Predictive Model Development and Evaluation with Unknown Spatial Units

Abstract

Existing spatial modeling techniques assume that geographic units of interest can be defined a priori. *While this may be valid in human geography, in the case of naturally occurring phenomena, it is often not possible to identify the spatial units of interest at the beginning of a study. This paper describes how an ecological system may be sampled and modeled and how the model can be evaluated. The sample is a grid-based approach in which the local density of the grid is related to the categories of the dependent variable. Because of its ability to handle* k-*categories of the dependent variable, discriminant function analysis is used to develop the model. Model evaluation involves an assessment of areal and locational accuracy, and examination of spatial autocorrelation among model residuals. Alternative methods for defining residuals are also explained.*

Introduction

In the development of geographic models to quantify, analyze, or project spatial phenomenon, one implicit assumption is that spatial units of interest can be defined *a priori*. For example, one may wish to develop a model which will "explain" or "predict" the amount of cattle present in the counties of a particular region. Therefore, all subsequent analyses must employ "county" as the smallest indivisible geographic unit. Thus, one will collect information for independent predictor variables — e.g., amount of pasture land, distance from a slaughterhouse — based on the county boundaries. This means that, if the model developed is used subsequently to estimate the amount of cattle present in an "area" for which these data are not available, that "area" must be a county.

While such a framework is widely used and accepted, it suffers from a number of limitations. If counties are irregularly shaped, measures of distances from a point feature such as a slaughterhouse may not be meaningful. More important is that counties may have "a lot" of pasture land but, because of its spatial distribution within a county, the number of cattle present in that county is relatively low. Similarly, the subdivision of the area into counties may be inappropriate for this analysis; cattle presence may be related most strongly to "prime pasture soil type" which is unlikely to be related to county boundaries.

This discussion is presented as a means of introducing the problems of applying such techniques to naturally occurring phenomena in general, and an ecological system in particular. In this study, archival aerial photographs (1939 to 1982) were available for a wildlife refuge area in central Missouri which had remained relatively undisturbed since 1939. It was desired to produce a statistical model which would describe the amount and location of change in the vegetative cover types of the area over time. Thus, the cover types Open, Cedar, and Forest had to be projected from 1939 forward to 1982. These represent the initial conditions, middle transition period, and climax stages of ecological succession, respectively, in the study region (Drew, 1942). The rate at which one vegetative type converts to another is thought to be dependent upon soil type, topographic exposure, and the distance from a Forest seed source (Huber, 1971).

If one were to follow conventional geographical modeling techniques, *a priori* one would identify vegetative type polygons present in 1939 as the spatial units of interest. In doing so, one imposes a number of limits on the analysis. First, one must quantify all independent variables relative to these polygons. In reality, the topographic aspect, for example, will change within each polygon. Even allowing the specification of the amount of each aspect for a given polygon will not be satisfactory as, for large polygons, it is the spatial distribution — not the amount — of each aspect that is relevant. Second, and more importantly, any model developed will not accommodate the sub-division of 1939 vegetative type polygons. Thus, as ecological succession occurs, instead of seeing a gradual encroachment of Cedar into Open areas, one will experience an "all at once" conversion.

Thus, it is desired to develop modeling approaches for natural systems which avoid these limitations. The purpose of this paper is to present and discuss a modeling methodology which may be more appropriate for natural systems than a more conventional methodology which demands that spatial units of interest be identified *a priori*. Specifically, the paper will present and discuss

- a spatial sampling scheme for predictive model development;
- techniques for the development and evaluation of predictive models, including accuracy assessments and evaluation of spatial autocorrelation among residuals; and
- the effect of spatial unit definition on the measurement of spatial autocorrelation.

Background

The following description of the study area, data, and nature of ecological succession is discussed in more detail in Lowell (1991); a synopsis of these items is presented here.

Study Area

The study area is a 932-ha wildlife refuge area — the Thomas W. Baskett Wildlife Research and Education Center (BWREC) — located in central Missouri in the United States. Until 1937, this area comprised a number of separately owned farms devoted primarily to row crops and pasture; a minor portion had remained forest. In 1937, the entire area was procured by the University of Missouri for the study of

Photogrammetric Engineering & Remote Sensing,
Vol. 59, No. 10, October 1993, pp. 1509–1515.

