

Quantitative tools for perfecting species lists

Michael W. Palmer^{1,*†}, Peter G. Earls¹, Bruce W. Hoagland²,
Peter S. White^{3,4}, Thomas Wohlgemuth⁵

¹*Department of Botany, Oklahoma State University, Stillwater, Oklahoma 74078, U.S.A.*

²*Oklahoma Biological Survey and Department of Geography, University of Oklahoma, Norman, Oklahoma 73019, U.S.A.*

³*Department of Biology, Campus Box 3280, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.*

⁴*North Carolina Botanical Garden, Campus Box 3375, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.*

⁵*Swiss Federal Institute of Forest, Snow and Landscape Research (WSL), 8903 Birmensdorf, Switzerland*

SUMMARY

A substantial body of literature has accumulated on the topic of the estimation of species richness by extrapolation. However, most of these methods rely on an objective sampling of nature. This condition is difficult to meet and seldom achieved for large regions. Furthermore, scientists conducting biological surveys often already have preliminary but subjectively gathered species lists, and would like to assess the completeness of such lists, and/or to find a way to perfect them. We propose several strategies for utilizing external data (such as might be obtained using GIS) to aid in the completion of species lists. These include: (i) using existing species lists to develop predictive models; (ii) using the uniqueness of the environment as a guide to find underrepresented species; (iii) using spectral heterogeneity to locate environmentally heterogeneous regions; (iv) combining surveys with statistical model-building in an iterative manner. We demonstrate the potential of these approaches using simulation and case studies from Oklahoma. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: biodiversity; floristics; inventory; mapping; predictive modeling; species richness

INTRODUCTION

'It is, I find, in zoology as it is in botany: all nature is so full, that that district produces the greatest variety which is the most examined' –Gilbert White, Natural History of Shelbourne, 1768 (cited in Kieran, 1957)

Biodiversity is widely recognized as important, both for its scientific study (Palmer, 1994; Rosenzweig, 1995) and its conservation (Wilson, 1988; McNeely, 1990; Ranta *et al.*, 1999). While interest in biodiversity has been high for decades, the ability to measure and monitor it has lagged far behind. Inventories of particular areas and some rare species distributions have greatly increased our knowledge in the last 25 years, yet large-scale floristic data (i.e. documented occurrences of plant species), though voluminous, are in many regions inadequate for assessing biodiversity. In particular,

*Correspondence to: Michael Palmer, Department of Botany, Oklahoma State University, Stillwater, Oklahoma 74078, U.S.A.

†E-mail: carex@okstate.edu

floristic data are subjectively gathered and notoriously incomplete (Palmer, 1995). The botanist is not to be blamed entirely for such flaws; rather, the flaws are consequences of plant distributions.

For large regions (e.g. parks, provinces, countries), it is unlikely that we will ever have complete species lists. This is partly because the species in a region change through time (Robinson *et al.*, 1994; McCollin *et al.*, 2000), but also because botanists cannot inspect all the individual plants in a region. Thus, our goal should be to complete species lists as efficiently as possible, given finite resources.

Quantitative approaches hold promise for estimating species richness: the number of species in an area (Palmer, 1990, 1991; Bunge and Fitzpatrick, 1993; Colwell and Coddington, 1994). Such approaches typically rely on objectively placed subsamples (e.g. quadrats) in which there is an exhaustive search for species. While objective, such approaches are labor-intensive, and only a tiny fraction of the study region can be sampled. Objective methods are likely to miss the rare or unclassifiable habitats that are likely to contribute the most to regional diversity. The error involved in extrapolation can be tremendous. Indeed, it is unlikely that such methods can outperform the guesses of experienced botanists.

We here define 'experienced botanist' in a rather narrow way: as someone who has developed some expertise in finding and listing species. The ideal experienced botanist would combine enthusiasm, intuition and natural history knowledge to direct his or her efforts. The intuition is itself developed and refined by experience. Variation in experience means that the efficiency of completing checklists is variable among botanists. Although it is beyond the scope of this article to evaluate, botanists (both amateur and professional) vary dramatically in ability. We suspect that teams of botanists, with varying ability, experience, expertise and intuition, will result in a fairly complete coverage of species. Furthermore, we cannot underestimate the educational and social value of teamwork. In what follows, the term 'experienced botanist' could refer to a team of botanists.

Experienced botanists generally have a strong intuition or 'educated guess' about where to direct one's efforts in collecting specimens. Botanists are drawn to unusual habitats, edges between vegetation types, environments suspected to be species rich, and geomorphic features such as bluffs, outcrops, streams, etc. Botanists often intensify their efforts when they detect new or scarce species, and move more quickly through areas in which the species are commonplace. However, the ability to visit 'interesting' places is sometimes compromised by inaccessibility, and hence 'convenient' locations (e.g. areas close to universities) are well-studied floristically, and in consequence may falsely appear to be rich in species.

The experienced botanist, using internal 'algorithms', will most likely outperform any objective methodology if the goal is to maximize species encountered. Although objective procedures are (ideally) repeatable, lead to unbiased estimates and allow rigorous analysis, it would be unwise to dismiss the efficient, yet subjective, contributions of the expert botanist. In this article we propose a few approaches to supplement the intuition of such a botanist, and thereby to find useful strategies to complete biodiversity information. We hope our brief sketches below will stimulate envirometricians to develop more robust and sophisticated techniques.

The utility of external information

Environmental variables. The spatial arrangements of species and biodiversity are decidedly non-random (Maurer, 1994; Rosenzweig, 1995). In particular, species have predictable (though perhaps complex and noisy; Gauch, 1982a,b) relationships to environmental gradients. Such relationships can occur at any spatial scale. The number of gradients that can influence species distributions is potentially infinite. Therefore, we always operate with incomplete information. Nevertheless,

expert opinion, plus knowledge from other systems, can be of value in selecting useful explanatory variables.

Not all variables will be equally useful for choosing locations for completing species lists. Direct measurements of important parameters (e.g. the chemistry of soils) may be as laborious as exhaustive species lists. However, indirect measures or interpolations of the environment (e.g. climatic variables, mapped soil types, classified vegetation type, elevation) may be instantaneously available in a Geographic Information System (GIS). While the procedures for generating data layers involve some subjectivity or 'art', they still have potentially predictive value for species distributions. There are different ways of generating such GIS data layers, e.g. from mapped or measured field data by interpolation, or from remotely sensed data with or without the use of ground-truthed information.

Other variables. Sometimes variables that are not related to 'environmental gradients' will nevertheless have predictive value. For example, species richness is usually correlated with area (Gleason, 1922; Williams, 1964; Palmer and White, 1994; Rosenzweig, 1995). Spatial coordinates (e.g. latitude, longitude) may be proxies for climatic factors or biogeographic history, but they can also be proxies for broad-scale biogeographic patterns: that is, it is possible that different regions have a different evolutionary and vicariance history. The conservation literature is overflowing with studies on 'indicator taxa'—i.e. (arguably) well known groups such as birds, butterflies and carabid beetles that can be used to predict the diversity of other groups (Rykken *et al.*, 1997; Prendergast and Eversham, 1997; Pearson and Carroll, 1998). Regardless of its nature or origin, any variable that can be used to predict species distributions, and that is readily available, is potentially useful for helping the botanist perfect species lists.

