SPECIAL FEATURE

The Statistics of Rarity[1]

Ecologists deal with rarity in many guises—species can be rare, particular interactions may be uncommon, and catastrophic events that reshape landscapes are, by definition, infrequent. Although ecologists often seek out abundant species or events for their investigations out of convenience, rare species are often of special central concern to conservation biologists, reserve managers, and legislators, and historical legacies of rare events are pervasive in ecosystems. Statistical analysis and modeling of rare events is often necessary, but as MacKenzie et al. point out in their contribution to this Special Feature, ''rare species [or events] are simultaneously the species for which strong inferences about state variables and vital rates are most needed and the species for which such information is most difficult to obtain.'' The standard set of statistical tools used by the majority of ecologists are difficult or inappropriate to use when analyzing rare species or events, either because assumptions such as normality or homoscedasticity do not hold, or because the required sample sizes are impossibly large. The overall goal of this Special Feature is to present a cross-section of techniques for sampling, quantifying, and modeling rarity.

Methods for analyzing data on rare species and events come from a variety of disciplines and were originally designed for addressing specific questions. Most of these questions, including problems related to quality control in manufacturing, frequency of flooding, and econometrics, are unrelated to ecological questions, but the mathematical and statistical tools transcend disciplinary boundaries. As a result, most of the papers in this Special Feature represent collaborations between ecologists and statisticians and reflect the growing need for interdisciplinary cooperation to effectively address complex ecological questions.

The first step in a study of rarity is collecting the data. Sampling rare species is challenging precisely because they are hard to find. The first three papers discuss sampling methods for rare species. The papers by Edwards et al. and Philippi focus on increasing the efficiency of sampling when the focal taxon is rare. Edwards et al. present methods for stratifying the sampling effort by modeling occurrences of common species known to be associated with the rare species of interest. Their interest is simply in determining whether or not a species occurs at a particular site. In contrast, Philippi addresses the problem of designing a sampling regime to more accurately estimate the abundance of a rare species. Measures of abundance are required to determine the legal status of a species (e.g., secure, threatened, or endangered), and the adaptive cluster sampling described by Philippi is more efficient than simple random sampling for estimating the abundance of sessile species that tend to be clumped. MacKenzie et al. discuss sampling and modeling methods for estimating the occupancy of a given site or patch by a mobile species. Occupancy can be used as a surrogate for abundance, especially in capture–recapture studies of mobile animals. Of particular importance is improvement of the accuracy of occupancy estimates when the probability of actually observing the species of interest is less than 1. The generality of all three of these approaches is reflected in the diversity of example organisms: lichens (Edwards et al.), herbaceous plants (Philippi), and giant wetas, gaurs, and salamanders (MacKenzie et al.).

Deriving precise estimates of rare events and expressing the uncertainty surrounding these estimates is challenging because the study of rare events often involves small sample sizes that provide little confidence in the estimates. Just as the papers by Edwards et al. and MacKenzie et al. take advantage of auxiliary data to improve sampling efficiency and estimates of occupancy, the paper by Dixon et al. uses auxiliary data to increase the precision (decrease the variance or uncertainty) in estimates of the frequency of rare events. Dixon et al. apply simple Bayesian methods with informed prior probability distributions, stratified data, regression with continuous

---

[1] Reprints of this 85-page Special Feature are available for $12.75 each, either as pdf files or as hard copy. Prepayment is required. Order reprints from the Ecological Society of America, Attention: Reprint Department, 1707 H Street, N.W., Suite 400, Washington, DC 20006.

covariates, and aggregated data from additional samples to the question of how frequently a sit-and-wait predator, the carnivorous plant *Darlingtonia californica*, captures prey. Large-scale disturbances are also infrequent events, and Katz et al. illustrate techniques for distinguishing the ‘‘signal’’ of extreme events from the ‘‘noise’’ of temporal variability. These techniques derive from the large body of statistical work devoted to the study of extremes, which were introduced to ecologists by Steve Gaines and Mark Denny in their 1993 paper, ‘‘The largest, smallest, highest, lowest, longest, and shortest: extremes in ecology’’ (*Ecology* **74**:1677–1692). Gaines and Denny focused on ‘‘light-tailed’’ distributions, in which the frequency of extreme events decreases at a relatively rapid rate, but many uncommon disturbances, including fires, floods, and large hurricanes, have ‘‘heavy-tailed’’ distributions, in which the frequency of extreme events decreases at a relatively slow rate. In their analysis of sedimentation rates at Nicolay Bay, Canada over ~500 years and in Chesapeake Bay since 1800, Katz et al. illustrate the utility of the statistics of extremes and show that these indicators of hydrological, climatological, and anthropogenic disturbances are heavy-tailed. The implication of their results is that some apparently rare events are less ‘‘rare’’ than would appear on first glance.

Armed with reasonable samples and reasonably precise parameter estimates, it is possible to model the effects of biotic processes or environmental drivers on the rare species or event of interest. Katz et al. show that the Nicolay Bay sedimentation record reflects only hydrological perturbations, whereas the Chesapeake Bay record reflects both climatic and anthropogenic disturbances. In a return to rare species, Cunningham and Lindenmayer use generalized linear models with link functions designed to account for excess zeros in presence/absence or abundance data to determine predictors of occurrence of Leadbeater’s possum in Australian eucalypt forests, and predictors of abundance of nesting frigatebirds and boobies in the Coral Sea.

The last two papers address issues of major importance to conservation planners and managers. Mao and Colwell discuss how to accurately estimate the species richness of a given assemblage when some (or many) of the species are rare. Two questions are of paramount interest. First, can rare species that go undetected in inventories be accounted for statistically? Second, when is an inventory complete? Their analysis, based on empirical data for breeding birds, seeds, and beetles, as well as from simulations, suggests that while it is relatively straightforward to estimate a lower bound for the number of species present in an assemblage, estimating the actual number of species present remains a challenging statistical problem. Managers cannot always wait for better data, however, and Doak et al. discuss how to build demographic models for rare species when few data are available. The counterintuitive conclusion from their simulations is that, for rare species, simple deterministic models of population growth and population viability analyses are preferable to data-hungry stochastic models, even though the former have known biases.

Many of the methods presented in these papers will be unfamiliar to ecologists raised on a steady diet of ANOVA, regression, and contingency tables. To encourage the use of these tools, the authors have published the statistical code and data sets accompanying their papers in *Ecological Archives*, or have made sophisticated packages freely available on their web sites. We hope that these methods will be used and that they will continue to evolve through productive collaborations between ecologists, statisticians, conservation planners, and managers.

—AARON M. ELLISON
Associate Editor-in-Chief
—ANURAG A. AGRAWAL
Special Features Editor

Key words: adaptive cluster sampling; Bayesian inference; detection probability; extremes; general linear models; maximum likelihood; parameter estimation; Poisson distribution; population viability analysis; precision; rarity; sampling.

Ecology, 86(5), 2005, pp. 1081–1090

MODEL-BASED STRATIFICATIONS FOR ENHANCING THE DETECTION OF RARE ECOLOGICAL EVENTS

THOMAS C. EDWARDS, JR.,[1,6] D. RICHARD CUTLER,[2] NIKLAUS E. ZIMMERMANN,[3] LINDA GEISER,[4]
AND JIM ALEGRIA[5]

[1]*U. S. Geological Survey Biological Resources Division, Utah Cooperative Fish and Wildlife Research Unit,*
Utah State University, Logan, Utah 84322-5290 USA
[2]*Department of Mathematics and Statistics, Utah State University, Logan, Utah 84322-3900 USA*
[3]*Department of Landscape Research, Swiss Federal Research Institute WSL, Zuercherstrasse 111,*
CH-8903 Birmensdorf, Switzerland
[4]*USDA Forest Service, Siuslaw National Forest, P.O. Box 1148, Corvallis, Oregon 97339 USA*
[5]*Bureau of Land Management Oregon State Office, 333 SW First Avenue, Portland, Oregon 97204 USA*

Abstract. A common concern when designing surveys for rare species is ensuring sufficient detections for analytical purposes, such as estimating frequency on the landscape or modeling habitat relationships. Strict design-based approaches provide the least biased estimates but often result in low detection rates of rare species. Here, we demonstrate how model-based stratification can improve the probability of detecting five rare epiphytic macrolichens (*Nephroma laevigatum*, *N. occultum*, *N. parile*, *Lobaria scrobiculataa*, and *Psuedocyphelaria rainierensis*) in the Pacific Northwest. We constructed classification tree models for four more common lichens (*L. oregana*, *L. pulmonaria*, *P. anomala*, and *P. anthraspis*) that are associated with the rare species, then used the models to generate strata for sampling for the five lichen species considered rare. The classification tree models were developed using topographic and bio-climatic variables hypothesized to have direct relationships to the presence of the modeled lichen species. When the expected detection rates using the model-based stratification approach was tested on an independent data set, it resulted in two- to fivefold gains in detection compared to the observed detection rates for four of the five tested rare species.

Key words: classification trees; epiphytic lichens; model-based stratification; Northwest Forest Plan; rare lichens; sampling; species rarity.

INTRODUCTION

Ecologists and conservationists often deal with rare or uncommon species. Reasons for interest in these species range from curiosity about the underlying ecological relationships that determine rarity (Goerck 1997) to the need for information about the species so that proper management and conservation strategies can be developed (Sheldon 1988, Maina and Howe 2000). A common concern when designing surveys for rare species is ensuring sufficient detections for analytical purposes (Green and Young 1993, Edwards et al. 2004), be they for estimations of frequency on landscapes (Alexander et al. 1997), or for use in explanatory or predictive models of suitable habitat or spatial distribution (Hill and Keddy 1992, Wiser et al. 1998).

One specific example where the species of interest appear to be ecologically rare, and for which there exists a strong need for estimates of frequencies for conservation purposes, is in the U.S. Pacific Northwest and a subset of that area defined by the Pacific Northwest Forest Management Plan (hereafter NWFP). The NWFP was a direct consequence of a court-mandated requirement to develop a scientifically credible conservation strategy for the Northern Spotted Owl *Strix occidentalis* (see Thomas et al. 1990, Murphy and Noon 1992) while also maintaining biodiversity on public lands within the area of the Plan. The Record of Decision leading to the Plan (ROD; U.S. Department of Agriculture, Forest Service, and U.S. Department of Interior, Bureau of Land Management 1994), and subsequent amendments (U.S. Department of Agriculture, Forest Service, and U.S. Department of Interior, Bureau of Land Management 2001), identify more than 350 species of concern in the Plan area that required management plans. These species included fungi, lichens, bryophytes, terrestrial mollusks, arthropods, a small number of vascular plants, and several terrestrial vertebrates. Collectively, these species and species groups were referred to as Survey and Manage species. All were hypothesized to have associations with old forests, and concern existed over the possible impacts of forest management on the persistence of these species.

One immediate realization was that insufficient information existed to ascertain the impacts of conservation decisions on Survey and Manage species. Most

---

Manuscript received 1 April 2004; revised 29 June 2004; accepted 1 July 2004; final version received 13 September 2004. Corresponding Editor: A. M. Ellison. For reprints of this Special Feature, see footnote 1, p. 1079.

[6] E-mail: tce@nr.usu.edu

of the species were known from relatively few sites, and their spatial distributions were unclear as well. Even less was known about basic life history attributes. In one sense, the Survey and Manage species are rare events in time and space. They are often associated with microsite characteristics that are themselves rare and patchy in time and space (lichens [McCune 1993, Esseen et al. 1996]). Many also exhibit irruptive population behavior, appearing only for brief periods or only once or twice over extended periods (fungi [Molina et al. 2001]), while in other cases, large distances may exist between individuals or populations (fungi [Dreisbach et al. 2002]). These characteristics make it difficult to design statistically valid sampling schemes, particularly when little is known about the species' distribution or habitat requirements that could be used to focus sampling efforts.

Sound study designs and analytical techniques for estimating species frequencies are well represented in the scientific literature (see Schreuder et al. 1993, Olsen and Schreuder 1997) and have been successfully applied to diverse resource issues (forest resources [Gregoire et al. 1995, Shiver and Borders 1996], animals [Sauer and Droege 1990, James et al. 1996], aquatic systems [Stow et al. 1998]). However, many of these approaches rely on design-based sampling efforts. While design-based approaches clearly provide the least biased estimates, they unfortunately are often of insufficient sampling intensity to ensure enough observations of rare species. Methods such as adaptive cluster sampling (Lohr 1999, Cutler et al. 2002, Philippi 2005) can increase sampling efficiencies, but themselves can be limited by the information needed for cluster starting points, or the need for immediate identification of the sample unit of interest so that the cluster can be identified and additional samples collected or measured.

One way to increase detection probabilities for rare species is to use presence–absence information from other, more abundant but related species to generate strata for sampling for the rare species. Strata could be derived from any number of available statistical models capable of relating a binary response to a set of predictor variables, such as generalized additive models, logistic regression, and classification trees (see Hoeting et al. 2000). Here, we examine the efficacy of a model-based stratification approach based on classification trees to enhance detection probabilities for rare epiphytic macrolichens in the Pacific Northwest. We fit classification trees for four common lichen species, using selected topographic and bio-climatic variables all hypothesized to have direct relationships to the presence of the common lichen species. The fitted trees were then used to define sampling strata for associated, rare lichen species. Specifically, we test whether model-based strata developed from more common, associated species can improve the detection probabilities of rare species. We next evaluate the efficacy of the stratification process for predicting occurrences of rare species on an independent data set within the same geographic area, comparing the gains in the model-based detection probabilities against those realized from a systematic random sample design.

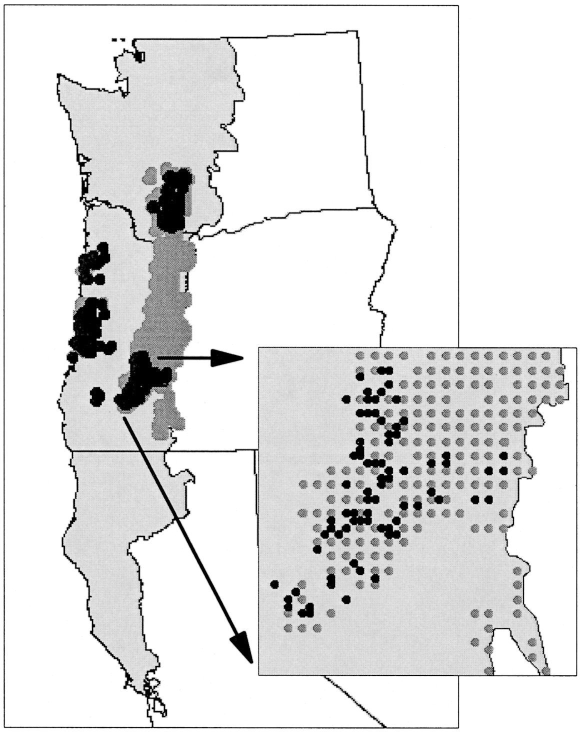

FIG. 1. Generalized location of the LAQ (lichen air quality) model building plots (medium gray) and PILOT validation plots (black) in relation to the area of the Northwest Forest Plan (light gray). The figure inset is a representation of the validation PILOT points (black) and LAQ model building points (medium gray).

METHODS

Study area and species

Data used in our analyses were collected in the NWFP area of the U.S. Pacific Northwest (Fig. 1). Two subset study areas within the plan boundary were used in our analyses. The first (hereafter LAQ) was used for model building purposes and included sampling sites on seven national forests: the Umpqua, Willamette, Mt. Hood, Deshutes, Gifford-Pinchot, Siuslaw, and Winema (Fig. 1). Epiphytic macrolichen species were surveyed from 1993 to 2000 as part of a larger effort using lichens to evaluate air quality in the Pacific Northwest (Geiser 2004). The second survey (hereafter PILOT) was conducted to provide data on rare, old-forest-associated species to the NWFP Survey and Manage program. These data served as an independent assessment tool for the models. The PILOT surveys included forests in three regions: the southern Washington Cascades (Gifford Pinchot National Forest); the Oregon Coast Range (Siuslaw National Forest and portions of

Table 1. Numbers (percentage) of detections of rare species that are co-located with more common species in the LAQ (lichen air quality) study area within the range of the Pacific Northwest Forest Plan.

	Common species†			
Rare species	*Lobaria oregana*	*Lobaria pulmonaria*	*Pseudocyphellaria anomala*	*Pseudocyphellaria anthraspis*
Lobaria scrobiculata		37/47 (78.7)	39/47 (83.0)	
Nephroma laevigatum		24/25 (96.0)	19/25 (76.0)	19/25 (76.0)
Nephroma occultum		10/13 (76.9)	13/13 (100.0)	
Nephroma parile		24/31 (77.4)	27/31 (87.1)	
Pseudocyphellaria rainierensis	8/9 (88.9)	8/9 (88.9)	7/9 (77.8)	

Note: Empty cells indicate species that were never co-located.

† Numerator = number of detections of the rare species that are co-located with the common species; denominator = total number of occurrences of the rare species in the LAQ study area.

Coos Bay, Eugene, and Salem Bureau of Land Management lands); and the Umpqua Basin (Umpqua National Forest, Roseburg BLM [Bureau of Land Management], and portions of the Willamette National Forest, and Eugene and Coos Bay BLM).

All sample sites in both the LAQ and PILOT study areas were surveyed at least once and samples of each detected epiphytic macrolichen were collected for subsequent identification. Crustose (e.g., *Calicium* spp. and *Chaenotheca* spp.) and ground-dwelling species (e.g., *Cladonia norvegica*, *Peltigera pacifica*) were not surveyed at all sites and were excluded from analysis. A total of 49 lichen species having designations ranging from rare to common were considered. The four most abundant species in the LAQ surveys that were also sampled in the PILOT surveys, *Lobaria oregana*, *L. pulmonaria*, *Psuedocyphelaria anomala*, and *P. anthraspis*, were used for modeling purposes. Each of these species was detected on >120 of the 840 LAQ sites, providing sufficient sample sizes for developing model-based strata. Cross-tabulations of these species with other, rarer species detected in both the LAQ and PILOT surveys revealed substantial overlap of five specific rare species with the four identified more common species (Table 1). Accordingly, we fit classification trees for each of these four common species, with a view to using the resulting predictive model for *L. pulmonaria* to predict the presences of *Nephroma laevigatum*, *L. scrobiculataa*, and *P. rainierensis*; the models for *L. oregana* and *P. anthraspis* to predict presences of *P. rainierensis*; and the model for *P. anomala* to predict presences for *N. occultum*, *L. scrobiculataa*, and *N. parile*.

Data structure and characteristics

Data in both study areas were collected on the current vegetation survey plots (CVS), a systematic grid overlaid on all Forest Service and BLM lands in the Pacific Northwest. Its principal application is the generation of estimates of forest resources (see Max et al. 1996). A total of 840 CVS plots were sampled in the LAQ, and an additional 300 sites in the PILOT study area. Sample sites were apportioned equally among the three areas comprising the PILOT study area. Presence and absence of each lichen species was recorded on a 1-acre (0.4-ha) plot centered on the central (#1) subplot on each CVS site for the LAQ survey (details in Edwards et al. 2004). Plot size for the PILOT surveys was 0.5 acres (0.2 ha). If the purpose of our study was to compare the estimated percent occupancy rates from the LAQ and PILOT surveys, the difference in the size of the sample units would be a concern given that larger plots will have higher probabilities of occupancy. However, the purpose of our analyses is to use the LAQ data to fit a model that can be used to predict PILOT survey sites that are more or less likely to be occupied by the rare lichens, with a view to developing a stratification for the PILOT sampling. For this application, it does not matter if the PILOT plot size is the same, larger, or smaller than the LAQ plot size.

All plot locations were intersected with spatially explicit maps of topographic and bio-climatic variables (Table 2) in a geographic information system (GIS). The selected topographic and bio-climatic variables were all hypothesized to have direct relationships to the presence of the modeled lichen species. Ninety-meter resolution topographic variables (slope, aspect and elevation) were obtained by resampling the 30-m resolution national elevation data set (NED) (Gesch et al. 2002). Bio-climatic variables were derived from the DAYMET 1-km daily gridded weather surfaces that have been reduced to 18-yr monthly and yearly climatological summaries (1981–1998; data is from DAYMET U.S. Data Center for Daily Surface Weather Data and Climatological Summaries, *available online*).[7]

DAYMET generates daily surfaces of temperature, precipitation, humidity, and radiation over large areas of complex terrain (Thornton et al. 1997, Thornton and Running 1999). It uses digital topographic models and observations of precipitation, maximum and minimum temperature from ground-based meteorological stations to generate other bio-climatic variables. Derivations for the modeled variables follow the DAYMET methodology described by Thornton et al. (1997), applied to monthly averages. Only the variable potential evapotranspiration (ETPJ), which is not part of the DAYMET

[7] ⟨http://www.daymet.org⟩

TABLE 2. Topographic and bio-climatic variables used to model the probability of presence for four common lichen species in the LAQ study area of the Pacific Northwest Forest Plan.

Variable type/ name	Description	Units
Topographic		
SLPE	percent slope	percent, 0–90
ASPE	aspect	degrees, 1–360
ELEV	elevation	m
Bio-climatic		
PREC	precipitation	cm
TMAX	maximum temperature	°C
TMIN	minimum temperature	°C
TAVE	monthly average temperature	°C
ETPJ	potential evapotranspiration	mm
RELH	relative humidity	percent
VPAM	ambient vapor pressure	Pa
VPSA	saturated vapor pressure	Pa
SFMM	monthly potential global radiation	kJ
TDAY	monthly average daytime temperature	°C
MIND	monthly moisture index	cm

Note: Classification tree models based on these variables were used to generate sampling strata for rare species.

procedures, was derived differently (see Jensen and Haise 1963). We down-scaled the 1-km DAYMET variables to 90-m resolution by first generating moving window regressions between the monthly averages of PREC, TMIN, and TMAX, and the topographic models, and then applying the regression parameters to the center cell of each 1-km window position. The regression parameters (i.e., lapse rates and intercept) were next interpolated to a 90-m resolution using inverse distance weighed interpolations, thus generating monthly 90-m resolution maps for each bio-climatic variable.

Preliminary analyses showed that correlations among the monthly values for the 11 sets of bio-climatic predictor variables were extremely high. Such extreme collinearity has implications for modeling. For example, two variables that essentially contain the same information can be selected in models for different species, implying differences in the models that are not real. The same phenomenon can occur even at different stages of the same classification tree model. To address the issue of collinearity, a principal components analysis was carried out on each of the 11 sets of monthly bio-climatic predictors.

In each case, the first principal component was an average of the 12 monthly measurements, while the second principal component was a contrast of values for six so-called summer months (April–September) to the six so-called winter months (October–March). For each set of 12 monthly variables, these two principal components explained over 95% of the variability, and in most cases the first two principal components explained over 99% of the variability in the sets of variables. Accordingly, for each set of monthly bio-climatic predictors we defined two new variables: (1) the average of the 12 monthly variables; and (2) the difference between the sum of the summer monthly values and the winter monthly values, divided by 12. Hereafter we use the variable suffix ''A'' to denote the average of the 12 monthly measurements, and the suffix ''D'' to denote the difference derived variable. Thus, TMINA is the average minimum temperature for the 12 months and PRECD is the difference between summer and winter precipitation.

Statistical modeling and assessment

We used classification trees (Breiman et al. 1984) to relate the LAQ lichen presences to the modeled bio-climatic and topographic predictor variables. Classification trees have several advantages over other discrimination techniques (e.g., GLMs, GAMs), principal among these being the ability to deal with nonlinearities in the predictor variables and assess interactions, and ease of interpretation. For the most part we followed the approach of De'ath and Fabricius (2000), but instead of allowing the tree size to be determined completely by cross-validation and the 1-SE rule, we set the complexity parameter at 0.02 instead of using the default value of 0.01. This relaxation in complexity results in smaller, easier to interpret classification trees that have cross-validated error rates at least as good as the larger trees obtained from the default value. The classification trees were fit using the rpart library of functions in the R statistical package (Ihaka and Gentleman 1996 and online resources).[8]

We next fit the LAQ-based classification trees to our spatially explicit predictors within the GIS and modeled the probability of presence of each the four common lichen species across the entirety of the NWFP region. This was accomplished by applying the classification tree to the GIS layers and generating new maps that portrayed, for each 90-m pixel, the proba-

---

[8] ⟨http://www.r-project.org/⟩

bility of each of the four common lichen species being present at that particular pixel. Proportions of presences from sample plots in the PILOT study area were then examined against predicted probability for that plot, allowing us to link the presence or absence of each of the four common species to an estimate of the probability of presence. Ideal concordance between the LAQ models and the common species in the PILOT area would be represented as a 45° line on a scatter plot of observed probabilities of presence against predicted probabilities of presence (see Ferrier and Watson 1997: Fig. 6.2). Data and a description of the process used to estimate expected values for each rare species under different sampling probabilities are found in Appendices A–D.

Three measures of model error were used to evaluate model performance. These measures were: (1) a resubstitution (model) error rate, calculated using the LAQ data and hence providing an indication of how well the classification tree fit the modeling data; (2) a 10-fold cross-validation (Manly 1997) estimate of error obtained by using 1000 subsamples of the original LAQ data; and (3) a prediction error rate for the PILOT surveys, based on a probability of presence threshold of $P > 0.5$.

Under the Survey and Manage program, actual detections as opposed to region-wide estimates were of paramount importance. To reflect this need and desire for additional observed occurrences of the rare species, we chose to allocate sample units proportionally to the estimated probability of occurrence of the associated common species in each stratum. Different researchers with different objectives could use a different objective function for assigning sample units to the strata and still use all the other methodology described in this section.

Results

Although the classification tree models are purely predictive in nature, it is interesting to note that the first split in each classification tree is on an average temperature index (Fig. 2) and that precipitation only occurs in one of the four classification trees (*L. oregana*). No topographic variables entered any of the models as significant predictors. Not unexpectedly, the LAQ model error estimates were lower than the LAQ cross-validated error (Table 3), ranging from a low of 10.2% to a high of 18.3%. Cross-validation error ranged from approximately 13% to 22%. Prediction errors were 15% to 27%, indicating that the LAQ models had, on average, about a one in four chance of being wrong. The cross-validated error rates were much closer to the observed error rates on the PILOT data, and likely better reflect the true error of the models.

LAQ models for the four common species, when evaluated as plots of the observed occurrences versus the predicted probabilities of occurrence, indicate all models had substantial predictive power (Fig. 3). Models for *L. pulmonaria*, *L. oregana*, and *P. anomala* are best, with the estimated values and their 95% confidence intervals encompassing the 45° line. The model for *P. anthraspis* is not as good, with observed values tailing off as the predicted probability of occurrence increases, indicating on over-prediction of potential locations for these species.

To evaluate the efficacy of the LAQ classification trees in determining a stratification for sampling of the rare, associated species, we calculated the estimated number of detections expected for each of the rare species had the predicted probabilities of presence obtained from the LAQ classification trees been used for stratification purposes. For almost all combinations of rare species and models of the more abundant species, the expected number of detections would have been higher under our model-based stratification than under the systematic design that characterized the PILOT surveys (Table 4). Gains in sampling efficiencies for four of the five species (*L. scrobiculata*, *N. laevigatum*, *N. occultum*, and *N. parile*) ranged from 1.2 to 5.0-fold, while results showed no difference for *P. rainierensis*. These increases in likely detections indicate that use of the LAQ models as the basis for developing sampling strata substantially increase the likelihood of detection of the rare lichens.

Discussion

The models of the four common lichen species used to develop the prediction strata for the rare lichens all showed similar error rates when evaluated on the LAQ data set. However, when tested on the independent PILOT data set, the *P. anthraspis* and *P. anomala* models performed best, followed by *L. pulmonaria*. The *L. oregana* model produced the highest error rate. Overall, the models were successful in generating clear improvements in estimated, or expected, detection rates for four of the five rare lichen species (Table 4). Only the model of *L. oregana* applied to *P. rainierensis* did not perform well. The most likely explanation for the low predictive extrapolation of this last model is the lack of spatially explicit depictions of stand age for use in building the associative models. In this case, *L. oregana* and *P. rainierensis* are both associated with old forest, but *rainierensis* is more likely to be found in 300+-yr-old stands. Such subtle differences in ecological relationships like this could easily confound model-based stratifications, increasing the likelihood of prediction error. This observation reinforces the importance of having a solid understanding of species ecology when model building, as well as the importance of testing any predictive model with an independent test data.

Existing knowledge on lichen ecology suggests that prediction error rates may be reduced substantially by including additional information about forest stand characteristics (e.g., stand age, size classes, hardwood shrubs), and local moisture content (McCune 1993, Sil-

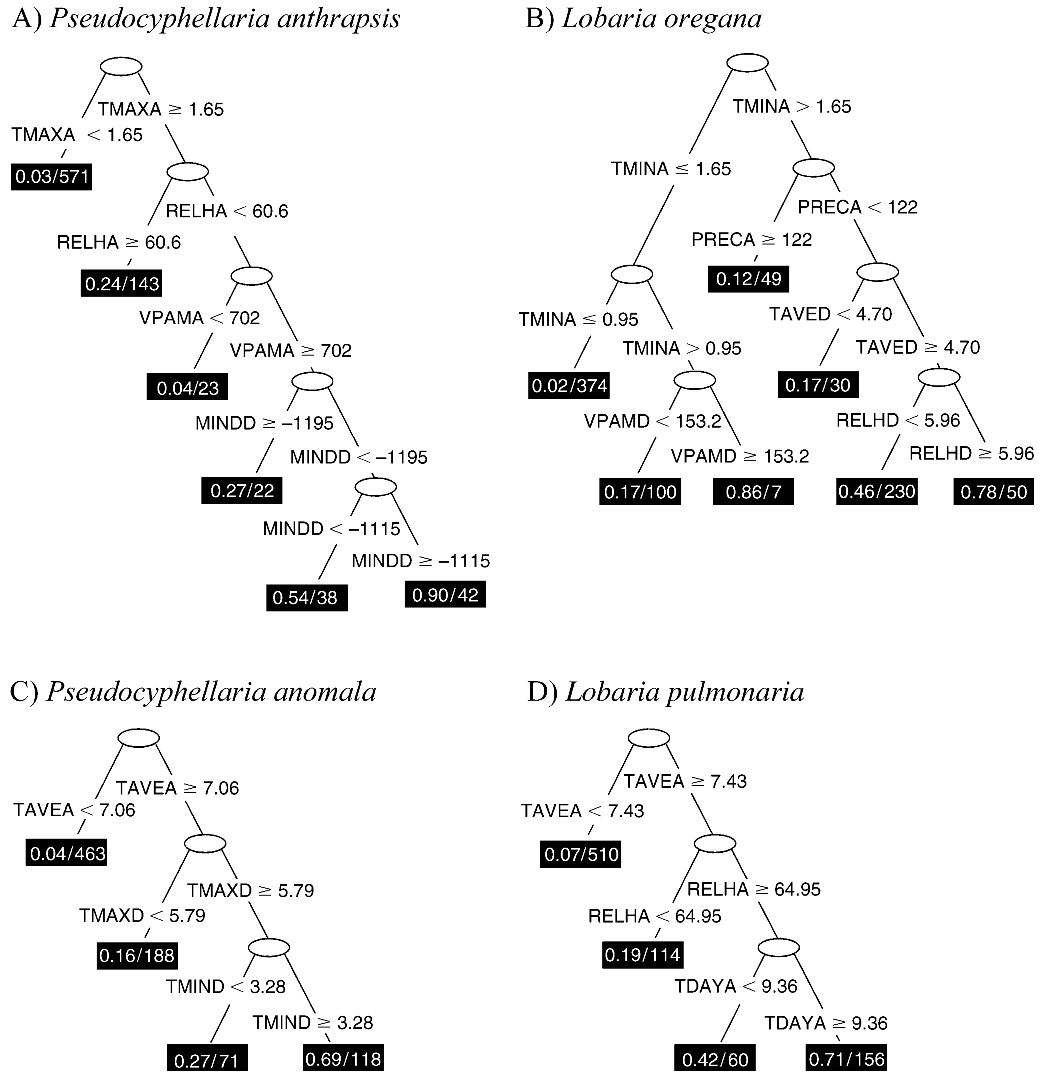

FIG. 2. Classification trees of the four modeled common lichen species used to develop prediction strata for the five rare lichen species in the area of the Northwest Forest Plan. See Table 2 for variable descriptions.

lett and Goslin 1999, Rosso et al. 2000, Peterson and McCune 2001) into models to be used for stratification. However, these types of variables were not available in spatially explicit formats for the entire study region, and hence could not be used for the purposes of stratification in advance of sampling. Spatially explicit depictions of these kinds of variables are simply unavailable in the vast majority of predictive modeling efforts, often forcing models to be built using surrogates (e.g., topography) that may or may not have rel-

TABLE 3. Measures of error (%) for classification tree models built for four species of lichen in the LAQ study area, and assessed using independent data collected in the PILOT study area, Pacific Northwest Forest Plan.

	LAQ stratification models		PILOT assessment
Species	Model error	Cross-validation error	Prediction error
Lobaria oregana	18.3	22.5	27.3
Lobaria pulmonaria	15.2	18.3	19.3
Pseudocyphellaria anomala	12.6	15.4	15.0
Pseudocyphellaria anthraspis	10.2	13.2	15.3

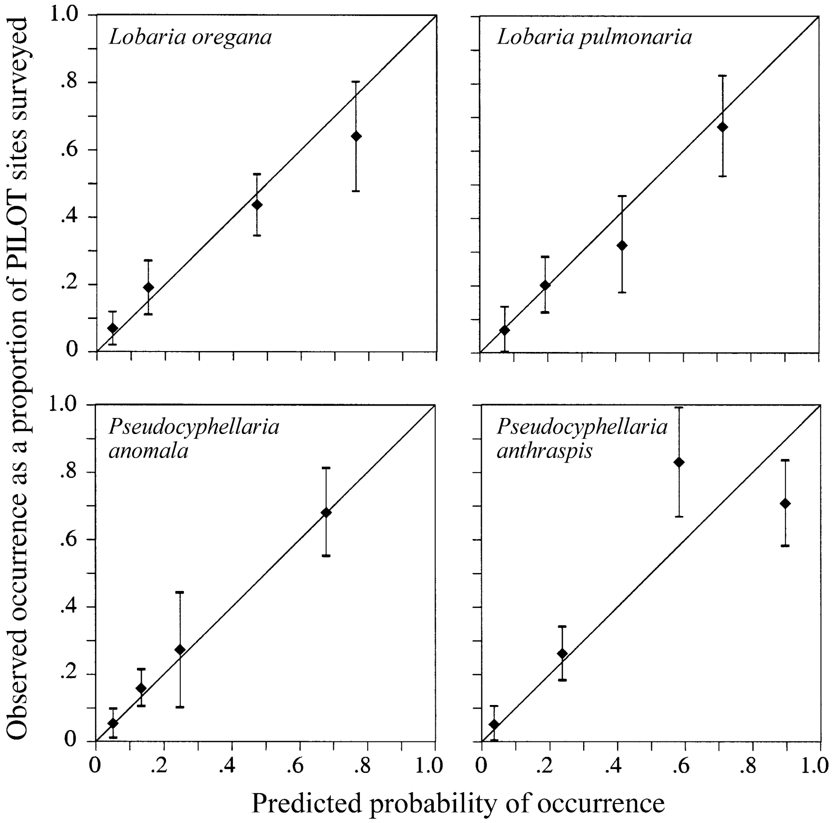

FIG. 3. Plot of the cumulative observed proportion of occurrences (plus 95% CI) against the probability of occurrence for four species of common lichen in the area of the Northwest Forest Plan. Predicted observations and 95% confidence intervals that encompass the expected 1:1 fit indicate models with high predictive capability.

evance to the distribution of the species being modeled. This shortcoming—the lack of strong linkage between species presence and ecologically relevant variables—is a common constraint faced by ecologists and cannot be satisfactorily resolved until spatially explicit predictions of such variables are more readily available (see Frescino et al. 2001).

There are several reasons why a species is or remains rare in a certain area, some of which can be captured by predictive models, while others cannot (see Gaston 1994 for an overview). Key patterns associated with rarity are (1) the main distribution of the species lies outside of the study area, ostensibly because the latter does not provide much suitable habitat (locally rare, otherwise abundant); (2) the species has a patchy, scattered distribution throughout its range, and populations are smaller and likely limited by competition and/or population constraints (generally and locally rare); (3) the species has a patchy distribution, but population size can be large at known sites (locally abundant, generally rare). Following the classification by Hanski (1982) and Collins et al. (1993), the first group would be classified as an ''urban'' type, the second as ''rural'' type, while the last group can be labeled as ''satellite'' type.

Accurate, predictive models for urban species that are dominant within their realized range are easier to construct than models for rural or satellite species. So-called urban species may not occur in many locations within the study area, but if the environment is suitable, they usually are present. Satellite type species may be restricted to few locations because there are actually only few suitable sites available. Often, however, dispersal limitations and other life history constraints limit this type from persisting, regenerating, or migrating to other suitable habitats. Predictive models may not easily predict the presences of satellite type species successfully, but the absences can be more accurately portrayed. This is because such species usually show a well defined ecological niche even though they do not occur on every suitable location. The most difficult group of species to model would be the rural type.

Because our approach uses associative models predicting rare from more common species, care must be exercised to ensure that the potential types of error are understood. For example, it is likely that rare species would have different niches than common species, by definition alone (Gaston 1994). In this case, the potential for commission error exists if the niche of the rare species is a subset of the more common species used

TABLE 4. Number of detections in the PILOT study area using simple random sampling (numerator), number of detections using the model-based stratifications (denominator), and model efficiency (value in parentheses).

	Common species			
Rare species	*Lobaria oregana*	*Lobaria pulmonaria*	*Pseudocyphellaria anomala*	*Pseudocyphellaria anthraspis*
Lobaria scrobiculata		13/26 (2.0)	13/36 (2.8)	
Nephroma laevigatum		19/23 (1.2)	19/48 (2.5)	19/60 (3.2)
Nephroma occultum		1/5 (5.0)	1/5 (5.0)	
Nephroma parile		7/14 (2.0)	7/16 (2.3)	
Pseudocyphellaria rainierensis	2/1 (0.5)	2/5 (2.5)	2/5 (2.5)	

Note: Model efficiency is estimated as the ratio of detections from the model-based stratification to those detected in the PILOT study area, Pacific Northwest Forest Plan. Empty cells indicate species that were never co-located.

to develop the stratification model. Although it is less likely that rare species' niches are larger than common species, it must be acknowledged that the potential for omission error also exists. Yet even though omission and commission errors occur with any predictive model, our use of models to develop sampling strata should overcome this problem. Any sampling based on models like ours should also include sampling effort allocated to the lowest probabilities of presence, thereby ensuring that sampling occurs across the spectrum of potential locations for the rare species. We acknowledge, however, that absences in the lower probability sites may still be confounded with low detection rates, such as is often the case for fungi (Dreisbach et al. 2002). We know of no simple resolution to this concern.

We argue that distinguishing different types of rarity may help to improve modeling of such rare events. To our knowledge, none of the cited studies for modeling rare species used any sort of formal model-based stratification design based on more common species. Wiser et al. (1998) demonstrated the difficulty of generating viable models from very limited observations only (10–24 presences), even though detailed, site-specific bio-geochemical predictors were additionally used to increase predictive capabilities. Engler et al. (2004) used two initial data sets of 46 and 77 occurrences of the rare species *Erynguim alpinum* L. to generate a model-based design improved their ability to identify new sites. A similar approach was performed by Sperduto and Congalton (1996). Such an approach is only a viable alternative if a minimal set of observations is available up-front for a realistic initial model. This was not the case for any of the tested rare species in our study.

