SPECIES-AREA RELATIONSHIPS INDICATE LARGE-SCALE DATA GAPS IN HERBARIUM COLLECTIONS

Justin K. Williams and William L. Lutterschmidt

Department of Biological Sciences, Sam Houston State University, Huntsville, Texas 77341-2116

Abstract: Species-area relationships (SAR) are useful in predicting species richness for a given geographical area. Using SAR and the state of Texas as a case study, we present a model that provides a quantifiable and objective approach for identifying large scale data gaps in species inventories and museum collections by comparing documented species richness (determined by herbarium records) to predicted species richness. For Texas our results indicate that 88% of the counties have documented species richness values that are below predicted values based upon our results from the proposed model. Many biological survey and inventory programs are funded to document species occurrence and richness. Such studies help identify species of concern and enhance species conservation efforts. Future species inventories may benefit from such predictive models in identifying regions of large scale data gaps.

Keywords: biodiversity, herbarium, inventory, mapping, predictive modeling, species richness, Texas.

Classic (Isley, 1972) and recent (Turner, 1998; Ertter, 2000; Heywood, 2001; Prather et al., 2004) articles have emphasized floristic studies and the need for continued collecting and cataloging of herbarium specimens. Unfortunately, this appeal for continued collecting has mostly been based on anecdotal evidence. Few articles attempt to quantify the current stagnation in botanical collections. Prather et al. (2004) provide the most recent and compelling evidence for large-scale information gaps by presenting data that show a temporal decline in herbarium collections over the last three decades. Prather et al. (2004) also identify regions with increasing and decreasing herbarium collections in the continental U.S.A. These geographical data, however, oversimplify spatial data and assume that specimen growth in a region's herbaria indicates an increase in that region's floristic inventory.

The species-area relationship (SAR) is regarded as "one of community ecology's few laws" (Schoener, 1976). SAR simply states that as area increases, species richness increases (Brown and Lomolino, 1998). Often SAR can be used to estimate species richness (S) for a given geographical area (A). Estimations of S are based on the formula $S = CA^z$ where z and C are constants varying with geographic location and taxa studied (MacArthur and Wilson, 1967). Such SAR have used geographical area to predict species richness of birds (Diamond and Mayr, 1976), earthworms (Judas, 1988), arthropods (Covarrubias and Elgueta, 1991), and stream fishes (Angermeier and Schlosser, 1989). These relationships have also been useful in determining floristic richness (McNeill and Cody, 1978; Buys et al., 1994; Palmer et al., 2002; Fridley et al., 2005). Although species-area analyses are commonly used and generally accepted for predicting species richness, there is little indication of its utility in identifying large scale data gaps in herbarium collections.

We present and discuss a model that provides a quantifiable and unbiased approach for identifying large scale data gaps in herbarium collections. By comparing documented species richness values (determined from herbarium records) with predicted species richness values (determined from the formula $S = CA^z$), we address the

TABLE 1. Known values of species richness for vascular plants and associated geographic area from published accounts.

Number of species	Area (km²)	Location	Citation
4839	677940.3	Entire state of Texas	Correll, D. S. & M. C. Johnston. 1970.; Turner, B. L. et al. 2003.
1498	2561.51	Travis Co., Tx	Carr. B. 2004.
1373	6221.18	Walker, Montgomery, & San Jacinto Cos., Tx	Nesom, G. L. & L. E. Brown. 1998.
1153	2038.33	Walker Co., Tx	Nesom, G. L. & L. E. Brown. 1998; Williams, J. K. (pers. obs.)
985	1217.3	Madison Co,. Tx	Neill, A. K. & H. D. Wilson. 2001.
666	2937.06	San Saba Co., Tx	Garner, P. M. B. 1975.
636	4019.68	Tom Greene Co., Tx	Eckhardt. R. F. 1975.
605	2768.71	McCulloch Co., Tx	Whisenant, S. G. 1982.
559	2362.08	Throckmorton Co., Tx	Cornelius, J. M. 1983.
457	3263.4	Coleman Co., Tx	Nixon, M. R. 1987.
495	80.9375	Lake Meredith, Carson Co., Tx	Phillips, J. W. 1997.
485	5.6721	Love Creek Nature Preserve, Bandera Co., Tx	Denny, G. 2002
485	2.6159	Little Thicket Nature Sanctuary, San Jacinto Co., Tx	Peterson, C. D. & L. E. Brown. 1983.
470	37.555	Ogallala ecotone on the Dempsey divide, Roger Mills Co., Ok	Freeman, C. C. et al. 2003.
459	15.6695	Big Lake Bottom wildlife management area, Anderson Co., Tx	Fleming, K. M. et al. 2002.
401	2.8231	Hickory Creek Unit of the Big Thicket National Preserve, Tyler Co., Tx	MacRoberts, B. R. et al. 2002.
229	64.75	Pantex Nuclear Facility, Carson. Co., Tx	Johnston, M. C. & J. K. Williams. 1995.

