*Diversity and Distributions* (1998) **4**, 49–61

**BIODIVERSITY RESEARCH**

# Estimating regional plant biodiversity with GIS modelling

LOUIS R. IVERSON and ANANTHA PRASAD *Northeastern Forest Experiment Station, USDA Forest Service, 359 Main Road, Delaware, OH, 43015, U.S.A. e-mail: iverson@trees.neusfs4153.gov*

**Abstract.** In this study, we analysed a statewide species database together with a county-level geographic information system (GIS) to build a model based on well-surveyed areas to estimate species richness in less surveyed counties. The Illinois Plant Information Network (ILPIN), a species-based database on all the vascular flora of Illinois, contains county distributions (totalling nearly 90,000) for each taxon and information on the taxonomy, ecology, biology, and ecodistribution. We compiled a statewide database with 112 variables on climate, landuse (current and historic), landscape pattern, soils and human population. We used a subset of this database to build a regression model for assessing native plant species richness for thirty-three botanically well-surveyed counties in Illinois. The best model was then used to predict the richness of the remaining sixty-nine less botanically surveyed counties. The model involved GIS (Arc/Info) and statistics (S-PLUS), including spatial statistics (S+SpatialStats). The resultant model had an $R^2$ of 0.80 and used the following variables: percentage of the county in cropland, the percentage with soils somewhat limiting for agriculture, the percentage of urban land, and the average size of farms. Although this particular model is not transferable to other locations without validation, the methodology shown here should be useful in estimating species richness patterns across regions where botanical sampling is heterogeneous.

**Key words.** Landscape ecology, spatial statistics/autocorrelation, multiple regression, biodiversity, geographic information system, Illinois Plant Information Network.

## INTRODUCTION

There is a growing need for studies aimed at integrating mechanistic explanations with large-scale patterns of diversity, as well as a need to understand the interaction between landscape configuration and patterns of species diversity (Lubchenco *et al.*, 1991; Holdgate, 1996). To achieve this better understanding, we must move from the conservation of single species only to the concept of ecosystem conservation, particularly as so many species are poorly known (Probst & Crow, 1991; Burton *et al.*, 1992; Franklin, 1993). For ecosystem conservation, data are needed on all elements of diversity, including the factors associated with increasing (or decreasing) species richness. In addition, it appears that the diversity at local scales is strongly limited by the regional diversity. Therefore, understanding species diversity in local assemblages requires knowledge of processes acting at larger spatial scales, including determinants of regional species richness (Caley & Schluter, 1997).

For this paper, we concentrate on vascular plant diversity, due to the availability of data. We recognize that this is only one component of biodiversity, and that the other components of biodiversity are also vitally important. Often there is a positive relationship between vascular plant diversity and that of the fauna, for example, but this is not always the case.

There also is a need for information systems that track the status and known information about flora (Bisby, Russell & Pankhurst, 1993). These systems can be of even greater utility when linked to a geographical information system (GIS) (Scott *et al.*, 1987; Davis *et al.*, 1990). Several such databases are being built by various public and private entities (e.g., the National Biological Service, the National Biodiversity Information Center, the Natural Resource Conservation Service (Peterson, 1993), The Nature Conservancy (Morse, 1993), individual states, Kartesz (1993), and the Flora of North American Editorial Committee (1993)). These databases vary in geographic extent, resolution or grain size, taxonomic extent, ancillary information, and degree of completion. All of the databases, however, fall short at this time due to the lack of comprehensive autoecological and synecological knowledge on all species.

Three approaches to databases are being used that help address issues of integrated analyses of botanical and ecosystem data (Morse *et al.*, 1981): (1) record point localities of individual plant sitings or collections, (2) record specimens by habitat or intrinsic features of the land, or (3) make synthetic summaries of plant collections (e.g. county records). In our study, we used the third approach with a county record database to derive relationships to species richness.

Examples are increasing where botanical data and environmental data are used together to derive new information. Flather, Joyce & Bloomgarden (1994) used a national, county-level database on rare species to assess species endangerment patterns across the United States. The Nature Conservancy uses the point-locality approach to catalogue critical elements of rare plants, animals, and communities; this approach is being used to assess threats and develop management strategies (e.g. Jenkins, 1985; Master, 1991). The Gap Analysis Program of the National Biological Service (Scott *et al.*, 1993; Kareiva, 1993) uses a GIS overlay approach to identify locations rich in species and to assess various preservation strategies, though Conroy & Noon (1996) state some shortcomings concerning this approach. Kartesz (1993) catalogued more than 15,000 taxa of North American vascular plants at a state level of spatial resolution. These data have been used for an 'envelopes' analysis of potential species changes under a changed climate (Morse *et al.*, 1993). Iverson & Prasad (1996) have done the same, at a county level, for about 100 tree species in eastern United States.

To improve on these studies, botanists and ecologists must be given the opportunity to enhance the distributional and habitat aspects of the databases as well as automate the known information. Meanwhile, there are a number of approaches being tested to estimate species richness in particular localities. The two main approaches include extrapolation (inferring richness based on subsamples within the area of interest) and interpolation (inferring richness based on comparisons with other areas) (Palmer, 1990, 1994). Remotely sensed data can be valuable in extrapolating estimates of biodiversity across the landscape (Walker *et al.*, 1992; Stoms & Estes, 1993; Podolsky, 1994). We use a modified interpolation method in this study.

Landscape spatial structure is another aspect of ecology that is now becoming understood as critical for an increasing number of species (Fahrig & Merriam, 1994; Tilman, 1994). Understanding and appropriately manipulating this structure is at least as important as preserves in protecting biodiversity. Landscape structure controls connectivity among patches in the landscape, which, in turn, influences the success of the preserves (Hudson, 1991; Franklin, 1993). Noss (1990) indicated that the regional landscape was one of four hierarchical levels of organization in monitoring biodiversity. In our study, we derived several metrics of landscape structure from a land-use database to test their importance in predicting biodiversity.

Illinois has a wide diversity of plant communities because of its geography, climate, and geologic history (Iverson *et al.*, 1989). Plant taxa in the state are abundant because of the mixing of species from a variety of floristic regions, which results in more than 3200 plant species (2309 native taxa) in Illinois, more than in any other midwestern state (Kartesz, 1992). The Illinois flora has been particularly well described and compiled for certain regions, but many counties have not been well surveyed. The Illinois flora is similar to that of several midwestern states, so any analysis of its flora likely will apply beyond the state's borders. However, the particular model developed here has been assessed only for Illinois. The model would need to be validated, and possibly reformulated, before applying elsewhere, but the basic approach could be the same.