A possible reason our effort worked so well is because all of our tested lichen species are cyanolichens, and they therefore have distinct and similar habitat requirements compared to green algal lichens. Green algal lichens, which are approximately 75% of the lichens in the Pacific Northwest, occupy wider ranges of habitats. Consequently the use of intrageneric species might not work so well in when applied to these lichens. In a study on coastal lichens, L. Geiser (*unpublished data*) documented the presence of *Ramalina farianacea* on nearly every plot, leading to an almost virtual overlap with two species, *R. thrausta* and *R. menziessi*, known to be rare. In such cases as this, it might be better to select a less common species that also has high association the targeted rare species, as we did here.

Our results suggest a model-based stratification design that predicts rare species from more common ones can improve detection likelihoods, especially if only very limited information is available. We expect that so-called ‘‘satellite’’ and ‘‘rural’’ type species would benefit from most from such a design, since their realized niches are difficult to detect if only very few observations are available for direct modeling an evaluation of the rare species. If one is interested in one to few rare species for management and evaluation, the best approach seems to be a combination of several approaches. Model-based stratification as presented here would logically be a first step in a sampling and modeling process.

ACKNOWLEDGMENTS

The authors acknowledge the contributions of the following people in field collection, identification and database entry: S. Berryman, M. Boyll, C. Derr, A. Ingersoll, D. Glavich, K. Gossen, A. Mikulin, J. Riley, and R. Ulrich. Special thanks to P. Halonen and B. Ryan (deceased) for help with difficult identifications, and to the many others who assisted with field collections. Field work was funded by the PNW Region Air Program. We are also grateful for the assistance from a large number of individuals working on Survey and Manage species in the Pacific Northwest. Specifically, we thank T. Brumley, N. Diaz, B. Rittenhouse, and N. Middlebrook for guidance on Forest Service and Bureau of Land Management Survey and Manage Issues. Two anonymous reviewers and T. B. Murphy helped improve the manuscript with their insights. Funding was provided by the State of Oregon Bureau of Land Management through a cooperative research arrangement with the USGS Forest and Rangeland Ecosystem Science Center (FRESC), Corvallis, Oregon, and the USGS Utah Cooperative Research Unit, Utah State University. Last, we thank A. Ellison and A. Agrawal for organizing the special features section on rarity, and for inviting our participation.

LITERATURE CITED

Alexander, H. M., N. A. Slade, and W. D. Kettle. 1997. Application of mark–recapture models to estimation of the population size of plants. Ecology **78**:1230–1237.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and regression trees. Wadsworth and Brooks/Cole, Monterey, California, USA.

Collins, S. L., S. M. Glenn, and D. W. Roberts. 1993. The hierarchical continuum concept. Journal of Vegetation Science **4**:149–156.

Cutler, R., T. C. Edwards, Jr., J. Alegria, and D. McKenzie. 2002. A sample design framework for Survey and Manage species under the Northwest Forest Plan. Proceedings of the Section on Statistics and Environment, 2001 Joint Statistical Meeting. American Statistical Association, Alexandria, Virginia, USA.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology **81**:3178–3192.

Dreisbach, T. A., J. E. Smith, and R. Molina. 2002. Challenges of modelling fungal habitat: when and where do you find chanterelles? Pages 475–481 *in* J. M. Scott, P. J. Heglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, and F. B. Samson, editors. Predicting species occurrence: issues of accuracy and scale. Island Press, Covello, California, USA.

Edwards, T. C.,Jr., R. Cutler, L. Geiser, J. Alegria, and D. McKenzie. 2004. Assessing rarity and seral stage association of species with low detectability: lichens in western Oregon and Washington forests. Ecological Applications **14**:414–424.

Engler, R., A. Guisan, and L. Rechsteiner. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology **41**:263–274.

Esseen, P-A., K-E. Renhorn, and R. B. Pettersson. 1996. Epiphytic lichen biomass in managed and old-growth boreal forests: effect of branch quality. Ecological Applications **6**:228–238.

Ferrier, S., and G. Watson. 1997. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. Environment Australia, New South Wales National Parks and Wildlife Service, Armidale, New South Wales, Australia.

Frescino, T. S., T. C. Edwards,Jr., and G. G. Moisen. 2001. Modelling spatially explicit forest structural attributes using generalized additive models. Journal of Vegetation Science **12**:15–26.

Gaston, K. J. 1994. Rarity. First edition. Chapman and Hall, London, UK.

Geiser, L. 2004. Manual for monitoring air quality using lichens on national forests of the Pacific Northwest. Technical Paper R6-NR-AQ-TP-1-04. USDA Forest Service, Pacific Northwest Region, Portland, Oregon, USA.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, and D. Tyler. 2002. The national elevation data set. Photogrammetric Engineering and Remote Sensing **68**:5–12.

Goerck, J. M. 1997. Patterns of rarity in the birds of the Atlantic forest of Brazil. Conservation Biology **11**:112–118.

Green, R. H., and R. C. Young. 1993. Sampling to detect rare species. Ecological Applications **3**:351–356.

Gregoire, T. G., H. T. Valentine, and G. M. Furnival. 1995. Sampling methods to estimate foliage and other characteristics of individual trees. Ecology **76**:1181–1194.

Hanski, I. 1982. Dynamics of regional distribution: the core and satellite species hypothesis. Oikos **38**:210–221.

Hill, N. M., and P. A. Keddy. 1992. Prediction of rarities from habitat variables: coastal plain plants on Nova Scotian lakeshores. Ecology **73**:1852–1859.

Hoeting, J. A., M. Leecaster, D. Bowden. 2000. An improved model for spatially correlated binary responses. Journal of Agricultural, Biological and Environmental Statistics **5**: 102–114.

Ihaka, R., and R. Gentleman. 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics **5**:299–314.

James, F. C., C. E. McCulloch, and D. A. Wiedenfeld. 1996. New approaches to the analysis of population trends in land birds. Ecology **77**:13–21.

Jensen, M. E., and H. R. Haise. 1963. Estimating evapotranspiration from solar radiation. Journal of Irrigation and Drainage Engineering-ASCE **89**:15–41.

Lohr, S. L. 1999. Sampling: design and analysis. Duxbury Press, Pacific Grove, California, USA.

Maina, G. G., and H. F. Howe. 2000. Inherent rarity in community restoration. Conservation Biology **14**:1335–1340.

Manly, B. F. J. 1997. Randomization, bootstrap and Monte Carlo methods in biology. Chapman and Hall, New York, New York, USA.

Max, T. A., H. T. Schreuder, J. W. Hazard, J. Teply, and J. Alegria. 1996. The Region 6 vegetation inventory and monitoring System. General Technical Report PNW-RP-493. USDA Forest Service, Pacific Northwest Research Station, Portland, Oregon, USA.

McCune, B. 1993. Gradients in epiphyte biomass in three *Pseudotsuga–Tsuga* forest of different ages in western Oregon and Washington. Bryologist **96**:405–411.

Molina, R., D. Pilz, J. Smith, S. Dunham, T. Dreisbach, T. O'Dell, and M. Castellano. 2001. Conservation and management of forest fungi in the Pacific Northwestern United States: an integrated ecosystem approach. Pages 19–63 *in* D. Moore, M. M. Nauta, S. E. Evans, and M. Rotheroe, editors. Fungal conservation. Issues and solutions. Cambridge University Press, New York, New York, USA.

Murphy, D. D., and B. R. Noon. 1992. Integrating scientific methods with habitat conservation planning: reserve design for northern spotted owls. Ecological Applications **2**:3–17.

Olsen, A. R., and H. T. Schreuder. 1997. Perspectives on large-scale natural resource surveys when cause–effect is a potential issue. Environmental and Ecological Statistics **4**:167–180.

Peterson, E. B., and B. McCune. 2001. Diversity and succession of epiphytic macrolichen communities in low-elevation managed conifer forests in western Oregon. Journal of Vegetation Science **12**:511–524.

Philippi, T. 2005. Adaptive cluster sampling for estimation of abundances within local populations of low-abundance plants. Ecology **86**:1091–1100.

Rosso, A. L., B. McCune, and T. R. Rambo. 2000. Ecology and conservation of a rare, old-growth-associated canopy lichen in a silvicultural landscape. Bryologist **103**:117–127.

Sauer, J. R., and S. Droege, editors. 1990. Survey designs and statistical methods for the estimation of avian population trends. Biological Report 90(1), U.S. Fish and Wildlife Service, Washington, D.C., USA.

Schreuder, H. T., T. G. Gregoire, and G. B. Wood. 1993. Sampling methods for multiresource forest inventory. John Wiley and Sons, New York, New York, USA.

Sheldon, A. L. 1988. Conservation of stream fishes: patterns of diversity, rarity, and risk. Conservation Biology **2**:149–156.

Shiver, B. D., and B. E. Borders. 1996. Sampling techniques for forest resource inventory. John Wiley and Sons, New York, New York, USA.

Sillett, S. C., and M. N. Goslin. 1999. Distribution of epiphytic macrolichens in relation to remnant trees in a multiple-age Douglas-fir forest. Canadian Journal of Forest Research **29**:1204–1215.

Sperduto, M. B., and R. G. Congalton. 1996. Predicting rare orchid (small-whorled Pogonia) habitat using GIS. Photogrammetric Engineering and Remote Sensing **62**:1269–1279.

Stow, C. A., S. R. Carpenter, K. E. Webster, and T. M. Frost. 1998. Long-term environmental monitoring: some perspectives from lakes. Ecological Applications **8**:269–276.

Thomas, J. W., E. D. Forsman, J. B. Lint, E. C. Meslow, B. R. Noon, and J. Verner. 1990. A conservation strategy for the northern spotted owl. Interagency Scientific Committee to Address the Conservation of the Northern Spotted Owl. Report No. 1990-791-171/20026. United States Government Printing Office, Washington, D.C., USA.

Thornton, P. E., and S. W. Running. 1999. An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. Agricultural and Forest Meteorology **93**:211–228.

Thornton, P. E., S. W. Running, and M. A. White. 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. Journal of Hydrology **190**: 214–251.

U.S. Department of Agriculture, Forest Service, and U.S. Department of Interior, Bureau of Land Management. 1994. Record of decision for amendments to Forest Service and Bureau of Land Management planning documents within the range of the northern spotted owl. ⟨http://www.or.blm.gov/roseburg/ROD_RMP/roseburg/appendices/appendixa.html⟩.

U. S. Department of Agriculture, Forest Service, and U.S. Department of Interior, Bureau of Land Management. 2001. Record of decision and standards and guidelines for amendments to the survey and manage, protection buffer, and other mitigation measures. Standards and guidelines. ⟨http://www.or.blm.gov/nwfpnepa/FSEIS-2000/ROD-SandG.pdf⟩.

Wiser, S. K., R. K. Peet, and P. S. White. 1998. Prediction of rare-plant occurrence: a southern Appalachian example. Ecological Applications **8**:909–920.

APPENDIX A

A table showing the expected number of detections on model-assisted sampling for *Pseudocyphellaria rainierensis* from the classification tree model of *Lobaria oregana* is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-057-A1.

APPENDIX B

A table showing the expected number of detections based on model-assisted sampling for *Pseudocyphellaria rainierensis* and *Nephroma laevigatum* from the classification tree model of *Lobaria pulmonaria* is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-057-A2.

APPENDIX C

A table showing the expected number of detections based on model-assisted sampling for *Nephroma occultum* and *Nephroma parile* from the classification tree model of *Pseudocyphellaria anomala* is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-057-A3.

APPENDIX D

A table showing the expected number of detections based on model-assisted sampling for *Pseudocyphellaria anthraspis* is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-057-A4.

Ecology, 86(5), 2005, pp. 1091–1100

ADAPTIVE CLUSTER SAMPLING FOR ESTIMATION OF ABUNDANCES WITHIN LOCAL POPULATIONS OF LOW-ABUNDANCE PLANTS

THOMAS PHILIPPI[1]

Department of Biological Sciences and Florida Coastal Everglades LTER, Florida International University, Miami, Florida 33199 USA

Abstract. Quantitative estimates of abundance for rare plants can be difficult, as the most widely used sampling techniques are ill-suited for rarity. Adaptive cluster sampling (ACS) can take advantage of the spatial clustering common in rare plant populations to provide more efficient unbiased estimates of population sizes than simple random sampling. When plants are found in a quadrat, all adjacent quadrats are adaptively added to the sample. Despite this biased sampling, the Horvitz-Thompson estimator for adaptive cluster sampling provides unbiased estimates of population means or totals, and variances of those estimates. Because ACS disproportionately samples quadrats with plants, it can provide additional efficiency whenever further attributes of rare individuals need to be assessed, such as demographic parameters or genotypes. Adaptive cluster sampling was performed on a population of *Aletris bracteata*, a moderately visible herbaceous species, in a savanna near Chekika, Everglades National Park, Florida. Both 1-m^2 and 4-m^2 quadrats provided reasonable estimates of the population size. The 1-m^2 sampling included 30–36% of the estimated total plants while sampling only 5% of the total area. The 4-m^2 sampling captured 78% of the estimated total population while sampling only 21% of the area.

Key words: adaptive cluster sampling; Aletris bracteata; *Chekika; population abundance; rare plant sampling.*

INTRODUCTION

The numbers of individuals occurring in populations is one fundamental piece of information for conservation and management of rare species. In terms of legal mandates, abundances are a major component of ‘‘status.’’ If several populations exist, population sizes allow prioritization for land acquisition or protection. Because processes ranging from demographic stochasticity to loss of genetic variation are functions of the numbers of individuals within breeding populations, the numbers of individuals within populations are crucial initial descriptors for population viability analysis (Soule 1986, Morris and Doak 2002). Conversely, exotic species are much easier to eradicate or manage if caught when still rare (Doren and Ferriter 2002), and accurate estimates of abundances are necessary to fund or oversee eradication efforts.

Censusing or estimating population sizes, however, is only one component of almost any rare species management program, especially rare plants. Management actions such as collection of seeds or cuttings or protection of individual plants may be at least as important as estimating population sizes. For exotic invasive species, management generally includes eradication as well as assessment. For conservation, additional information is often obtained for each individual discovered in the population. For plants, because the expected contribution to future population size (reproductive value) can vary by six or seven orders of magnitude among individuals, demographic status such as size and reproductive state are often recorded, and individuals tagged for estimates of survival. Survivorship and reproductive output can be combined to estimate size- or stage-specific reproductive value; those reproductive values can then be used as weights to provide a more informative population size than the unweighted number of individuals (Caswell 2001, Philippi et al. 2001). Tissue samples may be collected for genetic analysis to ascertain the genetic structure within the population. For plants and sessile animals, the locations of individuals may be mapped, allowing assessment of dispersal distances in conjunction with the genetic data. Finally, habitat attributes such as soil depth, shade, and nutrients and pH may be collected at each located plant and at random locations in order to delineate suitable microhabitat, both for possible management intervention and for subsequent model-based detection (Edwards et al. 2005). All of these components place a premium on the number of individuals encountered.

If rare plants are highly visible, their rarity may make nearly complete enumeration possible via systematic searching. Most rare plant species, however, require at least moderate effort to detect, especially when seedlings are included. Except for the most spatially constrained populations, such effort precludes searching

---

Manuscript received 7 April 2004; revised 9 August 2004; accepted 9 August 2004; final version received 15 November 2004. Corresponding Editor: A. M. Ellison. For reprints of this Special Feature, see footnote 1, p. 1079.

[1] E-mail: philippi@fiu.edu

the entire spatial extent of the population, and some form of sampling is required.

Unfortunately, low-abundance species present problems for sampling. Gaston (1994) pointed out that the sampling methods presented in most ecological methodology texts are aimed at abundant species, and do not address the special conditions of rare plants. Most notably, simple random sampling can lead to the vast majority of quadrats having no individuals, and thus large variances of estimates of population size, driven more by the binomial variation of numbers of quadrats with or without individuals than the variation in numbers of individuals among quadrats with individuals. Even moderate precision in estimates of population sizes may require substantial effort for little practical benefit (e.g., Evans and Viengkham 2001). Therefore, many rare plant surveys omit any quantitative sampling for estimation of population sizes in favor of expert guesses, and allocate the saved effort to other components (Keith 2000).

Alternative approaches to sampling exist that may be more applicable for rare plants. One approach in particular, adaptive cluster sampling, may be especially applicable, as it takes advantage of spatial clustering among individuals common in rare plant populations. Spatial clustering implies that once one plant is found, nearby locations are much more likely than random locations to also have plants. Many insects, fish, and birds with clumped prey use this information content from successful prey capture for adaptive foraging, modifying distances between probes and turning rules dynamically based on recent success rate (reviewed in Pyke 1984). Adaptive cluster sampling is a way for ecologists to use this information for their own benefit.

Sampling designs and estimators

To a field ecologist, nearly all of sampling theory can be thought of as special cases for applying ancillary information to provide more efficient estimates than simple random sampling—either more precise estimates for the same effort, or equally precise estimates with reduced effort. This growing collection of special cases can be organized by two attributes: the design of the sampling per se and the form of the inference from the sample to the population.

Sampling designs specify rules for which units of the population are included in the sample. Thompson and Seber (1996) distinguish three classes of sampling designs: conventional designs, adaptive designs, and nonstandard designs. In conventional designs, sample selection does not depend on the observed values, and thus the sample can be specified before any data are collected. Sample selection may depend on ancillary variables known beforehand for all units in the population, so stratified sampling, systematic sampling, sampling with probabilities proportional to a known attribute, as well as simple random sampling are all conventional designs. In adaptive designs, sample selection may also depend on the observed values of the sampled variable for units included in the sample. Sampling with stopping rules based on the observed values (e.g., sample randomly until 100 plants have been found) are adaptive designs. Nonstandard designs may also include unknown parameters or ancillary information from units not included in the sample, and thus are much more restricted in the forms of inferences they can support.

The two major approaches to making inferences from a sample to the population are design-based and model-based. Design-based inferences make few assumptions about properties of the population, and obtain their unbiasedness from the properties of the design itself. Such methods of inference are evaluated for unbiasedness and efficiency in terms of their expectation over repeated samples from the same fixed population. Model-based inferences assume a model for the distribution of values for the attributes of the population. Therefore, model-based inferences are evaluated for unbiasedness and efficiency in terms of expectations over repeated realizations of populations, given underlying population model parameters. Ratio estimators and spatial kriging are examples of model-based inferences. Edwards et al. (2005) explore model-based inferences for rare species in the context of detecting populations at much larger spatial scales. Both design-based and model-based inferences are possible with conventional designs. Similarly, adaptive designs might support either design-based or model-based inferences.

One important estimator for design-based inferences is the Horvitz-Thompson estimator:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i} \tag{1}$$

where $\hat{\mu}$ is the estimate of the population mean, N is the total number of sample units (quadrats) in the population, n quadrats are sampled, y_i is the number of plants in quadrat i, π_i is the probability that quadrat i is included in the sample, and the sum is across all units included in the sample (notation mostly follows Thompson 2002).

Under simple random sampling without replacement, the probability that any specified quadrat is included is the same for all quadrats, and is the number of possible samples that include quadrat i divided by the total number of possible samples. Thus, under simple random sampling, the Horvitz-Thompson estimator reduces to the more familiar estimator:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\left(\frac{n}{N}\right)} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}. \tag{2}$$

An estimator for the variance of the Horvitz-Thompson estimator is the following:

$$\widehat{\operatorname{var}}(\hat{\mu}) = \frac{1}{N^2}\left[\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{y_j y_k}{\pi_{jk}}\left(\frac{\pi_{jk}}{\pi_j \pi_k} - 1\right)\right] \tag{3}$$

where π_{jk} is the probability that both quadrat j and quadrat k are included in the sample (Thompson and Seber 1996).

The utility of the Horvitz-Thompson estimator is that it allows unbiased estimates of the population mean or total with unequal inclusion probabilities. Thompson (2002:54) notes in passing that to the extent that the inclusion probabilities π_i can be chosen to be proportional to the abundances y_i, the variance of this estimator will be reduced. While strict proportionality would require advance knowledge of the abundances, this observation provides insight into how many different design-based approaches improve on the efficiency of simple random sampling with equal inclusion probabilities. Sampling designs with unequal inclusion probabilities increase efficiency by using different forms of ancillary information to adjust the inclusion probabilities among units in the direction of the variation in abundances among units, so units with higher abundances are more likely to be included in the sample. For example, stratified random sampling increases efficiency when strata differ in densities (abundances per unit) and sampling intensity is higher in the strata with greater densities.

Adaptive cluster sampling

Adaptive cluster sampling (ACS) is an adaptive design that takes advantage of clustering of individuals within a population to adjust inclusion probabilities in favor of those quadrats with higher y_i. (Thompson 1990, 1991*a*, *b*, Thompson et al. 1992, Thompson and Seber 1996, Christman 2000*b*). The simplest example to visualize and the focus of this paper is spatial clustering. For many low-abundance plants and animals, restricted dispersal distances produces clumped spatial distributions, as in the points in Fig. 1. Therefore, once a quadrat is found to have plants present, quadrats in the neighborhood of that occupied quadrat are more likely than average to also have plants. Inclusion of such quadrats, if their inclusion probabilities can be computed, may reduce the variance of the Horvitz-Thompson estimator and thus increase the efficiency of the sampling. The criterion for sampling adjacent quadrats need not be the presence of a single plant, but could be the presence of more than some threshold number of plants. The only constraint is that the definition of neighbor must be transitive, so *i* is a neighbor of *j* implies that *j* is also a neighbor of *i*. The example below estimating *Aletris* by stage classes will take advantage of this flexibility. Further, the definition of neighbor need not be spatial adjacency, but could be based on degree of relatedness or participation as sexual partners, if the rare event were a genetic condition or sexually transmitted disease. A necessary condition for adaptive cluster sampling to be more efficient than simple random sampling is that neighbors have positive covariation of some form in the sampled attribute y_i.

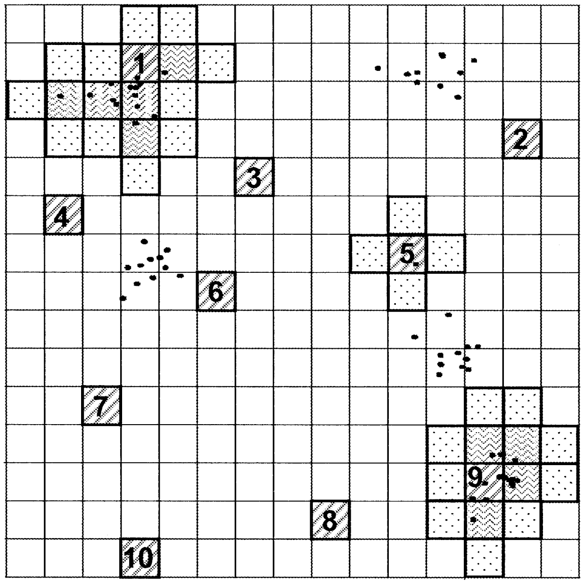

FIG. 1. Simulation example with 10 clusters, each with a Poisson mean of 10 individuals. The 10 initial quadrats are numbered and indicated with diagonal stripes, additional quadrats within intersected networks are indicated with wavy lines, and edge quadrats are stippled.

For illustration, consider a sampling design that starts with a simple random sample of quadrats in the population (diagonal striped and numbered quadrats in Fig. 1). Because of the low abundance of plants, most of these quadrats will have no plants. Assuming an inclusion threshold of any plants present, if a plant is found, the neighbors of that quadrat are added to the sample. In this example, the neighborhood is defined to be the four adjacent quadrats, as can be seen in the four stippled quadrats surrounding quadrat 5. (The neighborhood could also be defined as the 8 adjacent quadrats including diagonals.) If a neighboring quadrat also includes plants (wavy quadrats in Fig. 1), it is added to the network and its neighbors are recursively sampled as well. This procedure grows networks of sampled quadrats until each network is surrounded by a ring of empty quadrats (stippled quadrats in Fig. 1). The final result of this procedure is a sample with three types of quadrats. The first type is quadrats included in the original random sample (diagonal stripe), whether or not plants were found. The second type is quadrats with plants that were sampled not because they were included in the original draw, but because they were in the recursive neighborhood of a quadrat in the initial sample that included plants (wavy quadrats in Fig. 1). The third type is edge quadrats: quadrats without plants that were not in the initial sample, but were neighbors of quadrats with plants of either of the first two types. A network is either an initially sampled quadrat that had no plants, or an initial quadrat with plants plus all of the neighboring quadrats that also have plants, but a network does not include the associated empty edge quadrats. Networks thus are sets of quadrats such that if any quadrat in the network is sampled, all quadrats in the network are sampled.

Clearly, the simple average $\bar{y}$ of these sampled quadrats is a biased estimate of the population mean, as the point of this sampling design was to disproportionately include more quadrats with plants in them, and quadrats within large networks have increased probabilities of inclusion in the sample. In order to use the Horvitz-Thompson estimator, the inclusion probabilities π_i must be known for each sampled quadrat. Because quadrats may be included via more than one process (in the original sample, within a chain that had at least one quadrat included in the initial sample, or an edge quadrat adjacent to a network with at least one quadrat in the initial sample), the probability of inclusion is computed as 1 minus the fraction of possible sample draws that would not have included that quadrat (Thompson and Seber 1996):

$$\pi_i = 1 - \left[\binom{N - m_i - b_i}{n} \Big/ \binom{N}{n}\right] \tag{4}$$

where n is the number of quadrats in the initial sample, m_i is the number of quadrats in the network that included quadrat i, and b_i is the number of quadrats in networks for which quadrat i is an edge. N and n are known from the initial design, and m_i can be tallied for each sampled quadrat as long as the recursive sampling continued until the entire network was sampled. However, the sampling design does not provide enough information to calculate the probability of each quadrat being included in the sample because it was an edge of a network. Note in Fig. 1 that quadrat 6 could have been sampled as an edge quadrat for the network to its left, but the sampling procedure did not detect the existence of that cluster.

The solution is to exclude edge quadrats from the estimation (Thompson 1990, Thompson and Seber 1996). Then, the probability of inclusion in the estimator (instead of in the sample) for quadrat i can be calculated from the information obtained in the sampling. The probability of inclusion is equal for all quadrats in a network, so, switching subscripting from quadrats to networks, a_k, the probability of inclusion for network k is

$$a_k = 1 - \left[\binom{N - x_k}{n} \Big/ \binom{N}{n}\right] \tag{5}$$

where x_k is the number of quadrats found in network k. The Horvitz-Thompson estimator for this adaptive design is thus Eq. 1 with a_k substituted for π_i, and the number of individuals found in the entire network y_k rather than the numbers of individuals found in each quadrat y_i as follows:

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{\nu}\frac{y_k}{a_k}. \tag{6}$$

The estimator for the variance of the Horvitz-Thompson estimator also requires terms for the probabilities that both network j and network k were included in the sample:

$$a_{jk} = 1 - \left[\frac{\binom{N - x_j}{n} + \binom{N - x_k}{n} - \binom{N - x_j - x_k}{n}}{\binom{N}{n}}\right] \tag{7}$$

$$\widehat{\mathrm{var}}(\hat{\mu}) = \frac{1}{N^2}\left[\sum_{j=1}^{\nu}\sum_{k\neq j}^{\nu}\frac{y_j y_k}{a_{jk}}\left(\frac{a_{jk}}{a_j a_k} - 1\right)\right]. \tag{8}$$

The summations in both the Horvitz-Thompson estimator and estimated variance are over (unique) networks ν ($\leq n$). A given network may be intersected by more than one quadrat in the initial sample. For the Horvitz-Thompson estimators, that network is only included once; the estimators are based on the probabilities of inclusion or intersection of each sampled network.

A second unbiased estimator based on the Hansen-Hurwitz estimator exists for this adaptive sampling design, and is presented in Thompson and Seber (1996) and Thompson (2002). The Hansen-Hurwitz estimator is based on the numbers of times each network was intersected. Therefore, the summation is across initial sample quadrats, and networks intersected more than once are included multiple times in the computation. The second edition of Krebs' textbook (Krebs 1999) briefly presents the Hansen-Hurwitz estimator, which is simpler to compute than the Horvitz-Thompson estimator:

$$\tilde{\mu}_{HH} = \frac{1}{n}\sum_{i=1}^{n} w_i$$

$$\widehat{\mathrm{var}}(\tilde{\mu}_{HH}) = \frac{N - n}{Nn(n - 1)}\sum_{i=1}^{n}(w_i - \tilde{\mu}_{HH})^2 \tag{9}$$

where w_i is the mean quadrat abundance for quadrats in cluster i (y_i/x_i). However, in Thompson's examples and in Salehi's more extensive simulations (Salehi 2003), the Horvitz-Thompson estimator almost always had smaller variance than the Hansen-Hurwitz estimator. Christman (1997) and Smith et al. (1995) also found that the Horvitz-Thompson estimator had smaller variance than the Hansen-Hurwitz estimator. However, because of the computational simplicity of the Hansen-Hurwitz estimator, it has been used in analyses of confidence intervals (Christman 2000*a*) and the comparative efficiency of different sampling designs.

For the example in Fig. 1, the networks detected are listed in Table 1. The SAS macro for computing both estimators given in the Supplement requires a data set with one observation for each initial sample point and

TABLE 1. Tallied numbers of individuals per network (Y) from adaptive cluster sampling in Fig. 1.

Network	Y values	Sum of Y	Size (m_i)
1	1, 1, 1, 4, 5, 1	13	6
2	0	0	1
3	0	0	1
4	0	0	1
5	1	1	1
6	0	0	1
7	0	0	1
8	0	0	1
9	4, 2, 1, 4, 1	12	5
10	0	0	1

three variables. One variable is the identification of the network, unique for initial sample points except when two or more initial sample points intersect the same network. The second variable is the number of quadrats in that network (edge quadrats are not in the network), and is 1 for initial sample points without plants. The third variable is the number of plants found in that network, which is 0 for networks consisting of empty initial sample points only. The resulting estimators and variances are given in Table 2. The estimates are greater than the actual value, as this sample happened to hit three clusters with only 10 initial sample quadrats.

Aletris *example*

Aletris bracteata Northr. (Nartheciaceae) is an herbaceous perennial found in southern Florida and the Bahamas. *A. bracteata* is either an endangered species, or a white- to cream- flower color mutant or subspecies of the yellow-flowered *A. lutea*. *A. bracteata* is known from fewer than a dozen locations in southern Florida, all in savanna or glade vegetation with grasses, sedges, and other herbaceous species producing nearly 100% vegetative cover. *A. bracteata* grows as a nearly prostrate basal rosette up to 25 cm in diameter, and produces flowering stalks up to 40 cm tall. While flowers are visible at a distance of a few meters, rosettes and nonflowering individuals require close inspection and gentle moving of the grasses and sedges, and seedlings require careful searching of quadrats from hands and knees.

The known populations of *A. bracteata* differ greatly in their extent, abundance, and spatial patterning. The population in eastern Singeltary, Miami–Dade county, Florida, is abundant and compact enough that simple random sampling provided reasonable estimates of population size. Conversely, currently only three isolated individuals spread over roughly 0.5 km^2 are known from a nearby site, so neither simple random sampling nor adaptive cluster sampling will work for that population. At Chekika in eastern Everglades National Park, preliminary reconnaissance found scattered clusters of plants, suggesting that adaptive cluster sampling might be appropriate.

The general goals for the Chekika population were to estimate the total abundance, characterize the spatial distribution, collect flowers from individuals for laboratory analysis of flower color and morphology, and obtain demographic information for a sample of individuals to allow comparison to other populations. Additionally, plants were mapped and measured for comparison with post-fire demography if the site burns (a fresh fire line 25 m from the population was observed during sampling). While the diffuseness of these goals may reflect early stages of academic research compared to more concrete goals in rare species management, the multiplicity of goals is common to both situations. On two successive Saturdays during the flowering (and dry) season, different volunteer graduate students spent the afternoon sampling the same 40 × 100 m area containing several clusters of *Aletris bracteata*. For each trial, 150 random quadrat locations were generated without replacement, with southeast corner coordinates integer meters in UTM (even numbered meters for the 4-m^2 trial). These quadrat locations were sorted into order along the long axis, and printed on data sheets. A single 100-m tape was run as a centerline the long axis of the plot; distances perpendicular to the centerline were listed as ±20 m. Sample quadrats were located by running short tapes perpendicular to the centerline at the appropriate distances.

On the first trial, in one afternoon (∼4 h), three people were able to complete a sample of 150 initial 1-m^2 quadrats, while the fourth marked and collected GPS locations for all observed plants (including those detected but not part of the sampling effort). This trial intersected only two clusters of plants: one network of nine quadrats with a total of 33 plants and 16 additional edge quadrats; the other network of 16 quadrats with 32 plants and 18 edge quadrats (Table 3). On the second trial, in the same 4 h, seven people were able to complete a sample of 150 initial 4-m^2 (2 × 2 m) quadrats, plus record the number of leaves and length of longest leaf and flowering status for each individual, and record

TABLE 2. Horvitz-Thompson and Hansen-Hurwitz estimates of the number of plants in Fig. 1.

Parameter	Horvitz-Thompson	Hansen-Hurwitz
Total	139.4	131.2
Mean	0.619	0.583
Variance of estimated mean	0.0968	0.0987
Approximate 95% CI for total	(2.2–276.6)	(−7.3–269.8)

Note: The true number of plants is 96.

TABLE 3. Tallied data from adaptive cluster sampling of *Aletris bracteata* in Chekika, Everglades National Park, Florida, USA.

1-m^2 quadrats				4-m^2 quadrats			
Network	No. times intersected	No. quadrats	Y	Network	No. times intersected	No. quadrats	Y
1	1	9	33	1	1	3	14
2	1	16	32	2	3	20	114
3–150	1	1	0	3	1	6	33
				4–150	1	1	0

a soil depth in the center of every sampled quadrat. This trial intersected three clusters: 14 plants in three quadrats + seven edge, 114 plants in 20 quadrats + 20 edge, and 33 plants in six quadrats + 11 edge (Table 3). Both trials detected the 33 plants in one cluster, falling into a network of nine 1-m^2 or six 4-m^2 quadrats. The second cluster detected by both trials included 114 plants in the 4-m^2 trial, but only 32 plants were in the network defined by 1-m^2 quadrats. The three-quadrat, 14-plant network detected in the 4-m^2 trial was not intersected in the 1-m^2 trial.

The Horvitz-Thompson and Hansen-Hurwitz estimates for these trials are presented in Table 4. In both cases, the Horvitz-Thompson confidence intervals were somewhat narrower than the Hansen-Hurwitz intervals. The 4-m^2 trial produced much smaller confidence intervals for the Horvitz-Thompson estimator, but that is to be expected, as it included sampling of 21% of the population area while the 1-m^2 trial included only 5%. More to the point for the overall study of *Aletris bracteata*, the 4-m^2 trial yielded 161 of an estimated 207 plants (78%) with that 21% sampling, and even the 1-m^2 trial yielded 65 plants, 36% of the 1-m^2 estimate or 31% of the 4-m^2 estimate from only 5% of the area.

The demographic information collected in the second trial was used to compute separate estimates for flowering, nonflowering, and seedlings (Table 4). This illustrates an application of flexibility in the rule for network inclusion. Separate adaptive sampling for each life stage was not required. Rather, the single adaptive sampling, with presence of any plants as the inclusion rule, was used for all four estimates. Only the abundances y_i for each network differed for different stages (see the data file in the Supplement). The inclusion rule based on any life stage meets the transitivity requirement, and as long as included quadrats are also more likely than random quadrats to have individuals of each life stage, ACS can be more efficient than complete random sampling, even though some included quadrats (or even networks) contain no individuals of the life stage being estimated.

TABLE 4. Estimates and 1 SD of estimates (in parentheses) for *Aletris bracteata* in Chekika, Everglades National Park.

Parameter	Horvitz-Thompson	Hansen-Hurwitz
1-m^2 quadrats		
Total plants	183 (108)	151 (109)
4-m^2 quadrats		
Total plants	207 (49)	258 (88)
Flowering	99 (28)	114 (39)
Nonflowering	80 (19)	99 (34)
Seedlings	28 (5.4)	45 (18.3)

Notes: With 1-m^2 quadrats, 150 initial quadrats plus 23 additional network quadrats and 34 edge quadrats (207 total) detected 65 plants. With 4-m^2 quadrats, 150 initial quadrats plus 26 additional network quadrats plus 38 edge quadrats (212 total) detected 161 plants.

In addition to estimating total population size, the sampling also indicated that the Chekika population of *A. bracteata* was comprised of a few (perhaps 5–10) clusters of plants, with a wide range of cluster sizes (14, 33, and 114). This information suggests biological hypotheses for further investigation: are the smaller clusters comprised of single maternal genotypes (localized seed dispersal from individual source plants)? Is the large cluster merely a few single- maternal line clusters in spatial proximity, or does it have fundamentally different genetic structure than the smaller clusters (e.g., source–sink dynamics for the population)?

These two trials represent roughly equal person-hours (∼12) dedicated to estimating population size. Even though the random initial locations were sorted spatially on the data sheet, the time to locate and move to the next primary quadrat location exceeded the time required to thoroughly search either quadrat size. During the second trial over half of the person-hours were spent collecting the rudimentary demographic data. In this case, the larger quadrat size was more efficient in the broad sense of the overall project—while it only provided a slightly smaller variance about the population size estimate, it detected more than twice as many plants.

EXTENSIONS

Adaptive cluster sampling was developed largely with ecological (albeit animal) motivation (Thompson 1990, 1991*a*, *b*, Thompson et al. 1992). Many papers extending ACS have been published in ecological statistics journals such as *The Journal of Agricultural Biological and Environmental Statistics*, *Environmental and Ecological Statistics*, and *Environmetrics*; several key papers are published in *Biometrics. Environmental and Ecological Statistics* published a special feature on adaptive sampling in 2003 (Brown 2003, Chao 2003, Christman 2003, Di Battista 2003, Dryver 2003, Felix-Medina 2003, Salehi 2003, Smith et al. 2003, Su and Quinn 2003, Thompson 2003).

Applications of adaptive cluster sampling to birds (Smith et al. 1995), amphibians (Ishwar et al. 2001, Vasudevan et al. 2001), and marine and aquatic species (Lo et al. 1997, Conners and Schwager 2002, Hanselman et al. 2003, Smith et al. 2003) have been published. However, only one published paper applies ACS to rare plants. Acharya et al. (2000) sampled rare species of trees in Nepal. They used systematic plot location for the first stage sampling, and an inclusion rule based on the presence of an individual of a target species. They adaptively added plots separately for each target species, as unlike in the *Aletris* life-stage example above, the presence of an individual of one rare species was not predictive of the presence of other rare species in neighboring plots.

Other forms of initial sampling

The initial sampling need not be simple random sampling as in these examples. When movement within the population area is expensive relative to search, adaptive strip sampling (Thompson and Seber 1996, Pontius 1997) may be much more efficient. If plants are detectable over a narrow width while walking, initial samples may be chosen as contiguous parallel strips of quadrats, with random choice of origin of strips in the perpendicular direction. Whenever quadrats meet the inclusion threshold, surrounding individual quadrats are added to the sample. Computation of the co-inclusion probabilities α_{jk} for pairs of networks is more complex, as networks might be jointly included by the same strip or different strips. For larger areas, the area may first be divided into strata, then adaptive cluster sampling performed within each stratum (Thompson 1991*b*, Thompson and Seber 1996). When ancillary measures are to be kriged across the entire area, systematic initial sampling of regularly spaced plots may be appropriate.