following questions: 1) can species-area relations be used to predict plant diversity?, 2) using predicted species richness, can significant data gaps in herbarium records be geographically identified within a large-scale geographic region?, and 3) can predicted species richness be used to determine sampling effort and a threshold number of samples needed to eliminate data gaps in museum collections?

MATERIALS AND METHODS

A literature search was performed to identify published checklists and floras for regions of known area with defined boundaries within and bordering the state of Texas. In all 17, checklists and floras were found (Table 1). From these checklists one value represented the entire state of Texas, nine represented entire counties, and seven

represented smaller inventories collected within counties. Each study provides a value for species richness and geographical area smpled (which we converted to square kilometers). Both species richness and geographical area were log transformed and entered into a database. The database was imported into SPSS® version 10.1 and a linear regression was performed to determine the statistical relationship between species richness (dependent variable) and geographical area (independent variable). This analysis provided the theoretical slope (z) and intercept (C) for the formula $S = CA^z$. We then predicted species richness for each individual county in Texas by applying the above determined constants z and C to the Arrhenius (1921) log-log (log S = log C + z(log A)) model with A representing the area in square kilometers for each of the 254 counties in Texas.

Next we accessed cataloged herbarium specimens through the Flora of Texas Consortium (FTC; http://csdl.tamu.edu/FLORA/ftc/ftchome.htm) database and recorded the documented species richness (determined by the number of species collected and identified from each county to date) and the number of specimens reported from each county in Texas. We then ran a cubic regression analyses comparing documented and predicted species richness for each of the 254 counties with relation to area.

Lastly, using information gathered from the FTC we performed a linear regression to describe the relationship between the number of herbarium specimens (independent value) and documented species richness (dependant value).

Results and Discussion

The constants z (0.1553) and C (266) for vascular plants in Texas were determined using linear regression (Fig. 1) of geographical area and known species richness values cited from the 17 floristic inventories listed in Table 1. The determined

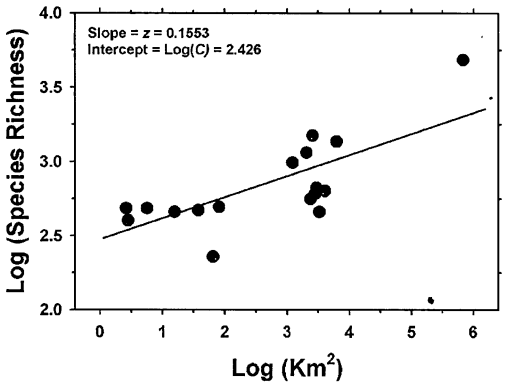

FIG. 1. Logarithmic relationship between species richness and geographic area. Dots = each of the 17 checklist and floras in Table 1 plotted for known log species richness and log geographical area. Solid line = the regression between log known species richness and log geographical area. The regression indicates a significant relationship between species richness and geographical area ($F = 19.60$, $df = 15$, $p < 0.001$, $r^2 = 0.567$) with area explaining nearly 57% of the variation in species richness. Constants z and C were obtained from this analysis for use in the species-area relationship formula $S = CA^z$.

value of z (0.1553) is consistent with the reported and accepted range of z values (0.12–0.17) for terrestrial plants within continents (MacArthur and Wilson, 1967). For a given square-kilometer in Texas, C indicates a species richness of 266. Using z and C in the formula log S = log C + z(log A), we addressed our first question and predicted species richness for each of the 254 counties in Texas. Our approach determined a statewide z and C value by plotting data for all checklists within the state of Texas (Fig. 1). Consequently, we most likely overestimated species richness in the northern counties and underestimated species richness in the southern counties. We used this approach because the checklists used to determine z and C are randomly scattered throughout Texas (Fig. 2). However, it is possible that predicted species richness for each county could be further modified by determining unique z and C constants for