How can we use such external information to help complete species lists? One of the simplest approaches is to select variables known or suspected of being correlated with species richness, and to choose new locations for surveys with values of those variables associated with high richness. In addition to seeking out such species-rich sites, the expert botanist will tend to seek out unusual or rare environments. It is important to realize that 'unusualness' is not only an attribute of nature, but also of past collection efforts. So if a botanist has concentrated past collection efforts in forests, grasslands will appear to be unusual. Hence, visiting a grassland is likely to result in the accumulation of new records, even if the grasslands are not particularly rich in species. Additional sampling effort in grasslands will make this habitat less unusual and thereby less attractive for future expeditions.

Spatial heterogeneity. In vegetation science and landscape ecology, remotely sensed images are typically highly processed and 'classified' in image processing software before information is extracted. However, we suspect that such processing results in much loss of information. With respect to finding new species, we are particularly interested in finding unusual (and therefore potentially poorly classified) habitats. We are also interested in transition zones—but it is difficult to ascertain whether a border between classes represents a true boundary or merely an arbitrary distinction within a gradually varying environment. Furthermore (almost by definition), classification degrades continuous, quantitative information into discrete classes. We believe that spatial patterns of reflectance in a remotely sensed image can be used, without classification, as a tool to guide the botanist. In particular, we propose that spatial variation (e.g. expressed as a standard deviation of reflectance) is likely to be correlated with spatial variation in the environment. Furthermore, this variation in the environment is likely to be correlated with species richness (and hence, with 'botanical interestingness'). The predicted link between spectral heterogeneity and biodiversity is termed the spectral variation

hypothesis (SVH; Palmer *et al.*, 2000). While it makes intuitive sense, the SVH has not yet been tested (with the exception of a variant in Gould, 2000).

Intensive vs. extensive. So far, we have not addressed how much effort is needed within a location. Several factors must be considered, including the amount of effort it takes to search within a location, the number of rare or inconspicuous species within a site, the number of taxonomically difficult species in a site, the effort it takes to move from one location to another, the environmental differences among sites, the beta diversity (or compositional differences) among sites, and the degree of nestedness (Wright and Reeves, 1992). All of these vary depending upon the biota and the study scale, so it is not clear whether generalizations or rules of thumb will emerge. However, we will approach the question of intensive vs. extensive studies with case studies later.

DATA SETS

The science of 'computer-assisted collection' is in its infancy. Unfortunately, it is nearly impossible to test the ideas discussed above with real floristic data. This is not only because floristic data are incomplete, but also because we have no verified, reliable method of judging such incompleteness. Simulated data offer one solution to this problem (e.g. Baltanás, 1992). However, in this article we utilize data from the Tallgrass Prairie Preserve that is at a finer scale than most botanists consider 'floristic', and employ a nested plot design to mimic sampling incompleteness. A second data set that is similar in design but dissimilar in vegetation (the Oosting Natural Area; Palmer and White, 1994; Palmer, 1995) yielded similar results to those of the Tallgrass Prairie Preserve; space precludes presenting these results here. We also present the results from a true floristic data set (Oklahoma counties).

Tallgrass Prairie Preserve

The 15 000 ha Tallgrass Prairie Preserve (hereafter, TGPP) is owned and managed by The Nature Conservancy, and is located in Osage County, Oklahoma, U.S.A. Although the preserve consists primarily of grassland, there are substantial areas of forests and wetlands (Hamilton, 1996). Palmer *et al.* (2000) summarize ongoing research on the preserve.

A total of 252 10 m × 10 m permanent plots have been established throughout the preserve. 127 of these have been located at the intersections of the 1 km UTM grid, while the remaining plots have been located subjectively to represent a variety of habitats. Each corner has a nested series of quadrats of increasing size (Figure 1). We recorded the presence of each vascular plant species within each quadrat. In this article, we use 'quadrat size' to mimic collection intensity, i.e. the species present in the collection of the four smallest quadrats implies a low collection effort, and the species only in the largest quadrat (the entire 10 m × 10 m plot) represent a strong effort.

For each plot, we gathered soils and analyzed them for major nutrients, and we estimated the slope and the per cent cover of rocks, water and woody plants. While it is unlikely that for a typical floristic study such variables would be readily available, it is very likely that some relevant information would be. Their inclusion in this case study illustrates the *principles* of the use of external variables.

Other external variables are derived from a GIS for the preserve. Data layers include soil type, burn history, grazing history, land use and elevation. We also used a 1m-resolution panchromatic aerial photograph taken in 1995 by the U.S. Geological Survey. We used this image to determine the