0099-1112/93/5910–1509$03.00/0

Kim E. Lowell
Centre de recherche en géomatique, Université Laval,
Pav. Casault, Cité Universitaire,
Ste-Foy, Québec G1K 7P4, Canada.

wildlife and its associated habitat. Only two types of disturbances have occurred on the area since that time. A number of grazing leases signed by individual farmers before 1937 were honored; the last of these expired in 1962. Disturbances related to forestry and wildlife research have also occurred, including the establishment of pine plantations as well as burning and clearcutting to improve and study wildlife habitat. Cartographic records of the areas affected have been maintained.

Data

Aerial photographs of the BWREC were available for a number of dates between 1939 and 1982 (Table 1); a U.S. Soil Conservation Service soils map was obtained; a U.S. Geological Survey Topographic map was acquired; and a map showing the distance to a Forest boundary in 1939 was produced using the 1939 vegetation base map and the buffering capabilities of a GIS. Each of the maps was treated as described in Table 1 for incorporation into a raster data model (11-m by 11-m cell size) registered to the Universal Transverse Mercator (UTM) projection. Topographic aspect was recoded into three classes: Cool (north, northeast, east), Warm (west, southwest, south), and Neutral (northwest, southeast, flat).

During subsequent analysis, the soils information was found to be spatially redundant with land cover in 1939. Further examination of this phenomenon showed that the soils map had been produced primarily from the 1939 aerial photographs also used in this study to map 1939 land cover. Thus, their redundancy is not surprising; the soils information was dropped from further consideration.

Ecological Succession

In Missouri, in the absence of disturbance, agricultural and pasture land will be invaded by eastern redcedar (*Juniperus virginiana* L.) and converted to this type after 15 to 20 years (Henning, 1937). This vegetative type will in turn be replaced by the climax oak-hickory forest type (Eyre, 1980). Three vegetative types representing these three stages of succession appeared on each land-cover map: Open, Cedar, and Forest. A fourth cover type — Disturbed — was also included, representing those areas which had been disturbed for grazing or research purposes at any time between 1939 and 1982. Thus, to project the 1939 cover type forward, it was necessary to model succession on four cover types based upon knowledge of how ecological succession could be expected to proceed over time on each.

By definition, Disturbed would remain Disturbed. Forest was expected to remain Forest but, because of minor photointerpretation differences and map registration and digitizing errors, it could not be expected that 100 percent of the areas mapped as Forest in 1939 would also be mapped as forest in 1982 and all years in between. It is known that Cedar can compete most successfully on limestone ridges and dry areas. Thus, the speed at which Open areas succeeded to Cedar was expected to be related to the underlying soil type and topographic aspect. The speed at which the pioneering Cedar yielded to the climax Forest was also expected to depend upon both soil and aspect, but also to the distance a given area was from a Forest seed source. This latter is true because the Forest type is comprised principally of oaks (*Quercus* spp.) and hickories (*Carya* spp.) which have relatively heavy seeds. These are dispersed only short distances by small mammals such as squirrels. Distance from a seed source was not considered a factor in the conversion of Open to Cedar as eastern redcedar is dispersed by being eaten by birds, remaining viable through the digestive tract, and being defecated when the bird is flying. This must be considered a spatially random process.

Sampling Scheme

A number of underlying assumptions must be made in employing the proposed sampling scheme. It is assumed that there are underlying polygons which are meaningful relative to the dependent variable (cover type in a given year) and which can be defined by combinations of the dependent and independent variables. These polygons are not definable *a*

TABLE 1. SUMMARY OF DATA SOURCES.

GIS layer	Base map source	Scale of map	Processing required to produce GIS layer
Land cover for 1939, 1950, 1956, 1962, 1970	Archival B+W stereo lead-off aerial photographs	1:10,000	Photo-interpretation and transfer to base map; digitizing; overlay with 1:10,000-scale map of Disturbed areas.
Soils	U.S. Soil Conservation	1:31,680	Digitizing.
Topography	U.S. Geological Survey 7.5-minute quad sheet	1:24,000	Digitizing one elevation point per 0.0081 ha; distance-weighted elevation interpolation; aspect extrapolation.
Forest boundary	1939 GIS land-cover map	11-m GIS cells	Screen-digitizing of forest/non-forest boundaries; generation of buffer zones.