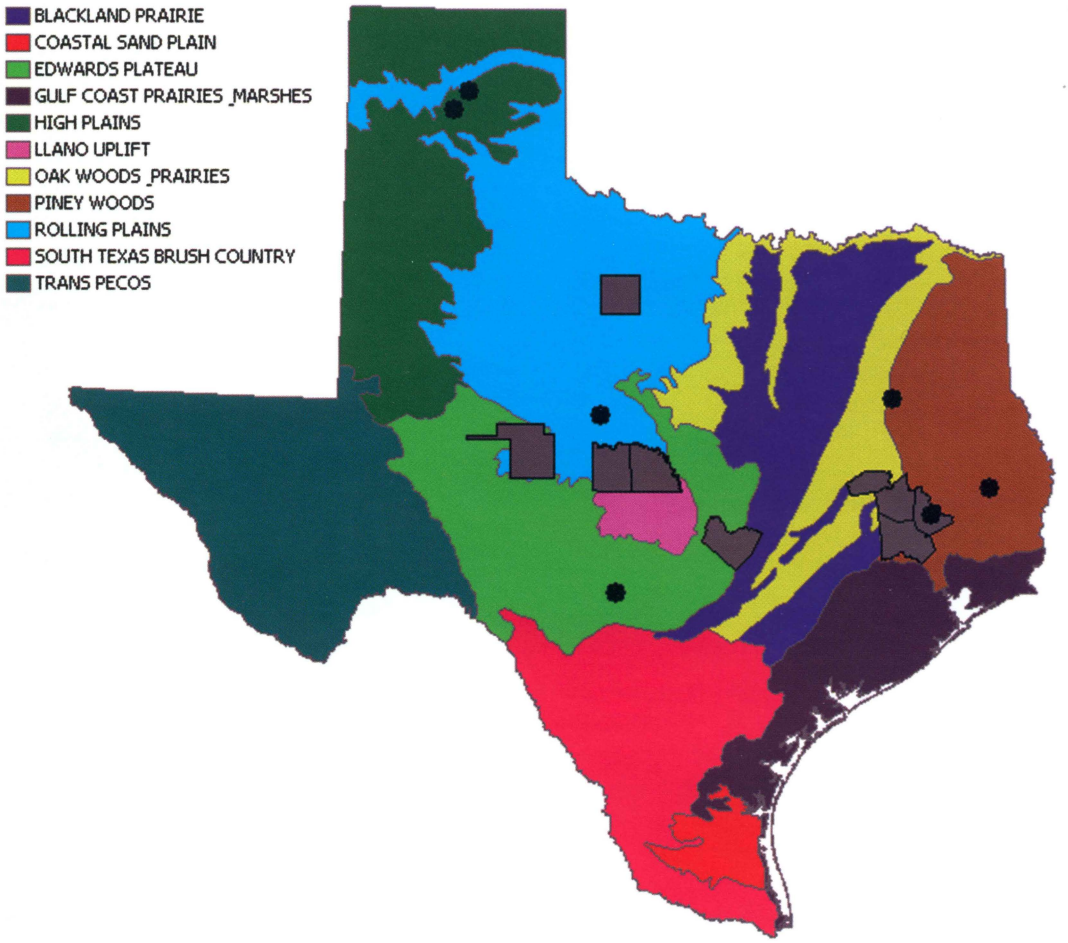

FIG. 2. Location of the local floras and checklists in Table 1 used to calculate *z* and *C* superimposed over the 11 different physiognomic regions in Texas. Gray polygons represent inventories of entire counties, black dots represent local inventories.

each of the 11 physiognomic regions of Texas and applying these values to the area of the counties within the specific physiognomic region. In order to determine unique *z* and *C* constants for each of the physiognomic regions, a minimum of three floristic inventories (within a known boundary) within each region needs to be performed and documented. Given that there are 11 physiognomic regions a minimum of 33 inventories need to be performed. To date there are 17. Ideally, more inventories performed per region would yield more optimal results. We also view the lack of checklists and floristic inventories across these physiognomic regions as a data gap.

To address our second question, we used cubic regression analyses to compare documented ($F_{(1, 250)} = 14.10$; $p < 0.001$; $r^2 = 0.145$) and predicted species richness ($F_{(1, 250)} = 5280.55$; $p < 0.001$; $r^2 = 0.984$) for each of the 254 counties with relation to geographical area (Fig. 3). Counties with documented species richness that approximate or exceed predicted species richness fall on or above the predicted species regression line; counties with under represented documented species richness fall below the predicted species regression line.