The objective of this paper is to propose a strategy, using the ILPIN database in conjunction with landscape, edaphic, socio-economic and climatic information, and to build a prediction model based on well-sampled areas to estimate species richness in less sampled areas. These estimates of 'species yet to find' can then be used to aid in targeting of inventories so that appropriate levels of sampling can be done in those areas. Eventually, such knowledge could also be helpful in identifying 'hot spots' of plant diversity for possible conservation efforts. In doing this research, we also hope to identifiy general landscape patterns and processes that are most threatening to vascular plant diversity, and which ones tend to be most compatible with such diversity.

## METHODS

### Illinois Plant Information Network (ILPIN) database

The Illinois Plant Information Network (ILPIN) was developed at the Illinois Natural History Survey in Champaign. Described in Iverson, Prasad & Ketzner

(1997), ILPIN was initiated to provide the taxonomy, distribution, ecology, and biology of the 3200+ species in Illinois. The database allows rapid summaries of various attributes of the state's flora. Example outputs of statewide floral summaries are given in Iverson *et al.* (1997). For this study, we used only two fields of information from ILPIN: (1) number of species present in each of the 102 counties in Illinois and (2) whether the species are native or exotic to Illinois.

For purposes of our study, we considered biodiversity to be synonymous with native species richness, or the number of vascular plant species known to be native to Illinois. A total of 2309 taxa were included in the analysis.

## Data for model

Many data sources were queried to acquire pertinent county-level data that could be related to the biodiversity summaries from ILPIN. These included extractions from national-level databases, extractions from state-level databases, and aggregations of higher resolution data.

### *County summaries*

Two national databases were used to extract county-level information for this study: the Geoecology databases (Olson, Emerson & Nungesser, 1980), and the ArcUSA database (Environmental Systems Research Institute, 1992). Variables extracted from Geoecology include general land use proportions and annual climatic variables; general land use and census information were extracted from ArcUSA (Table 1).

The Illinois Geographic Information System of the Illinois Natural History Survey at Champaign also was queried for estimates of forest land historically occupying each county, as compiled by Iverson *et al.* (1989) from United States Department of Agriculture (USDA) Forest Service data from 1985; the U.S. Geological Survey (USGS) land cover data; an estimate in 1924 by Telford (1926); and forest estimates by land surveyors in the early 1800's. The number of professional botanists who reside in each county was estimated from a directory compiled by Burton, Robertson & Dennis (1989).

### *Landscape metrics of land use/land cover data*

The USGS has mapped the nation's land use and land cover, based on late 1970's photography, and digitized it onto 1:250,000-scale topographic sheets. The data consist of two hierarchical levels of classification. Only the highest level of classification was used here. It consists of five classes with sufficient area for analysis: urban, agriculture, forest, water, and wetland (Anderson, Hardy & Roach, 1976). The minimum resolution on the original data was 4 ha (10 acres) for urban land and 16 ha for the other classes. These data were acquired in vector format for the sixteen quadrangles that enclose Illinois. The data were rasterized to a grain size of 200 m and converted to Arc/Info Grid GIS. Separate GIS files for each of the 102 Illinois counties were then prepared for landscape analysis.

The rasterized files were processed in a landscape metrics program, Landstat, developed by Kurt Riitters of the Tennessee Valley Authority (Riitters *et al.*, 1995). This program calculated a large number of metrics, a subset of which was used in this analysis (Table 1). The subset was chosen largely on the basis of Riitters *et al.*'s factor analysis of 55 metrics on 85 land use/land cover scenes.

## Selecting the subsample

Central to this paper is the notion that not all counties have been sampled with the same degree of intensity. Some have a relatively low number of species recorded which does not necessarily reflect the actual number of species in the county because all counties have not been adequately surveyed. Counties near major universities and arboreta generally have been much more thoroughly surveyed. Counties beyond the proximal range of universities and research centres tend to be undersampled (Palmer, 1994). A subset of better assessed counties was used to develop a predictive multiple regression model for the remaining counties. The subset consisted of thirty-nine counties selected from those with at least one professional botany-related individual recorded in a directory of systematists, ecologists, and field biologists (Burton *et al.*, 1989). Although this ignores the fact that botanists do travel outside their counties to survey, it did provide an objective way of selecting the subsample. We performed stepwise regression for the counties selected and eliminated six outliers based on residual analysis. We were satisfied that the thirty-three counties finally selected were the better sampled counties, and the number of native species reported for them was likely to be close to the population mean (Fig. 1).

**Table 1.** Variables used in assessment of vascular-plant biodiversity in Illinois.

| | |
|---|---|
| Data from Illinois Plant Information Network | |
| NATIVE | Number of native species reported per county |
| Data from Geoecology database (Olson *et al.*, 1980) | |
| *Land Use Information*: raw data converted to percentages based on county area; original data from 1967 Conservation Needs Inventory of USDA Soil Conservation Service | |
| P.FOREST.LAND | Percent forest land per county |
| P.URBAN.LAND | Percent urban land per county |
| P.GRAZE.PAST | Percent grazing and pasture land per county |
| P.CROPS | Percent cropland per county |
| *Climatic Information*: data based on weather-station data for 1941–70 | |
| ANNUAL.TEMP.AVE | Average annual temperature, °C |
| ANNUAL.RAIN | Average annual rainfall, cm |
| ANNUAL.EVAP | Average annual potential evaporation, cm |
| ANNUAL.MOIST | Average annual moisture index (ratio of above variables) |
| GROW.DAYS | Average annual growing season, days |
| Data from ArcUSA database (Environ. Syst. Res. Inst., 1992) | |
| *Environmental Attributes*: most attributes originally compiled by Olson *et al.* (1980) for Geoecology database; most data derived from Conservation Needs Inventory of USDA Soil Conservation Service | |
| CNTY.AREA | Total county area, ha |
| P.CNTY.FED | Percent federal land in county, 1977 |
| P.RUR.LND | Percent rural land in county, 1977 |
| P.URB.LND | Total urban land in county, 1977 |
| P.WATER | Total water in county, 1977 |
| *Soil Orders*: essentially all Illinois land falls in these two orders (data originally compiled for Geoecology database) | |
| P.ALFISOL | Percent land in county in alfisols |
| P.MOLLISOL | Percent land in county in mollisols |
| *Soil Limitations for Particular Uses* | |
| P.SOILS.OK | Percent land with few restrictions on land use |
| P.SL.SMLTS | Percent land with some restrictions on land use |
| P.SL.SVLTS | Percent land with severe restrictions on land use |
| P.SL.VSLTS | Percent land with very severe restrictions on land use |
| P.SL.WET | Percent land with wet/stony soils restricting land use |
| P.SL.RANGE | Percent land suitable for forest or range |
| P.SL.WILD | Percent land suitable for forest or wildlife habitat |
| P.SL.NO.AG | Percent land in county where cultivation is precluded |
| *Population Attributes*: data from Agricultural Product Inventory (U.S. Bureau of the Census, 1987), Public Law 94-171 (U.S. Bureau of the Census, 1991), Demographic and Health Attributes (U.S. Bureau of the Census, 1989) | |
| AVE.SIZE | Average farm size, in hectares, 1987 |
| LAND.BLD.A | Value of land, etc., dollars/hectare, 1987 |
| POP1990 | 1990 population for the county |
| POP90.SQMI | 1990 population/square mile |
| POP1980CR | 1980 population for the county |
| P.POP.CHG | Change in population, 1980–86 (%) |
| Data from Illinois Geographic Information System database (Iverson *et al.*, 1989) | |
| P.FOR.1985 | Percent land in forest, 1985 (Hahn, 1987) |
| P.FOR.1820 | Percent land in forest, 1820 (Anderson, 1970) |
| P.FOR. LUDA | Percent land in forest, 1980 (USGS) |
| P.FOR.1924 | Percent land in forest, 1924 (Telford, 1926) |
| BOTAN | Number of recorded botanists/county, 1989 (Burton *et al.*, 1989) |
| Selected landscape metrics calculated from U.S. Geological Survey land use/land cover data with Landstat program (Riitters *et al.*, 1995) | |
| *Data on Individual Type Classes*: each variable repeated for each of five classes (represented by x), where 1 = urban, 2 = agriculture, 4 = forest, 5 = water, and 6 = wetland. | |
| PER x | percent |
| AVE.PATCH.SIZE x | average patch size |
| NUM.PATCHES x | number of patches |
| PERIM/AREA x | perimeter/area ratio |
| CIRCUM.RATIO WAT | circumscribing circle ratio |
| *Image Averages*: includes all classes | |
| SHAN.EVENNESS | Shannon evenness |
| SHAN.CONTAGION | Shannon contagion |
| SUM.DIAGONAL | sum of diagonals of adjacency matrix |
| PROP.LARGE.PATCH | average proportion of area in patches >5 cells |
| FRACTAL.MASS.AVE | fractal mass average |
| PERIM/AREA.ALL | perimeter/area ratio, equal weights |
| PATCH. NORMAL.AREA | average patch normalized area, square model |
| RADIUS.GYRATION | radius gyration average patch |