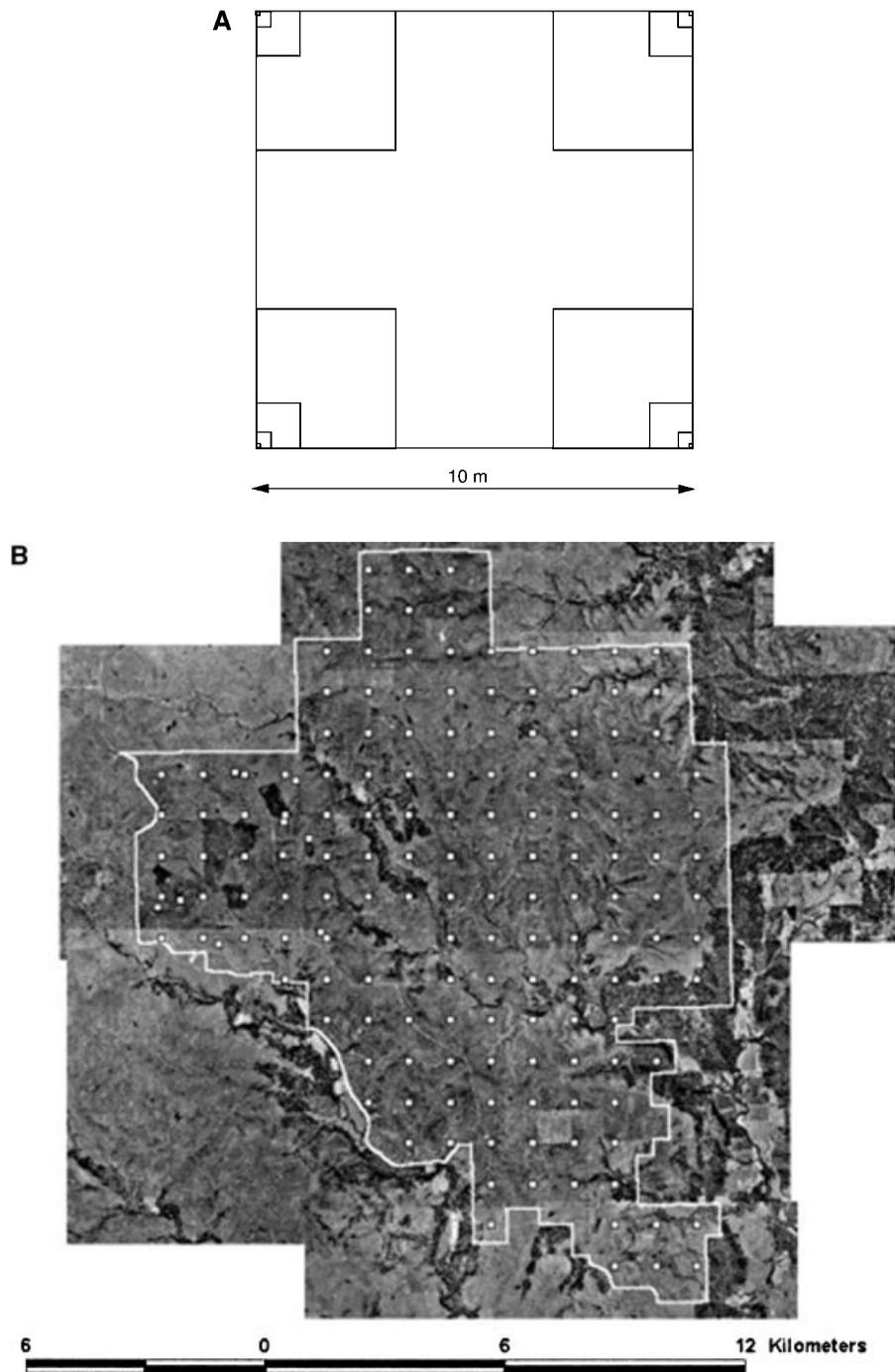


Figure 1. Field sampling in the Tallgrass Prairie Preserve (TGPP): (a) the design of the plots; (b) location of plots in the preserve

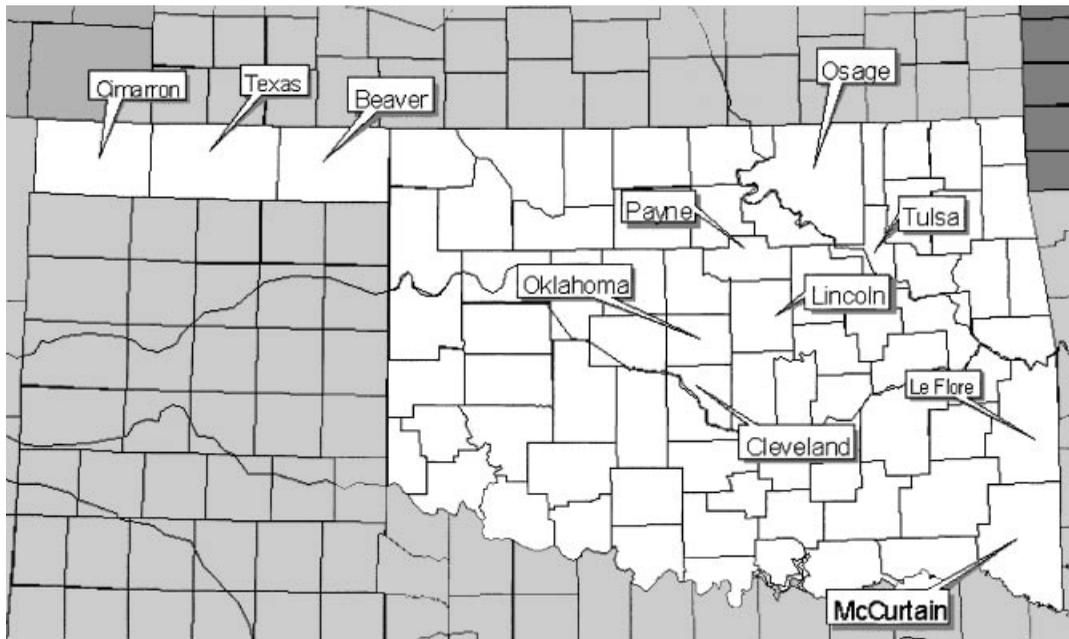


Figure 2. The counties in the State of Oklahoma, U.S.A. Counties to be discussed later are labeled

location of forests and to assess spectral heterogeneity. We implemented the GIS using ArcView and ArcInfo.

Woody Plants of Oklahoma

One of us (BWH) is developing a database of plant specimens within Oklahoma counties (Figure 2). The goal is to develop an interactive electronic floristic atlas of Oklahoma; the database currently contains 60 000 records. The woody plants are the most thoroughly documented. All label and annotation data on the herbarium sheet have been recorded. A review of the database indicates that not all counties have been studied equally, leading to an uneven distribution of completeness. Here, we use the total number of woody plant specimens as a crude proxy for collection intensity. For each county, we tabulated its area, population, mean annual precipitation and temperature, the percentage of land in various land use classes, and the percentage of land in different soil orders (alfisols, ultisols, inceptisols, vertisols, entisols and 'stony').

ANALYSES AND RESULTS

Extensive or intensive effort first?

We simulated three strategies for botanical exploration in the absence of external information for the TGPP data set. For each strategy, we assumed it took the same effort to stay in one location and explore the next largest quadrat as to move to another plot and to sample the smallest unstudied quadrat. This assumption is flawed (i.e. it actually takes more effort to move to a new location), but as we outline

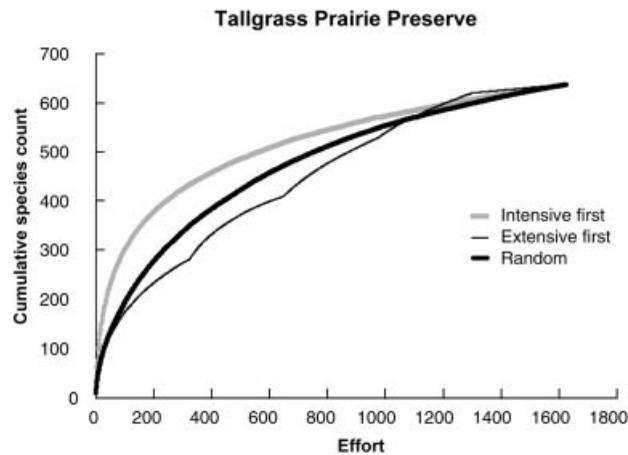


Figure 3. The cumulative species count as a function of cumulative effort for three different sampling strategies (discussed in the text) in the TGPP. Each curve represents the average of 100 random sequences

below our conclusions are conservative. For each strategy, we tallied the cumulative number of species encountered as a function of cumulative effort, and averaged the results of 100 random sequences.