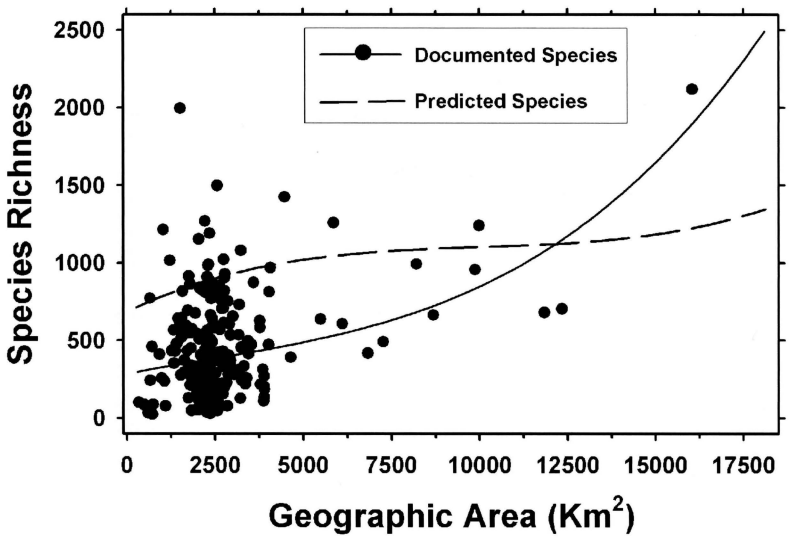

FIG. 3. Relationship between species richness and geographic area. Dots = each of the 254 counties in Texas plotted for documented species richness and geographical area. Solid line = the cubic regression between documented species richness and geographical area. Dashed line = the cubic regression between predicted species richness and geographical area. Counties (dots) near or above the predicted regression line (dashed line) indicate well collected counties that match or exceed predicted species richness.

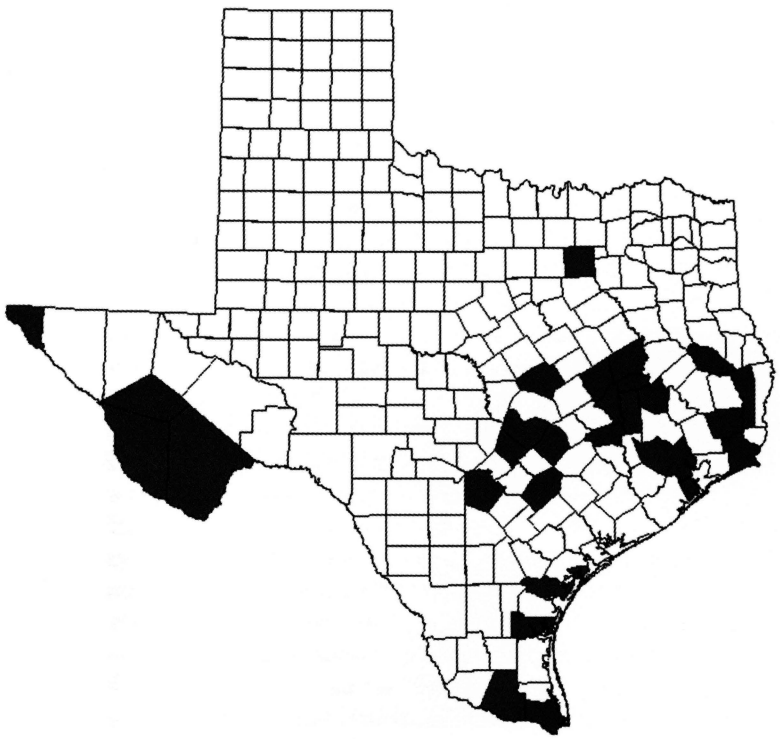

FIG. 4. Counties (shaded) where documented species richness approximates or exceeds predicted species richness. For the majority of counties in Texas documented species richness does not match predicted richness.

TABLE 2. Counties with documented species richness that approximate or exceed predicted species richness.