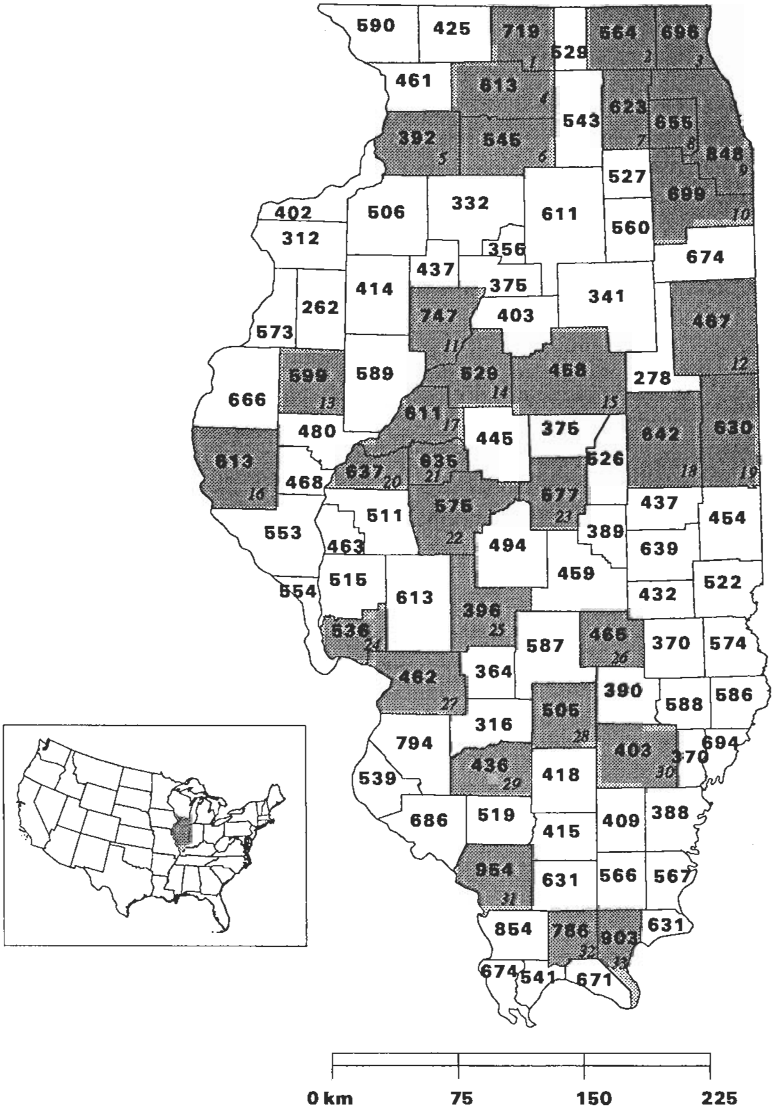

**Fig. 1.** Actual species richness of vascular plants, by county, according to ILPIN database (bold number). Shaded counties, labelled 1–33 in italics, represent the counties on which the regressions were based. County names are: 1, Winnabago; 2, McHenry; 3, Lake; 4, Ogle; 5, Whiteside; 6, Lee; 7, Kane; 8, Du Page; 9, Cook; 10, Will; 11, Peoria; 12, Iroquois; 13, McDonough; 14, Tazewell; 15, McLean; 16, Adams; 17, Mason; 18, Champaign; 19, Vermillion; 20, Cass; 21, Menard; 22, Sangamon; 23, Macon; 24, Jersey; 25, Montgomery; 26, Effingham; 27, Madison; 28, Marion; 29, Washington; 30, Wayne; 31, Jackson; 32, Johnson; 33, Pope.

## Model development

One of the difficulties is building a predictive model based on many possible causative factors is dealing with a lot of explanatory variables (112 in our case). The response surface (species richness) for so many explanatory variables is obviously too complex so the variables must be reduced before analysis. The number of observations, thirty-three in our case, further necessitated the reduction. Issues such as multicollinearity, interactions, skewed distributions, and spatial autocorrelation have to be addressed to reasonably fit the final model to the data. The search for a better model was an iterative process involving several exploratory data analysis steps. The analysis was done in S-PLUS (MathSoft, 1993).