The first strategy was random, i.e. at any given stage of the inventory, a plot was randomly chosen. If a plot was unsampled, we started with the smallest quadrat. If this plot was already sampled at a given quadrat size, we 'visited' the smallest unsampled quadrat and recorded the species present. This procedure continued until all quadrats in all plots were visited. For the second strategy ('extensive first') we again chose plots in a random order starting with the smallest quadrat size first, but we visit all of the plots at a given quadrat size before moving up to the next largest quadrat. The final strategy is 'intensive first'. Although we chose plots in a random order, we sequentially studied the quadrats in the plot from the smallest to the largest before choosing new plots. Figure 3 shows that the 'intensive first' strategy is superior to the random strategy, which in turn outperforms the 'extensive first' strategy. The differences in the curves in Figure 3 actually underestimate true differences between the strategies, since the true effort is the least in the 'intensive first' strategy (there is less traveling involved). Several interrelated factors appear responsible for the superior performance of the 'intensive first' strategy. First, the 'intensive first' strategy accumulates area at a faster rate than do the other strategies; the superior performance could be a simple manifestation of the species-area relationship. Secondly, the effects of grain are stronger than the effects of extent on the species-area relationship at these scales (Palmer and White, 1994). Finally, it is likely that there are a large number of infrequent species that are only likely to be sampled if a 'broad net' (i.e. a large quadrat or an intensive survey) is cast (Palmer, 1995).

Real floristic data are not only less structured than with these examples, but they typically cover regions orders of magnitude larger. It remains to be seen whether an 'intensive first' effort is a good general recommendation for floristic work, but we suspect that in most cases it will be. We recommend that botanists collect the frequent or commonplace species along with the infrequent or unusual. We have noticed that community dominants are often underrepresented in herbaria. Ultimately, we must stress that 'intense sampling' is relative: for large locations, complete inventories are nearly impossible—even if these 'locations' (e.g. counties) are merely subsets of a larger region (e.g. a state).

Using environmental information

Whether the initial efforts are extensive or intensive or some combination, it is likely that information concerning the environment can assist in selecting sites for intensive inventories.

Unusual environments. When an inventory is partially completed, information about the environment in unsampled or undersampled locations can potentially be used to guide future expeditions. In particular, we are interested in identifying locations with rare or unusual environments. Such locations are good candidates for finding new species. However, this approach assumes that there is a connection between the unusualness of the environment and the unusualness or rarity of the species present.

To assess the 'unusualness' of the environment, we first select a set of variables that are conceivably related to species composition, and standardize them (by subtracting the mean and dividing by the standard deviation). Although there are numerous conceivable indices for 'unusualness', we select three here. The first is the Euclidean distance to the nearest neighbor, the second is the Euclidean distance to the centroid and the third is the 'leverage', a measure of outlier status of independent variables in a regression (Draper and Smith, 1981). We note that these indices will be sensitive to the correlation structure of the variables, but this is not necessarily a drawback. If highly correlated variables are included, the effect is to weight them more than uncorrelated variables.

We developed an index of the rarity of species in a location as follows:

$$\text{Rarity}_k = [\sum_j I_{jk}(N - F_j)/N]/S_k$$

where I_{jk} is the incidence (1 or 0) of species j in location (e.g. plot or county) k , F_j is the total number of locations containing species j , N is the total number of locations and S_k is the number of species in location k . This index scales between zero (no infrequent species) and one (only infrequent species).

Rarity of species appears to be positively related to environmental unusualness for the TGPP (Figure 4) and Oklahoma woody plants (Figure 5). Euclidean distance to the centroid appears to be the best predictor of rarity for the TGPP. Nearest neighbor distance predicts rarity the best for the Oklahoma woody plants, although we stress that 'rarity' in this data set may be a function of collection intensity (illustrated later).

Sequential sampling of unusual environments. Since unusualness in the species appears to be related to the unusualness in the environment, it makes sense that we can use unusualness in the environment as a guide for selecting potential collection locations. Here, we propose sampling the environment in a sequential manner: After an initial inventory of one location, we choose the location, out of all available locations, that is most different from the initial one (as defined by one of the three 'unusualness' indices). After this second location is inventoried, we select the third location as the one that is most 'unusual' of the remaining ones. This continues until all locations are sampled. We should reiterate that 'unusualness' is as much a property of our database as it is of nature—an 'unusual' location during the beginning of the survey might be commonplace by the end of the survey.

Figure 6 shows that this sequential strategy results in a rapid accumulation of species in the TGPP. Except at the very beginning of a survey, the directed sampling attains a given number of species much earlier than for an undirected, or random, sampling. Thus, if we have good information about the environment at unsampled locations, we recommend a strategy in which 'unusual' locations have priority. We do not suspect that the rigid algorithm presented here will be very useful, because in practice some locations are more accessible than others—but we feel that a list of a suggested sites can be a useful tool.

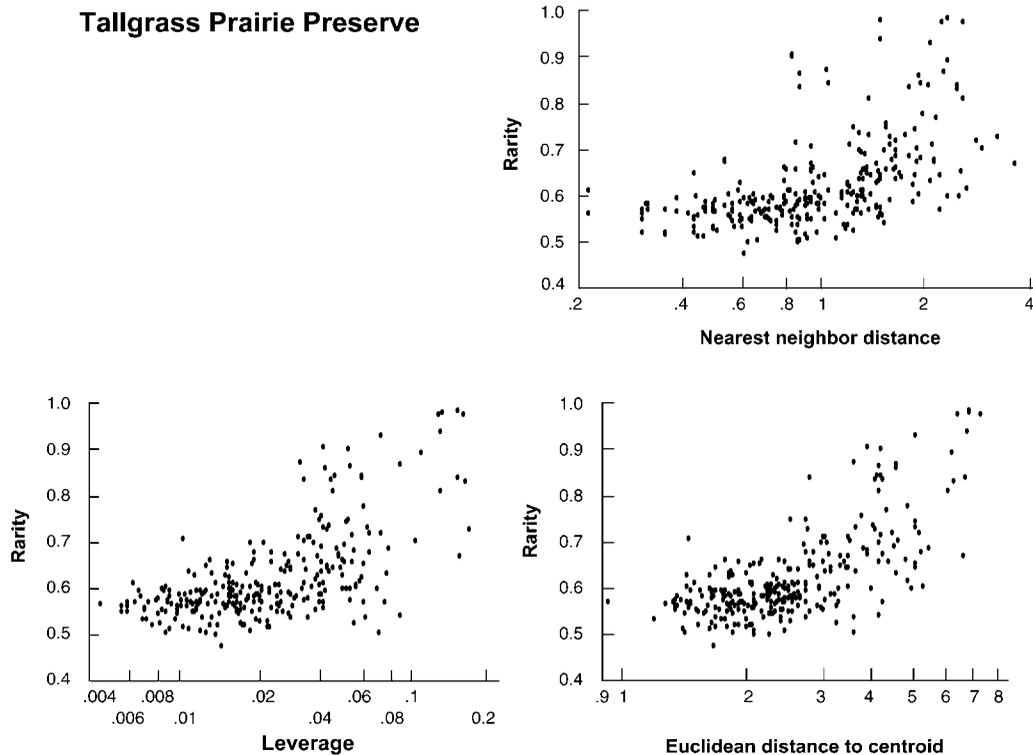


Figure 4. A plot of rarity of species (index given in the text) as a function of the unusualness in the environment in the TGPP. Unusualness is measured in three ways: nearest neighbor distance, leverage, and Euclidean distance to the centroid of the environmental variables

Residuals from richness regressions. The approaches heretofore described would be ideally suited for a floristic inventory of a region such as a park or preserve, for which there does not yet exist a species list. Another kind of floristic inventory is for a larger region (such as a state) in which the overall species list is fairly well known, but the lists of species within subdivisions of the region (such as counties) are incomplete. In this section, we describe an approach to accelerate the accumulation of records within the subdivisions.