	County in Texas	Documented sp. rich.	Predicted sp. rich.	Herbarium Specimens	Area km^2
1	ANGELINA	834	875	1728	2077.18
2	ARANSAS	771	731	2002	652.68
3	BASTROP	984	889	2394	2299.92
4	BELL	1023	914	3168	2745.4
5	BEXAR	1080	937	2548	3229.73
6	BRAZOS	1996	833	14300	1517.74
7	BREWSTER	2120	1202	13705	16039.87
8	CAMERON	1191	892	5461	2346.54
9	DALLAS	909	888	2241	2279.2
10	EL PASO	851	907	2296	2623.67
11	GALVESTON	1214	785	3530	1030.82
12	GONZALES	904	915	1999	2766.12
13	GRIMES	840	873	1746	2056.46
14	HARDIN	990	890	2460	2315.46
15	HARRIS	1425	986	3751	4478.11
16	HAYS	914	852	2625	1756.02
17	HIDALGO	968	971	2931	4066.3
18	JASPER	867	896	1955	2426.83
19	JEFF DAVIS	1259	1028	4719	5863.76
20	JEFFERSON	883	891	1853	2341.36
21	KLEBERG	857	886	2115	2255.89
22	LEON	928	915	2431	2776.48
23	MADISON	985	805	2430	1217.3
24	PRESISIO	1240	1116	3654	9987.04
25	ROBERTSON	1269	884	4867	2214.45
26	SAN PATRICIO	863	855	2425	1792.28
27	TRAVIS	1498	904	7351	2561.51
28	WALKER	1153	872	2727	2038.33
29	WASHINGTON	816	838	2040	1577.31

This cubic regression model allows one to identify counties that are well collected and those that are under collected. Our results indicate that only 29 (or 11.4%) of the 254 counties in Texas fall close to or above the predicted line and are, therefore, considered well collected (Fig. 3). The 29 well collected counties are listed in Table 2 and are presented spatially on a map of Texas (Fig. 4). Interestingly, all counties with documented species richness values approximating or exceeding predicted values have, or are neighboring, universities with systematic botany programs. A comparison between counties with and without herbaria (Fig. 5) indicate a significant difference between both the percent species representation (documented species richness/predicted species richness) ($t_{(253)} = -9.494$; $p < 0.001$) and mean herbarium specimens ($t_{(253)} = -10.156$; $p < 0.001$). Although not significant ($t_{(253)} = -0.492$; $p = 0.623$), counties with herbaria have more documented species per geographical area than non-herbaria counties (Fig. 6).

The implications for the above model may have broad interest. Apart from isolating geographical areas with paucity in collection, the model identifies and defines geographical areas with limited data on documented species richness and distribution. Detailed specimen collections are im-

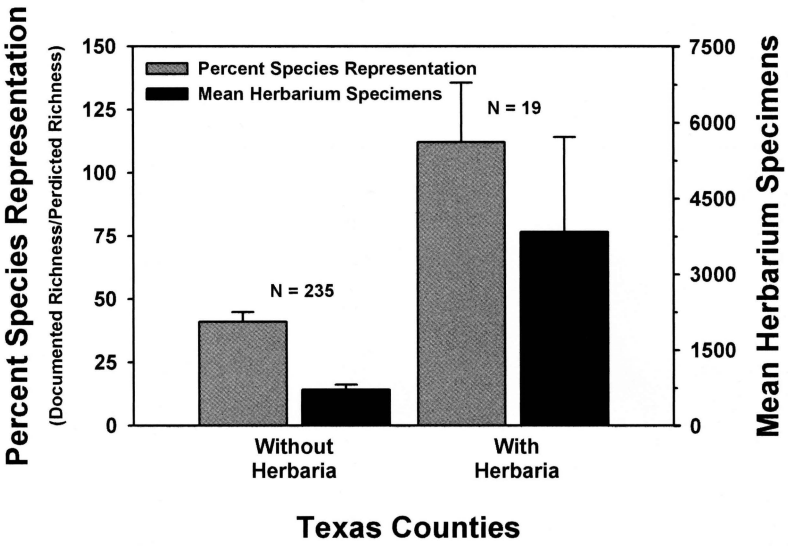

FIG. 5. The percent species representation (calculated by documented species richness divided by predicted species richness) by counties with and without herbaria are shown in gray. The mean number of herbarium specimens for counties with and without herbaria are shown in black. Caps over each bar indicate the 95% confidence interval.

portant for future conservation efforts and provide a historical perspective for increasing or decreasing species richness in a given area. Accurate records of species richness prior to disturbance events will also allow for an accurate evaluation of the disturbance and appropriate conservation measures.