First, because the data were compiled from several different sources, we eliminated obviously senseless variables that were deemed as non-causative. We then calculated a correlation matrix between the response variable and the remaining explanatory variables and eliminated variables whose correlation was <0.2 (positive or negative). We then examined the histograms of the remaining explanatory variables and transformed those that were skewed to a more normal distribution A stepwise regression was run on the transformed set, and the variables that were reported as significant were selected for the multiple regression model. We also ran a regression tree analysis (RTA) on the reduced-dataset and examined the tree structure to see if any important variables were missed in the stepwise procedure (Clark & Pergibon, 1992). The RTA picked POP90.SQMI as an important variable in addition to the variables that were already selected by the stepwise procedure. We added this variable to the model and tested it using the $C_p$ statistic to see if there was any improvement in the fit. This approach was used to see if there were any significant variables that were missed by using only the transformed stepwise approach (because a very large number of coefficient tests are conducted in the stepwise procedure, the probability is high that Type I or Type II errors may have occurred in the selection of the final variables (McClave & Dietrich, 1988)).

The final model with the variables selected by stepwise/RTA included interactions to achieve a better fit. However, the large number of variables that result from interactions can create a complex model that is hard to explain and is very sensitive to slight changes in data due to possible multicollinearity (Chatterjee & Price, 1977). So, for the variables selected by stepwise, we ran a backward elimination step regression, including interactions among variables (backward elimination does a better job in detecting multicollinearity (Mantel, 1970)). However, when we tested the final set for multicollinearity using Principal Components Analysis (Chatterjee & Price, 1977), we found that the interaction terms introduced significant multicollinearity making the regression coefficients unstable. So, we decided to use the model without interactions. The residual plot revealed an increase in variance with the $X$ axis indicating that the error variance is not constant. This could be because the number of native species per county (species richness) is a Poisson variable whose variance is equal to the mean. To stabilize the variance, we used square-root transformation of the response variable.

## Spatial data analysis

Because we are dealing with polygonal county data that can be treated as an irregular lattice, we needed to address the issue of spatial autocorrelation. A visual analysis of the dataset, mapped as number of species recorded for each of the thirty-three counties, indicated that there may be some spatial autocorrelation, which would warrant further spatial analysis (Fig. 1). If the response-surface is spatially autocorrelated, it is better to model it while including the small-scale variation due to interaction with neighbours (Haining, 1990; Cressie, 1991) in addition to the linear model (the small-scale variation is modelled by fitting an autoregressive or moving average covariance model). Also, if the spatial autocorrelation is significant, it may be caused by a trend in the data, which can be modelled by including a trend surface model using a polynomial based on data locations. Both components of the model (trend-surface and neighbour-interaction) interact, causing the model to be fit iteratively (MathSoft, 1996).

The length of the common boundary between two county polygons was used to weight the neighbour relationships (MathSoft, 1996). To test if there was significant spatial autocorrelation, we calculated the 'Moran' and 'Geary' statistics (Haining, 1990). We decided to calculate both the indices because they react differently to neighbour weights, with the Moran statistic yielding slightly better power (Griffith, 1995). Since we visually detected some spatial autocorrelation, we wanted to be sure that both the tests agreed. The Moran statistic is normally distributed under most conditions, with a mean of $-1(n-1)$ under the null hypothesis of no spatial autocorrelation where $n$ is the number of regions (Goodchild, 1986). Our subsample of thirty-three regions had a correlation coefficient of $-0.003$ which does not indicate positive spatial autocorrelation. To evaluate the significance of Moran's coefficient, we compare it to the two-tailed normal $P$-value (0.889 in this case), which is not significant at the county-level scale and our choice of neighbours and weights (Qi & Wu, 1996).

The Geary statistic is also normally distributed, with range and variance depending on neighbour weights. The mean of the Geary statistic under the null hypothesis of no spatial autocorrelation is 1; low values near 0 indicate strong positive spatial autocorrelation. In our study, the value of 0.73 indicates some positive spatial autocorrelation but is not significant (2-sided *P*-value of 0.35).

Both tests are dependent on the scale of analysis, choice of neighbours and neighbour weights. Since we did not have much control over the scale and choice of neighbours, we scaled the weights of each set of neighbours by the total boundary lengths and reran the tests. Even this analysis failed to detect any significant spatial autocorrelation at the county-level scale, again suggesting that including a spatial componemt to our linear model was not necessary. Nonetheless, we compared the fits of both the spatial regression model and the ordinary regression model to check if there were any major differences. As expected, there was little difference and the ordinary linear model was used.

### Model assumptions

We used the above model, based on well-surveyed counties, to predict the species richness for less surveyed counties. Because we are using the model to extrapolate only within the Illinois domain, we believe that we are not overextending the predictive power of the model. However, several assumptions have to be made in a model such as this.

1. The input data are assembled from various sources and at various spatial and temporal resolutions. For example, some data were aggregated to county level, whereas some were acquired at the county level. We assume that no spurious model outputs will result from these types of data incongruencies.
2. The explanatory variables were initially screened under a linear response criterion; we assume that nonlinear variables were not critical to the model.
3. We assume that one botanist working in a county means that particular county will be better sampled and botanical reports will be more accurate than that for other counties.
4. We assume that the high number of explanatory variables, often intercorrelated, will be properly handled by the stepwise methodology.
5. We assume that the range of variability within the thirty-three counties used in developing the model will approximate that in the remaining sixty-nine counties; that is, the data range of environmental characteristics selected by the model encompasses all the range of possibility within the Illinois domain.
6. We assume that the number of native species per county is a Poisson variable.

We also emphasize that the specific model developed here is not transferable, without testing, to other geographic locations. However, the methodology employed is transferable and we do anticipate that several midwestern states would have similar, but not identical, patterns.

## RESULTS AND DISCUSSION

### Relationships between biodiversity and explanatory variables

Many variables correlated with the number of native species recorded for the thirty-three counties used to develop the model (Table 2). In general, the number of native species increases with increasing forest land and/or decreasing cropland, an expected outcome. The number of species is also directly related to the number of patches in agriculture (NUM.PATCHES_AG), indicating that a large number of smaller farm fields probably allows proportionately more land to be available as refugia and corridors for species retention. The Shannon evenness index, an indicator of the proportional abundances of land use classes, was also positively correlated. This indicates that native species increase as the main land use classes (urban, agriculture, forest, water, and wetland) become more evenly distributed. Perimeter/area ratios of agriculture land were also positively related, indicating that more edge around agricultural fields will harbour more species. The proportion of soil suited for forest or wildlife (PL.SL.WILD) was also expectedly related as these lands are conversely unsuited for agricultural disturbances.