OPEN ISSUES

Optimal quadrat size

Quadrat size matters for most forms of sampling through the trade-off in costs between numbers and sizes of quadrats, and to a lesser extent through an effect on sample variances via means of occupied quadrats being proportional to quadrat area (Pielou 1957, Greig-Smith 1983, Elzinga et al. 1998). In adaptive cluster sampling, quadrat size relative to plant nearest neighbor distances also controls network size, and thus total effort for a given number of initial samples. Quadrats too small will not capture additional individuals within clusters; quadrats that are too large may capture the entire cluster in one or a few quadrats. Brown (2003) explored efficiency in terms of network sizes and found that smaller network sizes (two to four quadrats) and higher variance among quadrats within networks increased the efficiency of designs. However, her results were from simulations varying the cluster sizes in the population and the threshold for inclusion (1 for rare plants, but other fixed thresholds are possible for other applications), and not quadrat size per se. Given limited prior knowledge about the rare plants being sampled, quadrat sizes somewhat larger than the nearest neighbor distances within clusters may be a sufficient rule of thumb.

Confidence intervals

Confidence intervals generated from the standard deviations are problematic for rare plant applications of ACS, as they are based on asymptotics or large sample behavior. Such confidence intervals tend to be very large, and because they are symmetric, the lower bound is often smaller than the number of plants actually observed in the survey, and might even be negative (see Table 2). Christman and Pontius (2000) evaluated the performance of several approaches to bootstrap confidence intervals for the Hansen-Hurwitz estimator. They simulated distributions of rare ($\approx$7% of quadrats occupied) and uncommon ($\approx$15% occupied) plants, and initial samples of 20 or 40 quadrats out of a population of 400 quadrats. For rare plants and $n = 20$, all nominally 90% confidence intervals had more than 98% coverage (included the true value in more than 98% of the simulations). While the coverages were closer to the nominal values for the other three combinations, the tails tended to be asymmetric, and no single approach was best across all three combinations.

Detectability

For many species of herbaceous plants, detection is not perfect, even for close inspection of small quadrats. Some individuals may dieback and resprout the following year, so are extant and alive belowground, but have no aboveground presence during the census and thus are not detected (Gilbert and Lee 1980, Shefferson et al. 2001, Kery and Gregg 2003, Kery 2004). If the detection probability g can be estimated by other means (e.g., MacKenzie et al. 2005) and is constant across individuals, the correction under conventional sampling designs is straightforward upweighting by $1/g$ to estimate the true number of individuals in the population. The variance of the estimator corrected for imperfect detection is only slightly more complicated (e.g., Thompson and Seber 1996). For adaptive sampling, the detection probabilities also enter into the numbers of quadrats sampled in each intersected cluster, and thus the inclusion probabilities for sampled quadrats. (Thompson and Seber 1994, 1996) provide a correction for adaptive cluster sampling with imperfect detection conditioned on the observed values of y.

In general, the detectability of rare plants is likely to differ greatly between individuals of different life stages—large flowering individuals are highly visible, but seedlings have lower detection probabilities. If stage-specific detection probabilities can be estimated from mark-recapture or other approaches, it might be useful to use presence of reliably-detected stages for the cluster inclusion rule, count all seedlings within such networks, and then use the 1/*g* upweighting, but under what conditions (if any) that this might be beneficial has not been explored.

Monitoring

Conventional sampling designs provide the opportunity to repeatedly sample the fixed sample locations over time, and then analyze the data as repeated measures, separating the spatial variation from the temporal variation (e.g., Ryan and Heyward 2003). Adaptive designs do not provide the same opportunity, as inclusion probabilities in the second sampling bout would be based on values observed in the first bout. There may be cases where the gain in efficiency from adaptive sampling within each sampling episode is greater than the loss of power due to confounding temporal variation with spatial variation, but that need not be the case in general. It might be possible to implement ACS in a form of rotating panel design, where in subsequent years both previously detected clusters and new primary sample units are sampled (Jim Nichols, *personal communication*). Additionally, monitoring efforts for populations small enough to require adaptive cluster sampling are likely to require information on demographic transitions as well (e.g., Doak 2005), so the premium of detecting more plants may be important.

When is adaptive cluster sampling more efficient than other approaches?

The usual approach to maximizing sampling efficiency is to consider sampling in isolation, assign costs to initial setup, travel among quadrats, and inspection of each quadrat (often a function of both quadrat size and the number of individuals found), and then minimize the variance of estimators for a given total effort. For adaptive designs, the total effort is not known until the sample is drawn. Brown and Manly (1998) showed that simple stopping rules that could stabilize the total effort (e.g., continue drawing initial sample quadrats until the total effort [initial plus network plus edge quadrats] exceeds 200) preclude an unbiased estimator, although Christman and Lan (2001) developed an unbiased estimator for adaptive cluster sampling based on an approximately fixed number of occupied (meeting the threshold for inclusion) quadrats. Therefore, equal-effort comparisons between sampling designs are difficult, and only rough guidance is possible.

The general pattern is that ACS is efficient for rare, relatively tight clusters of individuals, where the spatial extent of the population is compact enough that feasible sampling effort will allow initial sampling to intersect several clusters. Conversely, alternatives are more efficient if clusters are diffuse, or if spatial variation in abundance is driven by habitat attributes at spatial scales larger than restricted dispersal. Christman (1997) compared the efficiency of adaptive cluster sampling to balanced sampling excluding contiguous units (BSEC), and found that adaptive cluster sampling was more efficient for rare and highly clustered populations, but BSEC was more efficient over a broader range of parameters she considered—moderately common and diffuse clusters. (Brown 1999) compared adaptive cluster sampling to two-phase adaptive sampling—breaking the population into regional strata, sampling within each stratum, then performing a second sampling stage in those strata with high densities in the primary stage. She found the two-phase adaptive sampling more efficient when there were more than 20 clusters, and when the clusters were diffuse. Christman (2000*b*) reviewed and compared several designs for quadrat sampling of rare, clustered plants and found that stratification (e.g., by habitat variables) that confined individuals to a single small stratum was the most efficient design if such habitat attributes were available.

For the case of a low-abundance population spread over a large spatial extent, even if plants are tightly clustered, ACS alone will not succeed, as it requires sufficient initial sampling to intersect several clusters. ACS would not help for *Aletris* in the western Singeltary population, where three individuals are known in 0.5 km^2, and, at most, 500 1-m^2 quadrats could be sampled (0.1%). Unless there are a large number of clusters, the chances of intersecting even one cluster are extremely small. Some form of model-based stratification is required to winnow the search area down to a size where several percent of the area can be sampled. At that point, decisions could be made among the various sampling designs.

Finally, for estimation of abundances of rare plants within populations, efficiency of the estimation cannot be considered in isolation. Plants need to be found, mapped, and measured for viability analyses (Doak 2005), and most of that effort will occur whether or not the total population size is estimated. Therefore, adaptive methods such as adaptive cluster sampling (or design-based stratification) that yield larger fractions of individuals included in the sample may provide the most efficient unbiased estimates of population sizes from the perspective of the combined investigation of that population.

Acknowledgments

This manuscript benefited from comments by Jenny Richards, Jim Nichols, Aaron Ellison, and an anonymous reviewer. *Aletris bracteata* in Everglades National Park is being studied by Wendy Wilsdon of FIU under NPS research permit EVER2003SCI0082 and Florida Department of Agriculture and Consumer Affairs permit 564, both issued to Jennifer Richards. FIU graduate students in my workshop on rare plant sampling spent one beautiful Saturday afternoon sampling *A. bracteata* for this example. C. Galen sparked my interest in quadrat sampling of rare plants 20 years ago.

Literature Cited

Acharya, B., G. Bhattarai, A. de Gier, and A. Stein. 2000. Systematic adaptive cluster sampling for the assessment of rare tree species in Nepal. Forest Ecology and Management **137**:65–73.

Brown, J. A. 1999. Comparison of two adaptive sampling designs. Australian and New Zealand Journal of Statistics **41**:395–403.

Brown, J. A. 2003. Designing an efficient adaptive cluster sample. Environmental and Ecological Statistics **10**:95–105.

Brown, J. A., and B. J. F. Manly. 1998. Restricted adaptive cluster sampling. Environmental and Ecological Statistics **5**:49–63.

Caswell, H. 2001. Matrix population models. Second edition. Sinauer Associates, Sunderland, Massachusetts, USA.

Chao, C. T. 2003. Markov chain Monte Carlo on optimal adaptive sampling selections. Environmental and Ecological Statistics **10**:129–151.

Christman, M. C. 1997. Efficiency of some sampling designs for spatially clustered populations. Environmetrics **8**:145–166.

Christman, M. C. 2000*a*. Bootstrap confidence intervals for adaptive cluster sampling. Biometrics **56**:503–510.

Christman, M. C. 2000*b*. A review of quadrat-based sampling of rare, geographically clustered populations. Journal of Agricultural Biological and Environmental Statistics **5**: 168–201.

Christman, M. C. 2003. Adaptive two-stage one-per-stratum sampling. Environmental and Ecological Statistics **10**:43–60.

Christman, M. C., and F. Lan. 2001. Inverse adaptive cluster sampling. Biometrics **57**(4):1096–1105.

Conners, M. E., and S. J. Schwager. 2002. The use of adaptive cluster sampling for hydroacoustic surveys. ICES Journal of Marine Science **59**:1314–1325.

Di Battista, T. 2003. Resampling methods for estimating dispersion indices in random and adaptive designs. Environmental and Ecological Statistics **10**:83–93.

Doak, D. F., K. Gross, and W. F. Morris. 2005. Understanding and predicting the effects of sparse data on demographic analyses. Ecology **86**:1154–1163.

Doren, R. F., and A. Ferriter. 2002. Weeds won't wait: the strategic plan for managing Florida's invasive exotic plants. South Florida Ecosystem Restoration Task Force and Working Group, Miami, Florida, USA.

Dryver, A. L. 2003. Performance of adaptive cluster sampling estimators in a multivariate setting. Environmental and Ecological Statistics **10**:107–113.

Edwards, T. C., D. R. Cutler, N. E. Zimmerman, L. Geiser, and J. Alegria. 2005. Model-based stratifications for enhancing the detection of rare ecological events. Ecology **86**:1081–1090.

Elzinga, C. L., D. W. Salzer, and J. W. Willoughby. 1998. Measuring and monitoring plant populations. BLM Technical Reference 1730-1. BLM/RS/ST-98/005+1730. ⟨http://www.blm.gov/nstc/library/techref.htm⟩

Evans, T. D., and O. V. Viengkham. 2001. Inventory time-cost and statistical power: a case study of a Lao rattan. Forest Ecology and Management **150**:313–322.

Felix-Medina, M. H. 2003. Asymptotics in adaptive cluster sampling. Environmental and Ecological Statistics **10**:61–82.

Gaston, K. J. 1994. Rarity. Chapman and Hall, London, UK.

Gilbert, N., and S. B. Lee. 1980. Two perils of plant population dynamics. Oecologia **46**:283–284.

Greig-Smith, P. 1983. Quantitative plant ecology. Third edition. University of California Press, Berkeley, California, USA.

Hanselman, D. H., T. J. Quinn, C. Lunsford, J. Heifetz, and D. Clausen. 2003. Applications in adaptive cluster sampling of Gulf of Alaska rockfish. Fishery Bulletin **101**:501–513.

Ishwar, N. M., R. Chellam, and A. Kumar. 2001. Distribution of forest floor reptiles in the rainforest of Kalakad-Mundanthurai Tiger Reserve, South India. Current Science **80**: 413–418.

Keith, B. D. A. 2000. Sampling designs, field techniques and analytical methods for systematic plant population surveys. Ecological Management and Restoration **1**:125–139.

Kery, M. 2004. Extinction rate. estimates for plant populations in revisitation studies: importance of detectability. Conservation Biology **18**:570–574.

Kery, M., and K. B. Gregg. 2003. Effects of life-state on detectability in a demographic study of the terrestrial orchid *Cleistes bifaria*. Journal of Ecology **91**:265–273.

Krebs, C. J. 1999. Ecological methodology. Second edition. Benjamin/Cummings, Menlo Park, California, USA.

Lo, N. C. H., D. Griffith, and J. R. Hunter. 1997. Using a restricted adaptive cluster sampling to estimate Pacific hake larval abundance. California Cooperative Oceanic Fisheries Investigations Reports **38**:103–113.

MacKenzie, D. I., J. D. Nichols, N. Sutton, K. Kawanishi, and L. L. Bailey. 2005. Improving inferences in population studies of rare species that are detected imperfectly. Ecology **86**:1101–1113.

Morris, W. F., and D. F. Doak. 2002. Quantitative conservation biology: theory and practice of population viability analysis. Sinauer Associates, Sunderland, Massachusetts, USA.

Philippi, T., B. Collins, S. Guisti, and P. M. Dixon. 2001. A multistage approach to population monitoring for rare plant populations. Natural Areas Journal **21**:111–116.

Pielou, E. C. 1957. The effect of quadrat size on the estimation of the parameters of Neyman's and Thomas's distributions. Journal of Ecology **45**:31–47.

Pontius, J. S. 1997. Strip adaptive cluster sampling: probability proportional to size selection of primary units. Biometrics **53**:1092–1096.

Pyke, G. H. 1984. Optimal foraging theory: a critical review. Annual Review of Ecology and Systematics **15**:523–575.

Ryan, D. A. J., and A. Heyward. 2003. Improving the precision of longitudinal ecological surveys using precisely defined observational units. Environmetrics **14**:283–293.

Salehi, M. 2003. Comparison between Hansen-Hurwitz and Horvitz-Thompson estimators for adaptive cluster sampling. Environmental and Ecological Statistics **10**:115–127.

Shefferson, R. P., B. K. Sandercock, J. Proper, and S. R. Beissinger. 2001. Estimating dormancy and survival of a rare herbaceous perennial using mark–recapture models. Ecology **82**:145–156.

Smith, D. R., M. J. Conroy, and D. H. Brakhage. 1995. Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. Biometrics **51**:777–788.

Smith, D. R., R. F. Villella, and D. P. Lemarie. 2003. Application of adaptive cluster sampling to low-density populations of freshwater mussels. Environmental and Ecological Statistics **10**:7–15.

Soule, M. E. 1986. Viable populations. Cambridge University Press, Cambridge, Massachusetts, USA.

Su, Z. M., and T. J. Quinn. 2003. Estimator bias and efficiency for adaptive cluster sampling with order statistics and a stopping rule. Environmental and Ecological Statistics **10**: 17–41.

Thompson, S. K. 1990. Adaptive cluster sampling. Journal of the American Statistical Association **85**:1050–1059.

Thompson, S. K. 1991*a*. Adaptive cluster sampling: designs with primary and secondary units. Biometrics **47**:1103–1115.

Thompson, S. K. 1991*b*. Stratified adaptive cluster sampling. Biometrika **78**:389–397.

Thompson, S. K. 2002. Sampling. Second edition. Wiley, New York, New York, USA.

Thompson, S. K. 2003. Editorial: special issue on adaptive sampling. Environmental and Ecological Statistics **10**:5–6.

Thompson, S. K., F. L. Ramsey, and G. A. Seber. 1992. An adaptive procedure for sampling animal populations. Biometrics **48**:1195–1199.

Thompson, S. K., and G. A. F. Seber. 1994. Detectability in conventional and adaptive sampling. Biometrics **50**:712–724.

Thompson, S. K., and G. A. F. Seber. 1996. Adaptive sampling. John Wiley and Sons, New York, New York, USA.

Vasudevan, K., A. Kumar, and R. Chellam. 2001. Structure and composition of rainforest floor amphibian communities in Kalakad-Mundanthurai Tiger Reserve. Current Science **80**:406–412.

SUPPLEMENT

A SAS macro for adaptive cluster sampling and *Aletris* data sets from the example are available in ESA's Electronic Data Archive: *Ecological Archives* E086-058-S1.

Ecology, 86(5), 2005, pp. 1101–1113

IMPROVING INFERENCES IN POPULATION STUDIES OF RARE SPECIES THAT ARE DETECTED IMPERFECTLY

Darryl I. MacKenzie,[1,6] James. D. Nichols,[2] Nicole Sutton,[3] Kae Kawanishi,[4]
and Larissa L. Bailey[2,5]

[1]*Proteus Wildlife Research Consultants, P.O. Box 5193, Dunedin, New Zealand*
[2]*U.S. Geological Survey, Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, Maryland 20708-4017 USA*
[3]*Biological Sciences, University of Waikato, Private Bag 3105, Hamilton, New Zealand*
[4]*Department of Wildlife and National Parks Peninsular Malaysia, KM 10, Jalan Cheras, 56100, Kuala Lumpur, Malaysia*
[5]*U.S. Geological Survey, Cooperative Fish and Wildlife Research Unit, Department of Zoology,*
North Carolina State University, Campus Box 7617, Raleigh, North Carolina 26795-7617 USA

Abstract. For the vast majority of cases, it is highly unlikely that all the individuals of a population will be encountered during a study. Furthermore, it is unlikely that a constant fraction of the population is encountered over times, locations, or species to be compared. Hence, simple counts usually will not be good indices of population size. We recommend that detection probabilities (the probability of including an individual in a count) be estimated and incorporated into inference procedures. However, most techniques for estimating detection probability require moderate sample sizes, which may not be achievable when studying rare species. In order to improve the reliability of inferences from studies of rare species, we suggest two general approaches that researchers may wish to consider that incorporate the concept of imperfect detectability: (1) borrowing information about detectability or the other quantities of interest from other times, places, or species; and (2) using state variables other than abundance (e.g., species richness and occupancy). We illustrate these suggestions with examples and discuss the relative benefits and drawbacks of each approach.

Key words: *abundance; detection probability; occupancy; rare species; species richness.*

Introduction

Estimation of abundance and related quantities for animal populations requires that investigators deal with two important sources of variation in resulting count data (Skalski and Robson 1992, Lancia et al. 1994, Thompson et al. 1998, Yoccoz et al. 2001, Pollock et al. 2002, Williams et al. 2002). The first source of variation involves spatial sampling. Investigators are frequently interested in inferences about areas so large that they cannot survey or count animals over the entire areas of interest. This problem requires that investigators select a sample of smaller areas on which to conduct survey efforts and that this selection be conducted in a manner that permits inference about the entire area of interest. That is, sample locations must be selected such that counts on these areas can be used to draw inferences about locations that do not appear in the sample (i.e., by using a probabilistic sampling scheme). This problem is not unique to the sampling of animal and plant populations, and ecologists and wildlife managers can use the various sampling designs developed by statisticians in other contexts (e.g., Cochran 1977, Thompson 1992).

The second source of variation involves the issue of detectability and the fact that even when efforts are made to count animals on sample areas, it is extremely unlikely that the investigator will count every animal in any such area. Instead, the investigator obtains count statistics that represent unknown fractions of animals present in the sampled area. Count statistics, C, reflect the number of animals observed, captured, heard, harvested or otherwise detected by one or more survey methods used by animal ecologists. Count statistics are best viewed as random variables with expectation equal to the product of the true number of animals (N) present in the sample location and the detection probability (p) associated with the count:

$$E(C) = pN. \tag{1}$$

Inference about N requires inference about, or knowledge of, p. For example, estimation of abundance is accomplished by

$$\hat{N} = \frac{C}{\hat{p}} \tag{2}$$

where $\hat{p}$ is the estimated detection probability associated with the specific count statistic. We note that the seemingly diverse abundance and density estimation methods appearing in books such as Seber (1982) and Williams et al. (2002) represent different ways of es-

---

Manuscript received 5 July 2004; accepted 26 July 2004. Corresponding Editor: A. M. Ellison. For reprints of this Special Feature, see footnote 1, p. 1079.

[6] E-mail: Darryl@proteus.co.nz

timating detection probability, but that the final step in abundance estimation involves Eq. 2.

We note that in addition to abundance, estimation of demographic rate parameters also requires consideration of detection probability. For example, reproductive rate is often estimated using age ratios, representing relative abundances of young to adult or young to adult females. If inferences about age ratios are based on raw counts, neglecting detection probabilities, then resulting ratios will reflect a confounded function of true relative abundances and relative detection probabilities (MacKenzie and Kendall 2002, Williams et al. 2002). Estimates of survival and movement probabilities are typically based on marked individuals released at one time and location and estimated to be alive at some later time in the same or a different location. If time- and location-specific detection probabilities are not incorporated into estimation models, then biased estimates of survival and movement can result (Williams et al. 2002).

Rarity will typically increase the difficulties in dealing with both spatial sampling and detectability in the estimation of abundance and related parameters. A species may be considered ''rare'' either because it occurs at very low densities across a broad range, or it could be locally abundant but not widely distributed across the landscape, with the consequence that the species may only occupy a small fraction of spatial sampling units (Gaston 1994). Rarity is also likely to be associated with low probabilities of detecting individuals even in sampling units that are occupied. When detection probabilities are known and not estimated then for a fixed count C, $\widehat{\mathrm{var}}(\hat{N})$ will be larger when p is small (as $\widehat{\mathrm{var}}(\hat{N}) = \widehat{\mathrm{var}}(C)/p^2$; from Eq. 2). Furthermore, detection probability is seldom known, and most methods for estimating p use the count data themselves (e.g., Seber 1982, Williams et al. 2002). Small sample sizes tend to produce large $\widehat{\mathrm{var}}(\hat{p})$ and, hence large $\widehat{\mathrm{var}}(\hat{N})$. Thus, observation-based methods such as distance sampling, multiple-observers, time at detection or temporal removal, and marked subsamples, as well as capture-based methods such as capture–recapture, catch-effort, and change-in-ratio, all perform best with sample sizes that frequently exceed those that can be obtained for rare species (see Williams et al. 2002, and references therein, for details of these various techniques).

There may also be other more practical impediments to estimating abundance for rare species. In order to obtain an accurate count of the number of unique individuals, it must be possible to identify individual animals. Capture–recapture or resighting methods require that individuals detected during one sample period can be recognized as such if they are detected in a subsequent period. If animals can not be reliably identified using natural colorations or patternings, then some form of mark will need to be applied to the animal, which may require animals to be captured and handled. However many rare species are listed as a protected or endangered and in some instances there may be resistance (from a variety of quarters) to handling and marking animals if there is the potential for negative impact on the animals.

In this paper, we focus on the detectability issue and consider approaches to obtaining reliable inferences about animal populations and communities when dealing with rare species and with detection probabilities <1. We discuss two general approaches. First, we briefly present our views on the concept of borrowing information about detectability and relevant state variables from other times, places, and classes of animals (even species). The other approach involves consideration of state variables other than abundance to characterize status and changes in the population. We believe that both approaches hold promise for permitting reasonable inferences about populations and communities of rare species.

BORROWING INFORMATION

In some situations, it may be appropriate to share or borrow information about population parameters for rare species from multiple data sources. The general concept is that by combining the data, where appropriate, more precise estimates of the parameters may be obtained. We describe two situations that are distinguished by the similarity of the data being combined. In one situation, data of exactly the same type are aggregated over times, locations, individuals, or species. In the second situation, different kinds of data are combined to yield inferences about a common parameter(s).

The first situation is exemplified by capture–recapture modeling, where it has become common to consider aggregation of capture history data over times, places and individuals (e.g., Seber 1982, Lebreton et al. 1992, Williams et al. 2002). The decision about the appropriate level of aggregation is considered as a problem in model selection, where the task is to select a level that describes the data reasonably well, with the smallest number of parameters and greatest degree of aggregation (Burnham and Anderson 2002). The data are considered jointly, and the model set includes general models with separate parameters for different times and locations, for example, and reduced parameter models in which parameters are constrained to be equal over times and/or locations. The competing models thus represent different levels of aggregation and different degrees to which data are shared. We note that there are other approaches to aggregation than the described approach of equating model parameters (assuming homogeneity) over times, places, and individuals. For example, modeling parameters as function of covariates represents a form of aggregation in which a common relationship (over times, places, individuals, etc.) is assumed between a covariate and the parameter of interest (for closed capture–recapture models, see Pollock et al. [1984], Huggins [1989, 1991], Alho [1990]). Another approach is to assume a distributional form for variation in a parameter of interest over individuals, for example, and then use a hierarchical modeling approach to estimate parameters of that distribution for an aggregation of individuals (e.g., Dorazio and Royle 2003). Both covariate modeling and hierarchical modeling assume some common form or relationship for the variation in a parameter over individuals, or times, or places, and the task is then to estimate the parameters that describe that form or relationship. As for the case of equality constraints on parameters, model selection approaches are needed to ascertain the appropriateness of different distributions or relationships linking the different data sets.

This general approach to sharing information is relatively common and was used by Boyce et al. (2001) to estimate the number of female grizzly bears (*Ursus arctos*) with cubs-of-the-year in the Yellowstone ecosystem. Capture frequency data from 1986–1998 were fit to truncated negative binomial distributions with various assumptions about constancy of parameters over time. Model selection results favored a model with a heterogeneity in ''sightability'' parameter constant over time, so that information across years was borrowed to yield realistic, annual estimates of abundance. Franklin et al. (2004) conducted capture–recapture studies of California Spotted Owls (*Strix occidentalis occidentalis*) at four sites in the Sierra Nevada mountain range. They conducted a meta-analysis in which they investigated the plausibility of additive location + time models for rate of population increase. The a priori hypothesis was that the different sites might experience different average rates of population increase, but that the sites were exposed to similar environmental conditions and should thus exhibit parallel changes in population size. Indeed, a model with additive location + time effects was appropriate for the data, leading to the interesting biological inference about similar environmental effects and also permitting more precise estimation of rate of population increase than was possible for any single location.

A form of aggregation that is not used very commonly by ecologists involves sharing information across species. For example, Nichols et al. (2000) used double-observer models to estimate detection probability and abundance from avian point count data. They grouped species a priori based on hypothesized similarity of detection probabilities. These models appeared to work well and were used to estimate abundance for relatively rare species for which data were inadequate to compute species-specific estimates of detection probability. M. W. Alldredge, K. H. Pollock, T. R. Simons, and S. A. Shriner (*unpublished manuscript*) have greatly expanded on this idea and have developed sets of models for avian point count data that incorporate various assumptions about similarity of species groups identified *a priori*. They present a unified model selection framework similar to that now used for aggregating over time and space. We believe that this approach of borrowing information from species believed to exhibit similarities in detection or population dynamics will prove to be very useful for rare species.

The second situation in which information is borrowed uses different types of data within a modeling framework that contains shared parameters. Dixon et al. (2005) use such an approach to improve estimates of capture efficiency of wasps by the insectivorous pitcher plant, *Darlingtonia californica*. They had two forms of available data, both of which contained some information about the capture rate of wasps by the pitcher plant. One data set permitted direct estimation of capture efficiency, whereas the other permitted estimation of a function of capture efficiency and visitation rate. Combination of the two data sets permitted use of the information about capture efficiency encoded in the second data set, and thus resulted in more precise estimates. Other examples of this type of approach to borrowing information include recent work on integrating animal count data with capture-recapture data (Besbeas et al. 2002, 2003). Information about survival rate comes from the data on marked animals, whereas information about population growth rate comes from the count data. Population growth rate can also be computed using survival and reproductive rates with a simple population projection model. Thus an integrated estimation approach incorporating the two data types with a population model permits estimation of reproductive rate in addition to population growth and survival (Besbeas et al. 2002, 2003). We expect such efforts to combine data from multiple sources to become more common in the near future (see Schnute 1994, Gallucci et al. 1996, Quinn and Deriso 1999, Elliott and Little 2000, Millar and Meyer 2000, Trenkel et al. 2000, Grove et al. 2002, White and Lubow 2002).

State Variables Other Than Abundance

As indicated above, despite the large number of potential methods for estimating detection probability of individual animals, and hence their abundance and density, rarity will typically translate into small sample sizes and imprecise estimates of relevant parameters. One approach to sample size problems using existing methods is to increase sampling effort. This approach usually increases expenses but is sometimes feasible nonetheless. As another approach, we suggest that, in many cases, it may be appropriate to consider alternative state variables: (1) the number of species present within the area (species richness; perhaps within some taxonomic or other group of interest), or (2) the proportion of the area occupied by a single species (occupancy). As with abundance, we believe that these are relevant state variables that provide important information about systems of ecological and conservation interest. For example, species richness conveys information about community structure and biodiversity, while occupancy provides information about species range, likelihood of extinction (e.g., Lande 1988), metapopulation dynamics (e.g., Hanski 1999), and abundance. In addition, vital rates associated with dynamics of these state variables can be identified and estimated. The sampling effort required to estimate these two variables and their associated vital rates will typically be smaller (often substantially so) than that required to estimate abundance and rates of birth, death, and movement. Below we expand upon these concepts, provide a brief overview of potential analytical methods and give examples where these alternative state variables have been used in practice.

Species richness

Species richness (the number of species inhabiting a predefined area) is a state variable representing an alternative to abundance that may be useful in studies of groups of rare species. This state variable is aimed more at the community level than at a single species, and is widely used in ecological investigations and in conservation initiatives (e.g., Barbault and Hochberg 1992, Scott et al. 1993, Mangel et al. 1996, Boulinier et al. 1998, 2001, Cam et al. 2002, Doherty et al. 2003*a*, *b*). As in most cases in animal sampling, it is unlikely that all species at a sampling location will be observed. Even if the investigator is simply interested in relative species richness (e.g., computed as the ratio of richness at points in time and/or space), it is unlikely that raw species counts will be useful, as the fraction of any species pool that is observed will likely vary according to such factors as habitat and differences in the local species composition. The problem is then how to estimate species richness in the face of imperfect detectability. Although the bulk of the work on this topic has been relatively recent, the estimation of species richness has been considered many times over the last decade, including reviews by Bunge and Fitzpatrick (1993), Colwell and Coddington (1994), Nichols and Conroy (1996), Williams et al. (2002), and new synthetic contributions by Dorazio and Royle (2005) and Mao and Colwell (2005). Therefore, here we simply provide a brief overview of the various sampling situations that permit estimation of species richness, with reference to key publications that can be sought for further details of the particular methods. We also note that the concept of estimating species richness could be considered as a form of borrowing information: inferences about rare and infrequently encountered species are made by borrowing information about detectability from the other species in the community.

One sampling approach is where a single (or small number) of locations are surveyed multiple times within a relatively short time period. At each survey a list is maintained of the species detected, hence at the conclusion of the sampling it is possible to construct a ‘‘capture history’’ for each species, denoting whether the species was detected within each survey. The sampling should be conducted over a reasonably short time period to avoid any potential change in species composition, i.e., the community is assumed to be closed to any species additions or deletions. A species is considered to be analogous to an individual in a single-species, closed population capture–recapture study. Based upon the detection histories of species detected at least once, the number of species that were never detected can be estimated, i.e., an estimate of the total number of species at that location can be obtained. There is a wide range of capture–recapture models that could be used (e.g., Otis et al. 1978, Williams et al. 2002), although generally we would suggest that models permitting heterogeneous detection probabilities will be needed because of the variation in abundances and individual detection probabilities associated with different species.

Another sampling approach involves geographic replication rather than temporal replication. An area of interest is defined, and interest is focused on the community of species (perhaps of a certain taxon or functional guild) associated with that area. Sample locations are randomly selected from the area of interest, and the investigator surveys each such location only once. A capture history for each species can then be formed by denoting whether the species was encountered at each location, and species richness can be estimated as above. The key difference between the two sampling approaches is that one uses temporal replication of surveys, while the other uses spatial replication. Naturally this has some influence on how the results are interpreted, but the mechanics of the estimation procedure are identical.

A third general design is one where the investigator has both temporal and spatial replication. The usual approach to estimation with such data involves two steps. Species richness is estimated for each sampling site using the temporal replication, and then means and other summary statistics are computed for all sites within strata of interest. However, a more integrated approach is to consider prior distributions of individuals and species over space and to estimate parameters of this spatial distribution (see Dorazio and Royle 2005).

A final sampling design for richness estimation involves the so-called empirical species abundance distribution. Sampling is conducted at a single location during a single short time interval so there is neither temporal nor geographic replication. However, the investigator now records the number of individuals for each detected species. So the data used for estimation are the so-called detection frequencies, f_i, indicating the number of species for which exactly one individual was detected (f_1), exactly two individuals (f_2), etc. These data can then be used with a limiting form of the jackknife estimator developed by Burnham and Overton (1979), or other appropriate methods (e.g., Norris and Pollock 1998, or see Bunge and Fitzpatrick 1993).

The previously cited reviews deal almost exclusively with estimation of species richness, itself, and there has been much less attention devoted to estimation of the vital rates influencing this state variable. Initial approaches to estimate local species extinction probability, number of local colonizing species, colonization rate, and species turnover have been developed by Nichols et al. (1998*a*, *b*; also see Williams et al. 2002). These methods have been used to draw inferences about community dynamics of forest birds exposed to forest fragmentation (Boulinier 1998, 2001), avian community dynamics at the edge and center of species ranges (Doherty et al. 2003*b*), and the relationship between avian community dynamics and sexual dichromatism (Doherty et al. 2003*a*). These analyses are especially relevant to the issue of rare species, as many of the species included in these different analyses were characterized by data that would have been inadequate for single-species inferences.

Finally, we note that an interesting quantity in some community studies is the fraction of the members of some identified species pool that are present in a specific community or location of interest. This problem was considered by Cam et al. (2000; also see Nichols 1998*b*, Chao et al. 2000). Here, we note that there are similarities between this problem and the estimation of the occupancy state variable discussed below, where the intent is to estimate the fraction of sites at which a species is present. By considering each species in the pool (or each species on any a priori reference list) as a ''site,'' the mechanics of the two problems are identical: the nondetection in the sampled community of a species on the list does not imply that the species was absent. Thus, we believe that some of the occupancy estimation and modeling described below (e.g., MacKenzie et al. 2002, 2003, Royle and Nichols 2003) may prove useful in community studies as well.

Occupancy

Occupancy has been used as an alternative state variable to abundance for many inferential purposes ranging from questions about habitat selection (e.g., Reunanen et al. 2002, Scott et al. 2002, Bradford et al. 2003) to questions about population dynamics and distribution (e.g., Hames et al. 2001, Barbraud et al. 2003, Martinez-Solano et al. 2003). Occupancy data are recognized to be especially useful for the study of rare species, although the historical emphasis for such species has been on use of occupancy as an index to abundance, rather than on occupancy as a state variable of interest in its own right (e.g., Diefenbach et al. 1994). When occupancy is viewed as the state variable of interest, focus shifts from ''how many individuals of the species are located at various locations across the landscape?'' to ''what fraction of the landscape does the species inhabit?'' For most species, the two questions should be closely related, as occupancy is simply the proportion of the landscape for which the local abundance distribution for the species of interest is >0. For territorial species, the number of individuals within an area will frequently be directly proportional to the fraction of the area occupied by the species. However, in other situations changes in abundance do not always result in a change in occupancy or species range (i.e., there is only a change in the density of the species). In such instances, occupancy may not be a useful surrogate state variable for abundance. However, we note that for questions dealing with topics such as metapopulation ecology (e.g., Hanski 1999) or geographic range (Brown 1995, Wikle 2003), occupancy is the state variable of primary interest regardless of the relationship between occupancy and abundance.

Occupancy is usually measured as the proportion of the area of interest where the species is present during the sampling. We use the term ''area'' in the sense of a statistical population, namely a collection of all the possible sampling units that we wish to make inference about. The sampling units themselves may represent naturally occurring discrete habitat patches (e.g., remnant patches of forest, islands, or ponds), or arbitrarily defined quadrats (henceforth we shall generically refer to sampling units as sites). The intent is to then survey a fraction of the total area (using an appropriate statistical method for selecting which sites to survey) and determine the presence or absence of the species at each site. However for most species, the investigator can never confirm species absence. There will generally be a non-negligible chance that the species was actually present, but due to chance, was undetected by the survey techniques. This issue of imperfect detectability (as in the case of abundance estimation) must be accounted for if robust inferences about occupancy are to be made. Comparisons of uncorrected, or naïve, estimates of occupancy can be strongly influenced by changes in species detectability, possibly leading to incorrect inferences about changes in occupancy. An observed difference between two naïve occupancy estimates may be due to differences in our ability to detect the species at the two times or places, rather than to a true difference in occupancy. In fact, Moilanen (2002) recently investigated the effect of various assumption violations with respect to metapopulation incidence functions (Hanski 1992, 1999) and found that changes in species detectability were the greatest contributor of bias to the estimated functions.

This issue of imperfect detectability has long been recognized. In an effort to minimize the probability of declaring the species falsely absent, many studies and monitoring programs for rare species conduct repeated surveys of sites over a reasonable timeframe, during which no changes in the occupancy states of sties are thought to occur. With this type of information (repeated surveys) it is possible to account for detection probabilities and obtain corrected, or unbiased, estimates of occupancy.

The development of analytical methods for estimating occupancy at a single point in time has been sporadic over the last 20 years (Geissler and Fuller 1987, Azuma et al. 1990), but more recently there have been a number of similar methods published (MacKenzie et al. 2002, Royle and Nichols 2003, Tyre et al. 2003, Stauffer et al. 2004, Wintle et al. 2004). Here, we shall briefly review the method presented by MacKenzie et al. (2002), of which Stauffer et al. (2004), Tyre et al. (2003) and Wintle et al. (2004) could be considered special cases. Royle and Nichols (2003) have developed a useful extension that allows for heterogeneity in detection probability between sites caused by differences in local abundance.

Despite this interest in occupancy estimation, until recently, few methods have been developed for estimating changes in occupancy, and the vital rates (local probabilities of extinction and colonization) that produce such changes, that explicitly account for imperfect detectability. Often in metapopulation studies, changes caused by local extinctions and colonizations of patches by the target species are of primary interest (e.g., Hanski 1992, 1999). Only recently has any attempt been made to allow for the possibility of false absences (Moilanen 2002), with arguable success. As noted above, Moilanen (2002) found that issues related to species detectability were the major contributor to bias in estimated incidence functions. Two recent papers (Barbraud et al. 2003, MacKenzie et al. 2003) use similar approaches to estimate local extinction and colonization probabilities directly, whilst allowing for imperfect detection. Given that the data have been collected from monitoring sites over several seasons or years, and within each season sites are surveyed more than once, probabilistic arguments are applied to form a model likelihood that can be used to obtain parameter estimates. The main difference between the two methods is that the approach of Barbraud et al. (2003) is conditional upon the first occasion at which the target species is detected at a site, whereas the approach of MacKenzie et al. (2003) is unconditional, allowing the proportion of sites occupied by the species each season to be estimated.

A general sampling scheme

From the area of interest, U sites are selected to be surveyed for the species using appropriate methods. These sites are surveyed multiple times each season for T seasons (e.g., years). In each survey, detections are regarded as truth (the species is never falsely recorded as being present), but nondetections are recognized as arising from both (1) true absence and (2) presence with nondetection (sometimes referred to as false absences). Within a season, sites are closed to changes in occupancy (i.e., sites are either always occupied or unoccupied by the species). This assumption may be relaxed provided that any changes occur completely at random in which case ‘‘occupancy’’ should be interpreted as ‘‘use.’’ Sites are repeatedly surveyed (possibly an unequal number of times), with the species either being detected or not detected. Between seasons, changes in occupancy may occur which are referred to as the processes of colonization and local extinction.

The sequence of detections and nondetections recorded at the sites form a detection history ($\mathbf{H}_i$ for site i). For example $\mathbf{H}_i = 101\,000$ represents data collected over two seasons, each with three surveys per season. In the first season, the species was detected in the first and third survey, but undetected in the second. In the second season, the species was never detected at the site during the surveys. The basic procedure for building a model is to develop a verbal description of the observed data, then translate that description into a mathematical equation representing the probability of observing the data using the defined model parameters.