TABLE 2. COVER TYPE AND SAMPLING SCHEME DESCRIPTION

1939 Cover Type	Expected Successional Variance	Total 1939 Area on the BWREC		Sample Points		Points per ha
		ha	%	No.	%	
Forest	Low	502	54	223	19	0.4
Cedar	Moderate	7	1	46	4	6.3
Open	High	282	30	782	69	2.8
Disturbed	0 (Zero)	141	15	90	8	0.6
	Totals	932	100	1141	100	

priori as it is impossible to determine exactly which independent variables and which categories of these will be related to the dependent variable. That is, if topographic aspect is related to ecological succession, perhaps only south-facing slopes affect ecological succession. Or it may be that north- and northeast-facing slopes are different from each other with respect to ecological succession and the rest are uniform.

Nonetheless, different polygons defined by the same variable/category combination can be expected to be similar to each other with respect to ecological succession. Furthermore, one of the dependent variables may be identifiable as a "dominant" or "controlling" variable. In this case, it seems fairly obvious that the cover type that one finds at a given location will be dependent on its cover type at the start of succession (1939). It makes sense, therefore, to start with such a dominant variable and describe the expected behavior of each of its categories relative to the dependent and independent variables. One then has a better chance of drawing a sample from each category of the dominant variable which is dense enough to have a good chance of representing all polygon combinations and their variability. This may necessitate a stratified sampling density, with the density in any given area being based on the categories of one or more variables.

The sampling scheme employed here is a regular grid whose spacing is dependent on the cover type in 1939 (Table 2); thus, it is a stratified sample with the stratification being based on the cover type in 1939. The sampling densities for each type were chosen primarily by intuitive means such as a rule-of-thumb of a minimum of 30 samples per cover type. Furthermore, in a raster GIS one cannot sample at a density higher than the grid size itself (11 m in this case). Note that, if the model to be fitted subsequently is likely to be cover type specific because of the use of dummy variables (see also Model Development section, Table 3), an oversampling of a given type will not affect the model description of succession on other types.[1]

The sampling intensity of those types having low expected ecological succession variance — Forest and Disturbed — is relatively low. These types covered a combined 69 percent of the BWREC in 1939, yet only 27 percent of the sample points resided on these types (Table 2). In absolute terms, there are a large number of points — 223 and 90 — to represent Forest and Disturbed, respectively. While this sample size does not ensure adequate representation of these types, given their low expected successional variability, it should be sufficient. Open was the most variable type with respect to ecological succession and also covered a relatively large percentage of the BWREC in 1939 (30 percent). Thus, the sample density for Open was moderately high (2.8 points per ha), and produced the greatest number of sample points — 782 — of any type. Cedar was the type that was the least-present in 1939. However, because Cedar had a moderate expected successional variability, it was sampled the most intensely at 6.3 points per ha (Table 2). Nonetheless, this yielded the least number of sample points for any type — 46 — because of the small amount of Cedar present in 1939[2].

Note that, while the sample drawn appears to "make sense" intuitively, no statistical constraints have been placed on it. Recall, however, that a basic premise of this sampling scheme is that there are underlying, yet undefinable, polygons in the system which require representation. Thus, the initial goal of the sample was to represent the possible combinations of

- 1939 cover type with four classes,
- topographic aspect with three classes,
- soils with four classes, and
- distance from Forest as a continuous variable.

Given (4 by 3 by 4) 48 possible variable combinations plus one continuous variable, a total sample size of 1141 points (Table 2) is likely to be adequate. It is acknowledged, however, that if 48 types are possible, ten or so will probably comprise around 90 percent of the area meaning that the minor types will probably be undersampled regardless of the sampling scheme. If the undersampled types are critical to a phenomenon being studied, users are well-advised to examine further the distribution of samples within variable combinations and possibly adjust it accordingly.

At each of the 1141 sample points, the values for each of the seven measurement years, and the variables aspect, soil type, and distance from Forest in 1939 were recorded. These were exported into an aspatial ASCII file for subsequent model development. The sample points were split into two data sets. The calibration data set contained approximately 80 percent of the samples and was used for model development; the remaining 20 percent was reserved for model validation.

Model Development

Because the dependent variable — cover type at a given year — is categorical, and because there are more than two categories of the dependent variable, discriminant function analysis (DFA) was employed to produce the predictive model. DFA produces a set of equations — one equation for each category of the dependent variable — which are linear combinations of the independent variables.

Stepwise DFA was employed to fit two models (Table 3) using the calibration data set. The first, which will subsequently be referred to as the "Full Model," estimated cover type at any given year using four independent variables and selected interactions thereof: 1939 cover type, aspect recoded into Warm, Cool, and Neutral slopes; distance from Forest in 1939; and the number of years elapsed since 1939. The second DFA model, which will subsequently be called the "Re-

[1]The stratified systematic sampling scheme was employed whose density varied with the land cover present at the year considered to be the beginning of ecological succession. It would also have been possible to make the grid density vary with additional variables by overlaying these and specifying a sample density for each variable combination. Moreover, it is not necessary that a systematic grid be employed for sampling; a random sample of a specified density could be extracted instead. This might be considered preferable because parametric statistical techniques assume a random sample. However, in order to ensure coverage of the entire territory, a grid-based sample was employed herein.