More difficult to interpret are the positive relationships between number of native species and the percent urban land (P.URB.LND) or population density (POP90.SQMI). Although these relationships are not strong, they could indicate one or more of three possibilities: (1) that more people in the vicinity means more botanical surveys, more surveyed parks, etc., and therefore, more species found and recorded; (2) that urban habitats generally allow more refugia for native species (including parks, etc.) than

**Table 2.** Correlation coefficients between variables and number of native species recorded per county, for thirty-three counties. Only variables with a coefficient $>|0.2|$ are given. Values $>|0.33|$ are significant at the 0.05 level.

| | | | |
|---|---|---|---|
| NUM.PATCHES AG | 0.59 | PERIM/AREA URB | 0.27 |
| P.CNTY.FED | 0.51 | P.POP.CHG | 0.26 |
| P.FOR.1985 | 0.44 | P.DIST.LND | 0.23 |
| P.FOR.LUDA | 0.43 | P.SL.WET | 0.22 |
| AVE.PATCH.SIZE FOR | 0.42 | | |
| SHAN.EVENNESS | 0.41 | PROP.LARGE.PATCH | −0.27 |
| PERIM/AREA_AG | 0.41 | PATCH.NORMAL.AREA | −0.27 |
| P.FOR.1820 | 0.41 | CIRCUM.RATIO_FOR | −0.28 |
| P.SL.WILD | 0.39 | P.SL.SMLTS | −0.29 |
| P.URB.LND | 0.37 | AVE.SIZE | −0.30 |
| POP90.SQMI | 0.35 | P.RUR.LND | −0.36 |
| P.FOR.1924 | 0.32 | AVE.PATCH.SIZE AG | −0.37 |
| P.FOREST.LAND | 0.32 | FRACTAL.MASS.AVE | −0.39 |
| BOTAN | 0.29 | P.SOILS.OK | −0.40 |
| PER URB | 0.29 | SHAN.CONTAGION | −0.41 |
| AVE.PATCH.SIZE URB | 0.29 | SUM.DIAGONAL | −0.43 |
| P.ENTISOL | 0.29 | PER AG | −0.54 |
| CIRCUM.RATIO WAT | 0.28 | P.CROPS | −0.61 |

agriculture-dominated habitats; or (3) that urban centres occur in regions of the state that have high habitat diversity nearby.

There was a positive relationship between the number of botanists and the species richness in the county, though not quite sigificant (Table 2). This trend indicates a tendency for better searching in counties where botanists reside, or that botanists tend to live in places with relatively higher plant diversity.

On the negative side, the percentage of agriculture is most related to decreasing native species, as are several related variables such as the percentage of soils suitable (P.SOILS.OK) or somewhat suitable (P.SL.SMLTS) for agriculture, the average size of farms (AVE.SIZE), the average patch size of agricultural land (AVE.PATCH.SIZE AG), the percentage of rural land (P.RUR.LND), and the proportion of large patches (PROP.LARGE.PATCH) (Table 2). A number of the fractal indices (SUM.DIAGONAL, FRACTAL.MASS.AVE, PATCH.NORMAL.AREA) also shows that as the land becomes more rectilinear shaped (e.g. big farm fields following human-moderated edges), less native species will be recorded. The negative relationship with Shannon's contagion index indicates that as the landscape exhibits a greater amount of clumping, less native species will be found. Interestingly, none of the climate variables in our analyses were significant. One would expect climate to operate at a scale encompassing several states, with fewer species in the states north and west of Illinois and greater diversity in the southeastern United States (Kartesz, 1992).

This general analysis of correlations aids in understanding the model which predicts the number of native species found per county in Illinois.

## Predicting biodiversity by county

The final model consisted of the following variables, with the coefficients and $P$ values listed in Table 3: P.CROPS, P.SL.SMLTS, P.URB.LND, and AVE.SIZE. It did not contain a spatial component because spatial autocorrelation did not significantly impact the relationships. Although there was evidence of some limited multicollinearity among P.CROPS, P.SL.SMLTS, and AVE.SIZE, the latter two variables did substantially increase the $R^2$ value for the model, and neither the residual analysis nor the principal components analysis reported significant multicollinearity.

A close relationship between actual species richness and that predicted by the model was found for the thirty-three counties (Fig. 2). The model was capable of explaining 80% of the variation ($R^2=0.80$) and was highly significant ($P$-value of 1.06e-09). The residual analysis did not indicate any outliers or major violations of linear model assumptions (Fig. 2).

**Table 3.** Model used to predict number of native species per county in Illinois.

| Coefficients | Value | Standard error | $t$ value | Pr $(>\lvert t\rvert)$ |
|---|---|---|---|---|
| (Intercept) | 27.4009 | 1.1464 | 23.9009 | 0.0000 |
| P.CROPS | −0.2796 | 0.0290 | −9.6489 | 0.0000 |
| P.SL.SMLTS | 0.1341 | 0.0268 | 5.0040 | 0.0000 |
| AVG.SIZE | 0.0259 | 0.0044 | 5.8787 | 0.0000 |
| log(P.URB.LND) | 0.8222 | 0.2141 | 3.8403 | 0.0006 |

Residual standard error: 1.31 on 28 degrees of freedom.
Multiple $R$-Squared: 0.809.
$F$-statistic: 29.65 on 4 and 28 degrees of freedom, the $P$-value is 1.06e-09.

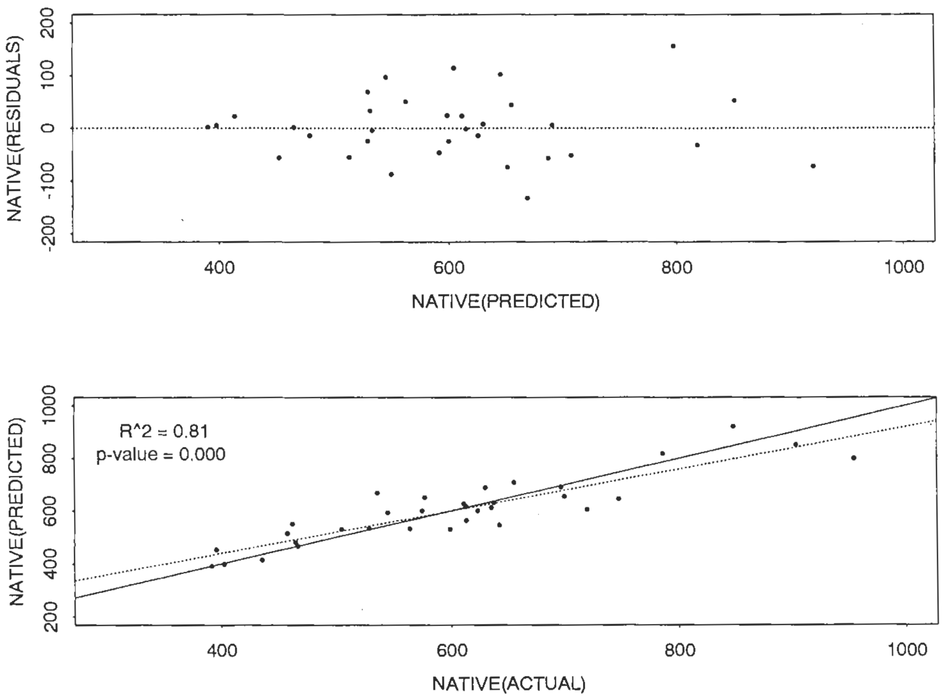

**Fig. 2.** Residuals and predicted vs. actual relationship of final model predicting number of species in thirty-three well-surveyed counties in Illinois. The solid line indicates the 45 degree separator (slope of 1), while the dashed line is the regression line for the displayed points.