The assumption of the approach is that sites that have been poorly explored will have fewer records (i.e. species recorded in them) than environmentally similar sites that have been well explored. If so, poorly explored sites should fall below the regression function. We propose that sites with negative residuals have high priority for botanical expeditions. Once such expeditions are made, not only will the data change, but so will the regression functions. This would lead to updated prioritizations for future expeditions.

Our simulation results (not shown due to space limitations) reveal that the regression approach works well for the Oosting Preserve (Palmer and White, 1994) but not for the TGPP. The reason for the poor performance in the prairie may be that diversity occurs in 'hot spots'. If so, it may be more profitable to search for new records in an area that is already demonstrated to reveal new and interesting species (i.e. a high *positive* residual). Thus, for the regression approach to work, we would

Oklahoma woody plants

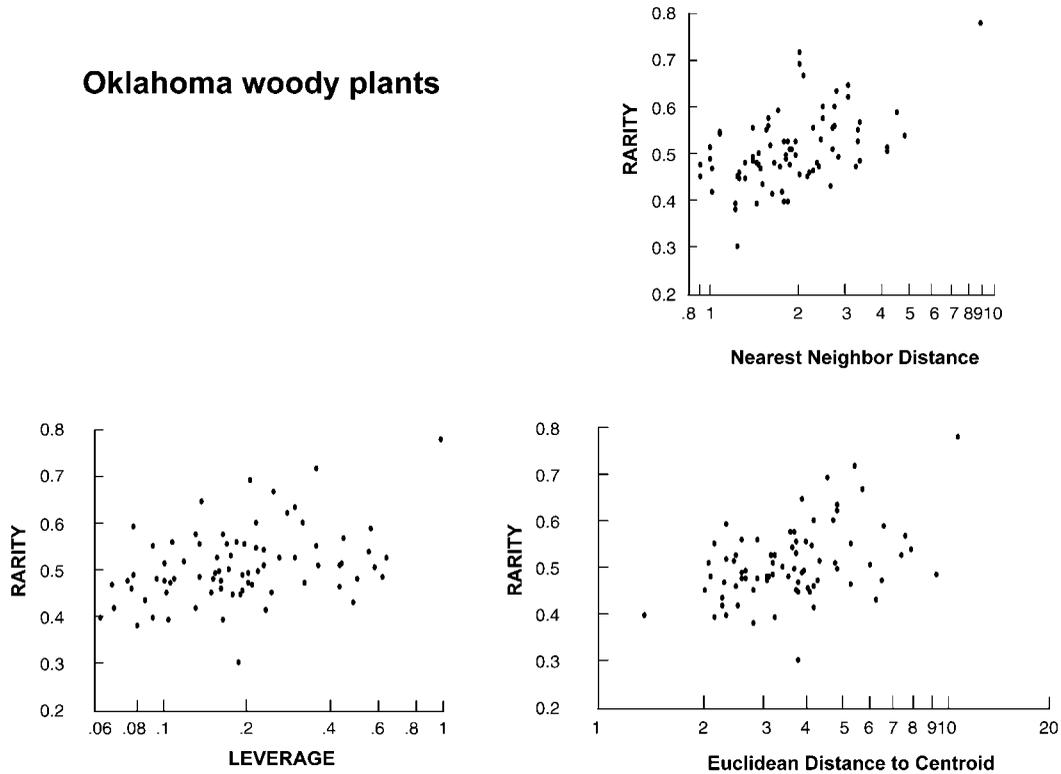


Figure 5. A plot of rarity of species (index given in the text) as a function of the unusualness in the environment in Oklahoma counties. Unusualness is measured in three ways: nearest neighbor distance, leverage, and Euclidean distance to the centroid of the environmental variables

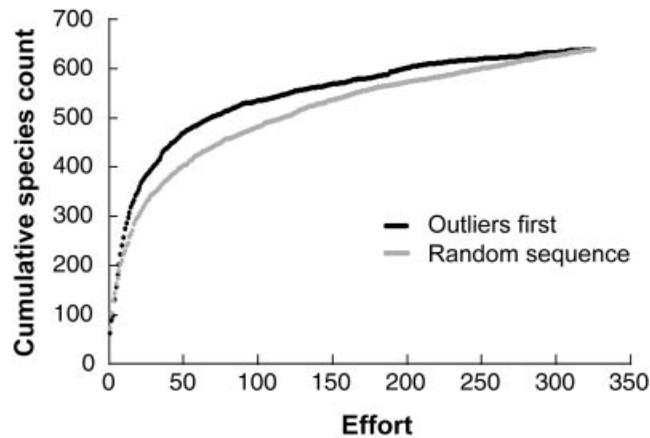


Figure 6. The cumulative number of species encountered as a function of cumulative number of quadrats sampled, for a random sequence and a sequence in which environmental outliers are sampled first in the TGPP. The outliers used here are selected on the basis of distance to the environmental centroid, although the other measures of 'environmental unusualness' give similar results

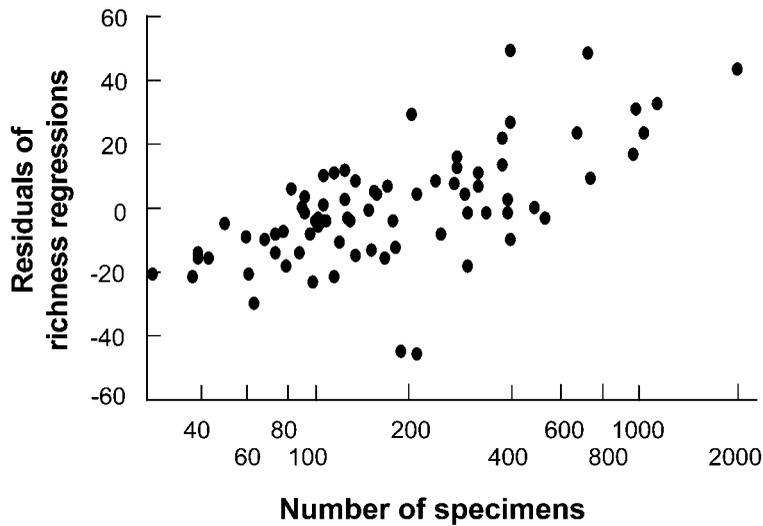


Figure 7. The relationship between the residuals of species richness from a multiple regression, and the number of specimens present in herbaria, for woody plant species in Oklahoma counties

need to find tools for distinguishing between regression errors caused by undersampling and those caused by intrinsic variation. Perhaps the solution lies in an analysis of complementarity (Williams *et al.*, 1996; Ney-Nifle and Mangel, 1999) or nestedness (Cam *et al.*, 2000a,b).