A single season model

MacKenzie et al. (2002) define ψ as the probability that a site is occupied by the species, and p_j is the probability of detecting the species (given presence) during the jth independent survey of a site. To illustrate model construction, consider the detection history 1001. A verbal description of this history would be that the species is present (as it was detected at least once), and it was then detected in the first and fourth surveys of the site, but not in surveys two and three. Translating this into a mathematical equation using the model parameters, the probability of observing this history could be expressed as

$$\Pr(\mathbf{H}_i = 1001) = \psi p_1(1 - p_2)(1 - p_3)p_4.$$

A similar expression can be obtained for detection histories from all sites at which the species is detected at least once. However, for sites where the species was never detected (e.g., $\mathbf{H}_i = 0000$), there are two possibilities that must be accounted for. The verbal description for this history would be, either the species was present but went undetected in the surveys (which has a probability of $\psi \prod_{j=1}^{4} (1 - p_i)$), or the species was genuinely absent from the site (with probability $(1 - \psi)$). As either reason for never detecting the species at the site is possible, the probability of observing this history is the sum of these two components, i.e.,

$$\Pr(\mathbf{H}_i = 0000) = \psi \prod_{j=1}^{4} (1 - p_j) + (1 - \psi).$$

Assuming that the detection histories for all sites are independent, the likelihood function for an estimation model takes the usual form:

$$L(\psi, p \mid \mathbf{H}_1, \ldots, \mathbf{H}_U) = \prod_{i=1}^{U} \Pr(\mathbf{H}_i).$$

Note that this requires both the probability of occupancy and each detection probability to be constant across all U sites. If this not the case, then the probabilities are said to be heterogeneous. While it is not possible to estimate site-specific occupancy probabilities, the probabilities may be modeled as a function of measured covariates using the logistic equation

$$\theta_i = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}$$

where θ_i is the probability of interest for site i, $\mathbf{X}_i$ is the row vector of covariate information for site i and $\boldsymbol{\beta}$ is the column vector of coefficients to be estimated. In general, occupancy may be a function of site-specific covariates that are constant throughout the season (e.g., habitat type), while detection probabilities may also be a function of covariates that change through the season (e.g., weather conditions).

However one potential source of heterogeneity that cannot be easily accommodated directly by a covariate is heterogeneity in detection probabilities caused by differences in the species abundance between sites. Royle and Nichols (2003) extend the above approach by arguing that the species is not detected only if none of the N_i individuals at site i are detected. This suggests the probability of detection can be written as a function of abundance, i.e., $p_{ij} = 1 - (1 - r_j)^{N_i}$, where r_i is the probability of detecting an individual of the species in the jth survey. As the N_i are unknown, this approach requires that a distribution for how N_i varies across sites be specified. Estimation under this model permits estimation of occupancy in the face of abundance-related heterogeneity.

An additional comment on the method for estimating occupancy detailed by MacKenzie et al. (2002) is that an equal number of surveys across all sites is not required. By having the ability to deal with ''missing observations,'' MacKenzie et al. (2002) developed a very flexible method that could be used to analyze data from a number of different designs that could be used in practice (e.g., repeatedly surveying only a subset of all sites or surveying only until the species is first detected). Finally, we note that recent work on single-season occupancy modeling includes development of a goodness-of-fit test for the above model (MacKenzie and Bailey 2004) and an approach for modeling and estimating possible dependencies in occupancy patterns (and detection probabilities) of multiple species (MacKenzie et al. 2004).

A multiple season model

MacKenzie et al. (2003) extended the single season model of MacKenzie et al. (2002) to multiple seasons by introducing two parameters that govern changes in the occupancy state of sites over time. These parameters can be viewed as the vital rates associated with occupancy dynamics. Let ε_t be the probability that a site occupied in season t is unoccupied by the species in season $t + 1$ (local extinction), and γ_t be the probability that an unoccupied site in season t is occupied by the species in season $t + 1$ (colonization). Therefore, a matrix of the probability of a site transitioning between occupancy states between seasons may be defined as;

$$\boldsymbol{\phi}_t = \begin{bmatrix} 1 - \varepsilon_t & \varepsilon_t \\ \gamma_t & 1 - \gamma_t \end{bmatrix},$$

where rows of $\boldsymbol{\phi}_t$ represent the occupancy state of the site at t (state 1 = occupied; state 2 = unoccupied), and columns represent the occupancy state at $t + 1$. For completeness, a row vector $\boldsymbol{\phi}_0$ may be defined as $\boldsymbol{\phi}_0 = [\psi_1 \quad 1 - \psi_1]$, where ψ_1 is the probability the site is occupied in the first season ($t = 1$).

To incorporate detection probabilities into the model, define a column vector $\mathbf{p}_{\mathbf{H},t}$ that denotes the probability of observing the portion of the detection history $\mathbf{H}_i$ relevant to season t, conditional upon occupancy state. For instance

$$\mathbf{p}_{101,t} = \begin{bmatrix} p_{t,1}(1 - p_{t,2})p_{t,3} \\ 0 \end{bmatrix}$$

$$\mathbf{p}_{000,t} = \begin{bmatrix} \prod_{j=1}^{3} (1 - p_{t,j} \\ \\ 1 \end{bmatrix}$$

where $p_{t,j}$ denotes the detection probability for visit j in season t. Note that whenever the species is detected at least once during a season, the second element of $\mathbf{p}_{\mathbf{H},t}$ will be zero because it is impossible to observe such a history if the site is in the unoccupied state. Similarly, the second element of $\mathbf{p}_{\mathbf{0},t}$ will always be 1, because the all zero history is the only observable outcome if the site is unoccupied.

For any given detection history, the probability of observing such an outcome can be expressed as, $\Pr(\mathbf{H}_i) = \boldsymbol{\phi}_0 \prod_{t=1}^{k-1} [D(\mathbf{p}_{\mathbf{H},t})\ \boldsymbol{\phi}_t]\ \mathbf{p}_{\mathbf{H},K}$, where $D(\mathbf{p}_{\mathbf{H},t})$ is a 2×2 diagonal matrix with the elements of $\mathbf{p}_{\mathbf{H},t}$ along the main diagonal (top left to bottom right), zero otherwise. For example, consider again the detection history $\mathbf{H}_i = 110\,000\,010$. The probability of observing this would be

$$\begin{aligned}
\Pr(\mathbf{H}_i &= 110\ 000\ 010) \\
&= \boldsymbol{\phi}_0 D(\mathbf{p}_{110,1})\boldsymbol{\phi}_1 D(\mathbf{p}_{000,2})\boldsymbol{\phi}_2 \mathbf{p}_{010,3} \\
&= [\psi_1 \quad 1 - \psi_1] \begin{bmatrix} p_{1,1}p_{1,2}(1 - p_{1,3}) & 0 \\ 0 & 0 \end{bmatrix} \\
&\quad \times \begin{bmatrix} 1 - \varepsilon_1 & \varepsilon_1 \\ \gamma_1 & 1 - \gamma_1 \end{bmatrix} \begin{bmatrix} \prod_{j=1}^{3} (1 - p_{2,j}) & 0 \\ 0 & 1 \end{bmatrix} \\
&\quad \times \begin{bmatrix} 1 - \varepsilon_2 & \varepsilon_2 \\ \gamma_2 & 1 - \gamma_2 \end{bmatrix} \begin{bmatrix} (1 - p_{3,1})p_{3,2}(1 - p_{3,3}) \\ 0 \end{bmatrix} \\
&= \psi_1 p_{1,1} p_{1,2}(1 - p_{1,3}) \\
&\quad \times \left[(1 - \varepsilon_1) \prod_{j=1}^{3} (1 - p_{2,j})(1 - \varepsilon_2) + \varepsilon_1\gamma_2\right] \\
&\quad \times (1 - p_{3,1})p_{3,2}(1 - p_{3,3}).
\end{aligned}$$

The central term in brackets on the final line represents the two possibilities for the species during the second season when it was not detected. Either the species (1) did not go locally extinct between seasons 1 and 2, was undetected in the three surveys during the second season, and continued to occupy the site into season three (with probability $(1 - \varepsilon_1) \prod_{j=1}^{3} (1 - p_{2,j}) (1 - \varepsilon_2)$); or (2) went locally extinct between seasons 1 and 2, and recolonized the site between seasons 2 and 3 (with probability $\varepsilon_1\gamma_2$).

Once the probability of observing each detection history has been established, the model likelihood can be formed as usual, i.e.,

$$L(\psi_1, \varepsilon, \gamma, \mathbf{p} \,|\, \mathbf{H}_1, \ldots, \mathbf{H}_U) = \prod_{i=1}^{U} \Pr(\mathbf{H}_i).$$

MacKenzie et al. (2003) describe how missing observations and covariate information can be included in the model, and also how the model can be reparameterized to get seasonal estimates of occupancy or estimates of the rate in change of occupancy.

EXAMPLES

To illustrate how the above occupancy estimation methods could be used in practice with respect to rare species, we now consider data collected for the following three species: (1) Mahoenui giant weta (*Deinacrida mahoenui*) from New Zealand, (2) gaur (*Bos frontalis*) from Malaysa, and (3) Blue-ridge salamander (*Eurycea wilderae*) from the eastern United States. Each example is used to demonstrate some aspects of the methods we describe above: single-season occupancy estimation, occupancy at multiple study areas with the concept of borrowing information, and multiple-season occupancy estimation. For brevity, only relatively simple analyses are presented here, but note that more complex analyses could be conducted to further investigate relationships between occupancy (and related parameters) and covariates that have been collected in the field. All analyses have been conducted using Program PRESENCE (*available online*).[7]

Mahoenui giant weta

Weta are one of the more unique and specialized groups of New Zealand insects. More than 70 endemic species of weta survive in New Zealand today. Weta are ancient species of the order Orthoptera (e.g., grasshoppers, crickets, and locusts) and remain almost unchanged from their ancestors of 190 million years ago. All weta are flightless and relatively large, and members of a subgroup called the giant weta are among the largest insects in the world. Most giant weta species are now endangered, with populations having been decimated by the introduction of mammalian predators. Currently, most species only survive on predator-free offshore islands or in protected reserves.

The Mahoenui giant weta (*Deinacrida mahoenui*) is endemic to the King Country on the North Island of New Zealand. The main naturally occurring population is restricted to a 240-ha block of reverting farmland at Mahoenui (near the town of Te Kuiti), which is designated as a scientific reserve, with a second population near Otangiwai 20 km to the east. The reserve is characterized by steep-sided gullies and is largely covered by dense gorse, *Ulex europaeus*, a perennial pest plant with sharp spiny stems and bright yellow flowers that can form dense thickets, originally introduced to New Zealand as a hedging plant by the early European settlers. The New Zealand Department of Conservation (DOC) uses goats and cattle to maintain the gorse habitat through browsing. Mahoenui giant weta use gorse as protection from predators and also as a food source.

In order to monitor the population, in March 2004 DOC began a pilot study to assess the effectiveness of using occupancy as the state variable of interest. While ideally DOC would like to monitor abundance of the species, the main impediments are (1) the weta only occur at low densities, hence few individuals are likely to be observed in any given survey; (2) they are not individually identifiable by natural markings; and (3) the weta are usually found in the brittle, dead foliage of a gorse bush, hence attempts to capture and mark them would likely destroy their apparently preferred habitat. Thus the decision was made to try occupancy rather than use unadjusted transect counts as had been done previously.

Between 23 and 27 March 2004, 72 circular plots of 3 m radius where surveyed for the Mahoenui giant weta. The plots were randomly positioned within the more accessible regions of the reserve. This means some caution must be used if the results are generalized to the entire reserve, but restricting the sample frame was determined to be reasonable given it was a pilot study. Each plot was surveyed between three and five times during the 5-d period. Three different surveyors were used and the study was designed such that each surveyor surveyed each site at least once. This allows surveyor-specific detection probabilities to be estimated.

Weta were detected at 35 of the 72 plots (a naïve occupancy estimate of 0.49), however often weta were only detected in one or two of the repeated surveys clearly indicating that detection probabilities are <1. Conceivably, there may be a number of plots where weta were indeed present but were simply never detected during the surveys. Here we use the single-season occupancy model described above to estimate the proportion of plots that may be occupied. We consider four simple models. In each case occupancy probability is assumed to be constant for all plots (denoted as $\psi(\cdot)$), and detection probability was either constant ($p(\cdot)$), different on each day ($p(t)$), varied by surveyor ($p(S)$) or varied by both day and surveyor ($p(t + S)$).

---

[7] ⟨http://www.mbr-pwrc.usgs.gov/software⟩

TABLE 1. Summary of models fit to Mahoenui giant weta example.

Model	AIC	ΔAIC	$\hat{\psi}$	$\mathrm{SE}(\hat{\psi})$
$\psi(\cdot)p(t+S)$	258.55	0.00	0.64	0.09
$\psi(\cdot)p(t)$	260.80	2.25	0.63	0.09
$\psi(\cdot)p(S)$	263.36	4.81	0.62	0.09
$\psi(\cdot)p(\cdot)$	265.79	7.24	0.62	0.09

Notes: Models are ranked in terms of Akiake's Information Criterion (AIC). The relative difference in values (ΔAIC), estimates of occupancy ($\hat{\psi}$), and its standard error ($\mathrm{SE}(\hat{\psi})$) are also given.

The four models are ranked for parsimony in terms of Akiake's Information Criterion (AIC; Burnham and Anderson 2002), with the results summarized in Table 1. In this instance, regardless of the structure used to model detection probability, estimates of occupancy (and its associated standard error) are similar, approximately 30% greater than the naïve estimate. This result is partially due to the design of the study so will not always be the case. There is some evidence that the surveyors differed in their ability to detect weta. If surveyors always sampled the same sites, a form of heterogeneity in detection probability would be introduced, resulting in underestimates of occupancy if not accounted for (Royle and Nichols 2003).

Malaysian gaur

The gaur (*Bos frontalis*) is a large (up to 1000 kg) Asian bovid found in scattered areas of suitable habitat (forests and associated grassy clearings) extending from India and Nepal eastward to Indochina and the Malaysian peninsula. The species is believed to have declined dramatically over the past several decades. It is classified as vulnerable by IUCN, as endangered by the U.S. Department of Interior, and is listed on appendix 1 of CITES (Nowak and Paradiso 1983).

Kawanishi studied tigers (*Panthera tigris*) and potential prey species in three extensive study areas within Taman Negara National Park in Peninsular Malaysia (Kawanishi and Sunquist 2004). A grid with nine 5 × 5 km cells was superimposed on each of the three study areas for the purpose of estimating occupancy by gaur. Each grid cell was sampled monthly for 5 mo using camera traps and ground surveys for animal sign. Virtually all travel was on foot, so the three areas were sampled sequentially over the period 1999–2001.

Gaur were detected in six, five, and one of the grid cells in each of the study areas, leading to naïve occupancy estimates of 0.67, 0.55, and 0.11 respectively. Applying a constant detection single-season occupancy model to the data from each study area separately, we obtain occupancy estimates of 0.86 and 0.60 for the two areas with adequate data. Only one gaur was detected at a single site in the third data set, leaving us uncomfortable with attempting to estimate occupancy separately for this area. In order to improve the estimates of occupancy and obtain some inference about the third study area, here we model the data from the three study areas within a single framework by sharing information about the probability of detecting gaur among the areas. We consider two models for these data: (1) detection probability is the same at all sites, or (2) detection probability at the third area is the average (on the logistic scale) of the other two areas. Note that a third model in which detection probabilities are different at each site is equivalent to fitting the occupancy model to each data set independently, which is the model that does not permit reasonable inference about the third area. For all models we wish to obtain area specific estimates of occupancy.

Table 2 presents a summary of the model selection procedure for each of these models. For the purpose of determining model weights and model averaged estimates of occupancy (Burnham and Anderson 2002), the third model has not been considered due to the concerns of estimator reliability. Note that the model averaged estimates of occupancy for study areas 1 and 2 are reasonably similar to those obtained when each data set is analyzed separately, however the estimate for area 3 appears much more realistic.

Blue-ridge salamander

The Blue-ridge salamander (*Eurycea wilderae*) is one of more than 75 salamander species found in the

TABLE 2. Summary of model selection procedure for the Malaysian gaur example.

			$\hat{\psi}$ ($\mathrm{SE}(\hat{\psi})$)		
Model	ΔAIC	*w*	Area 1	Area 2	Area 3
$\psi(\text{area})p(\cdot)$	0.00	0.62	0.79 (0.20)	0.66 (0.21)	0.13 (0.13)
$\psi(\text{area})p(\text{average})$	1.00	0.38	0.90 (0.29)	0.61 (0.19)	0.13 (0.13)
$\psi(\text{area})p(\text{area})$	1.28		0.86 (0.26)	0.60 (0.18)	1.00 (0.00)
Model-averaged estimates			0.83 (0.24)	0.64 (0.20)	0.13 (0.13)

Notes: Models are ranked in terms of Akiake's Information Criterion (AIC). Presented here are the relative difference in values (ΔAIC), AIC model weights (*w*), estimates of occupancy ($\hat{\psi}$), and its associated standard error ($\mathrm{SE}(\hat{\psi})$, in parentheses) for each of the three study areas. Model-averaged estimates are based upon the first two models only.

TABLE 3. Parameter estimates and summary of model selection procedure according to AIC for the Blue-ridge salamander (*Eurycea wilderae*) example.

Model	ΔAIC	w	$\hat{\psi}_{98}$	$\hat{\gamma}_{98}$	$\hat{\gamma}_{99}$	$\hat{\gamma}_{00}$	$\hat{\varepsilon}_{98}$	$\hat{\varepsilon}_{99}$	$\hat{\varepsilon}_{00}$	$\hat{p}_{98}$	$\hat{p}_{99}$	$\hat{p}_{00}$	$\hat{p}_{01}$
$\psi(98)\gamma(\cdot)\varepsilon(\cdot)p(\text{yr})$	0.00	0.68	0.87	0.16	0.16	0.16	0.22	0.22	0.22	0.52	0.41	0.28	0.28
$\psi(98)\gamma(\cdot)\varepsilon(\text{yr})p(\text{yr})$	2.95	0.16	0.87	0.17	0.17	0.17	0.24	0.26	0.06	0.52	0.41	0.29	0.26
$\psi(98)\gamma(\text{yr})\varepsilon(\cdot)p(\text{yr})$	3.75	0.10	0.85	0.32	0.16	0.13	0.22	0.22	0.22	0.54	0.41	0.28	0.28
$\psi(98)\gamma(\text{yr})\varepsilon(\text{yr})p(\text{yr})$	6.70	0.02	0.84	0.33	0.17	0.14	0.24	0.25	0.06	0.54	0.41	0.29	0.26
$\psi(98)\gamma(\cdot)\varepsilon(\cdot)p(\cdot)$	6.98	0.02	0.94	0.17	0.17	0.17	0.27	0.27	0.27	0.38	0.38	0.38	0.38
$\psi(98)\gamma(\cdot)\varepsilon(\text{yr})p(\cdot)$	10.18	0.00	0.93	0.19	0.19	0.19	0.26	0.35	0.21	0.39	0.39	0.39	0.39
$\psi(98)\gamma(\text{yr})\varepsilon(\cdot)p(\cdot)$	10.51	0.00	0.94	0.00	0.12	0.21	0.27	0.27	0.27	0.38	0.38	0.38	0.38
$\psi(98)\gamma(\text{yr})\varepsilon(\text{yr})p(\cdot)$	13.59	0.00	0.94	0.00	0.13	0.23	0.25	0.36	0.21	0.39	0.39	0.39	0.39
Model averaged estimates			**0.87**	**0.18**	**0.16**	**0.16**	**0.22**	**0.23**	**0.19**	**0.52**	**0.41**	**0.29**	**0.28**
Standard errors			0.08	0.14	0.11	0.11	0.07	0.08	0.09	0.06	0.05	0.05	0.05
Naïve estimates			**0.76**	**0.44**	**0.21**	**0.27**	**0.28**	**0.44**	**0.29**	**1.00**	**1.00**	**1.00**	**1.00**

Note: ΔAIC is the relative difference in AIC values from the top ranked model; w is the AIC model weight. Bold values are referred to in the text. Subscripts in column heads represent years 1998–2001.

southeastern United States. *E. wilderae* has a dual life phase, with a larvae period lasting 1–2 years in the southern Appalachians (Bruce 1988). Usually considered a stream-side salamander, *E. wilderae* is believed to undergo seasonal migration away from streams during warmer months and can often be found far from running water (Petranka 1998). Because *E. wilderae* potentially uses both aquatic (stream) and terrestrial habitats, it was proposed as a management indicator species (MIS) for National Forests in North Carolina (FY2002 Monitoring and Evaluation Report: National Forests in North Carolina; *available online*).[8] Recently, many proposed amphibian and reptile species have been removed from MIS lists by the Forests Service because count-based statistics for these species show high temporal and spatial variability, thus making inferences about relationships between population and habitat changes unreliable and difficult. *E. wilderae* shares this characteristic and may be considered ''rare'' because it occurs at low densities in terrestrial habitats. For example, in a capture–recapture study in Great Smoky Mountains National Park (GSMNP), Bailey et al. (2004*a*) sampled 15 × 15 m forest plots 14–16 times each of three spring sampling seasons. *E. wilderae* was detected on more plots than any other salamander species (16 out of 20 plots), but the average number of captured individuals per sampling occasion was less than 1. Abundance estimates are impossible for *E. wilderae* without borrowing information from other species, times or locations (see Bailey et al. 2004*a*). Alternatively, occupancy has been proposed as a more appropriate state variable for large-scale salamander monitoring programs in the southeastern United States (Bailey et al. 2004*b*).

We used the multiple season model presented above to estimate occupancy dynamics (vital rates) for *E. wilderae* populations found within a single watershed within GSMNP. In 1998, Hyde and Simons (2001) initiated a salamander study within the Roaring Fork Watershed (Mt. LeConte USGS Quadrangle) and Bailey et al. (2004*b*) continued to sample a subset of 39 sites for a total of four years (1998–2001); only these 39 sites are used in this analysis. Each sample site was sampled with both a natural cover transect (50 m long × 3 m wide) and a parallel coverboard transect consisting of five stations placed 10 m apart (see Hyde and Simons [2001] for details). Sites were located near trails and spaced approximately 250 m apart, beginning at a random point at least 250 m from each trail head. Sites were sampled once per month in 1998 (June–August), then once every two weeks from early April to late June for the remaining three years of the study. Relative abundance information was collected on all salamander species, but we use only detection/nondetection information for *E. wilderae* in this analysis.

We used MacKenzie et al.'s (2003) initial parameterization to estimate initial occupancy (1998), vital rates (colonization and extinction), and detection probabilities. A number of covariates were measured at each site (e.g., elevation and stream proximity) but here we only consider a suite of simple models where vital rates and detection probabilities are year specific or constant across time. We expected yearly differences in all multi-season parameters as rainfall levels varied considerably among years, with total April–June precipitation declining over the last three years of the study. Again for brevity, we assume the probability of detecting *E. wilderae* in a single survey of a site was constant within each year.

Table 3 presents the parameter estimates and summary of the model selection procedure for the eight simple models considered. Note that the models with constant detection probability among years (which is an implicit assumption when comparing naïve counts) constitute <3% of the total AIC model weights. That is, from this analysis there is very strong support for the hypothesis that detection probabilities vary over time suggesting reliable conclusions about the population can only be made if these detection probabilities

---

[8] ⟨http://www.cs.unca.edu/nfsnc/me2002/fy2002_me_report.pdf⟩

differences are explicitly accounted for. Considering the model averaged estimates, the proportion of sites occupied by *E. wilderae* in 1998 was 0.87, the probability of a site being colonized between years is ≈0.16, and the probability of *E. wilderae* going local-extinct from a site was ≈0.20. By using the fact that sites occupied in the next year are a combination of occupied sites this year where the species did not go locally extinct, and unoccupied sites that the species colonizes (i.e., $\psi_{t+1} = \psi_t(1 - \varepsilon_t) + (1 - \psi_t)\gamma_t$), we can determine the proportion of sites occupied in years 1999–2001. Using the model averaged parameters estimates we obtain the values 0.70, 0.59, and 0.54 respectively, suggesting a downward trend in occupancy over this time frame for *E. wilderae* that is consistent with our a priori expectations based on seasonal rainfall records. It is worth noting that estimated vital rates are lower than naïve estimates, suggesting that some of the apparent turnover is likely the result of nondetection rather than true colonization and extinction events.

DISCUSSION

It is particularly unfortunate that rare species are simultaneously the species for which strong inferences about state variables and vital rates are most needed and the species for which such information is most difficult to obtain. These dual concerns provide a formidable challenge to conservation biologists and wildlife biologists to obtain useful information on which to base management decisions in the face of substantial sampling difficulties. We are pleased at recent efforts to meet this challenge, including this Special Feature and the new edited volume by W. L. Thompson (2004).

We began this paper with a brief summary of statistical principles important in estimating abundance and related parameters of animal populations. The response of many biologists and managers to the poor sample sizes achieved in studies of rare species is to abandon these principles and simply base inference on raw count statistics. We do not believe that this response is satisfactory, so we discussed two general approaches to inference that we hope might be useful for rare species.

The borrowing of information is nothing new and indeed underlies all inferential statistics. Our suggestion is simply to extend the basic idea beyond the units of aggregation typically used in statistics (e.g., individuals, replicate locations) to possibly disparate times and locations, to different data sources, and even to different species. In the case of rare species, such aggregation may permit reasonable inference in situations where it would not be possible with more typical disaggregated treatments.

The use of state variables other than abundance is also not new. In particular, other investigators have used occupancy as a state variable of interest for rare species. However, until the last year or so, these past uses have ignored the issue of detection probability. We described a framework for drawing inferences about occupancy, changes in occupancy, and the vital rates responsible for such changes using models that properly incorporate detection probability. As illustrated by our examples, we believe that these methods hold great promise for use in studies of rare species.

In summary, we recognize that rare species present problems to biologists and managers who wish to study their populations. However, we do not believe that the appropriate response to such problems is to abandon reasonable inference methods. Instead, we recommend consideration of the methods provided in this Special Feature. In particular, we recommend consideration of borrowing information and using state variables such as occupancy as means of dealing with detection probability when studying rare species.

ACKNOWLEDGMENTS

We thank all the field crews who aided in the collection of the data used in this article; K. Langdon and the staff at Great Smoky Mountains National Park for their logistical support and T. R. Simons and E. J. Hyde, who provided valuable assistance in all facets of the salamander research. Our thanks also go to Avi Holzapfel, Evan Cooch, Aaron Ellison, and an anonymous reviewer whose comments challenged our thinking and improved earlier manuscript drafts.

LITERATURE CITED

Alho, J. M. 1990. Logistic regression in capture–recapture models. Biometrics **46**:623–635.

Azuma, D. L., J. A. Baldwin, and B. R. Noon. 1990. Estimating the occupancy of Spotted Owl habitat areas by sampling and adjusting for bias. General Technical Report PSW-124. U. S. Department of Agriculture, Forest Service, Pacific Southwest Research Station, Albany, California, USA.

Bailey, L. L., T. R. Simons, and K. H. Pollock. 2004*a*. Estimating detection probability parameters for plethodon salamanders using the robust capture–recapture design. Journal of Wildlife Management **68**:1–14.

Bailey, L. L., T. R. Simons, and K. H. Pollock. 2004*b*. Estimating site occupancy and species detection probability parameters for terrestrial salamanders. Ecological Applications **14**:692–702.

Barbault, R., and M. E. Hochberg. 1992. Population and community level approaches to studying biodiversity in international research programs. Acta Oecologica **13**:137–146.

Barbraud, C., J. D. Nichols, J. E. Hines, and H. Hafner. 2003. Estimating rates of extinction and colonization in colonial species and an extension to the metapopulation and community levels. Oikos **101**:113–126.

Besbeas, P., S. N. Freeman, B. J. T. Morgan, and E. A. Catchpole. 2002. Integrating mark–recapture–recovery and census data to estimate animal abundance and demographic parameters. Biometrics **58**:540–547.

Besbeas, P., J. -D. Lebreton, and B. J. T. Morgan. 2003. The efficient integration of abundance and demographic data. Applied Statistics **52**:95–102.

Boulinier, T., J. D. Nichols, J. E. Hines, J. R. Sauer, C. H. Flather, and K. H. Pollock. 1998. Higher temporal variability of forest breeding bird communities in fragmented landscapes. Proceedings of the National Academy of Sciences (USA) **95**:7497–7501.

Boulinier, T., J. D. Nichols, J. E. Hines, J. R. Sauer, C. H. Flather, and K. H. Pollock. 2001. Forest fragmentation and bird community dynamics: inference at regional scales. Ecology **82**:1159–1169.

Boyce, M. S., D. I. MacKenzie, B. F. J. Manly, M. A. Haroldson, and D. Moody. 2001. Negative binomial models for abundance estimation of multiple closed populations. Journal of Wildlife Management **65**:498–509.

Bradford, D. F., A. C. Neale, M. S. Nash, D. W. Sada, and J. R. Jaeger. 2003. Habitat patch occupancy by toads (*Bufo punctatus*) in a naturally fragmented desert landscape. Ecology **84**:1012–1023.

Brown, J. H. 1995. Macroecology. University of Chicago Press, Chicago, Illinois, USA.

Bruce, R. C. 1988. An ecological life table for the salamander *Eurycea wilderae*. Copeia 1988:15–26.

Bunge, J., and M. Fitzpatrick. 1993. Estimating the number of species: a review. Journal of the American Statistical Association **88**:364–373.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Burnham, K. P., and W. S. Overton. 1979. Robust estimation of population size when capture probabilities vary among animals. Ecology **62**:927–936.

Cam, E., J. D. Nichols, J. E. Hines, J. R. Sauer, R. Alpizar-Jara, and C. H. Flather. 2002. Disentangling sampling and ecological explanations underlying species–area relationships. Ecology **83**:1118–1130.

Cam, E., J. D. Nichols, J. R. Sauer, J. E. Hines, and C. H. Flather. 2000. Relative species richness and community completeness: avian communities and urbanization in the mid-Atlantic states. Ecological Applications **10**:1196–1210.

Chao, A., W.-H. Hwang, Y.-C. Chen, and C. Y. Kuo. 2000. Estimating the number of shared species in two communities. Statistica Sinica **10**:227–246.

Cochran, W. G. 1977. Sampling techniques. Wiley, New York, New York, USA.

Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B **345**:101–118.

Diefenbach, D. R., M. J. Conroy, R. J. Warren, W. E. James, L. A. Baker, and T. Hon. 1994. A test of the scent-station survey technique for bobcats. Journal of Wildlife Management **58**:10–17.

Dixon, P. M., A. M. Ellison, and N. J. Gotelli. 2005. Improving the precision of estimates of the frequency of rare events. Ecology **86**:1114–1123.

Doherty, P. F., Jr., T. Boulinier, and J. D. Nichols. 2003*a*. Extinction rates at the center and edge of species' ranges. Annales Zoologici Fennici **40**:145–153.

Doherty, P. F., Jr., G. Sorci, J. A. Royle, J. E. Hines, J. D. Nichols, and T. Boulinier. 2003*b*. Sexual selection affects local extinction and turnover in bird communities. Proceedings of the National Academy of Sciences (USA) **100**: 5858–5862.

Dorazio, R. M., and J. A. Royle. 2003. Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics **59**:351–364.

Dorazio, R. M., and J. A. Royle. 2005. Estimating size and composition of biological communities by modeling the occurrence of species. Journal of the American Statistical Association. *in press.*

Elliott, M. R., and R. J. A. Little. 2000. A Bayesian approach to combining information from a census, a coverage measurement survey, and demographic analysis. Journal of the American Statistical Association **95**:351–362.

Franklin, A. B., et al. 2004. Population dynamics of the California Spotted Owl (*Strix occidentalis occidentalis*): a meta-analysis. Ornithological Monographs **54**:1–54.

Gallucci, V. F., S. B. Saila, D. J. Gustafson, and B. J. Rothschild. 1996. Stock assessment: quantitative methods and applications for small-scale fisheries. Lewis, New York, New York, USA.

Gaston, K. J. 1994. Rarity. Chapman and Hall, London, UK.

Geissler, P. H., and M. R. Fuller. 1987. Estimation of the proportion of area occupied by an animal species. Proceedings of the Section on Survey Research Methods of the American Statistical Association **1996**:533–538.

Grove, N. E., J. R. Skalski, P. Zager, and R. L. Townsend. 2002. Statisical models for population reconstruction using age-at-harvest data. Journal of Wildlife Management **66**: 310–320.

Hames, R. S., K. V. Rosenberg, J. D. Lowe, and A. A. Dhondt. 2001. Site reoccupation in fragmented landscapes: testing predictions of metapopulation theory. Journal of Animal Ecology **70**:182–190.

Hanski, I. 1992. Inferences from ecological incidence functions. American Naturalist **139**:657–662.

Hanski, I. 1999. Metapopulation ecology. Oxford University Press, Oxford, UK.

Huggins, R. M. 1989. On the statistical analysis of capture experiments. Biometrika **76**:133–140.

Huggins, R. M. 1991. Some practical aspects of conditional likelihood approach to capture experiments. Biometrics **47**: 725–732.

Hyde, E. J., and T. R. Simons. 2001. Sampling Plethodontid salamanders: sources of variability. Journal of Wildlife Management **65**:624–632.

Kawanishi, K., and M. E. Sunquist. 2004. Conservation status of tigers in a primary rainforest of peninsular Malaysia. Biological Conservation **120**:329–344.

Lancia, R. A., J. D. Nichols, and K. H. Pollock. 1994. Estimating the number of animals in wildlife populations. Pages 215–253 *in* T. Bookhout, editor. Research and management techniques for wildlife and habitats. Wildlfie Society, Bethesda, Maryland, USA.

Lande, R. 1988. Demographic models of the Northern Spotted Owl (*Strix occidentalis caurina*). Oecologia **75**:601–607.

Lebreton, J.-D., K. P. Burnham, J. Clobert, and D. R. Anderson. 1992. Modelling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecological Monographs **62**:67–118.

MacKenzie, D. I., and L. Bailey. 2004. Assessing the fit of site occupancy models. Journal of Agricultural, Biological and Environmental Statistics **9**:300–318.

MacKenzie, D. I., L. L. Bailey, and J. D. Nichols. 2004. Investigating species co-occurrence patterns when species are detected imperfectly. Journal of Animal Ecology **73**: 546–555.

MacKenzie, D. I., and W. L. Kendall. 2002. How should detection probability be incorporated into estimates of relative abundance? Ecology **83**:2387–2393.

MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin. 2003. Estimating site occupancy, colonization and local extinction probabilities when a species is detected imperfectly. Ecology **84**:2200–2207.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology **83**:2248–2255.

Mangel, M., et al. 1996. Principles for the conservation of wild living resources. Ecological Applications **6**:338–362.

Mao, C. X., and R. K. Colwell. 2005. Estimation of species richness: mixture models, the role of rare species, and inferential challenges. Ecology **86**:1143–1153.

Martinez-Solano, I., J. Bosch, and M. Garcia-Paris. 2003. Demographic trends and community stability in a montane amphibian assemblage. Conservation Biology **17**:238–244.

Millar, R. B., and R. Meyer. 2000. Bayesian state-space modeling of age-structured data: fitting a model is just the beginning. Canadian Journal of Fisheries and Aquatic Sciences **57**:43–50.

Moilanen, A. 2002. Implications of empirical data quality for metapopulation model parameter estimation and application. Oikos **96**:516–530.

Nichols, J. D., T. Boulinier, J. E. Hines, K. H. Pollock, and J. R. Sauer. 1998*a*. Estimating rates of local extinction, colonization and turnover in animal communities. Ecological Applications **8**:1213–1225.

Nichols, J. D., T. Boulinier, J. E. Hines, K. H. Pollock, and J. R. Sauer. 1998*b*. Inference methods for spatial variation in species richness and community composition when not all species are detected. Conservation Biology **12**:1390–1398.

Nichols, J. D., and M. J. Conroy. 1996. Estimation of species richness. Pages 226–234 *in* D. E. Wilson, F. R. Cole, J. D. Nichols, R. Rudran, and M. Foster, editors. Measuring and monitoring biological diversity. Standard methods for mammals. Smithsonian Institution Press, Washington, D.C., USA.

Nichols, J. D., J. E. Hines, J. R. Sauer, F. W. Fallon, J. E. Fallon, and P. J. Heglund. 2000. A double-observer approach for estimating detection probability and abundance from point counts. Auk **117**:393–408.

Norris, J. L., III, and K. H. Pollock. 1998. Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. Environmental and Ecological Statistics **5**:391–402.

Nowak, R. M., and J. L. Paradiso. 1983. Walker's mammals of the world. Fourth edition. Volume II. John's Hopkins University Press, Baltimore, Maryland, USA.

Otis, D. L., K. P. Burnham, G. C. White, and D. R. Anderson. 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs **62**.

Petranka, J. W. 1998. Salamanders of the United States and Canada. Smithsonian Institution Press, Washington, D.C., USA.

Pollock, K. H., J. E. Hines, and J. D. Nichols. 1984. The use of auxillary variables in capture–recapture and removal experiments. Biometrics **40**:329–340.

Pollock, K. H., J. D. Nichols, T. R. Simons, and J. R. Sauer. 2002. Large-scale wildlife monitoring studies: statistical methods for design and analysis. Environmetrics **13**:1–15.

Quinn, T. J., II, and R. B. Deriso. 1999. Quantitative fish dynamics. Oxford University Press, Oxford, UK.

Reunanen, P., A. Nikula, M. Monkkonen, E. Hurme, and V. Nivala. 2002. Predicting occupancy for the Siberian Flying squirrel in old-growth forest patches. Ecological Applications **12**:1188–1198.

Royle, J. A., and J. D. Nichols. 2003. Estimating abundance from repeated presence absence data or point counts. Ecology **84**:777–790.

Schnute, J. T. 1994. A general framework for developing sequential fisheries models. Canadian Journal of Fisheries and Aquatic Sciences **51**:1676–1688.

Scott, J. M., F. Davis, B. Csuti, R. Noss, B. Butterfield, C. Groves, H. Anderson, S. Caicco, F. D'Erchia, T. C. Edwards, Jr., J. Ulliman, and R. G. Wright. 1993. Gap analysis: a geographic approach to protection of biological diversity. Wildlife Monographs 123.

Scott, J. M., P. J. Heglund, M. L. Morrison, J. B. Haufler, M. G. Rafael, W. A. Wall, and F. B. Samson, editors. 2002. Predicting species occurrences: issues of accuracy and scale. Island Press, Washington, D.C., USA.

Seber, G. A. F. 1982. Estimation of animal abundance and related parameters. Macmillan, New York, New York, USA.

Skalski, J. R., and D. S. Robson. 1992. Techniques for wildlife investigations: design and analysis of capture data. Academic Press, San Diego, California, USA.

Stauffer, H. B., C. J. Ralph, and S. L. Miller. 2004. Ranking habitat for marbled murrelets: a new conservation approach for species with uncertain detection. Ecological Applications **14**:1374–1383.

Thompson, S. K. 1992. Sampling. John Wiley and Sons, New York, New York, USA.

Thompson, W. L. 2004. Sampling rare or elusive species: concepts and techniques for estimating population parameters. Island Press, Washington, D.C., USA.

Thompson, W. L., G. C. White, and C. Gowan. 1998. Monitoring vertebrate populations. Academic Press, San Diego, California, USA.

Trenkel, V. M., D. A. Elston, and S. T. Buckland. 2000. Fitting population dynamics models to count and cull data using sequential importance sampling. Journal of the American Statistical Association **95**:363–374.

Tyre, A. J., B. Tenhumberg, S. A. Field, D. Niejalke, K. Parris, and H. P. Possingham. 2003. Improving precision and reducing bias in biological surveys by estimating false negative error rates in presence–absence data. Ecological Applications **13**:1790–1801.

White, G. C., and B. C. Lubow. 2002. Fitting population models to multiple sources of observed data. Journal of Wildlife Management **66**:300–309.

Wikle, C. K. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology **84**:1382–1394.

Williams, B. K., J. D. Nichols, and M. J. Conroy. 2002. Analysis and management of animal populations. Academic Press, San Diego, California, USA.

Wintle, B. A., M. A. McCarthy, K. M. Parris, and M. A. Burgman. 2004. Precision and bias of methods for estimating point survey detection probabilities. Ecological Applications **14**:703–712.