Actual versus predicted species richness for all 102 counties is shown in Fig. 3. Of the sixty-nine counties not in the original sample of thirty-three, forty-five had a predicted species richness greater than the actual number recorded for that county. Some counties have substantially fewer species recorded than predicted, indicating that there are considerably more taxa to be discovered and reported in many of those counties. Of course, error is associated with the regression, and the remaining twenty-four counties had a predicted species richness less than the actual recorded.

Predicted number of taxa is also mapped by county, along with the standard error (Fig. 4). Fig. 4 shows that thirty counties are predicted to have at least 100 species to discover (and record), and thirteen of those have more than 200 species to find. The model provides a mechanism to target locations of botanical

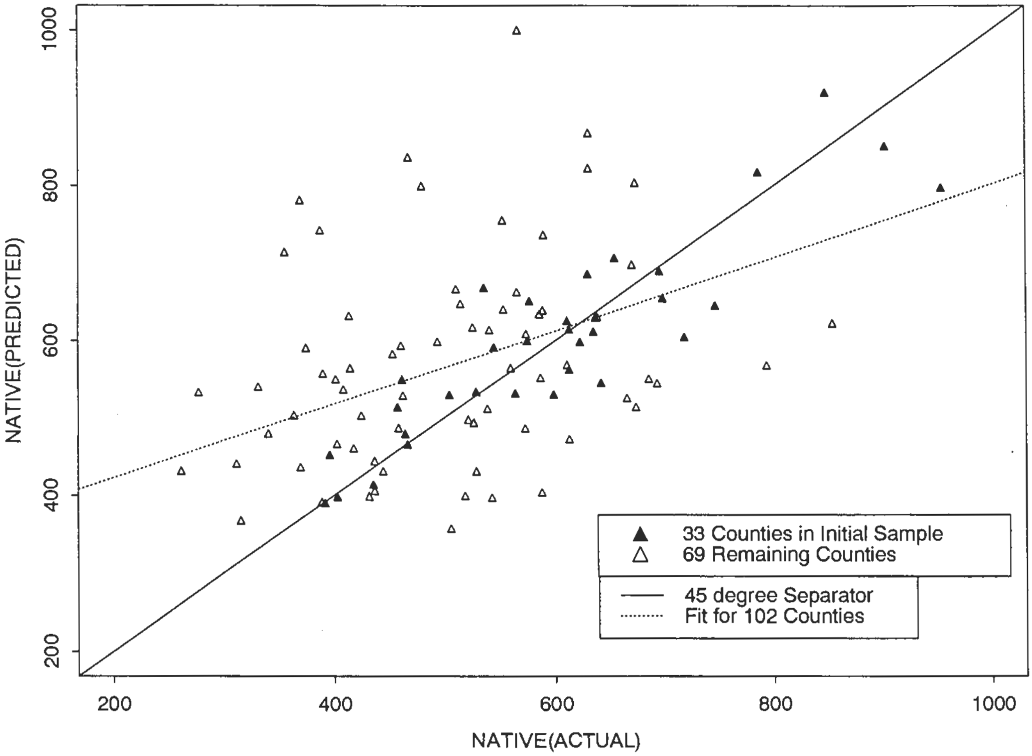

**Fig. 3.** Actual *v.* predicted number of native species for 102 Illinois counties.

searches. In Illinois, those regions seem to predominate in the extreme southeast, the west-central, the northwest, and the north-central portions.

## CONCLUSIONS

In the statistical model presented here, a regional perspective on the controlling variables related to species richness was obtained for Illinois. The use of the ILPIN dataset with other regional environmental data can yield information on the factors controlling biodiversity. Even though the correlated variables unveiled by the regression/RTA approach should be interpreted with caution (they may not always be causative *per se*), it does provide a method of evaluating what factors control species richness. The method is intended for use only in Illinois, though results would likely be similar for some other midwestern states. The methodology could be adopted, however, whenever there are well-inventoried locations scattered amongst less inventoried locations. This method should not be used to replace botanical surveys, but can be used to target future botanical forays. It should be noted that this method does not provide information on plant species abundances, dynamics, or viabilities, only the potential presence or absence of species.

Our analysis shows that several landscape patterns have a general negative influence on plant diversity: Most of these negative influences are related to the intensity and amount of agriculture occurring in the county (e.g. percent crop land, soils suitable for agriculture, large patches of agricultural land, and fractals indicating rectilinear shapes), which of course precludes the existence of most native species. The negative relationship with the contagion index indicates that as the landscape exhibits a greater amount of clumping, less native species will be found. Counties with high contagion probably have mostly agricultural fields which contribute most of the clumping.

As expected, species richness increases with the proportion of forest land in the county. It also is