Finally, we address our third question: can predicted species richness be used to

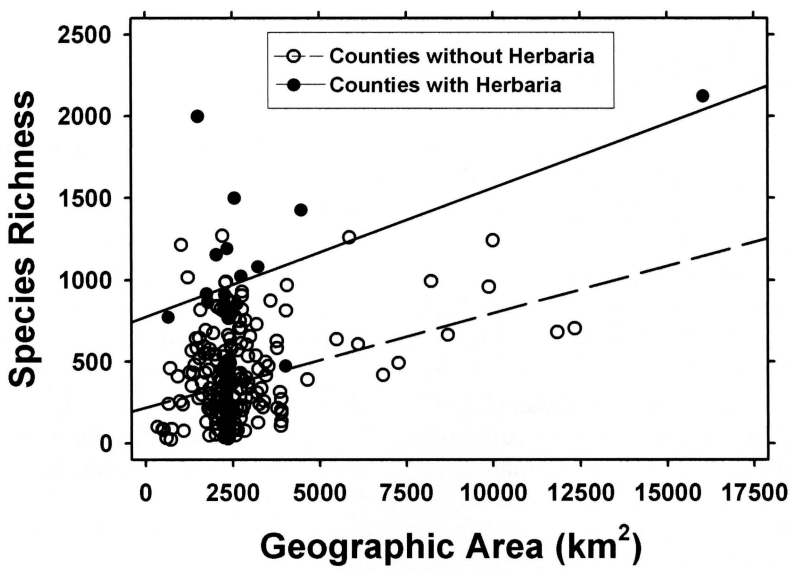

FIG. 6. Relationships between documented species richness and geographical area for counties with (closed circles and solid line) and without herbaria (open circles and dashed line).

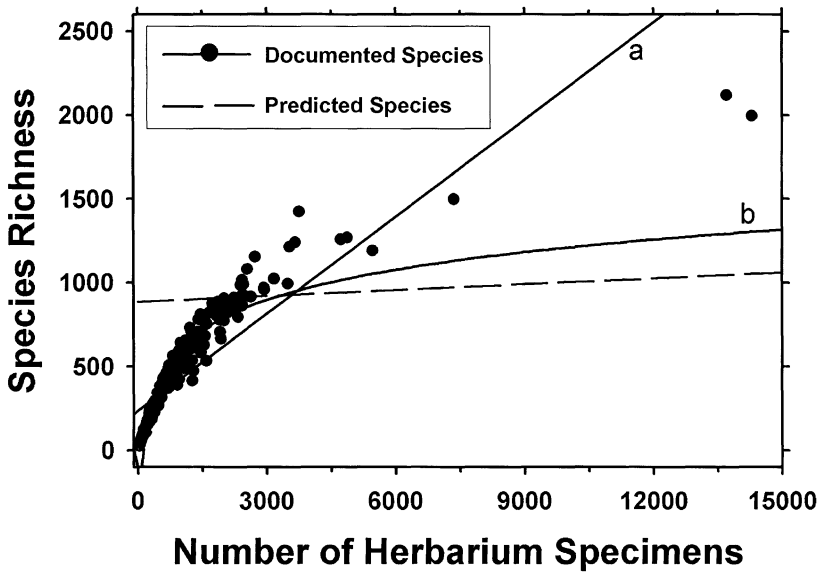

FIG. 7. Relationship between species richness and herbarium specimens. Dots = each of the 254 counties in Texas plotted for documented species richness and number of herbarium specimens. Solid line a = the linear regression between documented species richness and herbarium specimens. Solid line b = the logarithmic regression between documented species richness and herbarium specimens. Dashed line = the linear regression between predicted species richness and herbarium records. The point of intercept between the predicted species richness and documented species richness lines is an indication of optimal sampling effort. These lines intersect at approximately 3000 specimens after which collection effort (i.e., number of specimens) yields minimal increase in documented species richness.

determine sampling effort and a threshold number of samples needed to eliminate data gaps in herbarium collections? If sampling effort in a certain region nets little increase in documented species richness, sampling in different and new localities may prove more productive. The relation between the number of individual specimens sampled and the number of taxa documented was first suggested by Preston (1948) and has been referred to as the "collector's curve" (Colwell and Coddington, 1995). In addition, Miller and Wiegert (1989) utilized SAR to determine completeness in botanical exploration. However, their application was for only rare plants and relied heavily on hypothetical data.

