	0.002	0.120	1.202	1.997	2.475	2.226	1.201
3	5.524	0.003	0.039	0.551	0.908	1.091	0.893	0.364
4	6.516	0.453	0.239	0.002	0.101	0.253	0.343	0.168
5	4.816	0.210	0.101	0.020	0.191	0.488	0.713	0.641
6	2.833	0.010	0.002	0.021	0.093	0.214	0.377	0.466
7	4.533	0.337	0.266	0.100	0.033	0.001	0.014	0.036
8	2.691	0.074	0.052	0.007	0.003	0.039	0.137	0.282
9	2.975	0.121	0.090	0.042	0.013	0.001	0.046	0.192
10	3.258	0.228	0.200	0.107	0.060	0.011	0.013	0.194
11	1.416	0.016	0.011	0.001	0.000	0.007	0.048	0.209
12	1.841	0.047	0.042	0.023	0.008	0.001	0.008	0.103
13	0.992	0.002	0.003	0.002	0.001	0.001	<0.001	<0.001
14	1.275	0.010	0.008	0.008	0.007	0.005	0.004	0.002
15	1.133	0.007	0.007	0.006	0.005	0.003	0.003	0.002
16	2.125	0.096	0.070	0.064	0.059	0.052	0.051	0.042
17	0.992	0.010	0.010	0.003	0.003	0.003	0.002	0.001
18	0.992	0.010	0.010	0.004	0.004	0.003	0.002	0.002
19	0.283	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
20	0.992	0.010	0.010	0.005	0.001	0.031	0.159	0.589
Mean squared difference:		15.981	9.350	0.243	1.966	9.199	22.412	38.743

amount of the 20th percentile estimates in class 1 (74.6 percent) and the amount of the original data in class 1 (50 percent). This squared difference was multiplied by 0.5, which is the proportion of the values of the original data in class 1. Because this first class contains 50 percent of the data, any difference observed in this class is more important then one observed in, e.g., class 10, which contains less than 4 percent of the data. We can thus weight each difference by the percentage of original data in that class, and then determine the smallest weighted average squared difference between the actual data and the estimates in order to choose that as the percentile to report. In our example, the choice would be the 40th percentile.

FINAL POINTS

The indicator approach shows itself to be much less restrictive than traditional approaches such as parametric regression, or geostatistics under the multiGaussian assumption. It makes no assumptions about the underlying shape of the CCDF describing the random variable at any location, and it also allows for the incorporation of secondary data, as in the IKLVM procedure. The benefits of incorporating additional "soft information" into the estimation procedure become readily apparent when examining the resulting final maps from univariate IK and multivariate IKLVM (fig. 8). The amount of detail available in the regression-based map is dramatically higher than that found in the univariate-derived map. This is due to the ability of the technique to account for sharp changes in the landscape over short distances. Univariate IK, on the other hand, assumes that a smooth transition occurs between levels of the primary variable in the intervening spaces between the plots; therefore it fails to take into account the fine-scale features.

In conclusion, the random function model allows us to implement indicator geostatistical methods that can alleviate concerns about non-normal data distributions. The use of the indicator approach also allows us to define optimality criteria for reporting a final estimate or creating a map of an environmental variable. We have found that these approaches, especially IKLVM, are useful tools for modeling forestry data.

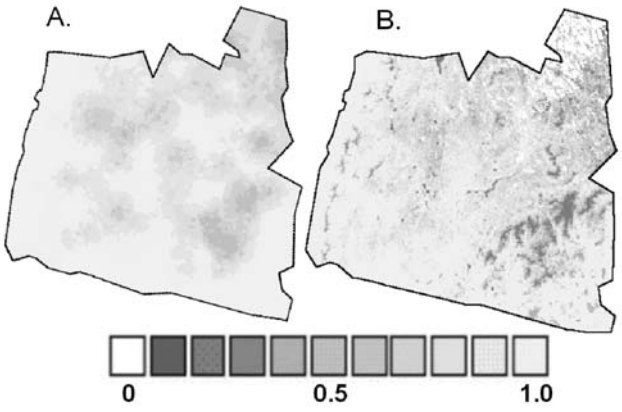

Figure 8—Comparison of the final univariate IK map (A) and the IKLVM multivariate map (B). The IKLVM map reveals much more of the fine scale spatial heterogeneity that exists across the landscape than does the IK map.

REFERENCES

Deutsch, C.V.; Journel, A.G. 1998. GSLIB: Geostatistical software library and user's guide. 2nd ed. New York: Oxford University Press. 369 p.

Goovaerts, P. 1997. Geostatistics for natural resources evaluation. New York: Oxford University Press. 483 p.

Hunner, G.; Reich, R.M.; Mower, H.T. 1998. Modeling forest stand structure using spatial statistics. In: Proceedings of the 2nd southern forestry GIS conference; 1998 October 27–29; Athens, GA: 103–120.

Isaaks, E.; Srivastava, R.M. 1989. An introduction to applied geostatistics. New York: Oxford University Press. 561 p.

Iverson, L.R.; Prasad, A.M.; Hale, B.J.; Sutherland, E.K. 1999. An atlas of current and potential future distributions of common trees of the eastern United States. Gen. Tech. Rep. NE-265. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northeastern Research Station. 41 p.

King, S.L. 2000. Sequential Gaussian simulation vs. simulated annealing for locating pockets of high-value commercial trees in Pennsylvania. Annals of Operations Research. 95: 177–203.

Lister, A.; Riemann, R.; Hoppus, M. 2000. Use of regression and geostatistical techniques to predict tree species distributions at regional scales. 4th international conference on integrating GIS and environmental modeling (GIS/EM4): problems, prospects and research needs. 2000 September 2–8; Banff, AB, Canada.

Majure, J.; Cressie, N.; Cook, D.; Symanzik, J. 1996. GIS, spatial statistical graphics, and forest health. Third international conference on integrating GIS and environmental modeling. 1996 January 21–26; Sante Fe, NM: Santa Fe, National Center for Geographic Information and Analysis.

Metzger, K.L. 1997. Modeling forest stand structure to a ten meter resolution using Landsat TM data. Fort Collins, CO: Colorado State University, M.S. thesis. 123 p.

Moeur, M.; Riemann Hershey, R. 1999. Preserving spatial and attribute correlation in the interpolation of forest inventory data. In: Lowell, K.; Jaton, A., eds. Spatial accuracy assessment: Land information uncertainty in natural resources. Chelsea, MI: Ann Arbor Press: 419–429.

Montgomery, D.C.; Peck, E.A. 1982. An introduction to linear regression analysis. New York: John Wiley. 504 p.

Myers, D.E. 1994. Statistical methods for interpolating spatial data. Journal of Applied Science and Computers. 1(2): 283–318.

Riemann Hershey, R; Ramirez, M.A.; Drake, D.A. 1997. Using geostatistical techniques to map the distribution of tree species from ground inventory data. In: Gregoire, T.; Brillinger, D.R.; Diggle, P.J., eds. Modeling longitudinal and spatially correlated data: methods, applications, and future directions. Lecture Notes in Statistics 122. New York: Springer Verlag: 187–198.

Zhu, Z. 1994. Forest density mapping in the lower 48 States: a regression procedure. U.S. Department of Agriculture, Forest Service. Res. Pap. SO–280. New Orleans, LA: Southeastern Forest Experiment Station. 11 p.