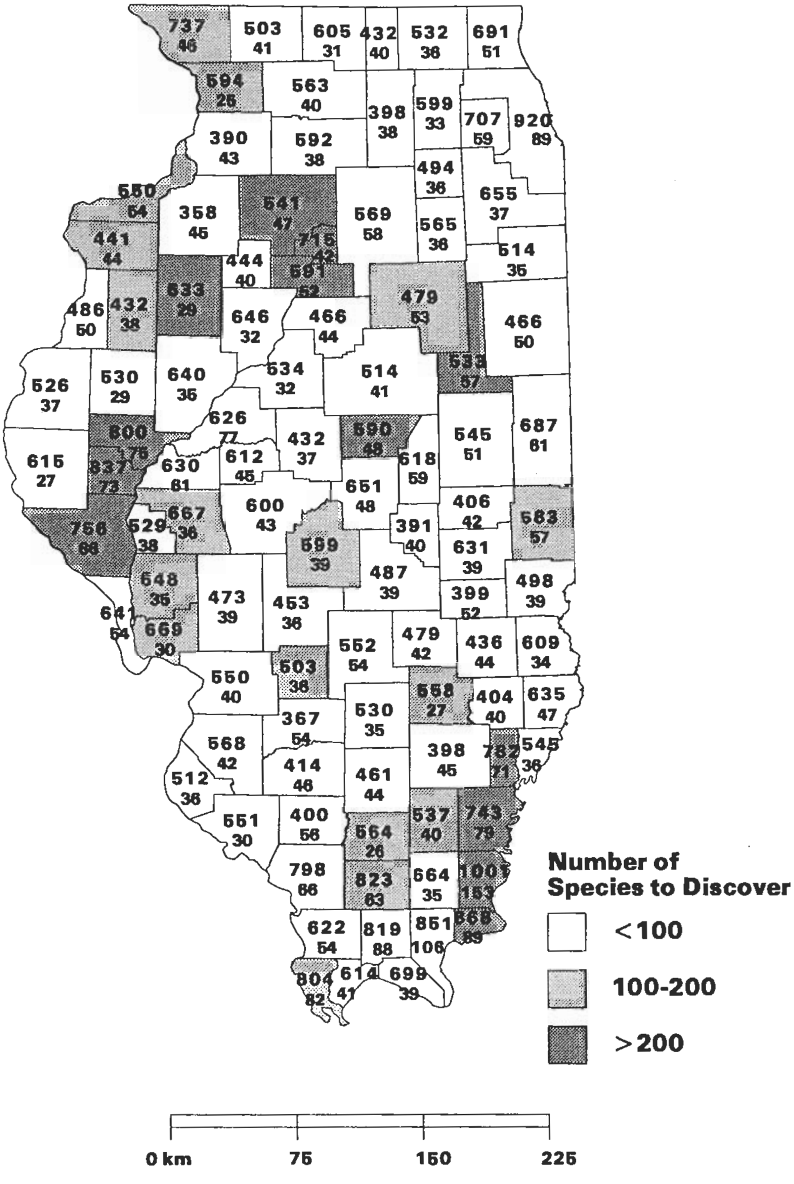

**Fig. 4.** Illinois counties with predicted (plus standard error) numbers of native taxa per county.

enhanced by conditions where agriculture is spatially distributed as small patches, or as irregular shaped patches. This spatial arrangement allows refugia and corridors to remain for the native flora. A relatively even distribution of the five main land uses also tends to favour higher plant species richness. For conservation efforts, these results argue for incentives to keep crop fields smaller and distributed among other land types.

The most severe limitation to methods such as ours is that only well-known areas can serve as a basis for building the methods. We endorse the national, state, and private organizations compiling databases to help address this issue. Collections must continue to be made in the field and the data must be computerized with coordinates suitable for GIS analysis. If a fully researched, adequately surveyed, and up-to-date database with fields similar to ILPIN's were available nationally (eventually globally), we would gain an accurate measure of floristic diversity that allows the use of efficient conservation methods.

## ACKNOWLEDGMENTS

We thank Kurt Riitters for providing the program for calculating landscape metrics, and the Illinois Natural History Survey and Illinois Department of Transportation for providing funding for ILPIN for 10 years. Kenneth Robertson, John Taft, Steve Sutherland, Larry Page, Thomas Jacob, Charles Warwick, Mary Buchanan, and two anonymous reviewers reviewed earlier drafts of this manuscript. Thanks to Mary Boda for making manuscript corrections.

The use of trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the U.S. Department of Agriculture or the Forest Service of any product or service to the exclusion of others that may be suitable.

## REFERENCES

Anderson, J.R., Hardy, E.E. & Roach, J.T. (1976) *A land-use classification system for use with remote-sensor data.* Circular 671. United States Geological Survey, Reston, VA.

Anderson, R.C. (1970) Prairies in the prairie state. *Trans. Illinois State Acad. Sci.* **63**, 214–221.

Bisby, F.A., Russell, G.F. & Pankhurst, R.J. (1993) *Designs for a global plant species information system.* 350 pp. Clarendon Press, Oxford.

Burton, P.J., Balisky, A.C., Coward, L.P., Cumming, S.G. & Kneeshaw, D.D. (1992) The value of managing for biodiversity. *For. Chron.* **68**, 225–237.

Burton, P.J., Robertson, K.R. & Dennis, A. (1989) *A diversity of Illinois systematists, ecologists, and field biologists.* Illinois Natural History Survey Special Publication 3, Champaign, IL.

Caley, M.J. & Schluter, D. (1997) The relationship between local and regional diversity. *Ecology,* **78**, 70–80.

Chatterjee, S. & Price, B. (1977) *Regression analysis by example.* John Wiley, New York.

Clark, L.A. & Pergibon, D. (1992) Tree-based models. *Statistical models in S* (ed. by T.J. Hastie), pp. 377–419. Wadsworth, Pacific Grove, California.

Conroy, J.J. & Noon, B.R. (1996) Mapping of species richness for conservation of biological diversity: conceptual and methodological issues. *Ecol. Appl.* **6**, 763–773.

Cressie, N. (1991) *Statistics for spatial data.* John Wiley, New York.

Davis, F.W., Stoms, D.M., Estes, J.E. & Scepan, J. (1990) An information systems approach to the preservation of biological diversity. *Int. J. Geo. Info. Sys.* **4**, 55–78.

Environmental Systems Research Institute, Inc. (1992) *ArcUSA 1:2m, user's guide and data reference.* Environmental Systems Research Institute, Inc., Redlands, CA.

Fahrig, L. & Merriam, G. (1994) Conservation of fragmented populations. *Cons. Biol.* **8**, 50–59.

Flather, C.H., Joyce, L.A. & Bloomgarden, C.A. (1994) *Species endangerment patterns in the United States.* General Technical Report RM-241, Rocky Mt. Forest and Range Experiment Station, USDA Forest Service. Ft. Collins, CO.

Flora of North America Editorial Committee. (1993) *Flora of North America north of Mexico.* Oxford University Press, New York.

Franklin, J.F. (1993) Preserving biodiversity: species, ecosystems, or landscapes? *Ecol. Appl.* **3**, 202–205.

Goodchild, M.F. (1986) *Spatial autocorrelation. Concepts and techniques in modern geography.* Geo Books, Norwich.

Griffith, D.A. (1995) Some guidelines for specifying the geographic weights matrix contained in spatial statistics models. *Practical handbook of spatial statistics* (ed. by S.L. Arlinghaus pp. 65–82. CRC Press, Inc., Boca Raton.

Hahn, J.T. (1987) *Illinois forest statistics, 1985.* Resource Bulletin NC-103. U.S.D.A. Forest Service, St. Paul, MN.

Haining, R. (1990) *Spatial data analysis in social & environmental sciences.* Cambridge University Press, Cambridge, MA.

Holdgate, M. (1996) The ecological significance of biological diversity. *Ambio,* **25**, 409–416.

Hudson, W.E. (1991) *Landscape linkages and biodiversity.* Island Press, Washington.