[2]The Cedar type produced the fewest samples (i.e., 46) of any type. If this number had been judged inadequate for model development, this could have been increased in two ways. First, the sampling intensity could have been increased. Instead of taking 6.3 points per ha (Table 1) or one point per 0.16 ha, one could have sampled more with the functional limit being only the cell size (0.0121 ha). However, such a move possibly would have caused so much spatial autocorrelation among sample points that results for the type would have been spurious. A second alternative might be preferable. The first year following 1939 for which a land cover map was available (1950) could have been used as a "base year." The Cedar areas present in 1950 which had not been present in 1939 could have been sampled. The same could also have been done for the following years. While this would have produced measurements for only (1982 − 1950 =) 32 years instead of the (1982 − 1939 =) 43 years of the life of the study, it would nonetheless have provided additional data for Cedar for certain time periods.

duced Model," estimated cover type at any given year as a function of 1939 cover and the number of years since 1939 only.

To estimate the cover type for an unknown GIS cell using either DFA model, one notes the value for each of the independent variables and calculates the discriminant score for each of the cover types using the equation calibrated for that type. The cell is then assigned to the cover type whose equation produces the largest discriminant score. This procedure was followed for each raster cell and each model to produce maps of estimated land cover for each of the years for which land cover was known (Table 1). The predictive ability of each DFA model was then evaluated using these Estimated maps.

Model Evaluation

To simplify the explanation of model evaluation procedures, results will be presented for 1982 only, because this is the final year of the study and because results for 1982 were fairly representative of results which occurred throughout the duration of the study.

Locational and Areal Accuracy

One way to evaluate predictive models which have been developed from, and applied to, a spatial system is through an error or confusion matrix (Congalton and Mead, 1983). That is, one uses a model to estimate land cover at some point when the actual land cover is known, the two maps are overlayed, and the areas on each cross-tabulated. An evaluative statistic such as the kappa coefficient k may then be calculated (Foody, 1992).

Such an approach is concerned principally with locational accuracy: are land-cover types predicted to be in the right place? Another concern in model evaluation, however, is whether or not a model estimates the amount of each cover type correctly. One may desire to know the amount of pasture land in a township and not be overly concerned with its distribution within the township. Thus, areal accuracy may be important: is the total amount of each cover type estimated "well?" To assess areal accuracy, one can determine how much of each cover type is present on the Estimated and Actual maps. These amounts can then be compared for all cover types using the Student's t statistic.

Note that an assessment of areal accuracy without an examination of locational accuracy can be extremely misleading. Clearly, if the locational accuracy is zero (i.e., nothing located correctly) but areal accuracy is 100 percent (i.e., the total amount of each type is exactly correct), one's model is performing extremely poorly because it does not appear to be using the underlying information available. Thus, while one would like the estimate to have high areal accuracy, it should not be considered to be as useful as locational accuracy for assessing model performance.

In this study, areal and locational accuracies were assessed in two ways. First, because sample points were drawn from the GIS database, it could be argued that these sample points are the spatial units of interest and that model evaluation should be conducted using these points. Such a supposition ignores the fundamental assumption that, in the ecological system modeled, there are underlying polygons on which ecological successional processes occur uniformly, yet which are undefinable *a priori*. Nonetheless, for comparative purposes it would be useful to examine these points. Second, after the development of the DFA models, the underlying polygons assumed to exist can be defined by virtue of the variables which entered the models. That is, in the Full model aspect was important but only on the Open type. Thus, aspect further defines polygons within those polygons designated as Open in 1939. Locational and areal accuracies were assessed for the sample points in the validation data set and the *a posteriori* polygons (area).

TABLE 3a. THE FULL DFA MODEL. ALL INCLUDED VARIABLES STATISTICALLY SIGNIFICANT ($a = 0.01$). VALUES ARE COEFFICIENTS FOR A VARIABLE FOR EACH COVER TYPE.