Yoccoz, N. G., J. D. Nichols, and T. Boulinier. 2001. Monitoring of biological diversity in space and time. Trends in Ecology and Evolution **16**:446–453.

Ecology, 86(5), 2005, pp. 1114–1123

IMPROVING THE PRECISION OF ESTIMATES OF THE FREQUENCY OF RARE EVENTS

PHILIP M. DIXON,[1,4] AARON M. ELLISON,[2] AND NICHOLAS J. GOTELLI[3]

[1]*Department of Statistics, Iowa State University, Ames, Iowa 50011 USA*
[2]*Harvard Forest, Harvard University, P.O. Box 68, Petersham, Massachusetts 01366 USA*
[3]*Department of Biology, University of Vermont, Burlington Vermont 05405 USA*

Abstract. The probability of a rare event is usually estimated directly as the number of times the event occurs divided by the total sample size. Unfortunately, the precision of this estimate is low. For typical sample sizes of $N < 100$ in ecological studies, the coefficient of variation (CV) of this estimate of the probability of a rare event can exceed 300%. Sample sizes on the order of 10^3–10^4 observations are needed to reduce the CV to below 10%. If it is impractical or impossible to increase the sample size, auxiliary data can be used to improve the precision of the estimate. We describe four approaches for using auxiliary data to improve the precision of estimates of the probability of a rare event: (1) Bayesian analysis that includes prior information about the probability; (2) stratification that incorporates information on the heterogeneity in the population; (3) regression models that account for information correlated with the probability; and (4) inclusion of aggregated data collected at larger spatial or temporal scales. These approaches are illustrated using data on the probability of capture of vespulid wasps by the insectivorous plant *Darlingtonia californica*. All four methods increase the precision of the estimate relative to the simple frequency-based estimate (absolute precision = 1.26, relative precision [CV] = 70%): stratification (absolute precision = 1.10, CV = 62%); regression models (absolute precision = 1.59, CV = 55%); Bayesian analysis with an informative prior probability distribution (absolute precision = 4.28, CV = 47%); and using temporally aggregated data (absolute precision = 6.75, CV = 36%). When informative auxiliary data is available, we recommend including it when estimating the probability of rare events.

Key words: *aggregation; Bayesian inference, coefficient of variation; estimators; precision; rare events; regression; sampling, stratification.*

INTRODUCTION

Rare events are important in ecology and evolution. Familiar examples include genetic drift in founding populations (Mayr 1963), seedling establishment in plant populations with low growth rates (Harper 1977), successful establishment of seedlings following long-distance dispersal (Clark et al. 2001), species extinction (Roberts and Solow 2003), and extreme meteorological events such as ice storms, wildfires, or hurricanes (Whelan 1995, Foster and Aber 2003). The ecology (Rabinowitz 1981) and biogeography (Jetz and Rahbek 2002) of rare species may be very different from that of common species, and the statistical distribution of rare species is a key prediction that distinguishes many models of community structure (Williams 1964, Hubbell 2001, Magurran 2003, Chave 2004).

Precisely estimating the probability of rare events is a statistical challenge. If the true probability of a discrete rare event is π, the standard frequentist estimate of this probability, $\hat{p}$, is calculated as the number of observations n of the rare event divided by the total number of observations (or trials) N (Gotelli and Ellison 2004): $\hat{p} = n/N$. The standard error of this estimate $\text{SE}_{\hat{p}} = \sqrt{\hat{p}(1 - \hat{p})/N}$, and its coefficient of variation $\text{CV}_{\hat{p}} = \text{SE}_{\hat{p}}/\hat{p}$.

If an event is truly rare ($\pi < 0.01$), its frequentist estimate $\hat{p}$ has reasonable precision ($\text{CV}_{\hat{p}} \leq 10\%$) only when the sample size N exceeds 1000 total observations or trials. For $N < 100$ total observations, typical for many ecological studies, $\text{CV}_{\hat{p}}$ may exceed 300%. When the precision is low, it can be difficult to detect trends in the frequency or differences between groups. Increasing the precision provides better estimates and higher power for statistical tests of trends and differences. The precision of an estimate can also be important for policy decisions (e.g., Lewison et al. 2004). In this article, we describe four methods that can provide more precise estimates of the probability of a rare event. All of these methods require auxiliary data, but obtaining this auxiliary data usually requires less effort or cost than obtaining larger samples of the rare event itself.

EXAMPLE DATA

We use data on the capture efficiency of wasps by the insectivorous pitcher plant, *Darlingtonia califor-*

Manuscript received 31 March 2004; revised 16 June 2004; accepted 12 July 2004; final version received 20 September 2004. Corresponding Editor: A. A. Agrawal. For reprints of this Special Feature, see footnote 1, p. 1079.

[4] E-mail: pdixon@iastate.edu

PLATE. 1. *Darlingtonia californica,* a rare carnivorous plant species endemic to the Siskiyou Mountains of Oregon and northern California, which grows in a threatened plant community type—serpentine fen. Photo credit: A. M. Ellison.

nica (Sarraceniaceae), to illustrate methods by which the precision of estimates of the probability of rare events can be increased (see Plate 1). Although prey capture by carnivorous plants provides nutrients required for successful sexual reproduction (reviewed in Ellison and Gotelli 2001), prey capture may be infrequent or rare (Zamora 1995, Zamora et al. 1998); most insects that enter pitcher-plant traps are not captured (Newell and Nastase 1998).

Like other pitcher plants in this family, *Darlingtonia* grows as a rosette of leaves that are modified to form pitcher-shaped traps (Arber 1941). These pitchers secrete copious nectar that attracts foraging insects, especially vespulid wasps (*Vespula atropilosa*) and ants (*Tapinoma sessile*). As part of a long-term study of the demography of *Darlingtonia*, we recorded the frequency with which *Darlingtonia* captures wasps and estimated the conditional probability of a successful capture: $\pi = P(\text{capture} \mid \text{visit})$. During July 2002, Ellison, Gotelli and their colleagues observed 753 *Darlingtonia* plants for one-half hour each, for a total of 376.5 plant-hours of observation (A. M. Ellison, R. J. Emerson, E. J. Farnsworth, N. J. Gotelli, C. M. Hart, H. R. Steinhoff, and S. E. Wittman, *unpublished data*). During this time, $N = 157$ wasps were seen to visit the pitchers, and $n = 2$ of these wasps were captured. For each visit, we also recorded the time a wasp spent in each pitcher, and we measured the orientation of the pitcher's opening (as degrees east of north). Assuming that the observed visits are a simple random sample of visits, the frequentist estimate of π is $\hat{p} = n/N = 2$ captures/157 visits $= 0.0127$. The estimated standard error for $\hat{p}$ is $\text{SE}_{\hat{p}} = \sqrt{\hat{p}(1 - \hat{p})/N} = 0.0089$. These estimates do not assume that the per-visit probability of capture is the same for all visits. When the sample is a simple random sample, heterogeneity in the population is irrelevant (Thompson 2002). Because the probability of capturing a wasp is very low and the total sample size is small, the precision of this frequentist estimate ($\hat{p}$) of capture probability by *Darlingtonia* is poor: $\text{CV}_{\hat{p}} = 70.2\%$ and precision (defined below) $= 1.26$.

MEASURES OF PRECISION OF A PROBABILITY

Precision ''refers to the dispersion of the observations'' (Marriott 1990). It can be quantified by at least four different, but related, measures (Table 1). The most familiar are measures of absolute precision, the standard error (SE) and variance (s^2). Because more precise estimates have smaller SEs, it is also common to define precision as $1/s^2$, especially in the Bayesian literature (Gelman et al. 1995:43).

When used to compare events that have different probabilities, absolute measures of precision have the counter-intuitive property that rarer events are more precisely measured. To illustrate this, consider an event (such as a visit of a wasp to a *Darlingtonia* pitcher) that occurs as an independent Poisson process over time with a constant rate of 0.1 visits/hour. If a plant is watched for a one-hour period, the probability of a visit during that hour is $P(\text{visit}) = 1 - e^{-0.1} = 0.095$. If 100

TABLE 1. Measures of absolute and relative precision when a proportion, p, is estimated from a simple random sample of n observations.

Measure	Absolute measures	Relative measures
Variability	Standard error (SE): $\sqrt{p(1-p)/n}$	Coefficient of variation: $\sqrt{(1-p)/(pn)}$
Precision	$1/\sigma^2 = (1/\mathrm{SE})^2$	$(1/\mathrm{CV})^2 = p^2/\sigma^2$

plants are watched, each for one hour, the SE of the visit probability is 0.029. If a plant is watched for a one-minute period, the probability of a visit during that one minute is $P(\text{visit}) = 1 - e^{-0.1/60} = 0.00167$. If 100 plants are watched, each for one minute, the SE of the one-minute-visit probability is much smaller, 0.00408. This apparent increase in precision is an artifact of a rarer event.

Measures of relative precision, including the CV or relative variance (Table 1), avoid this counterintuitive behavior by expressing the precision relative to the probability of the event. Rarer events are less precisely estimated, on a relative scale, than are more common events. Measures of relative precision are also unitless, unlike absolute measures of precision. To continue the example from the previous paragraph, when 100 plants are watched for an hour each, the CV is 0.029/0.095 = 32%. When 100 plants are watched for a minute each, the CV is much larger, 0.00408/0.00167 = 244%. The estimate from the shorter observation period is less precise, when measured using relative precision. The CV is one of many possible measures of relative precision. Others include the reciprocal of the CV, which is larger for more precise estimates, or the reciprocal of the CV^2, which is the relative analog of the Bayesian measure of precision ($1/s^2$).

HOW CAN THE PRECISION BE INCREASED?

Imprecise estimates of a probability are not unique to the *Darlingtonia* example. They are common whenever events are rare. When an event has a probability of occurring less than five times in a hundred trials ($P(\text{event}) = \pi = 0.05$), the coefficient of variation (CV) of estimates of this probability from samples of $N = 100$ are larger than 50%. The CV can exceed 300% when the event is very rare ($\pi < 0.01$; Fig. 1). On the other hand, when events are common, π can be estimated with high precision even with moderate sample sizes. If $\pi = 0.5$, a CV of 10% can be obtained with a sample size of $N = 100$.

Increasing the total sample size N increases the precision of the estimate of the probability of a rare event. For example, increasing N from 100 to 500 independent observations decreases the CV by a factor of $\sqrt{1/5}$. However, if an event is rare, a precise estimate (CV $\leq$ 10%) requires very large sample sizes. For example, if $\pi = 0.01$, a sample size of $N = 9900$ is required to achieve CV = 10%. Such large sample sizes may be expensive, difficult, or impossible to obtain.

Alternatively, the precision of the estimate of a rare event can be increased by combining the primary data (e.g., the observed numbers of visits and captures in the *Darlingtonia* dataset) with auxiliary data that provides additional information about the probability of the rare event. Auxiliary data may come from many different sources, of which we discuss four: prior information, stratified sampling, covariates, and aggregated data. We use the *Darlingtonia* data set to illustrate the methods by which auxiliary data can be used to improve the precision of point estimates of probability. We will compare methods using relative precision (CV) and absolute precision ($0.0001/s^2$). We use $0.0001/s^2$ instead of $1/s^2$ as a measure of absolute precision because the factor of 0.0001 converts the variance of a proportion to the variance of a percentage, which provides a more intuitive scale for interpreting absolute precision.

Incorporating prior information using Bayesian methods

Prior information about the probability π of a rare event can be derived from other studies of the same or related species, in the same or in different locations. Bayesian inference can be used to combine this prior information with the observed data (Ellison 1996). If probability estimates from the primary data are similar to those provided by the prior information, the combined estimate will have greater precision than the estimate based on the primary data alone.

Bayesian inference treats parameters, such as the probability π that *Darlingtonia* captures a wasp, as random variables described by statistical distributions (Barnett 1999, Ellison 2004). The distribution of each parameter summarizes both the expected value of the parameter and its variance. Bayesian inference uses the data (observations), along with information known about the parameter(s) before the data are analyzed (the prior probability distribution, or simply the prior) to construct a new distribution (the posterior probability distribution, or simply the posterior) that expresses what is known about the parameter after the data are analyzed.

The posterior is computed from the data and the prior using Bayes' Theorem (Ellison 2004):

$$f(\pi \mid C, V) = \frac{f(C \mid V, \pi) f(\pi)}{\displaystyle\int f(C \mid V, \pi) f(\pi)\, d\pi}. \qquad (1)$$

In Eq. 1, $f(\pi)$ is the prior, $f(C \mid V, \pi)$ is the likelihood

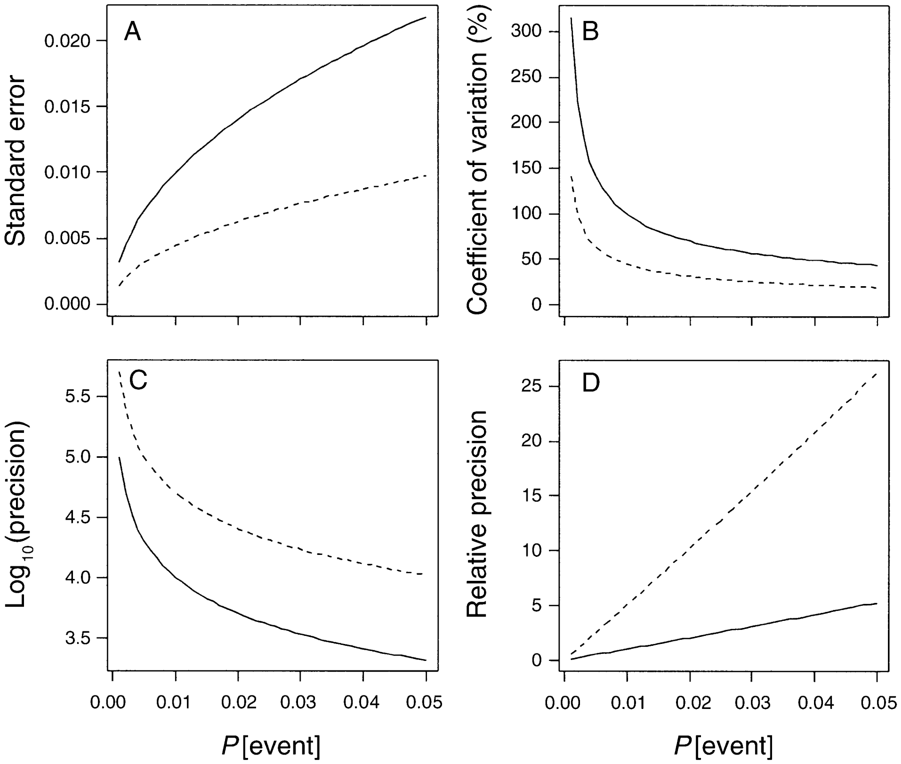

FIG. 1. The relationship between four measures of precision (Table 1) and the probability of an event when an event is rare (probability $\pi < 5\%$). (A) standard error (SE), (B) coefficient of variation (CV), (C) absolute precision (presented as a log value), (D) relative precision, $(1/\text{CV})^2$. Each measure of precision is calculated for sample sizes of $N = 100$ (solid line) and $N = 500$ (dashed line).

of the observations, and $f(\pi \mid C, V)$ is the posterior. The vertical bars indicate which quantities are considered fixed. That is, $f(C \mid V, \pi)$ is the probability distribution of C (the number of captures), conditional on the fixed values of V (the number of visits) and π (the capture probability). The integral in the denominator is a normalizing constant that ensures that the posterior distribution is a valid probability distribution (i.e., $0 \le f(\pi \mid C, V) \le 1$). Using Bayes' Theorem requires that the distributions of both the data and the prior be specified.

In the *Darlingtonia* data set, the data ($f(C \mid V, \pi)$) are the number of captures observed in a certain number of visits. A binomial distribution is commonly used to model count data when the outcomes (capture or not) are independent, the probability of success (capture) is the same for all visits, and where the number of success (captures) cannot exceed the number of visits (Gotelli and Ellison 2004).

When the data follow a binomial distribution, a Beta distribution,

$$\pi \sim \text{Beta}(\alpha, \beta) \tag{2}$$

is a convenient choice for the prior because the integral in the denominator of Eq. 1 can be evaluated analytically (Gelman et al. 1995). The values of the parameters α and β in the beta distribution (Eq. 2) summarize our knowledge of the capture probability before the data are analyzed. When $\alpha > 1$ and $\beta > 1$, the mean μ of the Beta(α, β) distribution equals $\alpha/(\alpha + \beta)$ and the mode equals $(\alpha - 1)/(\alpha + \beta - 2)$. Because the Beta distribution is skewed, the mode is the more appropriate measure of location. The variance σ^2 of a Beta(α, β) distribution is $\mu(1 - \mu)/(\alpha + \beta + 1)$. The posterior distribution given by Eq. 1 is also a Beta distribution (Gelman et al. 1995), and simulation of the posterior (e.g., with Markov chain Monte Carlo methods; Gilks et al. 1996) is not required. The parameters of the posterior depend on the parameters of the prior distribution (α, β) and the data (C, V):

$$\pi \mid C, V \sim \text{Beta}(\alpha + C, \beta + V - C). \tag{3}$$

The mode of the posterior is an updated estimate of the capture probability, and the standard deviation is an updated estimate of the variability:

$$\text{mode} = \frac{\alpha + C - 1}{\alpha + \beta + V - 2} \tag{4}$$

$$\text{SD} = \sqrt{\frac{(\alpha + C)(\beta + V - C)}{(\alpha + \beta + V)^2(\alpha + \beta + V + 1)}}. \tag{5}$$

The choice of prior distribution (i.e., of α and β) influences the posterior distribution, although the influence of the prior is small when V and C are large.

TABLE 2. Parameters of Beta distributions used as prior distributions in the Bayesian analysis of the *Darlingtonia* data, along with the resulting posterior distributions of the capture probability.

	Prior				Posterior			
Parameter	Mode	SD	α	β	Mode	SD	CV (%)	Precision
Data					0.0127	0.0089	70	12.5
Prior A	0.00931	0.0018	28	2873	0.0095	0.0018	19	31.5
Prior B	0.00931	0.0056	4.335	355.8	0.0104	0.0048	47	4.28
Prior C	0.00931	0.018	1.622	67.24	0.0117	0.0083	71	1.44
Prior D	0.00931	0.056	1.145	16.38	0.0124	0.010	80	0.99
Prior E	0.50	0.0833	1	1	0.0127	0.0108	85	0.86

Notes: All distributions except the uninformative prior (prior E) have a mode at the capture probability estimated by Newell and Nastase (1998) for the confamilial species *Sarracenia purpurea*. Prior A has a standard deviation (SD) equal to the sampling uncertainty reported by Newell and Nastase (1998). Priors B, C, and D have larger standard deviations to reflect uncertainty in the extrapolation across species and study sites. Prior E is the uninformative prior. The mode, SD, and CV reported in the first line (''Data'') of the table are the frequentist estimates for these parameters.

The prior distribution can be determined in many ways (Berger 1985). One is to use an uninformative prior: a prior for which any value of capture probability is equally likely. For a probability between 0 and 1, the uninformative prior is a uniform(0, 1) distribution which is equivalent to a Beta(1, 1).

Another approach is to use previous research to determine a prior distribution. Newell and Nastase (1998) estimated the per-visit probability of insect capture by a related pitcher plant, *Sarracenia purpurea*, to be 0.0093 (27 captures in 2899 visits with observed outcomes). If *S. purpurea* and *Darlingtonia* are assumed to have similar per-visit probabilities of insect capture, these data can be used to specify the prior distribution for the analysis of the *Darlingtonia* data. One approach is to do a Bayesian analysis of Newell and Nastase's data, using a noninformative hyperprior ($\alpha = 1$, $\beta = 1$) and the data ($C = 27$, $V = 2899$) in Eq. 3. The resulting posterior distribution, Beta(28, 2873) can be used as the prior for the *Darlingtonia* analysis. This distribution has a mode = 0.0093 and SD = 0.0018. There is some uncertainty introduced by extrapolating between species and between studies. This uncertainty can be expressed by increasing the standard deviation of the prior. Accordingly, we used three additional prior distributions with the same mode but with increasingly larger standard deviations (Table 2). If multiple prior data sets are available, the variability among the data sets can be used to estimate the parameters of the prior distribution (Birkes and Dodge 1993).

The posterior modes for the five choices of prior are given in Table 2. The posterior mode lies between the mode of the prior distribution and the capture probability estimated solely from the data. When the prior distribution has a small standard deviation (e.g., Newell and Nastase's prior [A] in Table 2), the posterior mode is very close to the prior mode (Fig. 2). As uncertainty in the prior increases, the posterior mode approaches the estimate based on the data (Table 2, Fig. 2).

The SD, CV, and precision ($0.0001/s^2$) of the posterior distribution summarize the uncertainty in the estimated capture probability. The improvement in precision gained by incorporating prior information depends on the SD of the prior and on the difference between the expectation of the prior (the probability of prey capture by *Sarracenia*) and the expectation of the primary data (the probability of prey capture by *Darlingtonia*). If the two species are very similar (priors A or B in Table 1) the Bayesian estimate is considerably more precise. When the two species are less similar (priors C or D in Table 1) or if the prior information is uninformative (prior E in Table 1), the Bayesian estimate is less precise than the estimate based on the primary data alone.

Stratified sampling

Stratification, dividing the population into more homogeneous strata, can lead to a more precise estimate of a proportion when heterogeneity in π is associated with identifiable characteristics of the events (Thompson 2002). Stratification can be used to estimate the probability of a rare event by dividing the population (e.g., all possible visits by wasps to *Darlingtonia*) into subgroups that have different capture probabilities. For example, one stratum may have a very small capture probability, another may have a slightly larger capture probability, and a third stratum may have a large capture probability. Stratification increases the precision of the estimated probability by removing the variability between strata. In this example we assume a simple random sample of observations within each stratum, but many other sampling designs could be used (Thompson 2002).

Strata cannot be defined on the basis of the response variable itself. In other words, it is not appropriate to define one stratum as those plants that captured a wasp ($N_1 = 2$) and the other as those plants that did not ($N_2 = 155$). Instead, strata should be defined a priori based on knowledge specific to the system. For example, the size of the plant or the orientation of the pitcher might be associated with the capture probability. As an illustration, we will use strata defined by the orientation of the pitcher. Two different definitions of strata will be used to illustrate the importance of between-strata heterogeneity in capture probabilities (Table 3). One

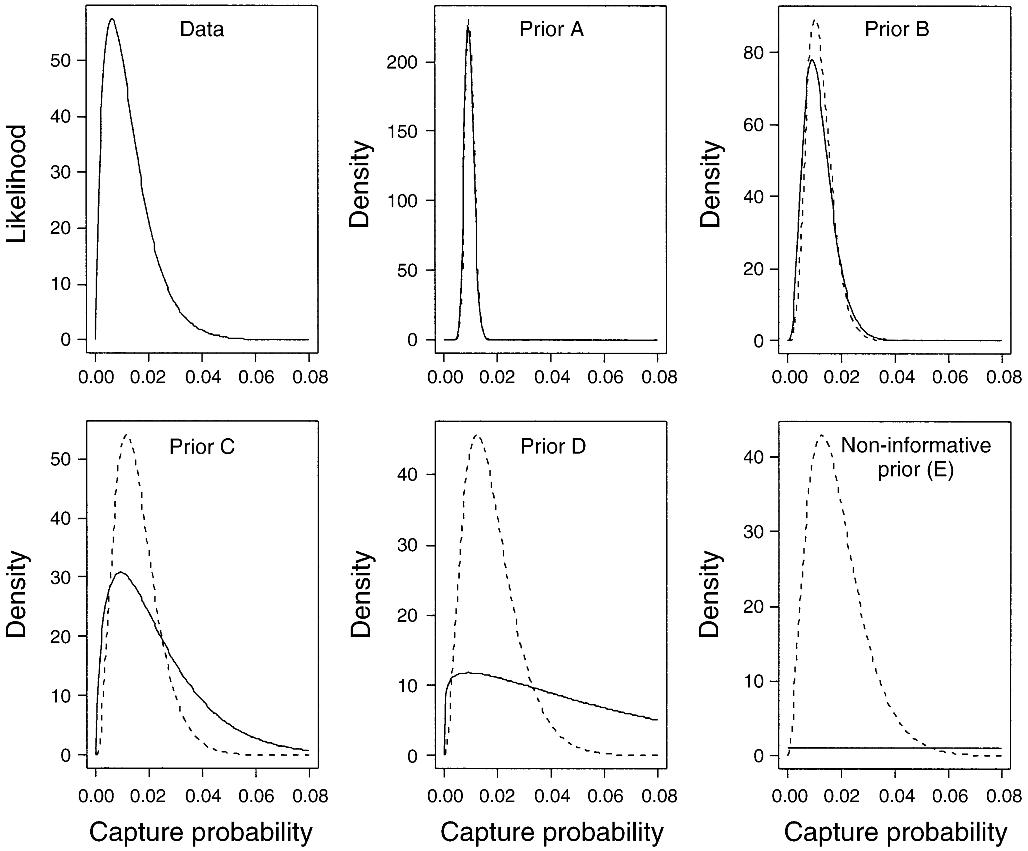

FIG. 2. Plots of the likelihood and prior (solid lines in density plots) and posterior (dashed lines in density plots) distributions for the five choices of prior distribution in Table 2. Note the different *y*-axis scales.

strata definition separates pitchers facing either 20° or 30° east of north from plants with all other orientations. The second strata definition separates those plants with orientations between 10° and 40° east of north from all other plants.

The estimated capture probability for the entire population from a stratified random sample is

$$\hat{p} = \frac{N_A\hat{p}_A + N_B\hat{p}_B}{N_A + N_B} \tag{6}$$

where N_A and N_B are the population sizes in the two strata, and $\hat{p}_A$ and $\hat{p}_B$ are the within-stratum estimates of the capture probability (Thompson 2002). When the event of interest only occurs in one stratum, the variance of the estimated probability is

$$\hat{s}^2(\hat{p}) = \left(\frac{N_A}{N_A + N_B}\right)^2 \left[\frac{\hat{p}_A(1 - \hat{p}_A)}{n_A}\right] \tag{7}$$

where stratum A is the stratum including all the events and n_A is the sample size of that stratum. This variance estimator assumes that the population size is large relative to the sample size. If this assumption is not appropriate, a finite population correction factor should be included in the variance estimate (see Thompson 2002 for details).

TABLE 3. Precision of the estimate of the probability of capture using two different stratifications of the data.

Parameter	Stratum A	Stratum B	Total
Stratification 1†			
Sample size (visits *V*)	9	148	157
Number of captures (*C*)	2	0	2
Capture probability ($\hat{p}$)	0.222	0	0.0127
$\mathrm{SE}_{\hat{p}}$	0.138	0	0.00792
$\mathrm{CV}_{\hat{p}}$			62.1%
Precision			1.59
Stratification 2‡			
Sample size (visits *V*)	19	138	157
Number of captures (*C*)	2	0	2
Capture probability ($\hat{p}$)	0.105	0	0.0127
$\mathrm{SE}_{\hat{p}}$	0.0070	0	0.00852
$\mathrm{CV}_{\hat{p}}$			66.9%
Precision			1.38

† Stratum A, plants with orientations of 20° or 30°; stratum B, all other plants.

‡ Stratum A, plants with orientations from 10° to 40°; stratum B, all other plants.

Estimating either $\hat{p}$ (Eq. 6) or its variance $s^2(\hat{p})$ (Eq. 7) requires knowledge of the relative sizes of the strata: $N_A/(N_A + N_B)$. The relative size of the strata may be estimated by independent criteria, such as stratum coverage or frequency in GIS databases. Because such information is lacking for the *Darlingtonia* population, we assume that the size of each stratum in the population is proportional to the size of each stratum in the sample: $N_A/(N_A + N_B) = 9/157$ for the first stratum (pitchers oriented either 20° or 30°), and 19/157 for the second (pitchers oriented between 10° and 40°).

Because stratum sizes were estimated from the sample itself, the capture probability $\hat{p}$ from either stratified sample (0.0127) is exactly the same as the estimate from the entire sample (Table 3). However, stratified sampling provides slightly more precise estimates of π (CV = 62.1% and 66.9% and absolute precision = 1.59 and 1.38 for the two definitions, respectively; Table 3) than do estimates based on the unstratified data (CV = 70.2%, absolute precision = 1.26). The first stratification (pitchers oriented either 20° or 30° vs. all others) is more precise than the second (Table 3) because the former has a larger between-strata difference in capture probability.

Stratification is especially useful when the probability of a rare event varies greatly among a small number of strata. However, if there are many strata, the number of observations per stratum is likely to be small and the stratum-specific probability will be poorly estimated.

Models incorporating covariates

Additional characteristics of the individuals may be measured. If these characteristics are associated with the rare event, they could be used either to stratify the observations (as in the previous approach) or to construct a model, e.g., a logistic regression model (Hosmer and Lemeshow 1989), to predict π for a specified set of covariates. The overall capture probability can be estimated by combining the model with information about the distribution of covariates in the population. The distribution can be enumerated when covariate information is available for all elements of the population or estimated from a simple random sample of the population. If the event is very rare (<10 events per covariate incorporated in the logistic regression; Van Belle 2002), logistic regression may not be useful for modeling the probability of very rare events.

In some cases, the event of interest is determined by an underlying continuous random variable. One example of this approach is the analysis of flood frequencies (Haan 2002). Floods are defined when water level exceeds a critical height for a specific patch of ground. The probability of flooding is the probability that the water level exceeds the critical height. Flood-frequency analysis uses a model for the distribution of water levels to estimate the probability of flooding (Hahn 2002).

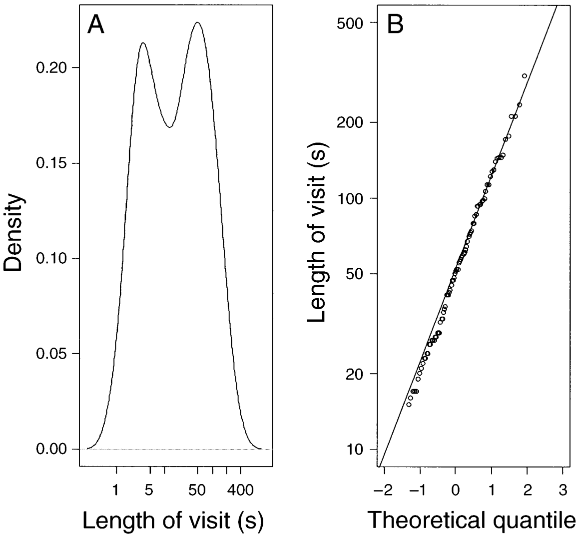

FIG. 3. (A) Probability density of visit length, estimated using a kernel smoother. The two modes are at 4 s and 50 s. The trough between the two peaks is centered at 13.5 s. (B) A log-normal quantile-quantile plot of the 86 visit lengths in the upper peak (visit lengths > 13.5 s). The theoretical quantiles were calculated after accounting for truncation (no value less than 13.5 s) and censoring (two captures with lengths > 307 s).

We used this last approach to estimate the probability that a wasp is captured by modeling the distribution of the length of time (visit lengths) that a wasp spends in a single pitcher. Visit lengths for noncaptured wasps ranged from a minimum of 1 s to a maximum of 307 s, with a median of 17 s. The empirical distribution of logarithmically transformed visit lengths is bimodal, with peaks at 4 and 50 s (Fig. 3A). The distribution of the logarithmically transformed values in the upper peak is very close to a normal distribution, as shown by a quantile-quantile plot (Fig. 3B). A two-component normal mixture model was fit to the log-transformed visit lengths by maximum likelihood. The two observed captures were considered censored observations, i.e., visit length >307 s. The upper peak was estimated to contain $\pi = 59.0\%$ of the visit lengths and have a normal distribution with $\bar{x} = 3.9$ and SD = 0.94.

The probability that a visit length exceeds s seconds is estimated using the normal cumulative distribution function, $\Phi(z)$:

$$\hat{P}(\text{visit length} > s) = \upsilon\left[1 - \Phi\left(\frac{\log s - \bar{x}}{\text{SD}}\right)\right] \qquad (8)$$

where υ is the probability that a visit is in the upper peak. Eq. (8) only applies to long visits where the contribution from the lower peak can be ignored.

Calculating the capture probability from this distribution requires specifying a critical visit length; any visit longer than that critical length is assumed to be a capture. This critical value could be determined from knowledge of wasp behavior and energetics. Lacking that information, we used a critical visit length of 307 s, the longest observed visit that did not result in a capture. The estimated capture probability $\hat{p}$ is the probability that a visit exceeds 307 s: $\hat{p} = 0.59[1 - \Phi(1.99)] = 0.0137$. The estimate $\hat{p}$ of π is very sensitive to the choice of critical visit length. For example, if the critical length is 360 s, the estimate $\hat{p}$ decreases to 0.0090.

Bootstrap resampling can be used to estimate the precision of $\hat{p}$ (Efron 1981, Dixon 2001). The bootstrapped standard error of the capture probability is estimated to be 0.0095, corresponding to a CV of 66% and precision of 1.10. The estimate from the threshold model is less precise than the frequentist estimate if precision is measured using an absolute measure ($0.0001/s^2$) and more precise if precision is measured by a relative measure (CV).

Using aggregated data from larger scales

The primary data used to estimate the probability of a rare event come from observations of individuals, such as detailed observations of 753 individual *Darlingtonia* plants. Such data provide information about both the number of events (e.g., captures) and the number of trials (e.g., visits). At larger spatial or temporal scales, we can obtain samples of entire populations and observe the total number of rare events over a given interval of time or space (e.g., Lawson and Williams 1994, Plummer and Clayton 1996). This sample yields the product of the rate of occurrence of the event × the number of trials (e.g., capture rate × visitation rate), We can glean indirect information about the rate at which the rare event occurs from this product. Combining the direct and indirect information using a statistical model provides a more precise estimate of the capture probability.

We collected aggregate data on the total number of wasps captured by *Darlingtonia* individuals at several nearby sites over one-hour and two-day periods (A. M. Ellison, R. J. Emerson, E. J. Farnsworth, N. J. Gotelli, C. M. Hart, H. R. Steinhoff, and S. E. Wittman, *unpublished data*). These aggregate data were much easier to collect; we simply counted the number of wasps trapped in each pitcher after one hour or two days, rather than collecting direct behavioral observations. However in the aggregate data, we only recorded the number of wasps successfully captured per pitcher; the number of visits to each pitcher by wasps was not recorded.

Direct observations of wasp behavior suggests that wasps are actively foraging at *Darlingtonia* pitchers only for a 4-h period (10:00–14:00 hours) each day, so the 2-d aggregate data were assumed to reflect all captures made during 8 h of wasp activity. In the aggregate data, a total of six wasps were captured in a total of 1416 plant-hours (162 plants in the 2-d sample = 1296 plant-hours + 120 plants in the 1-h sample). This aggregate information can be combined with the detailed data using a model that relates captures, visits and aggregated data to capture efficiency and visitation rate.

We again use a binomial random variable to model C as a function of π and V:

$$C \mid V \sim \mathrm{Bin}(V, \pi). \tag{9}$$

If visits are rare and independent of each other, the number of visits in the primary data (direct observation of visits and captures) follows a Poisson distribution:

$$V \sim \mathrm{Poiss}(\mu D) \tag{10}$$

where μ is the mean number of visits per plant hour and D is the total number of plant-hours of detailed observations.

The same model (Eqs. 9 and 10) applies to the aggregate data, except that we did not observe the number of visits V. A capture in the aggregated data represents two events: a wasp visits a plant, and then the wasp is captured. If the probabilities of visitation and capture are constant, W, the total number of captured wasps in the aggregate data also has a Poisson distribution:

$$W \sim \mathrm{Poiss}(\pi\mu A) \tag{11}$$

where A is the total number of plant hours of aggregated observations. Because the aggregated information, V, and W follow Poisson distributions, it is also convenient to use a Poisson distribution for the number of captures (cf. Eq. 9):

$$C \mid V \sim \mathrm{Poiss}(\pi V). \tag{12}$$

Note that the Poisson distribution approximates a binomial distribution when the counts of rare events (e.g., captures) are small (Gotelli and Ellison 2004).

The parameters π and μ in Eqs. 9–12 can be estimated using maximum likelihood (Appendix A). When captures are modeled using a Poisson distribution (Eq. 12), π and μ can be estimated using standard software for Poisson regression (Appendix B).

The estimated capture probability is $\hat{p} = 0.0107$, only slightly smaller than the estimate from using the detailed observational data alone (Table 4). However, incorporating the aggregate data increases the precision of this estimate; the CV is reduced to 36%, nearly 50% smaller than the CV of the estimate from only the detailed observational data (Table 4). The absolute precision is increased to 6.75, slightly more than five times the precision of the estimate from only the detailed observational data. To achieve an equally precise estimate using only direct observations of wasp foraging behavior would require just over 2000 plant-hours of continuous observation.

DISCUSSION

Auxiliary data come in many forms. We have illustrated four different methods of using auxiliary data to increase the precision of the estimate of the probability

TABLE 4. Summary of estimated capture probabilities and their coeffients of variation for five estimators of capture probability, ranked from least to most precise.

Estimator	Estimate ($\hat{p}$)	CV	Precision
Frequentist (proportion of captures)	0.0127	70%	1.26
Threshold model (visit > 307 s $=$ capture)	0.0137	66%	1.10
Stratification (best)	0.0127	62%	1.59
Bayesian (using Prior B, Table 1)	0.0104	47%	4.28
Aggregated data	0.0107	36%	6.75

of a rare event (Table 4). For the *Darlingtonia* data, the most precise and appropriate estimate of capture probability was estimated from pooling direct observations with temporally aggregated data. This method led to an estimate that was about twice as precise as the estimate derived from the direct observations alone (Table 4). Bayesian inference using informative, narrow priors yielded slightly less precise estimates. Stratification increased the precision only slightly, whereas modeling the distribution of visit lengths or Bayesian inference using informative priors with very large variance or uninformative priors decreased the precision of the estimate (see also Ellison 2004).

Which method is best? The appropriateness of a particular method can be judged by examining the assumptions and the choices that each requires. Bayesian inference assumes that information from previous studies is available and is relevant to the problem at hand. The relevance can be quantified by choosing the standard deviation of the prior distribution; a small standard deviation (i.e., high precision) implies that the prior is strongly informative, whereas a large standard deviation (i.e., low precision) implies little prior information or prior ignorance. If there is more than one previous study, the between-study standard deviation can be used as an estimate of the prior standard deviation, but if only one previous study is available, more care is needed in setting the precision of the prior, and the value may appear to be arbitrary. In the *Darlingtonia* example, as in many studies of rare events, the precision of the prior was important because when the sample size is small, the posterior will reflect more of the prior. In typical Bayesian analyses reported in the literature, data are more abundant, and the posterior reflects the likelihood of the data more strongly than the prior (Gelman et al. 1995, Ellison 1996, 2004).

Stratification requires strata that can be defined by characteristics other than the response variable. Stratification is most effective when event probabilities differ markedly between the strata. Correct use of stratification also requires that the sizes of each stratum in the sampled population are known. In the *Darlingtonia* example, we chose strata and estimated the sizes of the strata from the sample data. Realistic criteria and supporting information should be used to justify whatever strata are chosen.

Similarly, the modeling approach that incorporates covariates depends on a choice of a threshold value at which a rare event is said to have occurred. In some situations, such as analysis of flood frequencies (Haan 2002) or the probability of structural failure (Heffernan and Tawn 2004), the threshold can be identified clearly and objectively supported. In other situations, such as the *Darlingtonia* example, the threshold must be derived from the data (e.g., the length of a wasp visit designated as a capture was determined from the distribution of visit lengths). Deriving thresholds from the data must be done cautiously, and should be justified whenever possible using independent observations or methods.