For the Oklahoma woody plants data set, we can indirectly assess collection effort as the number of herbarium specimens per county. Since the residuals from the richness regressions are strongly correlated with the number of specimens (Figure 7), we strongly suspect that the richness regression approach is reasonable for Oklahoma as a whole.

Although our simulations adopt a fairly rigid approach, in which only one site (the one with the largest residual) is chosen per expedition, we do not advocate slavish adoption of such a procedure. Instead, we recommend using regression as one of many criteria (including logistical ones) to direct new expeditions.

Habitat diversity. In general, a high diversity of environments will lead to a high diversity of species. This phenomenon is so intuitively clear that it is difficult to dispute. Therefore, it is generally recommended (and appreciated by experienced botanists) that visiting a region with many habitats will result in a relatively long species list.

Spectral heterogeneity. According to the SVH, we expect spectral heterogeneity to be positively related to species richness. However, we do not observe this for 0.01 ha plots (Table 1). On the other hand, we do observe that the rarity index and the number of rare species are positively related to spectral heterogeneity, with the strongest correlations existing at the 10 m scale (the length of the sides of the plot). So although our data do not support the SVH at this fine scale, we believe that spectral heterogeneity holds promise for mapping biodiversity and for directing collection efforts. The weakness of the correlations may be due, in part, to inaccuracies in georeferencing, differences between the time of the image and that of the sampling, and inherent variation in richness of small plots. We strongly suspect that the SVH will prove useful for assessing species richness of

Table 1. Pearson correlations between spectral heterogeneity measured at three scales, and three measures of biodiversity for 187 0.01 ha plots in the TGPP. 'Infrequent' species are those found in no more than three plots

	2 m	10 m	30 m
Species richness	0.012	0.080	0.157*
Number of infrequent species	0.181*	0.229**	0.222**
Rarity index	0.351**	0.541**	0.418**

* $p < 0.05$.

** $p < 0.01$.

multispectral data and at scales of 1–1000 ha. The utility of the SVH at such scales will be the subject of future investigations.

MAPS FOR FINDING SPECIES

A map directing the botanist to potentially interesting locations is more likely to be useful (or to be intuitively grasped) than a numerical or text product. Fortunately, many of the analyses and principles discussed above lend themselves directly to mapping. A few examples are given below.

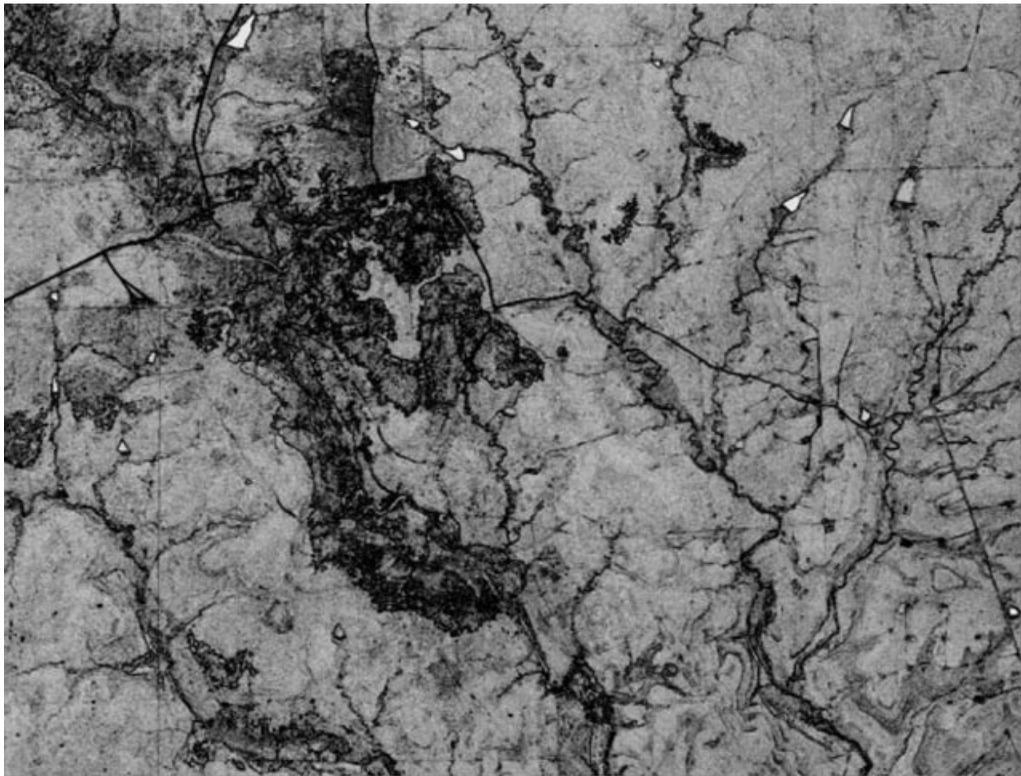


Figure 8. A map of spectral heterogeneity in the TGPP. Darker colors are more heterogeneous, as assessed by the standard deviation of reflectance of 100 1 m × 1 m pixels in a 10 m × 10 m grid

Tallgrass Prairie Preserve

Figure 8 illustrates the standard deviation of reflectance in a portion of the TGPP. Dark colors indicate locations where there is more spectral heterogeneity. The different color bands represent heterogeneity at different spatial scales. The map is essentially a map of 'botanical interestingness', in which rock outcrops, streams, pond margins, savannahs, forest edges, forest gaps, anthropogenic disturbances and wetlands—all sites likely to reveal new species—are clearly delineated. We predict, according to the Spectral Variation Hypothesis, that the darker regions in Figure 8 will also be relatively species-rich. We suspect use of more than one band of spectral information will result in even more useful maps.

We also produced maps of Euclidean distance to the environmental centroid for the TGPP (not shown, but see the discussion concerning Figure 4). Like Figure 8, these maps appeared to correspond with locations that were 'botanically interesting' but differed substantially in the details.

Figure 9 illustrates the predicted richness, number of infrequent species, and rarity, derived from regression equations. These maps also indicate potentially interesting sites for future investigation, but differ from Figure 8 in that they are based on prior floristic information. Updating the information will result in new regressions and hence new maps. The number of infrequent species and rarity will be especially sensitive to new data, since these variables are defined on the basis of the overall data set.