Variable[1]	Forest	Cedar	Open	Disturbed
Constant	−14.50	−13.32	−6.10	−2.52
Forest	29.46	24.24	14.37	0
Cedar	23.12	27.76	14.23	0
Cedar*Time	0.117	−0.070	0.003	0
Open*Cool	4.66	3.28	3.18	0
Open*Neutral	4.04	2.95	3.34	0
Open*Time	0.66	0.67	0.36	0
Open*Edge	0.012	0.018	0.022	0

TABLE 3b. THE REDUCED DFA MODEL. ALL INCLUDED VARIABLES STATISTICALLY SIGNIFICANT ($a = 0.01$).

Variable	Forest	Cedar	Open	Disturbed
Constant	−12.60	−11.99	−4.54	−2.52
Forest	25.68	21.17	11.12	0
Cedar	20.14	25.17	11.42	0
Cedar*Time	0.102	−0.080	−0.006	0
Open*Time	0.658	0.677	0.359	0

[1] Variables:
Forest, Cedar, Open—1 if Forest, Cedar, or Open in 1939; else 0.
Time—Number of years since 1939.
Cool—1 if aspect is north, northeast, or east; else 0.
Neutral—1 if aspect is northwest, southeast, or flat; else 0.
Edge—distance from a Forest in 1939.

In all cases, both the Full and Reduced models perform identically relative to areal and locational accuracies (Tables 4 and 5). This is somewhat surprising given that the models are not identical, and that all variables contained in each were statistically significant. This suggests that for the Full model some spatial dependence among model residuals was present. That is, variables entered the Full model which were statistically significant but which had no practical predictive ability. This will be examined in the following section.

Note that the use of only sample points underestimates the locational accuracy (70 percent — Table 4) compared to using all cells in the entire area (85 percent). Kappa coefficients (k) are also lower for the points than for the area. This reinforces the suggestion that, after using the data from the sample points to develop the DFA model(s), they are poorly suited for model evaluation if one is interested in the entire area (as is usually the case). Nonetheless, the sample point evaluation can be useful for identifying those cover types which are likely to be confused. For both points and areas, Forest and Cedar were often confused whereas, as expected, Disturbed was not confused with any other cover type.

For areal accuracy, both the points and areas give approximately the same results (Table 5) as the percent difference is comparable for Cedar, Open, and Disturbed and there is roughly the same amount of error in the total amount of each of these types. The sole important difference was found in the Forest type for which the percent error for area (16 percent) is considerably less than that for points (56 percent).

TABLE 4. LOCATIONAL ACCURACY FOR THE REDUCED AND FULL MODELS[1] POINTS AND AREAS.

FULL MODEL AND REDUCED MODELS (DATA POINTS — VALUES ARE NUMBER OF POINTS)

		Actual			
	Class	Forest	Cedar	Open	Disturbed
Estimated	Forest	64	3	0	0
	Cedar	90	135	4	0
	Open	0	0	0	0
	Disturbed	0	0	0	25

$\kappa = 0.364$
Overall accuracy (diagonal) = 70%

FULL MODEL AND REDUCED MODELS (AREA — VALUES ARE HECTARES)

		Actual			
	Class	Forest	Cedar	Open	Disturbed
Estimated	Forest	487	18	1	0
	Cedar	117	163	5	0
	Open	0	0	0	0
	Disturbed	0	0	0	141

$\kappa = 0.849$
Overall accuracy (diagonal) = 85%

[1] Despite the differences in the models, the locational accuracy was identical for points and areas.

TABLE 5. AREAL ACCURACY FOR THE FULL AND REDUCED MODELS[1] FOR POINTS AND AREAS.

			Difference	
Class	True	Estimated	No.	%
Full and Reduced Model (Data points)				
Forest	154 points	67 points	87	56
Cedar	138 points	229 points	−91	66
Open	4 points	0 points	4	100
Disturbed	25 points	25 points	0	0
Full and Reduced Model (Area)				
Forest	604 hectares	506 hectares	98	16
Cedar	181 hectares	285 hectares	−104	57
Open	6 hectares	0 hectares	6	100
Disturbed	141 hectares	141 hectares	0	0

[1] Despite the differences in the models, the areal accuracy was identical for points and areas.

Spatial Independence of Residuals

One assumption of parametric statistical techniques is that model residuals are randomly distributed. It has already been suggested that this is not the case for the Full model and that spatial autocorrelation is present among its model residuals. It remains to verify this. A number of measures of spatial autocorrelation have been developed for continuous and categorical data (Dacey, 1968; Cliff and Ord, 1973; Odland, 1988).