Finally, pooling of direct observations and aggregated data assumes that probability of the rare event is the same in both sets of data. Using Poisson distributions for both assumes that there is no between-year or between-site heterogeneity in the rate at which the rare events occur. This assumption of heterogeneity is almost impossible to test when the total number of events is small. In the *Darlingtonia* example, this assumption was reasonable because the two data sets were collected over the same years in the same general area, and each dataset (direct observations and temporally aggregated data) included observations collected from various sites and multiple years.

Estimating the probability π of an event from a series of independent observations is a very common activity in ecology and environmental science. The standard frequentist estimator of π, $\hat{p}$ = number of events n/number of observations N, is unbiased and straightforward to calculate. However, if the event is rare, the estimate is very imprecise if $N < 1000$. By incorporating other kinds of information, some of which may be from other studies, ecologists can increase the precision and the usefulness of these estimates. Ecologists should be alert for ways to incorporate auxiliary data to improve the precision of conventional statistical estimates.

ACKNOWLEDGMENTS

Data collection was supported by grants from the Packard Foundation (to A. M. Ellison) and the National Science Foundation (DEB 03-01361 to A. M. Ellison and DEB 03-01381 to N. J. Gotelli).

LITERATURE CITED

Arber, A. 1941. On the morphology of the pitcher-leaves in *Heliamphora*, *Sarracenia*, *Darlingtonia*, *Cephalotus*, and *Nephentes*. Annals of Botany **5**:563–578.

Barnett, V. 1999. Comparative statistical inference. Third edition. John Wiley and Sons, Chichester, UK.
Berger, J. O. 1985. Statistical decision theory and Bayesian analysis. Second edition. Springer-Verlag, New York, New York, USA.
Birkes, D., and Y. Dodge. 1993. Alternative methods of regression. John Wiley and Sons, New York, New York, USA.
Chave, J. 2004. Neutral theory and community ecology. Ecology Letters **7**:241–253.
Clark, J. S., M. Lewis, and L. Horvath. 2001. Population spread with variation in dispersal and reproduction. American Naturalist **157**:537–554.
Dixon, P. M. 2001. The bootstrap and the jackknife: describing the precision of ecological indices. Pages 267–288 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments, 2nd edition. Oxford University Press, Oxford, UK.
Efron, B. 1981. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika **68**:589–599.
Ellison, A. M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecological Applications **6**:1036–1046.
Ellison, A. M. 2004. Bayesian inference in ecology. Ecology Letters **7**:509–520.
Ellison, A. M., and N. J. Gotelli. 2001. Evolutionary ecology of carnivorous plants. Trends in Ecology and Evolution **16**: 623–629.
Foster, D. R., and J. D. Aber, editors. 2003. Forests in time: the environmental consequences of 1000 years of change in New England. Yale University Press, New Haven, Connecticut, USA.
Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. Bayesian data analysis. Chapman and Hall, London, UK.
Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Markov chain Monte Carlo in practice. Chapman and Hall/CRC, Boca Raton, Florida, USA.
Gotelli, N. J., and A. M. Ellison. 2004. A primer of ecological statistics. Sinauer Associates, Sunderland, Massachusetts, USA.
Haan, C. T. 2002. Statistical methods in hydrology. Second edition. Iowa State University Press, Ames, Iowa, USA.
Harper, J. L. 1977. The population biology of plants. Academic Press, New York, New York, USA.
Heffernan, J., and J. Tawn. 2004. Extreme values in the dock. Significance **1**:13–17.
Hosmer, D. W., and S. Lemeshow. 1989. Applied logistic regression. John Wiley and Sons, New York, New York, USA.
Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey, USA.
Jetz, W., and C. Rahbek. 2002. Geographic range size and determinants of avian species richness. Science **297**:1548–1551.
Lawson, A. B., and F. Williams. 1994. Armadale: a case study in environmental epidemiology. Journal of the Royal Statistical Society, Series A **157**:285–298.
Lewison, R. L., S. A. Freeman, and L. B. Crowder. 2004. Quantifying the effects of fisheries on threatened species: the impact of pelagic longlines on loggerhead and leatherback sea turtles. Ecology Letters **7**:221–231.
Magurran, A. E. 2003. Measuring biological diversity. Blackwell Press, Oxford, UK.
Marriott, F. H. C. 1990. A dictionary of statistical terms. Longman Scientific and Technical, London, UK.
Mayr, E. 1963. Animal species and evolution. Harvard University Press, Cambridge, Massachusetts, USA.
Newell, S. J., and A. J. Nastase. 1998. Efficiency of insect capture by *Sarracenia purpurea* (Sarraceniaceae), the northern pitcher plant. American Journal of Botany **85**:88–91.
Plummer, M., and D. G. Clayton. 1996. Estimation of population exposure in ecological studies. Journal of the Royal Statistical Society, Series B **58**:113–126.
Rabinowitz, D. 1981. Seven forms of rarity. Pages 205–207 *in* H. Synge, editor. The biological aspects of rare plant conservation. John Wiley and Sons, New York, New York, USA.
Roberts, D. L., and A. R. Solow. 2003. Flightless birds: when did the dodo become extinct? Nature **426**:245.
Thompson, S. K. 2002. Sampling. Second edition. John Wiley and Sons, New York, New York, USA.
Van Belle, G. 2002. Statistical rules of thumb. John Wiley and Sons, New York, New York, USA.
Whelan, R. J. 1995. The ecology of fire. Cambridge University Press, Cambridge, UK.
Williams, C. B. 1964. Patterns in the balance of nature and related problems in quantitative ecology. Academic Press, New York, New York, USA.
Zamora, R. 1995. The trapping success of a carnivorous plant, *Pinguicula vallisneriifolia*: the cumulative effects of availability, attraction, retention, and robbery of prey. Oikos **73**:309–322.
Zamora, R., J. M. Gomez, and J. A. Hodar. 1998. Fitness responses of a carnivorous plant in contrasting ecological scenarios. Ecology **79**:1630–1644.

APPENDIX A

A description of the likelihood function for combining detailed and temporally aggregated data is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-059-A1.

APPENDIX B

The SAS code used to fit a Poison regression to detailed and aggregate data is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-059-A2.

Ecology, 86(5), 2005, pp. 1124–1134

STATISTICS OF EXTREMES: MODELING ECOLOGICAL DISTURBANCES

RICHARD W. KATZ,[1,3] GRACE S. BRUSH,[2] AND MARC B. PARLANGE[2,4]

[1]*Environmental and Societal Impacts Group, National Center for Atmospheric Research, Boulder, Colorado 80307 USA*
[2]*Department of Geography and Environmental Engineering, Johns Hopkins University, Baltimore, Maryland 21218 USA*

Abstract. The potential advantage of extreme value theory in modeling ecological disturbances is the central theme of this paper. The statistics of extremes have played only a very limited role in ecological modeling, despite the disproportionate influence of unusual disturbances on ecosystems. An overview of this theory is provided, with emphasis on recent developments that both make more efficient use of the available data on extremes and enable applications that are more ecologically realistic. Consistent with the emphasis on scale in ecology, scaling properties of extremes are emphasized. It is argued that the existence of distributions whose extreme upper tail is ''heavy'' (i.e., decreases at a relatively slow rate) implies that ecological disturbances are sometimes regarded as more ''surprising'' than they ought to be.

The application focuses on modeling disturbances in paleoecology. Two examples are considered: the first, a sediment yield time series for Nicolay Lake in the high Arctic, reflects only the influence of hydrologic disturbances; the second, a sediment rate time series in the Chesapeake Bay, includes both climatic and anthropogenic influences. Strong evidence supports a heavy-tailed distribution for the Nicolay Lake sediment yield, but not necessarily for the Chesapeake Bay sediment rates. For each example, it is demonstrated how the statistics of extremes can readily incorporate information about covariates, such as large-scale atmospheric–oceanic circulation patterns or land use.

Key words: generalized extreme value distribution; generalized Pareto distribution; paleoecology; peaks over threshold; return level; return period; sediment accumulation.

INTRODUCTION

> *It seems that the rivers know the theory. It only remains to convince the engineers of the validity of this analysis.*
>
> —Emil J. Gumbel, 1891–1966

In this quote by Emil Gumbel, ''theory'' refers to the statistical theory of extreme values. Gumbel was a pioneer in the application of this theory, particularly in fields such as climatology and hydrology (Gumbel 1958). Now, several decades after this quote appeared, the engineers are indeed convinced of the theory's utility in water resources management, building design, etc. (e.g., Katz et al. 2002). Yet perhaps this quote would remain apt if the word ''engineers'' was replaced with ''ecologists.''

Extreme events, rare but not necessarily unprecedented, play an important role in ecology (Gutschick and BassiriRad 2003). Ecological disturbances are commonly associated with the occurrence of extreme events, such as an excursion of a climate variable like temperature outside of some range (e.g., above a relatively high, or below a relatively low, threshold). Compounding the problem is the specter of global climate change, with anticipated increases in the frequency of extreme events such as hot spells or intense precipitation (Folland and Karl 2001).

Given that their occurrence is by definition unusual, it has been a challenge for statisticians to devise appropriate methods for quantifying the likelihood and intensity of extreme events. Yet ecologists are largely unaware that, akin to the central limit theorem for averages, a specialized statistical theory is now available for extremes (e.g., Coles 2001). This lack of awareness exists despite at least one review article that has appeared in the ecological literature advocating the use of the statistical theory of extremes (Gaines and Denny 1993; see also Denny and Gaines 2000). This theory can model not just the frequency of rare ecological events (e.g., Dixon et al. 2005), but also their severity.

In the present paper, we build upon the review by Gaines and Denny (1993). An ecologically relevant example is provided with a compelling need for the application of extreme value theory; namely, an instance in which the distribution has a heavy tail (i.e., decreases at a relatively slow rate). Recent developments in the application of this theory, either beyond the scope of or too recent to be taken into account in Gaines and Denny (1993), are surveyed. Among other things, annual or diurnal cycles and trends, as well as more physically based variables such as El Niño events,

---

Manuscript received 29 March 2004; revised 25 June 2004; accepted 29 June 2004; final version received 22 July 2004. Corresponding Editor: A. M. Ellison. For reprints of this Special Feature, see footnote 1, p. 1079.

[3] E-mail: rwk@ucar.edu

[4] Present address: Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland.

can be readily incorporated into the theory as covariates. The peaks over threshold (or point process approach), an alternative to block (e.g., annual) maxima, is introduced to exploit more of the information available about the upper tail of the distribution.

The paper begins with a background section covering the origin and early applications of the statistics of extremes, ecological extremes, and the previous use of the statistics of extremes in ecology. An overview of the statistical theory of extreme values with emphasis on recent developments is then provided, followed by the results from a paleoecological application of modeling disturbances using sediment accumulation. The paper concludes with a discussion of the benefits of extreme value theory in modeling ecological disturbances, as well as the unresolved issue of how to model spatial extremes.

BACKGROUND

Historical origins of statistics of extremes

In the 1920s, a number of individuals simultaneously began deriving the statistical theory of extreme values. An early theoretical breakthrough was produced by the British statisticians R. A. Fisher and L. H . C. Tippett, who derived the limiting form of the distribution of the maximum or minimum value in a random sample (Fisher and Tippett 1928). Tippett immediately applied this theory to the strength of cotton yard, a situation in which the ''weakest link'' (i.e., a minimum value) governs failure (Box 1978). This application could be viewed as a precursor to the field of engineering reliability, in which structural failure is modeled statistically. That ecological stresses reflect statistical extremes, not averages or even variances, was aptly described in Gaines and Denny (1993).

In subsequent decades, extreme value theory found application in other areas in which extreme events naturally play an important role. The first book-length treatment on the statistics of extremes covered a number of applications, many related to engineering design (Gumbel 1958). Applications directly relevant to ecology have included environmental variables such as those in climate (e.g., temperature, precipitation, wind speed), hydrology (e.g., stream flow), and oceanography (e.g., sea level, wave height), with several of these variables being included in the examples in Gaines and Denny (1993).

Ecological extremes

From a scientific perspective, the importance of extreme events in ecology is well recognized. Focusing on plants, Gutschick and BassiriRad (2003) developed the thesis that extreme events ''play a disproportionate role in shaping the physiology, ecology and evolution of organisms.'' Despite fire being an integral component of ecosystems, large fire is a graphic example of a disturbance that can disrupt ecosystem-level processes (Moritz 1997). Perhaps the ultimate extreme event results in the extinction of a population, with extinction probabilities receiving much attention in the ecological literature (Ludwig 1996). Another topic with ecological implications is longevity, particularly the variation in life spans among different species of plants and animals (Carey 2003).

Paleoecology focuses on the reconstruction of the long-term history of ecosystems through the analysis of information, such as the pollen and seeds contained in sediment cores (Brush 1989). This approach is particularly effective at uncovering the influence of disturbances on ecosystems, such as those attributable to changes in land use (e.g., deforestation) or to heavy rains. For instance, sediments have been used to reconstruct temporal and spatial histories of fire regimes in ecosystems (Lynch et al. 2003). The closely related field of paleohydrology focuses on the use of fluvial features that reflect the hydrologic cycle, generally extreme hydrologic events such as paleofloods (Gregory and Benito 2003). Extreme value theory is routinely used to estimate flood frequency and intensity on the basis of measurements of precipitation or stream flow, occasionally in combination with other historical information (Stedinger and Cohn 1986). Yet this statistical theory has rarely ever been applied to the corresponding paleohydrologic or paleoecologic data (an exception is Lamoureux [2000]).

Statistics of extremes in ecology

Despite the review article by Gaines and Denny (1993), it remains difficult to find examples of the explicit application of the statistical theory of extreme values to ecology. Instead, assumptions about tail behavior are typically made which are more restrictive and difficult to verify. The conventional approach effectively assumes that the form of distribution, chosen on the basis of a fit dominated by the majority of the observations in the center of the distribution, necessarily fits the upper and lower tails of the distribution satisfactorily as well.

All the examples of the application of the statistics of extremes presented in Gaines and Denny (1993) happen to involve variables (i.e., minimum and maximum sea surface temperature, maximum wind speed, maximum ocean wave force, and maximum human life span) whose upper (or lower) tail is either unbounded, but ''light'' (i.e., decreases at a relatively rapid rate) or bounded (the shape of the tail of a distribution will be more formally characterized in the next section). For variables such as temperature (e.g., Brown and Katz 1995) and wind speed (e.g., Palutikof et al. 1999), this characteristic of a light or bounded upper tail is known to hold more generally. Nearly all the continuous statistical distributions (e.g., normal, exponential, gamma, lognormal) commonly used to model ecological and related variables have light upper tails.

Although the statistics of extremes certainly can be useful in applications involving light or bounded tails, its need is more compelling for variables which instead possess a heavy-tailed distribution (e.g., Katz et al. 2002). One such ecological example concerns fire disturbance, in which the distribution of the largest fire within a region is heavy-tailed (Moritz 1997), work evidently inspired by Gaines and Denny (1993). Similarly, Schoenberg et al. (2003) reviewed the use of a power law (or Pareto) distribution (i.e., a form of heavy-tailed distribution [Arnold 1983]) to fit wildfire size. Rates of population spread have also been described by distributions which are heavy tailed (Clark et al. 2001). Although not necessarily synonymous with damage to ecosystems, the distribution of economic damage from disturbances such as hurricanes can be heavy-tailed as well (Katz 2002*a, b*).

The search for universal scaling laws in ecology typically involves assuming a form of power-law distribution. For example, Ferriere and Cazelles (1999) showed how a power law relationship can arise in population dynamics, and Keitt and Stanley (1998) fitted a power law relationship in analyzing the dynamics of bird populations. The connection between such scaling laws and extreme value theory will be treated in a subsequent section.

STATISTICAL THEORY OF EXTREME VALUES

Overview of theory

We provide only a relatively short overview of the essential features of the statistical theory of extreme values (for more details, see Leadbetter et al. 1983, Coles 2001, Reiss and Thomas 2001). For convenience, extremes are discussed solely in terms of maxima (or upper tails) of distributions. Nevertheless, minima (or lower tails) are effectively encompassed through the relationship

$$\min(x_1, x_2, \ldots, x_T) = -\max(-x_1, -x_2, \ldots, -x_T) \qquad (1)$$

for a sample of data, $x_1, x_2, \ldots, x_T$.

A heuristic scaling argument provides some motivation for how the basic extreme value theory arises. It involves thinking about different ways to determine the maximum value in a sample of length $2T$, say data $x_1, x_2, \ldots, x_{2T}$. This highest value could be determined indirectly through separately taking the maximum of the first half and of the second half of the sample and then combining these two maxima; that is,

$$\max(x_1, x_2, \ldots, x_{2T}) = \max[\max(x_1, x_2, \ldots, x_T), \max(x_{T+1}, x_{T+2}, \ldots, x_{2T})]. \qquad (2)$$

This relationship constrains the possible form of the limiting distribution of the maximum, suitably normalized, of a sequence of random variables, say $X_1, X_2, \ldots, X_T$ (for the moment, assumed to be independent and identically distributed). Such a limiting distribution must satisfy the "max-stability property" (Leadbetter et al. 1983); namely, distributions for which the operation of taking the maximum of a finite sequence of independent and identically distributed random variables leads to an identical distribution, except for differing location and scale parameters (the concepts of "location" and "scale" will be explained shortly). This property is akin to the reason why the normal distribution arises in the central limit theorem; that is, the mean of normally distributed observations has exactly a normal distribution.

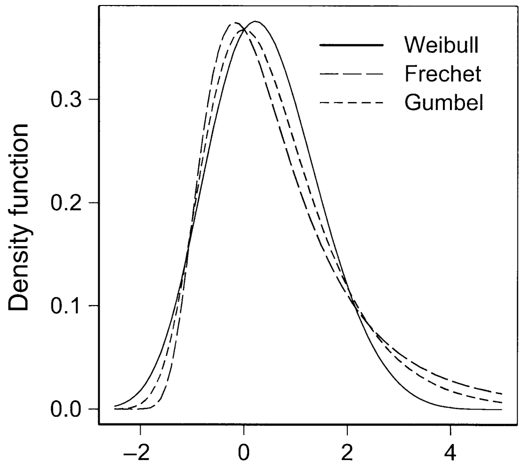

FIG. 1. Plots of the GEV (generalized extreme value) probability density function with $\mu = 0$, $\sigma = 1$, $\xi = -0.2$ (Weibull type), $\xi = 0$ (Gumbel), and $\xi = 0.2$ (Fréchet).

The max-stability property gives rise to a form of distribution known as the generalized extreme value (GEV), with cumulative distribution function

$$F(x; \mu, \sigma, \xi) = \begin{cases} \exp\{-[1 + \xi(x - \mu)/\sigma]^{-1/\xi}\}, & \\ \qquad 1 + \xi(x - \mu)/\sigma > 0 & \xi \neq 0 \\ \exp\{-\exp[-(x - \mu)/\sigma]\} & \xi = 0. \end{cases} \qquad (3)$$

Here μ is termed a location, $\sigma > 0$ a scale, and ξ a shape parameter. If the random variable X has a GEV distribution (Eq. 3), then the standardized variable $(X - \mu)/\sigma$ has a distribution that does not depend on either μ or σ, only on ξ. Like the mean and standard deviation of the more familiar normal distribution, the location parameter specifies where the distribution is "centered," the scale parameter its "spread." It should be noted that Gaines and Denny (1993) used a somewhat different, but equivalent parameterization.

The shape of the GEV distribution assumes three possible types (Fig. 1), depending on the value of ξ:

(i) $\xi = 0$, a light-tailed (or Gumbel) distribution;

(ii) $\xi > 0$, a heavy-tailed (or Fréchet) distribution;

(iii) $\xi < 0$, a bounded (or Weibull) distribution.

The type (i) distribution has an unbounded upper tail which decreases at a relatively rapid (i.e., exponential) rate. Although the type (ii) distribution also has an unbounded upper tail, it decreases at such a slow (i.e., power law) rate that its moments are infinite for all orders greater than $1/\xi$ (e.g., the variance is infinite if $\xi > 0.5$; the mean is infinite if $\xi > 1$). The type (iii) distribution has a finite upper bound at $x = \mu - (\sigma/\xi)$.

It is natural to focus on the extreme upper quantiles of the GEV distribution. Specifically, the ‘‘return level’’ associated with a ‘‘return period’’ of $1/p$ is the $(1 - p)$th quantile of the GEV distribution (e.g., when modeling annual maxima, $p = 0.01$ would correspond to a 100-yr return period), a concept used extensively in hydrology (e.g., Katz et al. 2002). This quantity can be obtained by inverting the distribution function of the GEV (Eq. 3), giving

$$F^{-1}(1 - p; \mu, \sigma, \xi) = \begin{cases} \mu - (\sigma/\xi)\{1 - [-\ln(1 - p)]^{-\xi}\} & \xi \neq 0 \\ \mu - \sigma \ln[-\ln(1 - p)] & \xi = 0 \end{cases} \qquad (4)$$

with $0 < p < 1$ (Coles 2001). The return level is linear in μ and σ, but highly nonlinear in ξ.

It is not actually necessary to assume that the observations be independent, with the limiting distribution of the maximum still being the GEV under a wide range of dependence conditions (e.g., for an autoregressive process) and the only effect being on the values of the location and scale parameters, μ and σ (Leadbetter et al. 1983). The assumption of identically distributed observations can also be relaxed, with nonstationarity being introduced through covariates. The parameters of the GEV distribution could be dependent on the time t; say,

$$\mu(t) = \mu_0 + \mu_1 t, \qquad \ln[\sigma(t)] = \sigma_0 + \sigma_1 t, \qquad \xi(t) = \xi, \qquad t = 1, \ldots, T \qquad (5)$$

the location parameter and the logarithm of the scale parameter (applying the logarithm to impose the constraint of $\sigma > 0$) being linear functions of time and the shape parameter being independent of time. More generally, a covariate z_t (e.g., the state of the El Niño phenomenon at time t) could be used instead of time t in Eq. 5.

As did Gaines and Denny (1993), we estimate the parameters of the GEV distribution by the method of maximum likelihood (see the Supplement). Although other types of parameter estimation techniques are sometimes used in fields such as hydrology, one advantage of the maximum likelihood approach is that covariates (as in Eq. 5) can be readily incorporated (Coles and Dixon 1999). In this way, annual or diurnal cycles and trends need not be removed before fitting the GEV distribution (as did Gaines and Denny [1993]). This approach has the advantage of allowing for cycles or trends in the upper tail not necessarily of the same form as in the center of the distribution (Smith 1989).

Recent developments

The GEV distribution (Eq. 3) is typically fitted to ‘‘block maxima,’’ such as the highest daily precipitation amount over an entire year. Although this approach is sometimes viewed as advantageous because it requires only a simplified data summary (Gaines and Denny 1993), it is disadvantageous because it does not make use of all of the information available about the upper tail of the distribution (e.g., the two highest daily precipitation amounts over the entire record might occur in the same year). An alternative approach whose idea originated in hydrology, called ‘‘peaks over threshold’’ (POT), attempts to exploit more of this information (Todorovic and Zelenhasic 1970).

We start by envisioning the process by which extremes occur, defining an extreme event in terms of exceeding a relatively high threshold, say $X > u$. Recalling that the Poisson distribution arises as an approximation to the binomial for rare events, it is reasonable to assume that the sequence of times the event occurs is governed by a Poisson process, say with rate parameter $\lambda > 0$. In this case, the number of occurrences in a time interval of length T, N_T say, has a Poisson distribution with mean λT; that is,

$$\Pr\{N_T = k\} = [(\lambda T)^k e^{-\lambda T}]/k!, \qquad k = 0, 1, \ldots . \qquad (6)$$

The other aspect of an extreme event is its intensity, termed the ‘‘excess’’ over the threshold, say $Y = X - u$. Consistent with the theory about the limiting distribution of the maximum being GEV, the distribution of the excess should have an approximate generalized Pareto (GP) distribution for a sufficiently high threshold (Pickands 1975). The GP has cumulative distribution function

$$F(y; \sigma^*, \xi) = \begin{cases} 1 - [1 + \xi(y/\sigma^*)]^{-1/\xi}, & \\ \qquad 1 + \xi(y/\sigma^*) > 0 & \xi \neq 0 \\ 1 - e^{-y/\sigma^*} & \xi = 0. \end{cases} \qquad (7)$$

Here $y > 0$, with $\sigma^* > 0$ being a scale and ξ a shape parameter. If the random variable Y has a GP distribution, then the rescaled random variable Y/σ^* has a distribution which does not depend on σ^*, only on ξ. As for the GEV, the scale parameter governs the spread of the distribution.

The shape parameter of the GP distribution has precisely the same interpretation as for GEV distribution. In this case, the three possible types (Fig. 2) are:

(i) $\xi = 0$, a light-tailed (or exponential) distribution;

(ii) $\xi > 0$, a heavy-tailed (or Pareto) distribution;

(iii) $\xi < 0$, a bounded (or beta) distribution.

The type (i) or exponential distribution has a ‘‘memoryless’’ (or lack of aging) property (e.g., Ross 1970); that is,

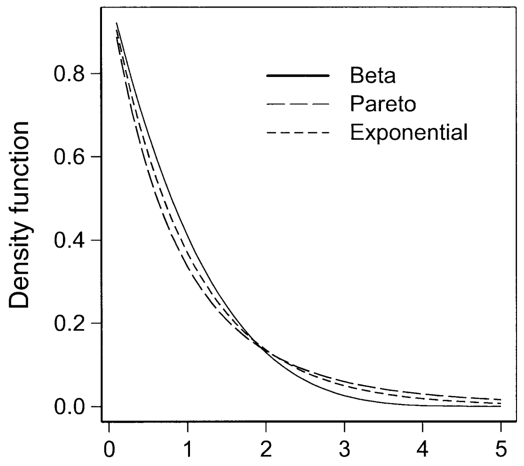

FIG. 2. Plots of the GP (generalized Pareto) probability density function with $\sigma^* = 1$, $\xi = -0.2$ (Beta type), $\xi = 0$ (exponential), and $\xi = 0.2$ (Pareto).

$$\begin{aligned}\Pr\{Y > y + y' \mid Y > y'\} &= \Pr\{Y > y\} \\ &= e^{-y/\sigma^*} \qquad y' > 0. \qquad (8)\end{aligned}$$

In other words, if Y represents life span and has an exponential distribution, then the conditional distribution of future survival is still exponential with the same scale parameter, no matter how long the individual has already survived. This type of distribution would be reasonable for the life expectancy of many species of adult birds in which mortality is primarily the result of accidents, not aging (Ehrlich et al. 1988).

As the exponential is the only memoryless distribution (Ross 1970), the GP distribution necessarily cannot retain this property for a nonzero shape parameter. Suppose the distribution of the excess Y is exactly, instead of only approximately, GP for some threshold u. As the threshold is increased, say to some value $u' > u$, then the distribution of the excess over the higher threshold would remain the GP with the identical shape parameter ξ. Only the scale parameter needs to be adjusted; that is,

$$\sigma^*(u') = \sigma^*(u) + \xi(u' - u) \qquad (9)$$

where the dependence of the scale parameter on the threshold u is made explicit by writing $\sigma^*(u)$ (Coles 2001).

The $(1 - p)$th quantile of the GP distribution can be obtained by inverting Eq. 7, giving

$$F^{-1}(1 - p; \sigma^*, \xi) = \begin{cases} (\sigma^*/\xi)(p^{-\xi} - 1), & \xi \neq 0, \\ \sigma^* \ln(1/p), & \xi = 0 \end{cases} \qquad (10)$$

(Coles 2001). Like the GEV, a GP quantile is linear in σ^*, but highly nonlinear in ξ. In practice, to obtain a return level (analogous to that for the GEV in Eq. 4), the probability p that enters into Eq. 10 would need to be adjusted to take into account the probability of an exceedance of the threshold (Coles 2001).

The fact that the upper tail of essentially any distribution must be approximately of the GP form (Eq. 7) has implications concerning the search for ecological scaling laws. In particular, any distribution must approximately satisfy a scaling law beyond a high threshold if one permits a generalized Pareto, instead of an ordinary Pareto, form. So, at least for the upper tail of a distribution, the only issue which would remain to be addressed in practice is how large a threshold suffices for a good approximation.

A close correspondence exists between the block maxima and POT approaches to the statistical modeling of extremes, with the maximum of a sequence of observations falling below a threshold if and only if there are no exceedances of the threshold. Thus, the POT approach can be used to indirectly fit the GEV distribution, potentially making use of more information about extremes than just block maxima.

A point process representation provides a formal theoretical justification for the POT approach to fitting the GEV distribution. It consists of a two-dimensional, non-homogeneous Poisson process, combining the Poisson process (parameter λ) for the times of exceedance of the high threshold and the GP distribution (parameters σ^* and ξ) for the excesses over the threshold (Leadbetter et al. 1983, Smith 1989, Davison and Smith 1990). The parameters, λ and σ^*, of the point process are related to the parameters of the GEV distribution, μ, σ, and ξ, by

$$\begin{aligned}\ln \lambda &= -(1/\xi)\ln[1 + \xi(u - \mu)/\sigma] \\ \sigma^* &= \sigma + \xi(u - \mu) \qquad (11)\end{aligned}$$

with the shape parameter being identical (Coles 2001).

The block maxima and POT approaches can involve a difference in time scales; for example, with time scale parameter $h \approx 1/365$ if the observations are daily and annual maxima are modeled. To convert the parameters of the GEV distribution, μ, σ, and ξ for time scale h (e.g., annual maxima of daily data), to the corresponding parameters of the GEV, μ', σ', and ξ for time scale h' (e.g., monthly maxima of daily data), it can be shown that

$$\sigma' = \sigma\delta^{\xi} \qquad \mu' = \mu + [\sigma'(1 - \delta^{-\xi})]/\xi \qquad (12)$$

where $\delta = h/h'$ and the shape parameter is unchanged (Coles 2001). Taking logarithms, $\ln(\sigma)$ varies linearly with $\ln(h)$, but $\ln(\mu)$ does not.

In practice, the POT approach requires the choice of a high threshold. No reliable automatic techniques for threshold selection are yet available, but some helpful diagnostics do exist (Coles 2001). The difficulty arises because of the trade-off between making the threshold high enough to ensure that the GP approximation is valid, but not so high that the number of exceedances is too small for accurate estimation of the parameters.

Another issue concerns the possible clustering of high levels, because parameter estimation techniques such as maximum likelihood assume that the time series of excesses over the threshold be independent. For simple forms of dependence such as an autoregressive process, any apparent clustering should gradually disappear as the threshold increases. But there is some evidence that variables such as temperature do exhibit clustering that does not vanish for high thresholds (Coles 2001). To deal with this problem, the chief option is ''declustering.'' Originally developed by hydrologists, this somewhat ad hoc and inefficient procedure requires defining clusters of high levels and using only the single highest value within a cluster (Todorovic and Zelenhasic 1970). Rather than simply being viewed as a nuisance to be eliminated, clusters might well be ecologically relevant for their own sake (e.g., reflecting a persistent hot or wet spell).

Like the block maxima approach, covariates can be incorporated into the POT approach (Eq. 5). By this means, features such as annual or diurnal cycles that are necessarily ignored in the block maxima approach can be modeled as well. Any such cycles need not be removed before analyzing extremes as did Gaines and Denny (1993). For instance, Katz et al. (2002) treated an example of daily precipitation extremes in which annual cycles in both the location and scale parameters of the GEV distribution are permitted.

But how could heavy tails arise in ecological applications? To explain such behavior, at least one chance mechanism exists that is quite plausible ecologically. Suppose that a variable Y has a distribution with an exponential (i.e., light) upper tail (Eq. 7), but that the inverse of the scale parameter (i.e., rate parameter $\nu = 1/\sigma^*$) of the exponential is itself a random variable with a gamma distribution (say with probability density function f_ν). That is, the conditional distribution of Y is given by

$$\Pr\{Y > y \mid \nu\} = e^{-\nu y}$$
$$f_\nu(\nu; \alpha, \beta) = [\beta\Gamma(\alpha)]^{-1}(\nu/\beta)^{\alpha-1}e^{-\nu/\beta} \qquad \alpha, \beta > 0. \tag{13}$$

Then the unconditional distribution of Y would have a heavy tail (e.g., Arnold 1983). That is,

$$\Pr\{Y > y\} = (1 + \beta\, y)^{-\alpha} \tag{14}$$

the type (ii) (or Pareto) form under a different parameterization than that in Eq. 7 (i.e., with shape parameter $\xi = 1/\alpha > 0$). This chance mechanism, by which an unconditional heavy tail is generated from a conditional light tail, is similar to other mathematical modeling with a long tradition of use in ecology (Pielou 1977). For instance, the gamma distribution for the rate parameter could reflect some additional source of random variation in an ecological variable (e.g., associated with an unobserved covariate).

APPLICATION TO PALEOECOLOGICAL DISTURBANCES

Nicolay Lake sediment yield

The first example consists of a nearly 500-yr annual time series of sediment yield in a pristine environment in the high Arctic, with presumably minimal anthropogenic influence (Fig. 3). These sediment yields were reconstructed from varved sediments at Nicolay Lake on Cornwall Island, Nunavut, Canada (Lamoureux 2000). Any disturbances in the time series are predominantly of a hydrologic origin, with high sediment yields being associated with summers in which intense rainfall occurs.

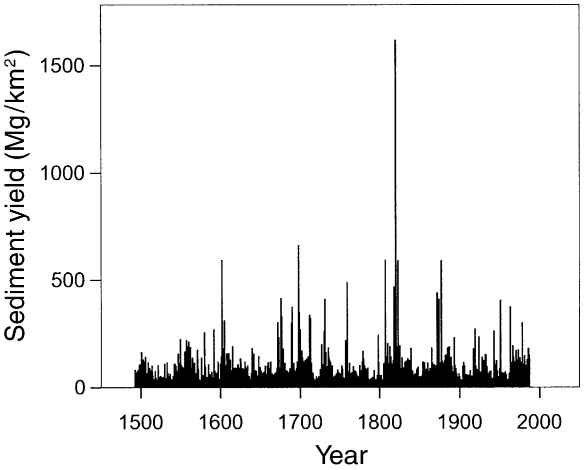

FIG. 3. Time series of annual sediment yield at Nicolay Bay, Nunavut, Canada, 1493–1987 (Source: Lamoureux 2000).

Lamoureux (2000) fitted a GEV distribution directly to the annual sediment yield time series (i.e., without taking block maxima). We repeat this analysis along with the alternative POT modeling approach. Table 1 summarizes the results of fitting a GEV distribution to this data by maximum likelihood. Consistent with Lamoureux (2000), the evidence of a heavy tail (i.e., $\xi > 0$) is very strong. A 95% confidence interval for the shape parameter ξ (all intervals based on the method of profile likelihood, Coles 2001) has a lower bound of about 0.3 or well above zero. As a diagnostic check, a quantile–quantile (Q–Q) plot for the fitted GEV distribution (Fig. 4a) is approximately linear, indicating that the assumed form of distribution is reasonable.

A more traditional approach in paleoclimatology or paleohydrology would be to assume a lognormal distribution; that is, the logarithmically transformed sediment yields are normally distributed (Rittenour et al. 2000). Like the GEV (or GP), a lognormal distribution is positively skewed; but unlike the GEV (or GP), it has a light upper tail (Leadbetter et al. 1983). The estimated 500-yr return level for the fitted lognormal distribution is 523.9 Mg/km^2, well below the highest and second highest observed values in the ~500-yr record of 1620.3 in 1820 and 663.7 in 1698. For the fitted GEV distribution, the estimated 500-yr return level is just under 1000 Mg/km^2 (i.e., $p = 0.002$ in Eq. 4), with the corresponding 95% confidence interval having an upper limit a bit smaller than the highest observed val-

TABLE 1. Parameter and return level estimates with standard errors (or confidence intervals) for GEV and GP distributions (as well as GEV via point process approach) fitted to time series of annual sediment yield (Mg/km^2) at Nicolay Bay, Nunavut, Canada, 1493–1987.

Parameter	Estimate	1 SE (or 95% CI)
GEV distribution		
Location μ	67.54	1.85
Scale σ	35.86	1.64
Shape ξ	0.384	(0.305, 0.470)
500-yr return level	987.6	(724.9, 1435.1)
GP distribution†		
Scale σ^*	47.82	5.98
Shape ξ	0.462	(0.277, 0.700)
500-yr return level	1168.9	(732.3, 2451.2)
GEV/POT†		
Location μ	62.84	6.28
Scale σ	30.64	6.33
Shape ξ	0.462	0.107

Notes: Data are from Lamoureux (2000). Abbreviations: GEV, generalized extreme value; GP, generalized Pareto; POT, peaks over threshold.

† Threshold $u = 100$ Mg/km^2.

ue and a lower limit somewhat greater than the second highest observation (Table 1). In other words, the GEV distribution fits the extreme upper tail of the data reasonably well, whereas the lognormal substantially underestimates the likelihood of the most extreme events.

Lamoureux (2000) made a convincing argument for why the Nicolay Lake sediment yield is a proxy for extreme hydrologic events, particularly for short spells of heavy rainfall (i.e., at most a few days in duration). Nevertheless, the fact that the GEV distribution appears to fit the annual sediment yield (without taking block maxima) well could be fortuitous. So the alternative POT approach, focusing only on the upper tail of the distribution of sediment yield, will be applied for comparison.

Table 1 includes the results of fitting the GP distribution to the excesses in annual sediment yield over a threshold of $u = 100$ Mg/km^2 (after a range of possible thresholds was examined). This threshold was exceeded in nearly 40% of the years (i.e., 189 out of 495). The estimated shape parameter is a bit higher than for the original GEV fit, with the 95% confidence interval being somewhat wider primarily due to the decreased sample size. Because of the heavier fitted tail, the estimated 500-yr return level (Eq. 10) is higher than for the original GEV fit, with a 95% confidence interval having a much higher upper limit well exceeding the highest observed value. The Q–Q plot for the fitted GP distribution (Fig. 4b) appears similar to that for the upper tail in the corresponding Q–Q plot for the GEV (Fig. 4a). In fitting the GP distribution, any clustering of the excesses over the threshold, a possibility suggested for the Nicolay sediment yield by Lamoureux (2002), has been ignored.

The point process approach can produce the equivalent fitted model in terms of the GEV parameterization, using the same threshold of $u = 100$ Mg/km^2 and a time scale parameter of $h = 1$. These parameter estimates and standard errors are also included in Table 1. They could also have been obtained indirectly from the fitted GP distribution and Poisson rate parameter λ for exceedance occurrence (estimated as $189/495 \approx 0.382$) using Eq. 11. As previously noted, the shape parameter estimates are necessarily identical. Because only the upper tail of the distribution is modeled, the GEV parameter estimates differ somewhat from those obtained by fitting the entire range of data.

It would be tempting to consider the North Atlantic Oscillation (NAO) as a covariate. This large-scale feature of the atmospheric circulation reflects a ‘‘see-saw’’ in pressure between the Azores High and the Icelandic Low (Hurrell et al. 2003). However, the NAO is well known to be ‘‘teleconnected’’ with climate in the upper latitudes of the Northern Hemisphere during the winter, but not necessarily in the summer as would be required for the Nicolay sediment yield. We use as a covariate, instead, an index of the Icelandic Low alone, the summer (i.e., June–August) mean sea level pressure at Rey-

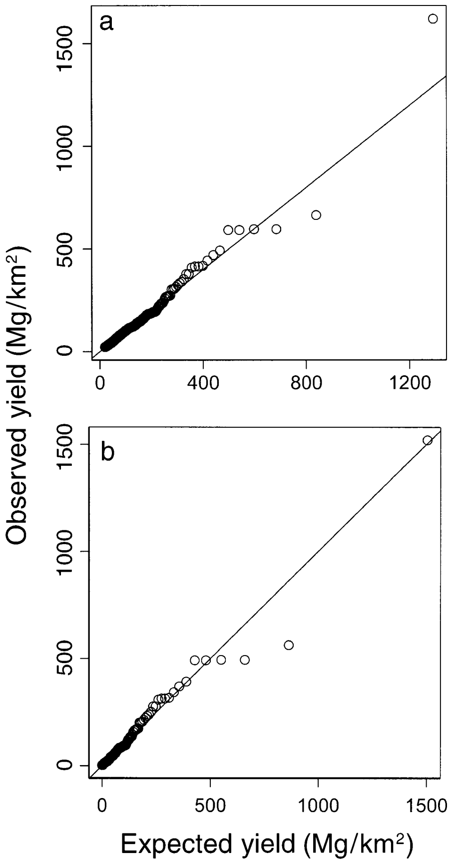

FIG. 4. Quantile–quantile plots for fit of (a) GEV and (b) GP distribution (threshold = 100 Mg/km^2) to time series of sediment yield at Nicolay Bay.