Here we present a simple statistical method for potentially determining optimal collection effort for documented species richness. A linear regression ($F_{(1,252)}$ = 21.73; $p < 0.001$; $r^2 = 0.079$) was used to describe the relationship between the number of herbarium specimens and documented species richness (Fig. 7, curve a). Although this relationship is statistically significant, only 7.9% of the variation in documented species richness is explained by the number of herbarium specimens. This is because "collector's curves" follow a logarithmic relationship where the rate at which new taxa are documented decreases with the number of specimens collected. Thus, the likelihood of finding a new taxon during the first 1000 specimens collected is much greater than while collecting the second 1000 specimens. A logarithmic regression ($F_{(1,252)}$ = 1366.24; $p < 0.001$; r^2 = 0.844) demonstrates this relationship where over 84% of the variation in documented species richness is explained by the number of herbarium specimens (Fig. 7, curve b).

Figure 7 also illustrates that documented species richness intersects with predicted species richness near 3000 herbarium records. However, for the next 3000 specimens added to the collection, there is only a net gain of 100 new documented species above the predicted species richness. Our methodology suggests that the intercept between the linear and logarithmic regressions indicates optimal sampling effort (barring bias) and the threshold number of samples needed to be collected in order to reach predicted species richness values for a locality. In the statistical model presented here, 3000 samples should be collected to approximate predicted species richness for each Texas county. Once 3000 specimens are collected within a county, additional sampling effort beyond this point will result in minimal gain in additional documented species richness. This methodology may aid in eliminating collecting redundancy in over sampled counties and increase sampling efforts in under sampled counties.

Despite the obvious explanation of under collecting, several other contributing factors may lead to low documented species richness values per county. These include collector bias and the fact that not all collections are inventoried and data-based in the FTC. Collector bias is difficult to test and is an innate aspect of collecting. Incomplete data-basing however, reflects another growing example of data gaps and can be rectified through inter-herbaria cooperation and increased funding. Nevertheless those counties identified as having documented species values greater than or equal to predicted species richness values are indeed well collected counties.

We welcome the application and testing of this approach to other biological collections. The further development of such models may aid in identifying data gaps within collections and may benefit future collecting efforts for species inventory.

ACKNOWLEDGEMENTS

We thank Bruce Hoagland (University of Oklahoma), Michael W. Palmer (Oklahoma State University), Jake Schaefer (University Southern Mississippi) and Beryl Simpson for comments and improvements to the paper. This research was partially funded through USDA grant 321-20-A164.

LITERATURE CITED

Angermeier, P. L. and I. J. Schlosser. 1989. Species-area relationships for stream fishes. Ecology 70: 1450–62.

Arrhenius, O. 1921. Species and area. J. Ecol. 9: 95–99.

Brown, J. H. and M. V. Lomolino. 1998. *Biogeography.* Sunderland, Massachusetts: Sinauer Associates, Inc.

Buys, M. H., J. S. Maritz, C. Boucher and J. J. A. Van Der Walt. 1994. A model for species-area relationships in plant communities. J. Veg. Sci. 5: 63–66.

Carr, B. 2004. Vascular flora of Travis County, Texas. Austin: Austin Chapter of the Native Plant Society of Texas. http://www.npsot.org/Austin/TravisCountyFlora

Colwell, R. K. and J. A. Coddington. 1995. Estimating terrestrial biodiversity through extrapolation. Pp. 101–118 *in* D.L. Hawksworth, (ed.). *Biodiversity- measurement and estimation.* London: Chapman & Hall.

Cornelius, J. M. 1983. The Vascular Flora of Throckmorton County, Texas. M.S. Thesis, San Angelo, Texas: Angelo State University.

Correll, D. S. and M. C. Johnston. 1970. *Manual of the Vascular Plants of Texas.* Renner: Texas Research Foundation.

Covarrubias, R. and M. Elgueta. 1991. Relación especies-area de artropodos en cimas de montañas. Acta Ent. Chilena 16:151–60.

Denny, G. 2002. The vascular flora of The Nature Conservancy of Texas Love Creek Nature Preserve, Bandera County, Texas. M.S. Thesis, Austin, Texas: University of Texas.