Iverson, L.R., Oliver, R.L., Tucker, D.P., Risser, P.G., Burnett, C.D. & Rayburn, R.G. (1989) *Forest resources of Illinois: an atlas and analysis of spatial and temporal trends.* 181 pp. Illinois Natural History Survey Special Publication 11, Champaign, IL.

Iverson, L.R. & Prasad, A. (1996) Modelling present and future tree importance values in the Eastern United States. *Third international conference on integrating environmental modelling and GIS.* National Center for Geographic Information and Analysis (CD-ROM and www site:http:/www.ncgia.ucsb.edu/conf/SANTE FE CD ROM/sfposters.html), Santa Fe, New Mexico.

Iverson, L.R., Prasad, A. & Ketzner, D.M. (1997) A summary of the Illinois flora based on the Illinois Plant

Information Network. *Trans. Illinois Acad. Sci.* **90** 41–64.

Jenkins. R.E. (1985) Information methods. why the Heritage Programs work. *Nature Conservancy News*, **35**. 21–23.

Kareiva. P. (1993) No shortcuts in new maps. *Nature* **365**. 292–293.

Kartesz. J.T. (1992) Preliminary counts for native vascular plant species of U.S. states and Canadian provinces. *Biodiversity Network News* **5**, 6.

Kartesz. J.T. (1993) *A synonymized checklist of the vascular flora of the United States, Canada, and Greenland* 2nd edn, 622 pp. Timber Press, Portland.

Lubchenco. J.. Olson. A.. Brubaker, L.. Carpenter. S., Holland, M.. Hubell, S.. Levin. S., MacMahon, J., Matson, P., Melillo. J., Mooney. H.. Peterson, C.. Pulliam, H.. Real, R.. Regal. P. & Risser. P. (1991) The sustainable biosphere initiative: an ecological research agenda. *Ecology*, **72**. 371–412.

Mantel. N. (1970) Why stepdown procedures in variable selection. *Technometrics*, **12**. 591–612.

Master. L.L. (1991) Assessing threats and setting priorities for conservation. *Cons. Biol.* **5**. 559–563.

MathSoft. (1993) *S-PLUS guide to statistical and mathematical analysis, vers. 3.2.* StatSci, Seattle, WA

Mathsoft. (1996) *S+Spatialstats user's manual, version 1.0.* MathSoft Inc., Seattle. WA.

McClave, J.T. & Dietrich. F.H. III. (1988) *Statistics* 4th edn. Dellen Publishing Co., San Francisco.

Morse. L.E. (1993) Standard and alternative taxonomic data in the multi-institutional Natural Heritage Data Center Network. *Designs for a global plant species information system* (ed. by F.A. Bisby. G.F. Russell and R.J. Pankhurst), pp. 69–79. Clarendon Press, Oxford.

Morse. L.E.. Henifin, M.S.. Ballman. J.C. & Lawyer, J.I. (1981) Geographical data organization in botany and plant conservation: a survey of alternative strategies. *Rare plant conservation: geographical data organization* (ed. by L.E. Morse, and M.S. Henifin), pp. 9–29. The New York Botanical Garden. Bronx. N.Y.

Morse. L.. Kutner. L.S., Maddox, G.D.. Honey. L.L., Thurman, C.M., Kartesz. J.T. & Chaplin. S.J. (1993) *The potential effects of climate change on the native vascular flora of North America: a preliminary envelopes analysis.* Electric Power Research Institute. TR-103330, Palo Alto.

Noss, R.F. (1990) Indicators for monitoring biodiversity: a hierarchical approach. *Cons. Biol.* **4**. 355–364.

Olson. R.J.. Emerson. C.J. & Nungesser. M.K. (1980) *Geoecology: a county-level environmental data base for the conterminous United States.* Publication No. 1537. Oak Ridge National Laboratory Environmental Sciences Division. Oak Ridge, TN

Palmer, M.W. (1990) The estimation of species richness by extrapolation. *Ecology*, **71**, 1195–1198.

Palmer, M.W. (1994) How should one count species? *Natural Areas Journal*, **14**, 22–28

Peterson, J.S. (1993) U.S. interagency botanical data applications, needs, and the PLANTS database. *Designs for a global plant species information system* (ed. by F.A. Bisby, G.F. Russell, and R.J. Pankhurst). pp. 13–19. Clarendon Press, Oxford.

Podolsky, R. (1994) A method for estimating biodiversity directly from digital earth imagery. *Earth Observ.* June, 30–36.

Probst. J.R. & Crow, T.R. (1991) Integrating biological diversity and resource management. *J. For.* **89**. 12–17.

Qi. Y. & Wu. J. (1996) Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices. *Landscape Ecol.* **11**, 39–49.

Riitters, K.H.. O'Neill. R.V., Hunsaker, C.T.. Wickham. J.D.. Yankee. D.H.. Timmins. S.P., Jones. K.B. & Jackson, B.L. (1995) A factor analysis of landscape pattern and structure metrics. *Landscape Ecol.* **10**, 23–39.

Scott. J.M.. Csuti. B.. Joacobi, J.D. & Estes, J.E. (1987) Species richness. A geographic approach to protecting future biological diversity. *BioSci.* **37**, 782–788.

Scott. J.M.. Davis, F.W.. Csuti. B.. Noss, R.. Butterfield. B.. Groves, C.. Anderson. H.. Caicco. S.. D'Erchia, F.D., Edwards. T.C., Ulliman. J. & Wright. R.G. (1993) Gap analysis: a geographic approach to protection of biological diversity. *Wildl. Monogr.* **123**, 1–41.

Stoms. D.M. & Estes. J.E. (1993) A remote sensing research agenda for mapping and monitoring biodiversity. *Int J. Rem. Sensing*. **14**. 1839–1860.

Telford. C.J. (1926) Third report on a forest survey of Illinois. *Illinois Nat. Hist. Surv. Bull.* **16**, 1–102.

Tilman. D. (1994) Competition and biodiversity in spatially structured habitats. *Ecology*, **75**. 2–16.

U.S. Bureau of the Census. (1987) *Census of Agriculture. 1987. on CD-ROM.* Data Access and Use Staff. Data User Services Division, U.S. Bureau of the Census. Washington.

U.S. Bureau of the Census. (1989) *County and city data book, 1988: Files on CD-ROM technical documentation.* Data Access and Use Staff, Data User Services Division. U.S. Bureau of the Census. Washington.

U.S. Bureau of the Census. (1991) *Public Law. 94–171.* Washington.

Walker. R.E., Stoms. D.M., Estes. J.E. & Cayocca. K.D. (1992) Relationships between biological diversity and multi-temporal vegetation index data in California. *Technical Papers of the 1992 Annual Meeting of ASPRS/ACSM*, pp. 562–571. American Society of Photogrammetry and Remote Sensing, Bethesda. MD.