TABLE 2. Parameter estimates and standard errors for GEV distribution (along with negative log likelihood), fitted to time series of sediment yield (Mg/km^2) at Nicolay Bay, 1822–1987, with and without June–August mean pressure (in hPa, minus 1000) at Reykjavik, Iceland, as covariate.

Parameter	Estimate	1 SE
GEV distribution†		
Location μ	73.44	3.36
Scale σ	37.71	2.86
Shape ξ	0.322	0.071
GEV (covariate)‡		
Intercept μ_0	84.18	10.67
Slope μ_1	−1.10	1.04
Scale σ	37.44	2.85
Shape ξ	0.329	0.071

Note: Data are from Jones et al. (1997).
† Negative log likelihood = 895.054.
‡ Negative log likelihood = 894.494.

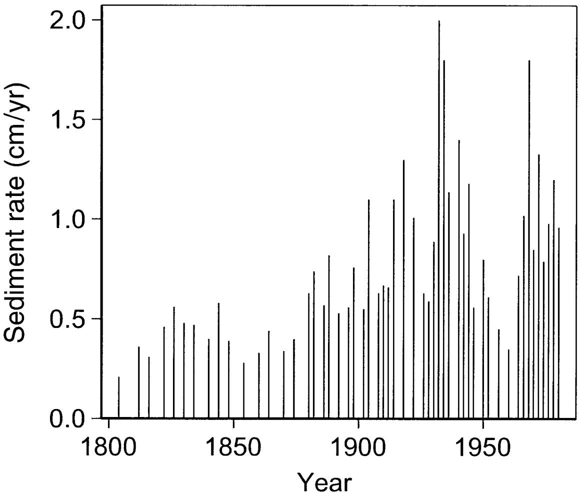

FIG. 5. Time series of sediment rate at Furnace Bay, Maryland, USA, 1804–1980.

kjavik, Iceland. Because long time series indices of the NAO are reconstructed from proxy information such as sediment yield (Cook et al. 2002), we restrict the analysis to the period starting in 1822 for which pressure measurements are available (Jones et al. 1997).

For this considerably shorter time series of sediment yield (i.e., 166 yr), Table 2 compares the fitted GEV distribution with and without the Icelandic pressure covariate, say a random variable denoted by Z. Given a value of the pressure covariate, $Z = z$, the conditional distribution of sediment yield is assumed GEV with a location parameter depending linearly on z; that is, $\mu(z) = \mu_0 + \mu_1 z$. As anticipated, the incorporation of the pressure covariate does not significantly improve the fit. Specifically, comparing the minimized negative log likelihood function with and without the constraint of the slope parameter $\mu_1 = 0$, a P value ≈ 0.290 is obtained for likelihood ratio test. If it was important to focus on the upper tail of the distribution of sediment yield, then the pressure covariate could be introduced via the point process approach instead.

Chesapeake Bay sediment rate

The second example consists of a shorter time series of sediment rates for an estuary in which the influences of climatic and anthropogenic factors are confounded (Fig. 5). These sediment rates were derived from a core at Furnace Bay in the upper Chesapeake Bay, Maryland, USA (Brush 1989). To obtain average sediment rates, the core was divided into 2 cm-wide layers and dated on the basis of pollen profiles. Because of lower temporal resolution before 1800, we only analyze the sediment rates during 1804–1980, a total of 53 measurements with temporal resolution ranging from 2 to 8 yr. Ecologically related disturbances include heavy rainfall, sometimes associated with hurricanes, as well as land use activities such as deforestation associated with agricultural practices.

The interpretation of the Furnace Bay sediment rates is more difficult than for Nicolay Lake, because they do not reflect exclusively extreme hydrologic events. In addition, even if extreme disturbances do play an important role, their influence is attenuated by the limited temporal resolution. Notwithstanding these complications, we repeat the analysis approach used in the Nicolay Lake example.

First, a GEV distribution is fitted to the sediment yield time series without taking block maxima. Table 3 includes the parameter estimates and standard errors, with the estimated shape parameter being positive but barely larger than its standard error. The Q–Q plot for the fitted GEV distribution (not shown) looks adequate. Unlike the Nicolay Lake example, a lognormal distribution would have produced reasonable extreme upper quantile estimates for the Furnace Bay data.

Table 3 also includes the results of fitting the GP distribution to the excesses in sediment rate over a threshold of $u = 0.5$ cm/yr, reducing the effective sample size from 53 to 38. Now the estimated shape pa-

TABLE 3. Parameter estimates with standard errors for GEV and GP distributions (as well as GEV via point process approach) fitted to time series of sediment rate (cm/yr) at Furnace Bay, Maryland, USA, 1804–1980.

Parameter	Estimate	1 SE
GEV distribution		
Location μ	0.567	0.042
Scale σ	0.264	0.033
Shape ξ	0.164	0.131
GP distribution†		
Scale σ^*	0.497	0.117
Shape ξ	−0.181	0.173
GEV/POT†		
Location μ	0.330	0.096
Scale σ	0.528	0.151
Shape ξ	−0.181	0.173

† Threshold $u = 0.5$ cm/yr.

TABLE 4. Parameter estimates and standard errors for GEV distribution (along with negative log likelihood and BIC statistic), with possible shifts in parameters after 1880, fitted to time series of sediment rate (cm/yr) at Furnace Bay, 1804–1980.

Parameter shift	$\mu(1)$	$\mu(2)$	$\sigma(1)$	$\sigma(2)$	$\xi(1)$	$\xi(2)$	$-\ln(L)$	BIC
None	0.567	0.567	0.264	0.264	0.164	0.164	18.165	48.241
μ	0.406	0.693	0.215	0.215	0.145	0.145	6.725	29.332
μ and σ	0.360	0.745	0.100	0.276	0.022	0.022	0.011	19.873†
μ, σ, and ξ	0.374	0.733	0.105	0.267	−0.239	0.110	−0.894	22.034

† Denotes minimum.

rameter is reversed in sign, but still about the same magnitude as its standard error. Despite only a fairly small proportion of the measurements being excluded, even weak evidence for a heavy upper tail has vanished. Again a Q–Q plot (not shown) appears satisfactory. The parameter estimates for the equivalent GEV distribution, based on the point process approach, differ quite a bit from those for the GEV directly fitted to the entire range of data.

An apparent shift in the level of sediment rates occurred around 1880 (Fig. 5), roughly the time of a transition from developing to commercial agriculture (Brush 1989). We incorporate this possible shift by allowing the parameters of the GEV distribution to change after 1880; that is, dealing with two sets of parameters, $\mu(i)$, $\sigma(i)$, $\xi(i)$, for the ith time period ($i = 1, 2$). By dividing the data set into two subsets, such a model could be fitted a single GEV distribution at a time. However, a more flexible approach involves fitting a GEV distribution with covariates. It is straightforward to allow only some of the parameters to vary, and other forms of trend in the parameters can be modeled as well.

Table 4 shows the results of fitting GEV distributions with various constraints on the parameter shifts, ranging from no change in any parameter (i.e., a single GEV distribution) to shifts in all three parameters (i.e., two completely different GEV distributions). Because several candidate models are considered, the Bayesian information criterion (BIC) is used to identify the best model (Schwarz 1978). We define the quantity

$$\mathrm{BIC}(k) = -2 \ln L(k) + k \ln(T) \qquad (15)$$

where $L(k)$ denotes the maximized likelihood function for a candidate model, requiring the estimation of k parameters, and T the sample size. This quantity is evaluated for each candidate model, the one with the smallest BIC value being selected.

For this example, the preferred model is the one in which both the location and scale parameters, but not the shape parameter, shift between the two time periods (Table 4). If tests of significance were employed instead, the same conclusions would have been reached (e.g., a likelihood ratio test, comparing the model in which no parameters are varied with the one in which only μ is varied gives a P value $< 1 \times 10^{-7}$; comparing the one in which only μ is varied with the one varying both μ and σ, a P value ≈ 0.00025). The optimal model has an estimated shape parameter of virtually zero, suggesting the apparent positive shape parameter with no parameter shifts might be an artifact of neglecting the shift in land-use practice. Whether the lack of evidence in favor of a heavy tail reflects a real difference between Furnace Bay and Nicolay Bay in the processes which influence sedimentation or is just an artifact of the measurement process at Furnace Bay cannot be readily ascertained.

In this example, an extreme value model has been adjusted by using time as a surrogate for land use practice. Covariates other than time could be introduced into the model as well. Some of the heaviest rainfall events in the Chesapeake Bay region are associated with remnants of tropical storms (Brush 1989). The frequency of hurricanes in the North Atlantic is well known to be modulated by the El Niño phenomenon (Gray 1984), with the preference for landfall along the Atlantic Coast, as opposed to the Gulf of Mexico, being influenced by the NAO (Elsner et al. 2000). But the very low frequency of hurricane-induced precipitation events within a single water basin, along with the coarse time resolution of the sediment rate time series, would mitigate against such covariates actually improving the fit in practice.

DISCUSSION

Some potential advantages of extreme value theory in modeling ecological disturbances have been pointed out. The reliance on more conventional statistical methods can result in labeling ecological disturbances as more ‘‘surprising’’ than they ought to be. As one example, strong evidence has been presented in support of a heavy-tailed distribution (i.e., decreasing at a power law rate) for the Nicolay Lake sediment yield. Although not identifiable, the origin of this heavy tail is consistent with the heavy tail that precipitation amount tends to possess (Katz et al. 2002), as well as with sediment yield being an integrator over a catchment. By not exaggerating how unusual they are, the proper treatment of extreme events (e.g., large fire) as an integral part of ecosystems would be facilitated.

It has been demonstrated how the statistics of extremes can readily incorporate information about covariates, such as patterns in large-scale atmospheric–oceanic circulation or shifts in land use associated with agricultural practices. For instance, the Furnace Bay sediment rate was shown to be better fitted by an extremal model whose parameters shift corresponding to the transition from developing to commercial agriculture. More generally, through adjustment to reflect the dynamic nature of ecosystems, the degree of ecological realism of extreme value models can be enhanced.

The issue of scale, both temporal and spatial, can be paramount in ecology. We have only focused on the temporal component in our treatment of extremes. In particular, an apparently unappreciated connection between the existence of power laws in ecology and statistical extreme value theory has been identified. Regarding spatial patterns in extremes, Gaines and Denny (1993) observed consistency in the parameter estimates of extreme value distributions for certain variables at different spatial locations. They postulated the ''possible existence of underlying principles governing these phenomena.'' While it may be difficult in practice to determine whether such consistency is attributable to a bona fide ecological invariance principle or just to the statistical theory of extreme values, it would certainly be useful to examine more systematically how ecological extremes vary spatially. In hydrology, ''regional analysis'' has long been used to obtain more accurate estimates of flood probabilities, exploiting the fact that some extremal parameters only gradually vary within a region (e.g., Hosking et al. 1985).

In principle, fully spatial-temporal modeling of extremes can be attempted through the extension of the concept of max-stability to max-stable processes, but applications so far are rather limited and unrealistic (Coles 1993). Some empirical work on how extremes scale spatially, particularly in the hydrologic context of flood estimation, has detected apparent universal behavior but connections to extreme value theory have not yet been made (Smith 1992).

Acknowledgments

We thank Scott Lamoureux for providing the data on Nicolay Lake sediment yield. This research was partially supported by NSF grants DMS-9815344 and DMS-0355474 to the NCAR Geophysical Statistics Project. The National Center for Atmospheric Research is sponsored by the National Science Foundation. We gratefully acknowledge the comments of Associate Editor-in-Chief Aaron Ellison and two anonymous referees.

Literature Cited

Arnold, B. C. 1983. Pareto distributions. International Co-operative Publishing House, Fairland, Maryland, USA.

Box, J. F. 1978. R. A. Fisher: the life of a scientist. John Wiley and Sons, New York, New York.

Brown, B. G., and R. W. Katz. 1995. Regional analysis of temperature extremes: spatial analog for climate change? Journal of Climate **8**:108–119.

Brush, G. S. 1989. Rates and patterns of estuarine sediment accumulation. Limnology and Oceanography **34**:1235–1246.

Carey, J. R. 2003. Longevity: the biology and demography of life span. Princeton University Press, Princeton, New Jersey, USA.

Clark, J. S., M. Lewis, and L. Horvath. 2001. Invasion by extremes: population spread with variation in dispersal and reproduction. American Naturalist **157**:537–554.

Coles, S. G. 1993. Regional modelling of extreme storms via max-stable processes. Journal of the Royal Statistical Society, Series B **55**:797–816.

Coles, S. 2001. An introduction to statistical modeling of extreme values. Springer-Verlag, London, UK.

Coles, S. G., and M. J. Dixon. 1999. Likelihood-based inference for extreme value models. Extremes **2**:5–23.

Cook, E. R., R. D. D'Arrigo, and M. E. Mann. 2002. A well-verified, multiproxy reconstruction of the winter North Atlantic Oscillation index since A.D. 1400. Journal of Climate **15**:1754–1764.

Davison, A. C., and R. L. Smith. 1990. Models for exceedances over high thresholds. Journal of the Royal Statistical Society, Series B **52**:393–442.

Denny, M., and S. Gaines. 2000. Chance in biology: using probability to explore nature. Princeton University Press, Princeton, New Jersey, USA.

Dixon, P. M., A. M. Ellison, and N. J. Gotelli. 2005. Improving the precision of estimates of the frequency of rare events. Ecology **85**:1114–1123.

Ehrlich, P. R., D. S. Dobkin, and D. Wheye. 1988. The birder's handbook. Simon and Schuster, New York, New York, USA.

Elsner, J. B., K.-B. Liu, and B. Kocher. 2000. Spatial variations in major U.S. hurricane activity: statistics and a physical mechanism. Journal of Climate **13**:2293–2305.

Ferriere, R., and B. Cazelles. 1999. Universal power laws govern intermittent rarity in communities of interacting species. Ecology **80**:1505–1521.

Fisher, R. A., and L. H. C. Tippett. 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society **24**:180–190.

Folland, C. K., and T. R. Karl. 2001. Observed climate variability and change. Pages 99–181 *in* J. T. Houghton et al., editors. Climate change 2001: the scientific basis. Cambridge University Press, Cambridge, UK.

Gaines, S. D., and M. W. Denny. 1993. The largest, smallest, highest, lowest, longest, and shortest: extremes in ecology. Ecology **74**:1677–1692.

Gray, W. M. 1984. Atlantic seasonal hurricane frequency. Part I: El Niño and 30 mb quasi-biennial oscillation influences. Monthly Weather Review **112**:1649–1668.

Gregory, K. J., and G. Benito, editors. 2003. Paleohydrology: understanding global change. John Wiley and Sons, New York, New York, USA.

Gumbel, E. J. 1958. Statistics of extremes. Columbia University Press, New York, New York, USA.

Gutschick, V. P., and H. BassiriRad. 2003. Extreme events as shaping physiology, ecology, and evolution of plants: toward a unified definition and evaluation of their consequences. New Phytologist **160**:21–42.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood. 1985. An appraisal of the regional flood frequency procedure in the UK Flood Studies Report. Hydrological Sciences Journal **30**:85–109.

Hurrell, J. W., Y. Kushnir, G. Ottersen, and M. Visbeck, editors. 2003. The North Atlantic Oscillation: climatic significance and environmental impact. Geophysical Monograph 134, American Geophysical Union, Washington, D.C., USA.

Jones, P. D., T. Jónsson, and D. Wheeler. 1997. Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and south-west Iceland. International Journal of Climatology **17**:1433–1450.

Katz, R. W. 2002*a*. Techniques for estimating uncertainty in climate change scenarios and impact studies. Climate Research **20**:167–185.

Katz, R. W. 2002*b*. Stochastic modeling of hurricane damage. Journal of Applied Meteorology **41**:754–762.

Katz, R. W., M. B. Parlange, and P. Naveau. 2002. Statistics of extremes in hydrology. Advances in Water Resources **25**:1287–1304.

Keitt, T. H., and H. E. Stanley. 1998. Dynamics of North American breeding bird populations. Nature **393**:257–260.

Lamoureux, S. 2000. Five centuries of interannual sediment yield and rainfall-induced erosion in the Canadian High Arctic recorded in lacustrine varves. Water Resources Research **36**:309–318.

Lamoureux, S. 2002. Temporal patterns of suspended sediment yield following moderate to extreme hydrological events recorded in varved lacustrine sediments. Earth Surface Processes and Landforms **27**:1107–1124.

Leadbetter, M. R., G. Lindgren, and H. Rootzén. 1983. Extremes and related properties of random sequences and processes. Springer-Verlag, New York, New York, USA.

Ludwig, D. 1996. Uncertainty and the assessment of extinction probabilities. Ecological Applications **6**:1067–1076.

Lynch, J. A., J. S. Clark, N. H. Bigelow, M. E. Edwards, and B. P. Finney. 2003. Geographical and temporal variations in fire history in boreal ecosystems of Alaska. Journal of Geophysical Research **108**:A8152, doi: 10.1029/2001JD000332.

Moritz, M. A. 1997. Analyzing extreme disturbance events: fire in Los Padres National Forest. Ecological Applications **7**:1252–1262.

Palutikof, J. P., B. B. Brabson, D. H. Lister, and S. T. Adcock. 1999. A review of methods to calculate extreme wind speeds. Meteorological Applications **6**:119–132.

Pickands, J. 1975. Statistical inference using extreme order statistics. Annals of Statistics **3**:119–131.

Pielou, E. C. 1977. Mathematical ecology. Second edition. John Wiley and Sons, New York, New York, USA.

Reiss, R.-D., and M. Thomas. 2001. Statistical analysis of extreme values: with applications to insurance, finance, hydrology, and other fields. Extended second edition. Birkhäuser, Basel, Switzerland.

Rittenour, T. M., J. Brigham-Grette, and M. E. Mann. 2000. El Niño-like climate teleconnections in New England during the late Pleistocene. Science **288**:1039–1042.

Ross, S. M. 1970. Applied probability models with optimization applications. Holden-Day, San Francisco, California, USA.

Schoenberg, F. P., R. Peng, and J. Woods. 2003. On the distribution of wildfire sizes. Environmetrics **14**:583–592.

Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics **6**:461–464.

Smith, J. A. 1992. Representation of basin scale in flood peak distributions. Water Resources Research **28**:2993–2999.

Smith, R. L. 1989. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. Statistical Science **4**:367–393.

Stedinger, J. R., and T. A. Cohn. 1986. Flood frequency analysis with historical and paleoflood information. Water Resources **22**:785–793.

Todorovic, P., and E. Zelenhasic. 1970. A stochastic model for flood analysis. Water Resources Research **6**:1641–1648.

SUPPLEMENT

The R source code for fitting extreme value distributions is available in ESA's Electronic Data Archive: *Ecological Archives* E086-060-S1.

Ecology, 86(5), 2005, pp. 1135–1142

MODELING COUNT DATA OF RARE SPECIES: SOME STATISTICAL ISSUES

Ross B. Cunningham[1] and David B. Lindenmayer

Centre for Resource and Environmental Studies, The Australian National University, Canberra, A.C.T., 0200, Australia

Abstract. Most species abundance data show that a small number of species contribute the vast majority of individuals to a community. Thus, most taxa in a community are uncommon or rare. Yet such species will frequently be of ecological, conservation, or management interest. Data for uncommon or rare species will be presence/absence data or counts of abundance that contain a greater number of zero observations than would be predicted using standard, unimodal statistical distributions. Such data are generally referred to as zero-inflated data and require specialized methods for statistical analysis. Statistical approaches to modeling zero-inflated data include nonstandard mixture models; two-part, conditional models; and birth process models. In this paper, we briefly summarize two of these methods and illustrate the two-part, conditional approach to the problem of modeling count data with extra zeros. An advantage of this approach includes separate fits and separate interpretations of both components of count data; that is to say, the presence/absence component and the abundance component (given presence) can be analyzed separately. This can be valuable not only for simplicity, but also such a two-step method may assist ecological understanding in cases where the basis for species presence might be separated from the underlying reasons affecting the population size of that species at those sites where it is present.

We present two case studies of the application of the two-part conditional model for modeling count data with extra zeros. One deals with modeling relationships between counts of the rare and endangered arboreal marsupial, Leadbeater's possum (*Gymnobelideus leadbeateri*) and habitat variables in the wet eucalypt forests of southeastern Australia. The other is an analysis of data obtained from a monitoring study of seabird nesting from the Coral Sea off northeastern Australia. Finally, we briefly discuss some inferential and practical issues in developing designs and models for presence/absence data (which is the first component in the two-part conditional approach) when observed occurrences are low (e.g., <5%).

Key words: *count data; habitat analysis; Leadbeater's possum; monitoring; over dispersion; rare species; sea birds; statistical modeling; zero-inflated data.*

Introduction

Species occurrence can be broadly described by three key qualitative factors: abundance of a species within a community (common or rare), habitat specificity (common or specialized), and geographic range (widespread or restricted). The full factorial expansion of these factors gives eight combinations (Table 1). Seven of these describe the various types of rarity as defined by Rabinowitz et al. (1986). Note that species in cell 1 (the top row) are widespread and common and do not qualify as rare species.

Many of the forms of rarity in Table 1 indicate that in most cases, a ''rare species'' will typically not be numerically abundant in the majority of communities in which it is a member. This proposition has some congruence with the ideas of Preston (1962) and MacArthur and Wilson (1967) both of whom show that within almost all ecological communities, a few species are common and the vast majority are uncommon or rare. This phenomenon has been known in ecology for a long time (see Williams 1944, reviewed by Gaston 1994). Yet, rare and uncommon species will often be those of substantial management and conservation interest (Meffe and Carroll 1997, Fagan et al. 2002, Hartley and Kunin 2003), in part because they may be among the most extinction-prone taxa in an assemblage.

Frequency data arising from studies of the abundance of rare phenomena will often have special nonstandard features. For example, a common characteristic of count data of rare species is that there are many more zeros than would be expected on the basis of the nonzero data (see Fig. 1 for an example). Indeed, depending on the scale and type of a given field study, this could be true of all classifications in Table 1 except those in cell 1 (i.e., common and widespread taxa that would not be considered rare). Data with many zeros

Manuscript received 30 March 2004; revised 17 June 2004; accepted 23 June 2004; final version received 21 July 2004. Corresponding Editor: A. M. Ellison. For reprints of this Special Feature, see footnote 1, p. 1079.

[1] E-mail: rbc@cres.anu.edu.au

TABLE 1. Table of species distributions and patterns of abundance classified by abundance, habitat specificity, and geographic range (modified from concepts outlined in Cody [1986], Rabinowitz et al. [1986], and New [2000]).

Abundance of species within a community	Habitat specificity	Geographic range	Description
Common	common	widespread	widespread, occurs in a wide range of habitats and is abundant in those habitats (and therefore cannot be considered rare)
Common	common	restricted/localized	highly localized distribution but occurs across a range of habitats and is abundant in places where it occurs
Common	rare/specialized	widespread	widespread, but occurs in few habitats, and is common in places where it occurs
Common	rare/specialized	restricted/localized	highly localized distribution and occurs in few habitats, but is common in places where it occurs
Rare	common	widespread	widespread and occurs across a range of habitats but is scarce in places where it occurs
Rare	common	restricted/localized	highly localized distribution, occurs across a range of habitats but is scarce in places where it occurs
Rare	rare/specialized	widespread	widespread, but occurs in few habitats, and is scarce in places where it occurs
Rare	rare/specialized	restricted/localized	highly localized distribution, occurs in few habitats, and is scarce in places where it occurs

pose statistical challenges because key distributional assumptions (e.g., normality, homoscedasticity, and others) are not fulfilled for standard statistical analyses.

Zero counts can arise in ecological data for two reasons; either they are inevitable, (known as structural or necessary zeros), which arise when presence is not tenable (an example might be the occurrence of the Lion [*Panthera leo*] in Australia), or they are random or accidental zeros, which arise due to sampling (e.g., see Green and Young 1993), where conditions are potentially suitable but absence is observed. There are many possible reasons why absence may be observed ranging from non-detection when present, to absent due to several unidentifiable factors, even though all identifiable factors would suggest presence. A simple example might be that of a target species being removed by a predator (Elton 1927, Kavanagh 1988). In most applications, there is usually insufficient knowledge to distinguish the two types of zeros.

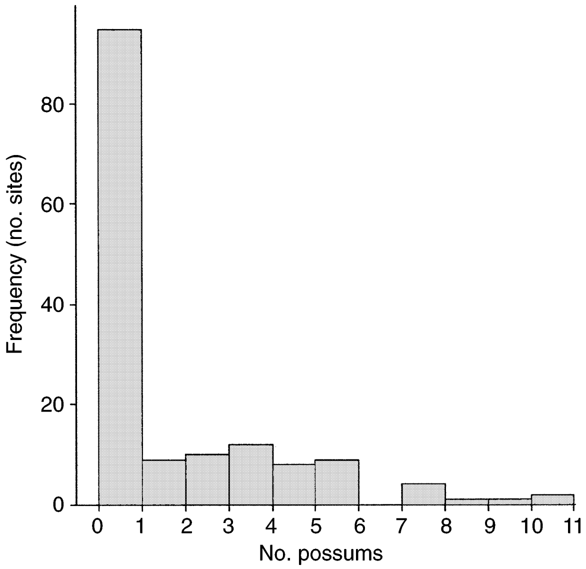

FIG. 1. Frequency graph of the abundance of Leadbeater's possum from 151 field survey sites, each measuring 3 ha in size (see Lindenmayer et al. [1991] for further details).

Statistical modeling provides a powerful framework for modeling presence and/or abundance in terms of a set of possible set of explanators or covariates (Morrison et al. 1992, Welsh et al. 1996, Burgman and Lindenmayer 1998). Such models are commonly used to identify important environmental variables that may explain patterns of distribution and abundance, to provide a concise description of the data, to allow the study of individual cases, and provide a means for constructing valid prediction intervals for new cases (Guisan and Zimmerman 2000). In recent years, there has been considerable activity by statistical scientists and others to develop methodologies for modeling count data with many zeros (Lambert 1992, Ridout and Demetrio 1992, Heilbron 1994, Welsh et al. 1996, Faddy 1998; M. S. Ridout, C. G. B. Demetrio, and J. Hinde, *unpublished manuscript*). These methods are not known or only poorly known by the majority of ecologists. Given this, in this paper we:

1) Briefly outline some of the current approaches to modeling count data with extra zeros. A more detailed review of that approach is given by Ridout et al. (M. S. Ridout, C. G. B. Demetrio, and J. Hinde, *unpublished manuscript*).

2) Provide two illustrations of the use of one of these methods—the two-part approach:

a) a statistical model for count data of Leadbeater's possum (*Gymnobelideus leadbeateri*), a species with restricted distribution, which has limited habitat specificity, but can be relatively common where it does occur.

b) an analysis of data arising from a long-term monitoring program of seabird populations (particularly Frigatebirds [*Fregata minor* and *F. ariel*] and the Red-footed Booby [*Sula sula*]) in the Coral Sea off north-eastern Australia.

3) Discuss some statistical estimation and inference issues that arise when studying rare species and outline some of the practical problems when attempting to model data which exhibit a very low frequency of occurrence.

BRIEF OUTLINE OF APPROACHES FOR MODELING COUNT DATA WITH EXTRA ZEROS

In the section below, we consider several approaches to the relatively widespread problem of modeling data characterized by extra zeros and where extra-Poisson variation may be present. The more accessible approaches are: (1) use standard unimodal distributions for discrete data with extra dispersion; (2) use nonstandard mixture models to account for the extra zeros; and (3) employ two-part models. Another somewhat different approach, not presented here, has been developed and promoted by Faddy (1997), uses ideas associated with birth process models.

Standard unimodal distributions for discrete data with extra dispersion

The baseline model for count data is the Poisson model. Allowing for extra-dispersion (i.e., variance $>$ mean) in a standard Poisson model or fitting a negative binomial model may be a simple way of dealing with data with moderate numbers of extra zeros. These approaches are widely known and the Poisson regression model is a special case of the generalized linear model (McCullagh and Nelder 1989).

For illustrative purposes, let us assume we are dealing with counts of a given animal for a number of sites. Then from the baseline Poisson model, the predicted number of sites with no animals for a total of n sites is $ne^{-\lambda(z)}$, where $\lambda(z)$ is the Poisson mean for the set of covariates z. However, the fit is usually poor because, as is usual in studies of rare species, there are typically many more observations with no animals than would be expected from this model.

The negative binomial distribution can be derived as a standard mixture of Poisson distributions. However as it only has only has one mode, as does the Poisson, it does not deal with the extra-zero problem even though it allows for extra dispersion. Other distributions arising as a mixture of Poisson distributions, such as the Neyman Type A (see Dobbie and Welsh 2001*a*, *b*) can have more than one mode, including a mode at zero, provide a more elegant solution to the problem.

Nonstandard mixture models to account for the extra zeros

Another approach to the problem is to model the response variable as a mixture of a Bernoulli distribution and a Poisson or negative binomial distribution. For the Poisson case, this mixture model, with covariates, is defined as follows:

Given a response vector of counts y_i, $i = 1, \ldots, n$ are independent and

$$\begin{aligned} y_i &= 0 && \text{with probability } 1 - \pi(x) \\ y_i &\sim \text{Poisson}\,[\lambda(z)] && \text{with probability } \pi(x) \end{aligned}$$

so that

$$P(Y = 0 \mid x, z) = 1 - \pi(x) + \pi(x)e^{-\lambda(z)}$$

$$P(Y = r \mid x, z) = \frac{\pi(x)e^{-\lambda(z)}\lambda(z)^r}{r!} \qquad r = 1, 2, \ldots .$$

Here $\pi(x)$ is the probability that the number of animals on a site has a Poisson distribution and, given that the number of animals on a site has a Poisson distribution, $\lambda(z)$ is the mean number of animals on the site. Both π and λ may depend on the same, or possibly a different, set of covariates x and z, respectively. This has become known as the zero-inflated (ZIP) Poisson model (Lambert 1992, Welsh et al. 1996). The simplest form of the model occurs if the covariates x and z coincide. Modification of the standard ZIP by replacing the Poisson distribution with a negative binomial distribution is relatively straightforward and is discussed by Lambert (1992) and Welsh et al. (1996).

ZIPs may provide insight into processes or mechanisms that may have generated the zero data i.e., distinguish between structural and random zeros. However, it will usually be unknown as to whether they distinguish the two types correctly.

Two-part, conditional models

The previous models are based on single distributions, a mixture of distributions or, in the case of ZIPS, an extreme form of mixture. Parameters of the resulting distributions will usually not be independent of each other and so interpretation will be difficult. In the most interesting case, that is regression modeling, parameters of the distribution will typically depend on covariates. This makes interpretation even more complicated.

Here, we consider an alternative to the mixture of distributions idea. Consider a response that has two states: one in which no animals occur and another in which animals occur with varying levels of abundance. If we are only concerned about modeling state 1 (that is, whether any animals occur at a site), then linear logistic modeling is commonly applied. Given that animals are observed, the number of animals recorded can be modeled by a truncated discrete distribution such as the Poisson or negative binomial distribution (Grogger and Carson 1991). We refer to the complete model as the two-part, conditional model. In this case, the components are orthogonal and so the model separates processes that determine whether or not an animal is present from the processes determining the number of animals, given they are present. In the Poisson case, this is defined as follows:

Suppose the counts y_i, $i = 1, \ldots, n$ are independent and

$$y_i = 0 \qquad \text{with probability } 1 - p(x)$$

$$y_i \sim \text{truncated Poisson } [\lambda(z)] \qquad \text{with probability } p(x)$$

so that

$$P(Y = 0 \mid x) = 1 - p(x)$$

$$P(Y = r \mid x, z) = \frac{p(x)e^{-\lambda(z)}\lambda(z)^r}{r![1 - e^{-\lambda(z)}]} \qquad r = 1, 2, \ldots.$$

Here $p(x)$ is the probability of observing at least one animal on a site and, given that there is at least one animal, $\lambda(z)$ is the parameter of the truncated Poisson distribution which describes the number of animals observed.

Note that if we substitute $1 - \pi(x) + \pi(x)e^{-\lambda(z)}$ for $1 - p(x)$ in the above formula we have the ZIP parameterization. However, in the regression context, π and p are different parameters and so the two-part, conditional model and the ZIP model are not equivalent.

A major advantage of two-part, conditional models over ZIPs is that both model fitting and model interpretation of the components can be done separately. This also leads to computational advantages. These models treat both structural and random zeros together.

CASE STUDIES IN MODELING OF COUNT DATA WITH EXTRA ZEROS USING TWO-PART MODELS

*Case study #1. The relationships between the abundance of Leadbeater's possum (*G. leadbeateri*) and key habitat variables*

Our first case study highlights an application of two-part, conditional modeling for use in analysis of the habitat requirements of a rare species. The target taxon is Leadbeater's possum (*Gymnobelideus leadbeateri*), which is a rare and endangered species virtually restricted to the montane ash forests of the Central Highlands of Victoria, southeastern Australia (Lindenmayer et al. 1991). As the species is of considerable management concern because of its occurrence in some of the most valuable wood production forest in Australia (Lindenmayer 2000), it is important for conservation as well as ecological reasons to identify factors which explain occurrence and/or local population sizes. *G. leadbeateri* conforms to cell four (in Table 1) in terms of the type of rarity it exhibits. The species' distribution is highly restricted and virtually confined to the Central Highlands of Victoria—an area of 60 × 80 km. Within this region, particular structural and floristic conditions of wet montane ash eucalypt forests provide suitable habitat for *G. leadbeateri*. Where such suitable habitat does occur, it may support sets of loosely linked colonies of two to 12 animals (Lindenmayer 2000).

As outlined above, the study of species with rarity characteristics like *G. leadbeateri* (Rabinowitz et al. 1986; see Table 1) will often lead to the collection and analysis of data which consists of counts with a high frequency at zero (Gaston 1994). Indeed, a histogram of counts of *G. leadbeateri* from 151 survey sites showed there are many more zeros than would be expected from standard statistical distributions for count data (Fig. 1).

Two-part, conditional Poisson and negative binomial models, as well as a Poisson mixture model (ZIP), were fitted to the data gathered on *G. leadbeateri* (Welsh et al. 1996). Covariates considered included forest age, slope, aspect, tree canopy height, crown cover, a score for degree of decorticating bark, basal area of acacia, the number of shrubs, and the number of trees with hollows (log transformed) (see Lindenmayer et al. 1991 for further details). These data are given in the Appendix.

The selection of significant explanatory covariates ($P < 0.05$) for each model was undertaken by assessing the magnitude of the changes in deviance for both components, and the magnitude of the ratio of parameter estimates divided by their standard errors. These statistics are distributed (approximately) as chi-square and Student's *t*, respectively. Methods for model selection are well known (e.g., see Nicholls 1989, 1991) and have been described in detail by McCullagh and Nelder (1989).

The two-part conditional Poisson model for data on *G. leadbeateri* restricting the linear predictors to have common covariates was:

Component 1: $p = P\{\text{at least one animal present}\}$ and $\text{logit}(p) = -2.178 + 0.857$ lstags (SE $= 0.251$)

Component 2: $\lambda =$ mean abundance of animals (given presence) and $\log(\lambda) = 0.572 + 0.321$ lstags (SE $= 0.103$)

where lstags is the $\log_e$(number of trees with hollows on the site + 1).

As the parameters for this model are orthogonal, interpretation of the components of the model can be made separately. Thus, where animals occur, abundance increases by approximately 0.32% for a 1% increase in the number of trees with hollows and the odds of recording at least one animal at a site is increased by approximately 0.86% for a 1% increase in the number of trees with hollows.

Fitting the conditional negative binomial model to explore the possibility of extra-Poisson variation gave the following results:

Component 2: λ = mean abundance of possums and $\log(\lambda) = 0.485 + 0.344$ lstags (SE $\approx$ 0.181).

Our estimate of the extra variation parameter was 0.143, with an estimated standard error of 0.161. Thus there was no evidence of extra-Poisson variation so the truncated Poisson model was considered appropriate.

The ZIP model for the data on *G. leadbeateri* restricting the linear predictors to have common covariates only was:

Component 1: $p = P\{\text{at least one animal present}\}$ and logit $(\pi) = -1.912 + 0.772$ lstags (SE = 0.257)
Component 2: λ = mean abundance of animals and $\log(\lambda) = 0.573 + 0.320$ lstags (SE = 0.102).

The Poisson mean abundance for the state where animals occur, increases by 0.32% for a 1% increase in the number of trees with hollows, as for the truncated model. In addition, the odds of a site being classified as state 2 (the Poisson model) increase by 0.77% for each 1% change in the number of trees with hollows.

If the case is considered where the probability of absence does not depend on covariates (i.e., logit(π) = 0.422), then the parameters for the Poisson mean model are 0.453 and 0.37, respectively. This differs from the values given above where the probability of absence depends on the number of trees with hollows. Thus, for the ZIP model, the parameters must to be interpreted together, and hence interpretation can be difficult when the covariates affect π and λ differently. Combining the two components leads to a change in abundance, which is not simply proportional to lstags.

When the covariates are allowed to differ in the two components of the model the significant factors include lstags, basal area of wattle, slope, degree of decorticating bark and number of shrubs (Welsh et al. 1996). These results were similar to those obtained using the ZIP model.

Our models showed the species was most likely to occur on sites with numerous trees with hollows and a high basal area of *Acacia* spp. Trees with hollows contain nest sites for the species (Lindenmayer et al. 1991). Sap produced by *Acacia* spp. trees is an important food source for *G. leadbeateri* (Lindenmayer et al. 1994). Mean abundance of *G. leadbeateri* was highest on areas of flatter topography supporting numerous trees with hollows, few shrubs and large quantities of decorticating bark (Welsh et al. 1996). Thus, there were some major differences in the explanatory variables for the models for presence/absence and models for abundance, given presence.

Case study #2. Monitoring the abundance of the Frigatebird nests on North East Herald Cay

In contrast to the previous case study on *G. leadbeateri* that focused on the application of two-part conditional modeling to habitat analysis, our second one below highlights the use of the same approach, but in the context of ecological monitoring as illustrated through ongoing work on seabird nesting.

Australia's Coral Sea Island Territory supports extensive seabird rookeries of great ecological significance, with 13 seabird species recorded breeding in the area. While some of these species such as the Red-footed Booby (*Sula sula*), Lesser Frigatebird (*Fregata ariel*), Great Frigatebird (*Fregata minor*), and Red-tailed Tropicbird (*Phaethon rubricauda*) have an extensive distribution outside of Australian waters, they are uncommon within Australia. The islands and cays of the Coral Sea are important in that they contain a significant proportion of the region's breeding populations (Baker et al. 2004). The species of seabirds conform to cell 3 in Table 1. That is, they are widespread but they have very specialized nesting requirements. However, where they do occur, the numbers of birds can be large.