Woody plants of Oklahoma

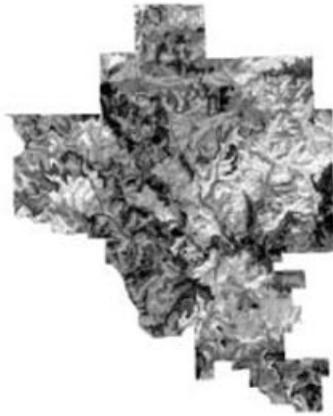
Figure 10 shows the observed number of species (number of records), the number of species predicted according to the regression equation and the number of herbarium specimens per county. The number of records is high in eastern Oklahoma counties (as in Le Flore and McCurtain County; Figure 2), where precipitation is high. A positive link between precipitation and floristic richness has been recorded in other locations (e.g. Linder, 1991). However, the number of records is also high in counties with high population density (Oklahoma, Tulsa), with major research universities (Payne, Cleveland) or with natural areas that have lured botanists in the past (Cimarron County, which contains Black Mesa, a Pliocene lava flow, and Osage County, the location of the TGPP). Thus, the number of records to a large degree reflects the number of specimens (Figure 10).

The counties with the strongest negative residuals occur in the areas that have not been well collected (e.g. Texas, Beaver and Lincoln counties) or in eastern counties with moderate collection intensity. It is possible that a thorough survey of the three or four counties with the strongest negative residuals will lead to a reassessment of the factors responsible for woody plant diversity, as well as a reprioritization of the counties worthy of later botanical exploration.

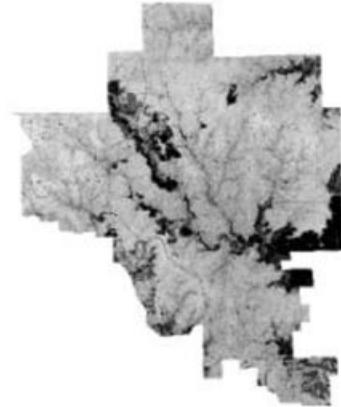
CONCLUSIONS

The techniques described in this article are designed to help the practicing botanist complete floras. They are intended to supplement, rather than replace, intuition and experience. Our evaluation of these techniques confirms, with a few exceptions, their promise. A number of factors challenge future development of these tools. We need better ways (other than number of specimens) to assess collection effort. We also need to develop more sophisticated statistical models—for example, an asymmetric error distribution might be more appropriate than a symmetric one for biodiversity data

Species richness



Rarity



Number of infrequent species

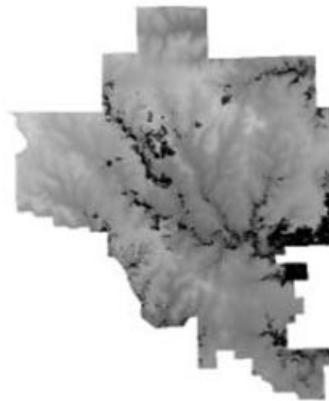


Figure 9. Predicted values from stepwise regressions in the Tallgrass Prairie Preserve, displayed geographically (darker = larger values). Independent variables selected for the richness regression include aspect, soil type and the standard deviation of reflectance at 30 m. Independent variables for the rarity regression include spectral standard deviation at 2 m and 10 m, elevation, soil type and woody cover squared. Independent variables for the number of infrequent species (defined as the number of species found in no more than three plots) include elevation and woody cover squared

(Palmer, 1995). Despite advances in technology, computer speed remains a limiting factor in image processing.

One of the biggest future challenges is the implementation of the iterative approaches described above, particularly for large regions. For a large region like a country or a state, the expertise is widely distributed and the compilation of data is difficult. We propose developing a centralized system in which maps of unusualness, rarity and floristic incompleteness are made generally available over the Internet. When new specimens are deposited in participating herbaria, the regression models, and hence the maps, can be instantaneously updated. Thus, the community of experts can simultaneously

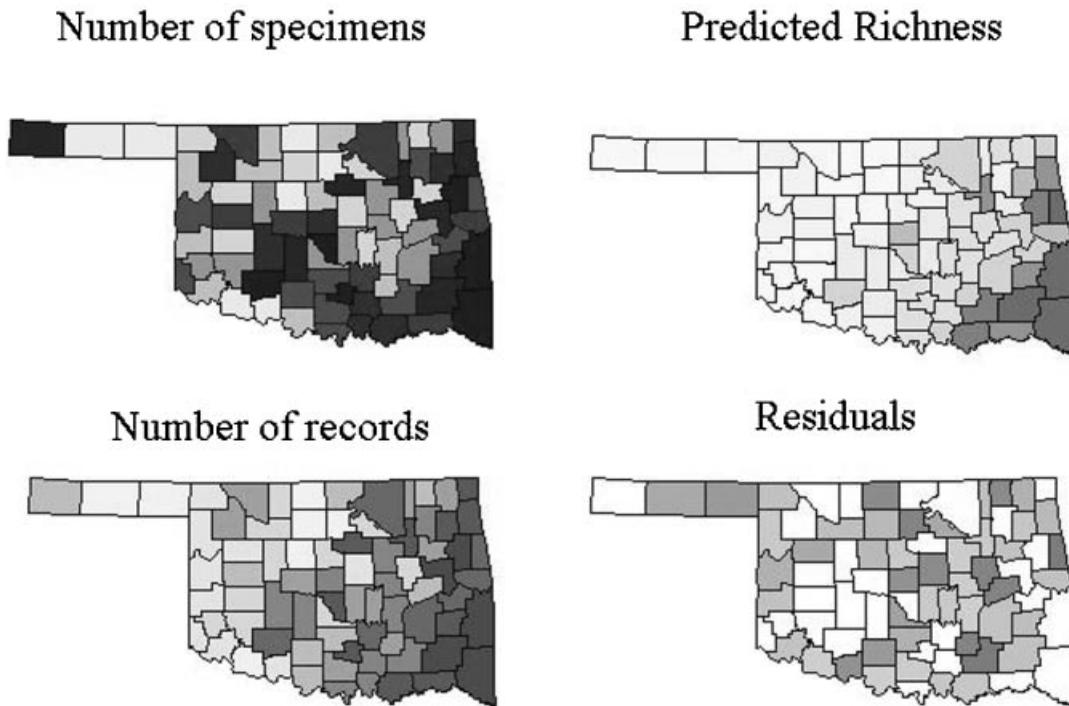


Figure 10. Woody plant species of Oklahoma: maps of the number of recorded species, predicted number of species, number of herbarium records and residuals from the regression equation

participate in completing an atlas of biodiversity. While this approach might raise some quality control issues, it might also serve to encourage the participation of dedicated amateurs.

We recognize that directed expeditions might lead to biased, even if more complete, results. For most conservation purposes, documented (and correct) records of occurrence are highly desirable whether or not the methodology is biased. However, we acknowledge that some form of objective sampling may be necessary for critical analysis, such as on the effects of climate change.

Although more information is generally better than less information, resources for gathering information are limited. Conservation decisions often need to be made when data are inadequate. We must therefore develop techniques for acting with incomplete information (Polasky *et al.*, 2000).