- The Join-Count statistic is used when map classes are categorical and data are nominal or ordinal. In a binary system such as residuals of "correct/incorrect" (i.e., "black/white"), the Join-Count statistic compares the actual number of "black-black (BB)," "white-white (WW)," and "black-white (BW)" links against the expected number in a random spatial system. The actual number of each is tested for significant differences from the expected number using the Student's *t* statistic. An excessive number of BB and WW links indicates non-random clustering or positive spatial autocorrelation. An excessive number of BW links indicates non-random dispersion or negative spatial autocorrelation.
- Both Moran's *I* and Geary's Contiguity Ratio are used when map classes are interval or ratio data. Though based on different methodology, for a spatial system both calculate a statistic with a well-defined distribution. As with the Join-Count statistic, this statistic can be compared to the critical value one would expect if the links in the system were randomly distributed and can be tested for statistical significance using the Student's *t* statistic to indicate significant positive or negative spatial autocorrelation.

To measure spatial autocorrelation among model residuals it would seem to be simply a matter of developing a map of residuals from the Estimated and Actual maps and applying the appropriate test. This is not as straightforward as it appears, however. The spatial units on which to measure spatial autocorrelation must be determined and a method for residual calculation must be determined.

There are three ways that one may define the spatial units. First, one may consider only the sample points themselves. In doing this, for the purpose of measuring spatial autocorrelation, one must still determine which points are adjacent. This is done by generating a Voronoi diagram composed of Thiessen polygons around each point.[3] Note, however, that it is the sample point location only which is used to determine the model residual, and not the entire Thiessen polygon associated with each point; these latter are used only to determine contiguity among points for the purpose of measuring spatial autocorrelation. Second, one may consider the Thiessen polygons to be the spatial units of interest rather than considering only the central point used to generate each. Third, one may define polygons by using the DFA model(s) *a posteriori*. Recall that it was assumed that there are underlying polygons which are meaningful relative to ecological succession but which could not be identified with certainty *a priori*. However, by virtue of a variable having been included in one of the DFA models it can be said to be "important" — at least statistically. Thus, the polygons of interest for the Reduced model can be defined by the boundaries of the land-cover classes in 1939. For the Full model, they will further be defined by three topographic aspects, and the distance from a Forest in 1939. All three approaches to spatial unit identification were examined in this study.

After definition of the spatial units, it also remains to determine the residual for each. For the points, the only residual possible is binary: correct/incorrect. To obtain this information, the land cover on the Estimated and Actual maps is noted for each sample point and a value of "correct" is assigned if they agree; otherwise, they are labeled "incorrect."

[3]The Thiessen polygon for any given point will be defined by the perpendicular bisectors between the point and its immediate neighbors.

Determining the residual for polygons is more difficult as it is not likely that the cover type will be uniform for each polygon on the Actual map. Note that each polygon on the Estimated map will be uniformly covered as each is uniform relative to the variables in a given DFA model. Thiessen polygons will also be uniform on the Estimated map as the cover type of the sample point is assigned to the entire polygon. There are three ways to determine a polygon residual; all three require that the polygons defined be overlayed on the Actual map and the true cover types for each polygon be tabulated.

- In a Plurality residual rule, if the most common land cover on the Actual map is the same as the cover for the polygon on the Estimated map, the polygon is considered "correct." Thus, a binary residual is produced. However, with four map classes, as little as 26 percent of a polygon on the Actual map may be covered with the Estimated cover and still be considered "correct."
- A Majority residual rule can be employed in which at least 50 percent of an Actual polygon must be covered with the Estimated type to be considered correct. Again, this produces a binary residual.
- Instead of using binary measures of "correct/incorrect," for each polygon the amount of area that is correct may be recorded as the Percentage. That is, if the polygon on the Estimated map is covered by Forest and 60 percent of the same polygon is covered by Forest on the Actual map, the residual would be 0.60.

Each of these three methods of defining residuals was examined for each of the three methods of defining spatial units, and spatial autocorrelation was measured for the Full and Reduced models using the appropriate statistic (Table 6).

The measures of spatial autocorrelation for the Full model consistently show that its residuals are not spatially independent and have a tendency to group (positive spatial autocorrelation). The same is true for the Reduced model and/or when the sample points or Thiessen polygons generated from the points are considered. This suggests that the same ecological succession which occurs at one point also occurs at neighboring points. This reinforces the idea that there are meaningful polygons relative to ecological succession even though they were not defined *a priori*. The points being positively spatially autocorrelated implies that these are within such a polygon and that it is not correct to consider them to be the individual spatial units of interest.