Diamond, J. M. and E. Mayr. 1976. The species-area relation for birds of the Solomon Archipelago. Proc. Natl. Acad. Sci. U.S.A. 73: 262–266.

Eckhardt, R. F. 1975. The vascular flora of Tom Greene County, Texas. M.S. Thesis, San Angelo, Texas: Angelo State University.

Ertter, B. 2000. Our undiscovered heritage: past and future prospects for species-level botanical inventory. Madroño 47: 237–252.

Fleming, K. M., J. R. Singhurst and W. C. Holmes. 2002. Vascular flora of Big Lake Bottom Wildlife Management Area, Anderson County, Texas. Sida 20:355–371.

Freeman, C. C., C. A. Morse and J. P. Thurmond. 2003. The vascular flora of the Ogallala ecotone

on the Dempsey Divide, Roger Mills County, Oklahoma. Sida, 20:1217–1245.

Fridley, J. D., R. K. Peet, T. R. Wentworth and P. S. White. 2005. Connecting fine- and broad-scale patterns of species diversity: species-area relationships of southeastern U.S. flora. Ecology 86: 1172–1177.

Garner, P. M. B. 1979. The vascular flora of San Saba County, Texas. M.S. Thesis, San Angelo, Texas: Angelo State University.

Heywood, V. 2001. Floristics and monography—an uncertain future? Taxon 50: 361–380.

Isley, D. 1972. The disappearance. Taxon 21: 3–12.

Johnston, M. C. and J. K. Williams. 1995. Floristic survey of the Pantex plant site, Carson County, Texas. Amarillo: U.S. Department of Energy.

Judas, M. 1988. The species-area relationship of European Lumbricidae (Annelida, Oligochaeta). Oecologia **76**: 579–587.

MacArthur, R. H. and E. O. Wilson. 1967. *The Theory of Island Biogeography.* Monographs in Population Biology. Princeton: Princeton Univ.

MacRoberts, B. R, M. H. MacRoberts and L. E. Brown. 2002. Annotated checklist of the vascular flora of the Hickory Creek Unit of the Big Thicket National Preserve, Tyler County, Texas. Sida 20: 781–795.

McNeill, J. and W. J. Cody. 1978. Species-area relationships for vascular plants of some St. Lawrence River islands. Canad. Field-Naturalist 92: 10–18.

Miller, R. I. and R. G. Wiegert. 1989. Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. Ecology 70: 16–22.

Neill, A. K, and H. D. Wilson. 2001. The vascular flora of Madison County, Texas. Sida 19: 1083–1121.

Nesom, G. L, and L. E. Brown. 1998. Annotated checklist of the vascular plants of Walker, Montgomery, and San Jacinto counties, east Texas. Phytologia 84: 107–153.

Nixon, M.R. 1978. The Vascular Flora of Coleman County, Texas. M.S. Thesis, San Angelo, Texas: Angelo State University.

Palmer, M. W., P. G. Earls, B. W. Hoagland, P. S. White and T. Wohlgemuth. 2002. Quantitative tools for perfecting species lists. Environmetrics 13: 121–137.

Peterson, C. D. and L. E. Brown. 1983. *Vascular Flora of the Little Thicket Nature Sanctuary San Jacinto County, Texas.* Houston: Outdoor Nature Club.

Phillips, J. W. 1997. An annotated checklist of the plants of Lake Meredith Recreation Area, Carson County, Texas. Panhandle Archeological Society 7: 1–52.

Prather, L. A., O. Alvarez-Fuentes, M. H. Mayfield and C. J. Ferguson. 2004. The decline of plant collecting in the United States: a threat to the infrastructure of biodiversity studies. Syst. Bot. 29: 15–28.

Preston, F. W. 1948. The commonness and rarity of species. Ecology 29: 254–283.

Schoener, T. W. 1976. The species-area relationship within archipelagoes: Models and evidence from island birds. Proceedings of the XVI International Ornithological Congress 6: 629–642.

Turner, B. L. 1998. Plant systematics: beginnings and endings. Aliso 17: 189–200.

Turner, B. L., H. Nicholls, G. Denny and O. Doron. 2003. *Atlas of the Vascular Plants of Texas,* Sida, Bot. Misc. 24. Fort Worth: Botanical Research Institute of Texas.

Whisenant, S. G. 1982. The vascular flora of McCulloch County, Texas, USA. Texas J. Sci. 33:197–220.