Field biologists have often recognized a distinction between the 'art' and the 'science' of their work. A quantitative approach has often been the hallmark of the 'scientific' aspect. We believe that this view is outdated: quantitative approaches, coupled with expert opinion and intuition, can help bridge the gap between art and science. Finally, we write at a time when field biology and systematics is on the decline in institutions of higher education (Lee, 2000). As a result, understanding and supplementing intuitive approaches to inventory is critical. In the midst of the biodiversity crisis, efforts to integrate the diverse scientific approaches to biodiversity are urgently needed. Otherwise, extinction may prevent the completion of species lists.

ACKNOWLEDGEMENTS

We thank Jayne Salisbury and the OSU Spatial and Environmental Information Clearinghouse for advice and use of GIS facilities. We thank Robert Hamilton for access to the TGPP, logistical support, and access to data. José Ramón Arévalo, Chris Chase, Chris Sellers, Holly Ewing, Dai-Jun Zhang, Karol Ezell, Sonia Jaiswal, David Nemcok, Adam Ryburn, Linda Gatti-Clark, Mary Carmen Cobo, John Wilson, Jerad Linneman, Steven Thompson and numerous volunteers provided field assistance at the TGPP. Adam Ryburn supplied environmental data for Oklahoma Counties. The Oklahoma Water Resources Center, Oklahoma State University College of Arts and Sciences, The Nature Conservancy, The Phileology Trust and The EROS Data Center all provided financial, material and/or logistical support for earlier phases of this research.

REFERENCES

- Baltanás A. 1992. On the use of some methods for the estimation of species richness. *Oikos* **65**: 484–492.
- Bunge J, Fitzpatrick M. 1993. Estimating the number of species: a review. *J. Am. Stat. Assoc.* **88**: 364–373.
- Cam E, Nichols JD, Hines JE, Sauer JR. 2000a. Inferences about nested subsets structure when not all species are detected. *Oikos* **91**: 428–434.
- Cam E, Nichols JD, Sauer JR, Hines JE, Flather CH. 2000b. Relative species richness and community completeness: birds and urbanization in the mid-Atlantic states. *Ecol. Appl.* **10**: 1196–1210.
- Colwell RK, Coddington JA. 1994. Estimating the extent of terrestrial biodiversity through extrapolation. *Phil. Trans. Roy. Soc. Lond. Ser. B.* **345**: 101–118.
- Draper NR, Smith H. 1981. *Applied Regression Analysis*. 2nd edn. Wiley: New York.
- Gauch HG Jr. 1982a. *Multivariate Analysis and Community Structure*. Cambridge University Press: Cambridge.
- Gauch HG Jr. 1982b. Noise reduction by eigenvalue ordinations. *Ecology* **63**: 1643–1649.
- Gleason HA. 1922. On the relation between species and area. *Ecology* **3**: 158–162.
- Gould W. 2000. Remote sensing of vegetation, plant species richness, and regional biodiversity hotspots. *Ecol. Appl.* **10**: 1861–1870.
- Hamilton RG. 1996. Using fire and bison to restore a functional tallgrass prairie landscape. *Transactions of the 61st North American Wildland and Natural Resources Conference*: 208–214.
- Kieran J (ed.). 1957. *Treasure of Great Nature Writing*. Hanover House: Garden City, New York.
- Lee MSY. 2000. A worrying systematic decline. *Tr. Ecol. Evol.* **15**: 346.
- Linder HP. 1991. Environmental correlates of patterns of species richness in the south-western Cape Province of South Africa. *J. Biogeogr.* **18**: 509–518.
- Maurer BA. 1994. *Geographic Population Analysis: Tools for the Analysis of Biodiversity*. Blackwell Scientific Publications: Oxford.
- McCollin D, Moore L, Sparks T. 2000. The flora of a cultural landscape: environmental determinants of change revealed using archival sources. *Biol. Cons.* **92**: 249–263.
- McNeely JA. 1990. *Conserving the World's Biodiversity*. IUCN: Gland, Switzerland.
- Ney-Nifle M, Mangel M. 1999. Species-area curves based on geographic range and occupancy. *J. Theor. Biol.* **196**: 327–342.
- Palmer MW. 1990. The estimation of species richness by extrapolation. *Ecology* **71**: 1195–1198.
- Palmer MW. 1991. Estimating species richness: the second-order jackknife reconsidered. *Ecology* **72**: 1512–1513.
- Palmer MW. 1994. Variation in species richness: towards a unification of hypotheses. *Folia Geobot. Phytotax.* **29**: 511–530.
- Palmer MW. 1995. How should one count species? *Nat. Areas J.* **15**: 124–135.
- Palmer MW, White PS. 1994. Scale dependence and the species-area relationship. *Am. Nat.* **144**: 717–740.
- Palmer MW, Wohlgemuth T, Earls P, Arévalo JR, Thompson SD. 2000. Opportunities for long-term ecological research at the Tallgrass Prairie Preserve, Oklahoma. Pages 123–128. In *Cooperation in Long Term Ecological Research in Central and Eastern Europe: Proceedings of the ILTER Regional Workshop*, Lajtha K, Vanderbilt K (eds). 22–25 June, 1999, Budapest, Hungary. Oregon State University: Corvallis, OR.
- Pearson DL, Carroll SS. 1998. Global patterns of species richness: spatial models for conservation planning using bioindicator and precipitation data. *Cons. Bio.* **12**: 809–821.
- Polasky S, Camm JD, Solow AR, Csuti B, White D, Ding R. 2000. Choosing reserve networks with incomplete species information. *Biol. Cons.* **94**: 1–10.
- Prendergast JR, Eversham BC. 1997. Species richness covariance in higher taxa: empirical tests of the biodiversity indicator concept. *Ecography* **20**: 210–216.
- Ranta P, Tanskanen A, Niemelä J, Kurtto A. 1999. Selection of islands for conservation in the urban Archipelago of Helsinki. *Finland. Cons. Bio.* **13**: 1293–1300.

- Robinson GR, Yurlina ME, Handel SN. 1994. A century of change in the Staten Island flora: ecological correlates of species losses and invasions. *Bull. Torr. Bot. Club* **121**: 119–129.
- Rosenzweig ML. 1995. *Species Diversity in Space and Time*. Cambridge University Press: Cambridge.
- Ryken JJ, Capen DE, Mahabir SP. 1997. Ground beetles as indicators of land type diversity in the Green Mountains of Vermont. *Cons. Bio.* **11**: 522–530.
- Williams CB. 1964. *Patterns in the Balance of Nature*. Academic Press: New York.
- Williams P, Gibbons D, Margules C, Rebelo A, Humphries C, Pressey R. 1996. A comparison of richness hotspots, rarity hotspots, and complementary areas for conserving diversity of British birds. *Cons. Bio.* **10**: 155–174.
- Wilson EO (ed.). 1988. *Biodiversity*. National Academy Press: Washington, DC.
- Wright DH, Reeves JH. 1992. On the meaning and measurement of nestedness of species assemblages. *Oecol.* **92**: 416–428.