When considering the residuals for the DFA-defined polygons of the Full and Reduced models, it becomes evident that positive spatial autocorrelation is replete among the residuals of the Full model. This suggests that polygons that have been called different statistically are, in fact, the same relative to ecological succession. The apparent absence of spatial autocorrelation among the residuals of the Reduced model indicates that neither aspect nor distance from a Forest were necessary to describe the ecological succession on the BWREC. (The reader is reminded that an evaluation of Locational and Areal accuracy suggested the same thing.) Given that aspect and distance from a Forest were statistically significant, this suggests that the positive spatial autocorrelation caused an underestimate of the variance of the system and led to spurious tests of significance. A similar phenomenon has also been documented in the presence of temporal autocorrelation in aspatial modeling (Ferguson and Leech, 1978).

TABLE 6. SPATIAL AUTOCORRELATION AMONG RESIDUALS AS CALCULATED IN A VARIETY OF WAYS. SEE TEXT FOR EXPLANATION OF EACH; VALUES ARE STUDENT'S *t*.

Residual calculation/measure		Full Model	Reduced Model
DFA Polygons/Plurality			
Join-Count Statistic:	BB	14.70**	1.83
	BW	−31.52**	0.99
	WW	7.14**	−2.59**
DFA Polygons/Majority			
Join-Count Statistic:	BB	19.66**	1.83
	BW	−34.44**	1.05
	WW	9.76**	−2.58**
DFA Polygons/Percentage			
Geary's Ratio		18.51**	−0.25
Moran's *I*		25.24**	−1.50
Points/(Correct/Incorrect)			
Join-Count Statistic:	BB	22.19**	28.59**
	BW	−36.38**	−45.97**
	WW	38.81**	47.03**
Thiessen Polygons/Plurality			
Join-Count Statistic:	BB	8.56**	8.56**
	BW	−11.03**	−11.03**
	WW	8.95**	8.95**
Theissen Polygons/Majority			
Join-Count Statistic:	BB	10.59**	10.59**
	BW	−15.69**	−15.69**
	WW	15.44**	15.44**
Theissen Polygons/Percentage			
Geary's Ratio		8.85**	8.85**
Moran's *I*		14.78**	14.78**

"**" indicates significantly different from zero ($\alpha = 0.01$).

Discussion and Conclusions

In natural systems, others have attempted to develop sampling schemes which account for spatial autocorrelation *a priori*. In particular, Pereira and Itami (1991) sought to develop a wildlife habitat model for red squirrel. For the independent variables in their study, Moran's *I* was calculated for the nearest cell (cells were square and 70.7 m on a side), then recalculated using the second-order neighbor cells, then the third, etc. Because spatial autocorrelation had diminished sufficiently by the seventh lag (495 m) for all independent variables, each seventh cell was sampled. A logistic regression model was then developed under the assumption that samples were independent. While such an approach is useful, spatial autocorrelation of model residuals was not evaluated *a posteriori*. Conversely, in this study, spatial autocorrelation was not evaluated *a priori* due to the difficulty of identifying polygons which were uniform relative to selected variables which were also related to ecological succession. If this could have been done, an *a priori* assessment of spatial autocorrelation may have been useful. Note that the results of such an assessment may vary widely as spatial autocorrelation is likely to be related to the characteristics of the system under study and the phenomenon being analyzed, and, in a raster GIS, the size of the cells.

Discriminant function analysis was employed for model development because of the presence of a categorical dependent variable. Logistic regression is another technique which can accommodate categorical variables by producing a model which estimates the probability of the occurrence of a binary event. Despite the limitation of being applicable to binary dependent variables only, logistic regression has been shown to be more robust for analyses involving categorical and continuous independent variables (Press and Wilson, 1978). Thus, if the analytical questions consider only questions of binary phenomena — "Forest or not forest?" rather than "Which of four cover types?" — then logistic regression may be a preferable alternative for model development.

In measuring spatial autocorrelation among model residuals, a simple contiguity matrix was employed. That is, polygons/points were considered only as "touching/not touching" and the magnitude of the connection was not considered. Others, however, have suggested that connections should be weighted by the length of the common boundary between two types (Cliff and Ord, 1981; Goodchild, 1986). It is doubtful that this would have changed the results significantly for the points as these were systematically located and there was relatively little variability in the length of common boundaries. It is possible that results would have changed for the polygons although unpublished work by the author concerning boundary length weights and spatial autocorrelation suggest that this is not the case in a forest system. Nonetheless, it may be worthwhile to examine this is the future.

Model evaluation showed that the model which produced the best fit of the sample points statistically was not necessarily the best model spatially. In this study, both a Full model – which was statistically optimal – and a Reduced model performed equally well spatially over the entire area relative to locational and areal accuracies. Analysis of spatial autocorrelation among residuals showed that the Full model identified factors which were statistically important for modeling ecological succession but which actually had no practical predictive ability. Spatial autocorrelation analysis also suggested that, if one uses the original sample points as the spatial units of interest, positive spatial autocorrelation is likely to be present. This supports the basic premise of this paper that, in a system where spatial units which are uniform relative to a dependent variable cannot be identified *a priori*, sample points can describe these underlying polygons; the positive spatial autocorrelation among model residuals was indicative that same-type polygons had been sampled.

This paper has attempted to provide a framework for the development and evaluation of models for spatial systems in which spatial units cannot be identified *a priori*. Often, the first decision made in a spatial modeling analysis is what are to be the units of interest. These are often defined based on a perception of data availability and the perception that modeling cannot proceed unless the spatial units are defined first. It is rarely recognized that this decision will profoundly affect subsequent results as all subsequent modeling will be constrained by the spatial model chosen. It is argued that one reason *a priori* identification of spatial units has been conducted is that in many cases this is appropriate. However, in many other cases this is not true and the same methodology has been followed because an alternative has not been presented. It is the intention of this author that this paper presents such an alternative. In doing so, practitioners should recognize that one does not necessarily have to identify spatial units *a priori* and that one can let statistically significant model variables do this. In doing so, however, a careful model evaluation – including verification of spatial independence of model residuals – must be conducted.

Acknowledgments

The author gratefully acknowledges the Association of Québec Forest Industries and Canadian Natural Sciences and Engineering Research Council for funding this work, and the suggestions of three anonymous reviewers which helped improve its presentation.

References

Cliff, A.D., and J.K. Ord, 1973. *Spatial Autocorrelation*, Pion, London, 178 p.

———, 1981. *Spatial Processes: Models and Applications*, Pion, London, 266 p.

Congalton, R.G., and R.A. Mead, 1983. A quantitative method to test for consistency and correctness in photo-interpretation, *Photogrammetric Engineering & Remote Sensing*, 49:69-74.

Dacey, M.F., 1968. A review of measures of contiguity for two and *k*-color maps, *Spatial Analysis: A Reader in Statistical Geography* (B.J.L. Berry and D.F. Marbel, editors), Prentice Hall, New Jersey, pp. 479-495.

Drew, W.B., 1942. *The Revegetation of Abandoned Cropland in the Cedar Creek Area, Boone and Callaway Counties, Missouri*, Agriculture Experiment Station Research Bulletin 344, University of Missouri-Columbia, 52 p.

Eyre, F.H. (editor), 1980. *Forest Cover Types of the United States and Canada*, Society of American Foresters, Washington, 148 p.

Foody, G.M., 1992. On the compensation from chance agreement in image classification accuracy assessment, *Photogrammetric Engineering & Remote Sensing* 58:1459-1460.

Ferguson, I.S., and J.W. Leech, 1978. Generalized least squares estimation of yield functions, *Forest Science*, 24:27-42.

Goodchild, M.F., 1986. *Spatial Autocorrelation*, CATMOG 47, Geo Books, Norwich, United Kingdom, 43 p.

Henning, W.L., 1937. *Zoological Reconnaissance of the Ashland Area with Special Reference to the Vertebrates*, Unpublished Masters Thesis, University of Missouri-Columbia, 153 p.

Huber, J.H., 1971. *Upland Old-field Succession Modeling in Mid-Missouri*, Unpublished Masters Thesis, University of Missouri-Columbia, 130 p.

Lowell, K.E., 1991. Utilizing discriminant function analysis with a geographical information system to model ecological succession spatially, *International Journal of Geographical Information Systems*, 5:175-191.

Odland,J., 1988. *Spatial Autocorrelation*, Sage, Beverly Hills, 87 p.

Pereira, J.M.C., and R.M. Itami, 1991. GIS-based habitat modeling using logistic multiple regression: a study of the Mt. Graham red squirrel, *Photogrammetric Engineering & Remote Sensing*, 57:475-1486.

Press, S.J., and S. Wilson, 1978. Choosing between logistic regression and discriminant analysis, *Journal of the American Statisticians Association*, 73:699-705.