Given the status of Frigatebirds and other seabirds within Australian waters, a long-term monitoring program of nesting success was established (Baker et al. 2004). One of the areas chosen for study is the Coringa-Herald National Nature Reserve on North East Herald Island where 11 transects set at 100-m intervals were established in 1992. For each transect, quadrats measuring 10 × 10 m were marked. A total of 415 quadrats was established across all habitats, and vegetation mapped for all transects and quadrats. The monitoring program has been designed to allow detection of ''shifts'' in nesting patterns on the island, and permits estimates of change in nest density from year to year (see Welsh et al. 2000).

An essential step in estimating the number of nests constructed by different species of seabirds on North East Herald Cay in each year is to relate the nest counts to other relevant variables such as the number of nests in the previous year, the transect identity, and the quadrat number. The data on seabird nests contains a large number of quadrats with zero counts and is possibly also characterized by extra dispersion (Fig. 2). A two-part, conditional model based on the truncated negative binomial distribution accounted for both the extra zeros and possible extra-Poisson variation was found to be compatible with the data. The key covariate was the number of nests in the previous year. Welsh et al. (2000) provides a detailed account of the complete analysis.

Given that the seabird nest data were collected in a regular spatial pattern, it is possible that counts were spatially correlated. In the initial analysis, this problem was dealt with in a sequential way by examining residuals for spatial dependence. Since then, Dobbie and

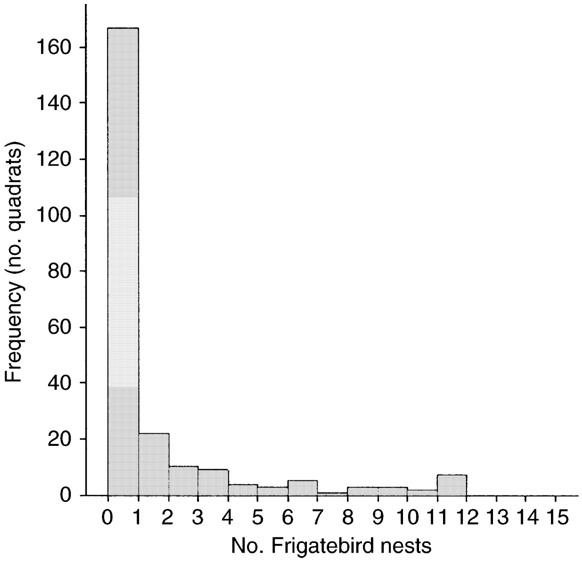

FIG. 2. Frequency graph of the number of Frigatebird (*F. ariel* and *F. minor*) nests on 236 100-m^2 quadrats established in the Coringa-Herald Nature Reserve on North East Herald Island in the Coral Sea.

Welsh (2001*a*) have developed direct methods for modeling spatially correlated zero-inflated count data using the conditional, two-part approach.

SOME OTHER STATISTICAL ISSUES IN THE DESIGN AND ANALYSIS OF STUDIES OF THE OCCURRENCE OF RARE SPECIES

It is well known that when detection probabilities of a species are very low, the ability to be able to establish the statistical significance of an effect tends to be low. That is, there is low statistical power. Fig. 3 (Nicholls and Cunningham 1995) shows the effect of increasing rarity (i.e., low probability of detection) on the standard error of log (relative risk) while maintaining a constant sample size. Here relative risk is a measure of the extent to which a site having a particular attribute is more (or less) likely to have a species present than a site without the attribute.

The effect is that for data on rare species, a change in a factor of interest (e.g., a ‘‘treatment’’) may result in a large relative change in odds of a given species being present, but this may not translate into a statistically significant effect. This is particularly evident as occurrence falls below 5%. Nicholls and Cunningham [1995] provide an example in the context of predicting the distribution of the koala and give details of these calculations.

In many instances when data are scant, particularly in terms of a very low frequency of occurrences (e.g., $<5\%$), numerical computation problems arise in model estimation and the fitting of data. This arises because, on scales appropriate for analysis such as the linear logit scale, an attempt is being made to estimate parameters that are extremely large and negative, i.e., zero, on the natural scale. In essence, there is insufficient information for estimation and this is reflected in large standard errors and/or computational difficulties.

DISCUSSION

Although the various types of rarity have been well discussed in ecology (Rabinowitz et al. 1986; see Table 1), the term ‘‘rarity’’ is used very loosely in the ecological literature. For example, Gaston (1994) listed many studies where the concept of rarity was used but its definition was different in almost all cases. From a statistical perspective, the problems and definitions of rarity and rare species need to be clearly formulated and stated so that progress can be made in solving relevant problems in ecology. This paper has discussed some approaches to modeling data having an excess of zeros; data that commonly arise in studies of rare and uncommon species. Recent advances in statistical methods, as briefly illustrated and discussed in this paper, can assist in the modeling of data on rare and uncommon species, thus providing a powerful, general framework for estimation and inference. Moreover, we believe that, depending on the scale and type of field study in question and the biology of the taxon in question, the modeling methods outlined in this paper would have potentially useful application for a wide range of the types of rare species discussed in detail by Rabinowitz et al. (1986) and outlined in the various cells in Table 1. Such flexibility is illustrated in this paper by both the differences in type of rarity between *G. leadbeateri* and *F. ariel* and *F. minor* and the different applications of modeling to data gathered for them—habitat analysis (i.e., comparative inference) vs. estimating population size (i.e., point estimation)

The methods outlined in this paper for the study of rare and uncommon species are increasingly important for two reasons. First, they extend the number of species in an assemblage which can be subject to the same

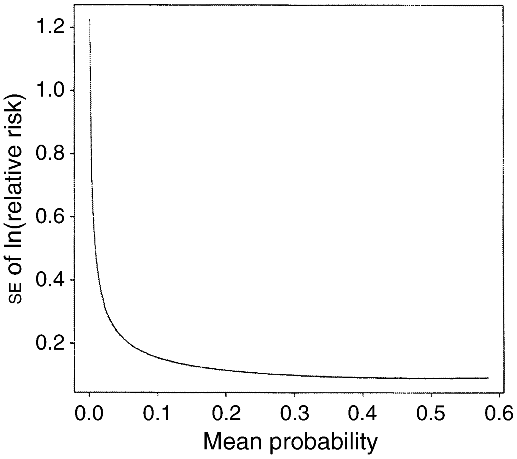

FIG. 3. Relationship between the standard error of ln(relative risk) and rarity.

sound, and flexible statistical analyses available for taxa characterized by extensive presence and abundance data. This is valuable because uncommon and rare taxa can comprise a substantial proportion of the species in most assemblages. In this context, flexible statistical analyses may help quantify what makes some species rare (e.g., quantifying habitat conditions that are themselves rare; as in the case of *G. leadbeateri*) which could make a valuable contribution to biodiversity conservation. Moreover, resource management practices are most often concerned with the conservation and persistence of rare and uncommon species. Hence, for example, the methods outlined here may assist in the quantification of the effects of an experimental treatment or environmental impact on rare or uncommon species.

A second reason why the statistical methods outlined in this paper may be important is that some approaches for modeling zero-inflated data may provide insight into ecological processes/mechanisms, which may have generated the data. For example, through being able to separately model presence-only and conditional abundance (given presence) components of count data, it may (in some circumstances) be possible to identify separate factors affecting the size of a colony of a rare species, given presence, from those factors affecting presence of a species. This was certainly the case for *G. leadbeateri* where several explanatory variables were important in accounting for variation in colony size once conditions (nest trees) are suitable for the presence of the species on a site (Lindenmayer 1996). Indeed, recent efforts in forest management have (1) established a zoning system to ensure that areas that are suitable for the presence of the species are not logged (Macfarlane et al. 1998) and, (2) aimed to harvest forests in new and more environmentally sensitive ways that recreate forest structural conditions which maximize on-site population sizes (D. B. Lindenmayer and R. B. Cunningham, *unpublished data*). These new silvicultural methods are targeted toward flat terrain where it is more operationally feasible to do them and also where our two-part modeling work has shown that population sizes of *G. leadbeateri* will be higher if other suitable habitat attributes can be maintained or created.

Both ZIPs and two-part conditional models tend to be sensitive to the choice of underlying distribution. Gurmu (1997) has suggested some robust, semi-parametric alternatives. Where the nonzero part is not easily modeled by a specific probability distribution, a standard ordinal regression model (see McCullagh 1980) may be a suitable alternative. However, in that case, the covariates for the zero and non-zero parts of the model must be the same.

Our final comment relates to the importance of interdisciplinary work in scientific research. We believe the collaborative approach we have taken to our work, whereby we have combined expertise in ecology and statistics has resulted in outcomes (both in quality and quantity) that exceed the sum of what could have been achieved individually. Such collaborative approaches are increasingly important because both the science of ecology and the science of statistics are extensive and complex and are changing rapidly making it impossible for any one person to keep abreast of new innovations in both fields. Perhaps this is neatly illustrated by the early origins of the ZIP regression approaches applied in the ecological case studies in this paper. The methods were originally developed for problems in econometrics and manufacturing and appeared in a literature read by few (if any) ecologists.

Acknowledgments

The authors thank A. Welsh, M. Dobbie, A. O. Nicholls, and C. Donnelly for past collaborative work on which this paper was based. We also thank A. Ellison for encouraging this paper to be written and coordinating a compilation of studies on the statistics of rarity. Comments by A. Ellison, A. O. Nicholls, and an anonymous referee greatly improved earlier versions of this paper.

Literature Cited

Baker, G. B., M. Double, M. Holdsworth, and M. Hallam. 2004. Seabird monitoring program—Coral Sea Island territory. Report on 2003 field season and update of Herald Cays longitudinal data sets for the period 1992 to 2003. Unpublished report to Department of Environment and Heritage, Canberra, Australia.

Burgman, M. A., and D. B. Lindenmayer. 1998. Conservation biology for the Australian environment. Surrey Beatty, and Sons, Chipping Norton, Sydney, Australia.

Cody, M. L. 1986. Diversity, rarity, and conservation in Mediterranean climate regions. Pages 122–152 *in* M. Soulé, editor. Conservation biology: the science of scarcity, and diversity. Sinauer Associates, Sunderland, Massachusetts, USA.

Dobbie, M. J., and A. H. Welsh. 2001*a*. Modeling correlated zero-inflated count data. Australian and New Zealand Journal of Statistics **43**:431–444.

Dobbie, M. J., and A. H. Welsh. 2001*b*. Models for zero-inflated count data using the Neyman type A distribution. Statistical Modeling **1**:65–80.

Elton, C. S. 1927. Animal ecology. Methuen, London, UK.

Faddy, M. 1997. Extended Poisson process modeling, and analysis of count data. Biometrical Journal **39**:431–440.

Faddy, M. 1998. Stochastic models for analysis of species abundance data. Pages 33–40 *in* D. J. Fletcher, L. Kavalieris, and B. F. Manly, editors. Statistics in ecology, and environmental monitoring 2: decision-making, and risk assessment in biology. University of Otago Press, Dunedin, New Zealand.

Fagan, W. F., P. J. Unmack, C. Burgess, and W. L. Minckley. 2002. Rarity, fragmentation, and extinction risk in desert fishes. Ecology **83**:3250–3256.

Gaston, K. J. 1994. Rarity. Chapman and Hall, London, UK.

Green, R. H., and R. C. Young. 1993. Sampling to detect rare species. Ecological Applications **3**:351–366.

Grogger, J., and R. Carson. 1991. Models for truncated counts. Journal of Applied Econometrics **6**:225–238.

Guisan, A., and N. E. Zimmerman. 2000. Predictive habitat distribution models in ecology. Ecological Modelling **135**: 147–186.

Gurmu, S. 1997. Semi-parametric estimation of hurdle regression models with application to medical utilization. Journal of Applied Econometrics **12**:225–268.

Hartley, S., and W. E. Kunin. 2003. Scale dependency of rarity, extinction risk, and conservation priority. Conservation Biology **17**:1559–1570.

Heilbron, D. C. 1994. Zero-altered, and other regression models for count data with added zeros. Biometrical Journal **36**:531–547.

Kavanagh, R. P. 1988. The impact of predation by the Powerful Owl *Ninox strenua* on a population of the Greater Glider *Petauroides volans*. Australian Journal of Ecology **13**:445–450.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics **34**: 1–14.

Lindenmayer, D. B. 1996. Wildlife and woodchips: Leadbeater's possum as a testcase of sustainable forestry. University of New South Wales Press, Sydney, Australia.

Lindenmayer, D. B. 2000. Factors at multiple scales affecting distribution patterns, and its implications for animal conservation—Leadbeater's possum as a case study. Biodiversity and Conservation **9**:15–35.

Lindenmayer, D. B., R. B. Cunningham, and C. F. Donnelly. 1994. The conservation of arboreal marsupials in the montane ash forests of the Central Highlands of Victoria, south-eastern Australia. VI. Tests of the performance of models of nest tree, and habitat requirements of arboreal marsupials. Biological Conservation **70**:143–147.

Lindenmayer, D. B., R. B. Cunningham, M. T. Tanton, H. A. Nix, and A. P. Smith. 1991. The conservation of arboreal marsupials in the montane ash forests of the Central Highlands of Victoria, south-east Australia. III. The habitat requirements of Leadbeater's possum, *Gymnobelideus leadbeateri* McCoy, and models of the diversity, and abundance of arboreal marsupials. Biological Conservation **56**:295–315.

MacArthur, R. H., and E. O. Wilson. 1967. The theory of island biogeography. Princeton University Press, Princeton, New Jersey, USA.

Macfarlane, M. A., J. Smith, and K. Lowe. 1998. Leadbeater's possum recovery plan, 1998–2002. Department of Natural Resources, and Environment, Government of Victoria, Melbourne, Australia.

McCullagh, P. 1980. Regression models for ordinal data. Journal of the Royal Statistical Society, Series B **42**:109–142.

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Chapman, and Hall, London, UK.

Meffe, G., and C. R. Carroll. 1997. Principles of conservation biology. Sinauer, Sunderland, Massachusetts, USA.

Morrison, M. L., B. G. Marcot, and R. W. Mannan. 1992. Wildlife habitat relationships: concepts, and applications. University of Wisconsin Press, Madison, Wisconsin, USA.

New, T. R. 2000. Conservation biology. An introduction for southern Australia. Oxford University Press, Melbourne, Australia.

Nicholls, A. O. 1989. How to make biological surveys go further with generalised linear models. Biological Conservation **50**:51–75.

Nicholls, A. O. 1991. Examples of the use of generalised linear models in analysis of survey data for conservation evaluation. Pages 54–63 *in* C. R. Margules and M. P. Austin, editors. Nature conservation: cost effective biological surveys, and data analysis. CSIRO Australia, Melbourne, Australia.

Nicholls, A. O., and R. B. Cunningham. 1995. Strategies for evaluating wildlife management. Koala Conservation in the south-east forests. *In* S. Cork, S. Feary, and C. Mackowski, editors. Proceedings of an expert workshop. National Parks and Wildlife Service, Sydney, NSW, Australia.

Preston, F. W. 1962. The canonical distribution of commonness, and rarity. Ecology **43**:185–215,410–432.

Rabinowitz, D., S. Cairns, and T. Dillon. 1986. Seven forms of rarity, and their frequency in the flora of the British Isles. Pages 184–204 *in* M. E. Soulé, editor. Conservation biology: the science of scarcity, and diversity. Sinauer Associates, Sunderland, Massachusetts, USA.

Ridout, M. S., and C. G. B. Demetrio. 1992. Generalised linear models for positive count data. Revista de Metematica s Estatistica **10**:139–148.

Welsh, A. H., R. B. Cunningham, and R. L. Chambers. 2000. Methodology for estimating the abundance of rare animals: seabird nesting on North East Herald Cay. Biometrics **56**: 22–30.

Welsh, A. H., R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmayer. 1996. Modeling the abundance of rare animals: statistical models for counts with extra zeros. Ecological Modeling **88**:297–308.

Williams, C. B. 1944. Some applications of the logarithmic series, and the index of diversity to ecological problems. Journal of Ecology **32**:1–44.

APPENDIX

A table showing Leadbeater's possum abundance data and a selection of habitat covariates is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-061-A1.

Ecology, 86(5), 2005, pp. 1143–1153
© 2005 by the Ecological Society of America

ESTIMATION OF SPECIES RICHNESS: MIXTURE MODELS, THE ROLE OF RARE SPECIES, AND INFERENTIAL CHALLENGES

CHANG XUAN MAO[1,3] AND ROBERT K. COLWELL[2]

[1]*Department of Statistics, University of California, Riverside, California 92521 USA*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269-3043 USA*

Abstract. We examine the role of rare species in the problem of estimating within-habitat species richness based on sampling data. Richness estimation can be modeled realistically for abundance-based and incidence-based data using Poisson or binomial mixtures, respectively. The problem can be reduced to estimation of the odds of the probability of a species remaining undetected in the sample or sample set. Within this rigorous statistical framework, we explore existing methods of richness estimation and assess their limitations. We do this by modeling the addition of increasing numbers of rare, undetected species to a reference assemblage, assessing the power of different methods to distinguish the modified species assemblages from the reference assemblage. (We use empirical example data sets for birds, seeds, and beetles as reference assemblages.) By considering the contributions of rare species and the role of undetected species for a fixed sampling effort, we show why the problem of richness estimation is so difficult, and we discuss what statistical models can provide.

Key words: conditional inference; mixture models; nonparametric maximum likelihood estimation; one-sided inference; richness estimation; richness extrapolation; singleton species; species accumulation curve; transient species; zero-truncated binomial; zero-truncated Poisson.

INTRODUCTION

The concept of species rarity in ecology and biogeography takes many forms, depending primarily upon spatial and temporal scale (Fisher et al. 1943, Preston 1948, Rabinowitz 1981, Gaston 1994, Magurran 2004). We will consider rarity at the habitat level, in the context of alpha diversity (Whittaker 1972) and the estimation of local species richness from sampling data.

With the exception of thorough biotic surveys in isolated, species-poor habitats, it is routine in species inventory work to find that, even after intensive sampling, some species are represented by only one or two individuals (singletons or doubletons) or are detected in only one or two samples in a replicated sample set (uniques or duplicates) (Colwell and Coddington 1994). Often, enlarging the sample (or sample set) yields additional individuals of these rare species, moving them into higher abundance or occurrence classes, but at the same time reveals additional species that now represent new singletons and doubletons or uniques and duplicates. These are the workings of Preston's demon, the moving ''veil line'' between detected and the undetected species as sample size increases (Preston 1948).

For example, in a multiyear study of the insect herbivores of a selected set of plant species in New Guinea, Novotny and Basset (2000) found that 278 of the 1050 insect species recorded (26%) were singletons, based on more than 80 000 individual insects. When a single host-plant species was considered, 45% of the leaf-chewing or sap-sucking insect species were singletons. Recent microbial inventories, made possible by molecular tools, have revealed an astonishing richness of species (however defined), with even more astonishing proportions of singletons, sometimes reaching more than 90% (Hughes et al. 2000, Falkowski and de Vargas 2004).

What do rare species mean, in inventory data? Biologists have long suspected that rare species in many assemblages are a mix of genuinely rare, ''persistent'' or ''resident'' species and ''transient'' or ''occasional'' species that may be common elsewhere (Magurran and Henderson 2003). In the most extensive local inventory of tropical ants on record, Longino et al. (2002) carried out a replicated, multimethod sampling campaign over many years at a Costa Rican rain forest site. Fig. 1 shows the results in the form of a species accumulation curve (a sample-based rarefaction curve [Gotelli and Colwell 2002]). Even when all samples are pooled, the numbers of uniques and duplicates are barely declining, as rarer and rarer occurrence classes come to light. Longino et al. (2002) considered the 51 uniques (12% of the total 437 species) individually and classified them as geographical edge species (14 species, known to be common elsewhere but rare at the site), methodological edge species (six species, probably common at the site, but not susceptible to the survey methods

---

Manuscript received 9 July 2004; revised and accepted 5 August 2004. Corresponding Editor: A. M. Ellison. For reprints of this Special Feature, see footnote 1, p. 1079.

[3] E-mail: cmao@stat.ucr.edu

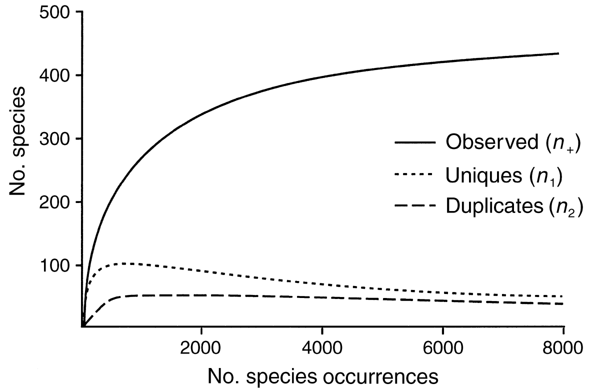

FIG. 1. Species accumulation (sample-based rarefaction) curve for rain forest ants. The lower curves show the number of species detected in only one sample (uniques) and in exactly two samples (duplicates). Note that rare species continue to accumulate even after more than 8000 samples have been pooled (Longino et al. 2002).

used), globally rare species (25 species, known elsewhere but not common anywhere), or globally unique species (known only from a single sample at the site, and nowhere else on earth).

Whatever the explanation for their rarity, the presence of substantial numbers of rare species in sampling data suggests that the inventory is incomplete—that the true species richness for the study habitat includes undetected species. What are the prospects for accounting for those undetected species statistically? Can we tell if an inventory is complete, or nearly so? In this paper, we explore these questions in the context of a statistically rigorous sampling model.

A FRAMEWORK OF MIXTURE MODELS

Two kinds of data

Consider an assemblage with a true richness of S species. The ith species has a true relative abundance ϕ_i for $i = 1, 2, \ldots, S$, with $\Sigma_{i=1}^{S} \phi_i = 1$. In the estimation of the species richness S, two general classes of sampling data may be distinguished: abundance-based data and incidence-based data. The simplest form of abundance-based data is a single, multispecies sample, in which the number of individuals from each species found in the sample is recorded. The number of individuals from the ith species Y_i will be treated as a Poisson random variable with a mean parameter λ_i, called the detection rate. The detection rates λ_i depend on the relative abundances ϕ_i, the probability of an individual being detected when it is present, and the sample size (the number of individuals), which, in turn, is a function of the sampling effort.

Incidence-based data, in the simplest case, consist of a set of multispecies samples (from timed observations, quadrats, traps, lures, seines, dredge hauls, mist nets, or other replicated sampling units) for which only the detection or nondetection of species in each sample is available. Let p_i be the detection probability of the ith species. Let Y_i be the number of samples in which the ith species is detected, which is a binomial random variable. We will use the detection odds $\lambda_i = p_i = (1 - p_i)$ for incidence-based data. The detection odds λ_i depend on the relative abundances ϕ_i, the probability of a species being detected when it is present, the sampling design (e.g., quadrat size), and on nonrandom, species-specific aggregation or disaggregation or individuals among samples (Colwell et al. 2004).

The Y_i will be called frequencies for both kinds of data. If $Y_i = 0$, the ith species does not appear in the sample (for abundance-based data) or sample set (for incidence-based data). (In statistical terms, an incidence-based sample set is treated as a single sample, but we will use the term ''sample set'' for such data, to conform to ecological terminology.) We will assume that if the detection is imperfect, then the effect of imperfect detection does not vary across individuals or species. Thus, for a fixed sampling effort, the detection rates/odds λ_i depend only on the relative abundances ϕ_i, in the sense that a large/small ϕ_i corresponds to a large/small λ_i. The homogeneous case means that the ϕ_i and the λ_i are identical, which is rarely, if ever, true. The heterogeneous case, in which the ϕ_i and the λ_i are allowed to vary across species, will be the focus of this article.

An empirical data set can be summarized in terms of the counts n_x, where n_x is the number of frequencies Y_i that equal x. Thus n_0 is the number of undetected species, n_1 is the number of singletons, and n_2 is the number of doubletons, etc. (For incidence-based data, the n_1 are often called uniques and the n_2 are called duplicates [Colwell and Coddington 1994]). The observed richness $n_+ = \Sigma_{x \geq 1} n_x$ is the number of species detected in the sample (for abundance-based data) or sample set (for incidence-based data), from among the S species actually present.

For example, Janzen (1973) collected an abundance-based data set on tropic beetles; Norris and Pollock (1998) analyzed an abundance-based avian dataset (1995 census data for the Wisconsin route of the North American Breeding Bird Survey); and Butler and Chazdon (1998) recorded the species of tropical plants emerging from seed-bank samples, which we treat here as a replicated, incidence-based sample set. In the beetle data, $n_+ = 78$ species were detected and the nonzero observed counts are $n_1 = 59$, $n_2 = 9$, $n_3 = 3$, $n_4 = 2$, $n_5 = 2$, $n_6 = 2$, $n_{11} = 1$; (Chao and Shen [2003] also analyzed this classic data set). In the bird data, $n_+ = 72$ species were detected and the nonzero observed counts are $n_1 = 11$, $n_2 = 12$, $n_3 = 10$, $n_4 = 6$, $n_5 = 2$, $n_6 = 5$, $n_7 = 1$, $n_8 = 3$, $n_9 = 2$, $n_{10} = 4$, $n_{12} = 1$, $n_{13} = 1$, $n_{14} = 1$, $n_{15} = 2$, $n_{16} = 1$, $n_{18} = 2$, $n_{25} = 1$, $n_{29} = 1$, $n_{30} = 1$, $n_{32} = 1$, $n_{39} = 1$, $n_{44} = 1$, $n_{53} = 1$, $n_{54} = 1$. In the seed-bank data, $n_+ = 34$ species were detected among 121 soil samples and the nonzero observed counts are $n_1 = 3$, $n_2 = 2$, $n_3 = 3$, $n_4 = 3$, $n_5 = 1$, n_6

$= 5$, $n_7 = 1$, $n_8 = 1$, $n_9 = 3$, $n_{10} = 1$, $n_{11} = 2$, $n_{13} = 1$, $n_{17} = 1$, $n_{24} = 2$, $n_{43} = 2$, $n_{47} = 1$, $n_{52} = 1$, $n_{61} = 1$.

The mixture models

In the model for abundance-based data, Y_i is a Poisson random variable with detection rate λ_i. In the incidence-based model, Y_i is a binomial random variable with detection odds λ_i. Let T be the number of samples in an incidence-based data set. A Poisson/binomial density can be written as

$$g(y;\lambda) = \begin{cases} \dfrac{\lambda^y}{y!e^{\lambda}} & \text{(abundance-based)} \\[2ex] \dbinom{T}{y}\dfrac{\lambda^y}{(1+\lambda)^T} & \text{(incidence-based)} \end{cases}$$

where $y \geq 0$ is a possible value of Y_i. Note that we use the term density for the probability mass function throughout this paper. We will assume that $\lambda_1, \lambda_2, \ldots, \lambda_S$ arise as a sample of identically and independently distributed random variables from a latent distribution $H(\lambda)$. For incidence-based data, this is equivalent to the assumption that the detection probabilities p_i also arise as a random sample from a latent distribution. Such a model is called a mixture model because the unconditional distribution of the frequencies Y_i is a mixture of Poisson or binomial distributions.

In a finite mixture model, the latent distribution H is discrete over G values of parameter λ (the support points ξ_k), with mixing weights π_k, $k = 1, 2, \ldots, G$. This means that if one selects one of the λ_i randomly, then the probability that it equals ξ_k is π_k. There is another biological interpretation of a finite mixture model. An assemblage consists of G groups of species, called homogeneous groups. Within each group, the species share the same detection rate/odds, called the group detection rate/odds. The mixing weight is the group relative richness: the number of species in group k divided by S, the total species richness in the assemblage. Fig. 2 illustrates an assemblage of 20 species with four homogeneous groups.

Thus, we treat the frequencies Y_i as a sample from a Poisson or binomial mixture, which we can specify as

$$g_H(y) = \pi_1 g(y;\xi_1) + \pi_2 g(y;\xi_2) + \cdots + \pi_G g(y;\xi_G).$$

where each term in the summation represents the contribution to the mixture from one of the G homogeneous species groups. The model $g_H(y)$ includes not only detected ($Y_i > 0$), but undetected ($Y_i = 0$) species. To model real data, we need a density for only the detected species. Given n_+ observed species, the density of those $Y_i > 0$ is a zero-truncated mixture of Poisson or binomial distributions:

$$g_H(x)/[1 - g_H(0)] \qquad x \geq 1.$$

This can be written as a mixture $f_Q(x)$ of truncated Poisson/binomial densities as follows:

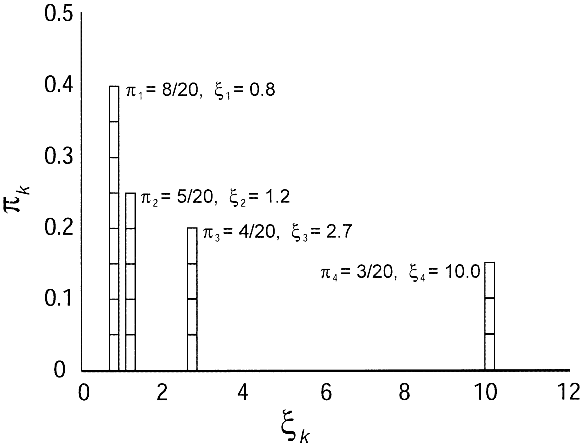

FIG. 2. Four homogeneous groups comprising an assemblage of $S = 20$ species. The group detection rates/odds ξ_k and the group relative richness π_k are indicated in the figure.

$$g_H(x)/[1 - g_H(0)] = f_Q(x) \qquad x \geq 1$$

where Q is a derived latent distribution (taking zero truncation into account) that has the same support points as the latent distribution H, but requires adjusted mixing weights ϖ_k. To be specific, we define

$$\begin{aligned} f_Q(x) = {} & \varpi_1 f(x;\xi_1) + \varpi_2 f(x;\lambda_2) + \cdots \\ & + \varpi_G f(x;\xi_G) \qquad x \geq 1 \end{aligned} \tag{1}$$

$$\begin{aligned} f(x;\lambda) &= \frac{g(x;\lambda)}{1 - g(0;\lambda)} \\ &= \begin{cases} \dfrac{\lambda^x}{x!(e^{\lambda} - 1)} & \text{(abundance-based)} \\[2ex] \dbinom{T}{y}\dfrac{\lambda^x}{(1+\lambda)^T - 1} & \text{(incidence-based)} \end{cases} \end{aligned} \tag{2}$$

$$\varpi_k = \frac{S\pi_k[1 - g(0;\xi_k)]}{S[1 - g_H(0)]} = \frac{1 - g(0;\xi_k)}{1 - g_H(0)}\pi_k. \tag{3}$$

As a probability density, $f(x;\lambda)$ is defined only for $x \geq 1$. For notational convenience, we will use $f(0;\lambda) = g(0;\lambda) = (1 - g(0;\lambda))$, which is not a probability, but it is the odds of a species being undetected if this species has detection rate/odds λ. Note that $f(0;\lambda)$ goes to infinity as λ goes to zero at the same rate as $1/\lambda$, because it is clear that

$$\begin{aligned} B &= \lim_{\lambda\to 0} \lambda f(0;\lambda) \\ &= \begin{cases} \lim\limits_{\lambda\to 0} \dfrac{\lambda}{e^{\lambda} - 1} = 1 & \text{(abundance-based)} \\[2ex] \lim\limits_{\lambda\to 0} \dfrac{\lambda}{(1+\lambda)^T - 1} = \dfrac{1}{T} & \text{(incidence-based).} \end{cases} \end{aligned} \tag{4}$$

This fact will be used in later discussion. As seen from the first equality in Eq. 3, the mixing weight ϖ_k is the group relative expected observed richness in the sense

TABLE 1. An example assemblage from 1995 census data for the Wisconsin route of the North American Breeding Bird Survey.

	k				
Parameter	1	2	3	4	5
ξ_k	2.1328	7.2628	13.6913	30.3418	48.7918
π_k	0.5407	0.2117	0.1362	0.0646	0.0469
ω_k	0.5717	0.1974	0.1269	0.0602	0.0437

Notes: The total species richness in this assemblage is $S = 77$. We present the support points ξ_k and mixing weights π_k of the latent distribution H and ω_k of the derived latent distribution Q. The derived latent distribution Q is the nonparametric maximum-likelihood estimate of the bird data set. H and S are calculated from Q and the observed richness.

that it is the ratio of the expected number of detected species in the kth homogeneous group over the total expected number of detected species from the assemblage. We will identify latent distribution H with the set of parameters $\xi_1, \xi_2, \ldots, \xi_G$ and $\pi_1, \pi_2, \ldots, \pi_G$ and identify the derived latent distribution Q with the set of parameters $\xi_1, \xi_2, \ldots, \xi_G$ and $\varpi_1, \varpi_2, \ldots, \varpi_G$. Table 1 is an example assemblage.

The abundance structure (for abundance-based data) or incidence structure (for incidence-based data) of an assemblage is determined by the total species richness S, the number of homogeneous groups G, the group relative richness π_k, or equivalently the group relative expected observed richness ϖ_k, and the group detection rates/odds ξ_k. Because these parameters are allowed to vary, finite mixture models can provide a good approximation to many real assemblages.

There are parametric mixture models (Fisher et al. 1947, Burnham 1972, Ord and Whitmore 1983, Sichel 1997, Dorazio and Royle 2003). For example, H can be a gamma distribution $H(\lambda; a; s)$ or a beta distribution $H(p; a, b)$ (on the detection probability p), where

$$H(\lambda;\, a,\, s) = \int_0^\lambda \frac{1}{s^a\Gamma(a)} z^{a-1} e^{-z/s}\, dz$$

$$H(p;\, a,\, b) = \int_0^p \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} z^{a-1}(1-z)^{b-1}\, dz.$$

One can find a discrete latent distribution H that approximates a continuous gamma/beta distribution. Most importantly, the mixture $f_Q(x)$ of zero-truncated Poisson/binomial densities from the continuous distribution and that from its discrete approximation can be very close; see Fig. 3.

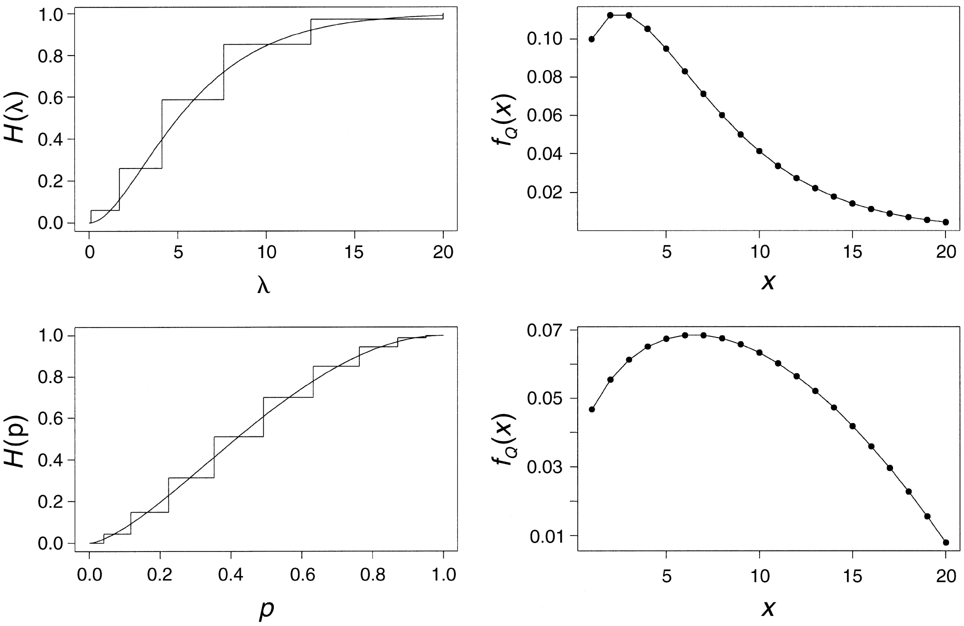

FIG. 3. A continuous latent distribution approximated by a discrete latent distribution. The top left panel presents a gamma distribution $H(\lambda; 2, 3)$ and a discrete latent distribution. The bottom left panel presents a beta distribution with $H(p; 1.5, 2)$ and a discrete latent distribution. The mixtures of the gamma distribution and its discrete approximation are presented in the top right panel. The mixtures of the beta distribution and its discrete approximation are presented in the bottom right panel. For the Poisson case, only $f_Q(x)$ with $x \leq 20$ are plotted; for the binomial case, $T = 20$.

The probability of observing a particular set of frequencies Y_i (a particular pattern of abundances or presences), called the joint density of the frequencies Y_i, is

$$\prod_{i=1}^{S} g_H(y_i) = \prod_{y\geq 0} g_H^{n_y}(y).$$

There are $S!/(\Pi_{x\geq 0}\, n_x!)$ ways to get the same counts n_x with $x \geq 0$ from the Y_i (note that this number includes the count for undetected species n_0). Because $n_0 + n_+ = S$, the density of the counts, as the full likelihood of the parameters S and H, becomes

$$L_1(S, H) = \frac{S!}{(S - n_+)!\prod_{x\geq 1} n_x!} g_H^{S-n_+}(0) \prod_{x\geq 1} g_H^{n_x}(x).$$

The observed counts n_x with $x \geq 1$ are called sufficient statistics because they contain all the information available for us to make statistical inference. The conditional likelihood of the observed counts given the observed species richness depends only on the latent distribution Q,

$$\begin{aligned} L_2(Q) &= \frac{n_+!}{\prod_{x\geq 1} n_x!} \prod_{x\geq 1} f_Q^{n_x}(x) \\ &= \frac{n_+!}{\prod_{x\geq 1} n_x!} \prod_{x\geq 1} \left[\frac{g_H(x)}{1 - g_H(0)}\right]^{n_x}. \qquad (6) \end{aligned}$$

Note that $L_1(S;\, H) = L_2(Q)L_3(S;\, H)$, where $L_3(S;\, H)$ is the density of n_+:

$$L_3(S, H) = \frac{S!}{(S - n_+)!\,n_+!} g_H^{S-n_+}(0)[1 - g_H(0)]^{n_+} \qquad (7)$$

which is called the marginal likelihood.

Predicting the number of undetected species

The key to estimating the true richness S from the sample data lies in modeling the undetected species, n_0. If we knew the latent distribution H, estimation of the species richness would be straightforward. The maximum likelihood estimator for S is the integer part of

$$\hat{S} = n_+ + n_+ g_H(0)/[1 - g_H(0)] \qquad (8)$$

which maximizes the marginal likelihood $L_3(S;\, H)$ in Eq. 7 and linearly depends on H only through the odds of a species being undetected:

$$\alpha = g_H(0)/[1 - g_H(0)].$$

We can write α as $\alpha(Q)$, because it is easy to show that

$$\alpha(Q) = \sum_{k=1}^{G} \varpi_k f(0;\, \xi_k). \qquad (9)$$

Note that $n_+\alpha$ for a fixed α predicts the number of undetected species n_0. Because α is unknown, the problem is reduced to estimation of α.

Statistical estimation methods

Various statistical methods have been applied toward estimation of the species richness S. Bunge and Fitzpatrick (1993) presented a comprehensive review; also see Colwell and Coddington (1994). Here we will summarize various methods from the point of view of mixture models, and in particular, we will include new developments in the last ten years. Each procedure can be thought to provide an estimator for α, the odds of a randomly selected species being undetected. If a procedure produces an estimator $\hat{S}$ for S directly, then we can write an estimator for α as $\hat{\alpha} = \hat{S}/n_+ - 1$.

Although Q is unknown, $f_Q(x)$ can be estimated by the sample proportion $\hat{f}(x) = n_x/n_+$, where $\hat{f}(x)$ is called the empirical density. If a parameter is a function of the $f_Q(x)$, then we can estimate it by replacing $f_Q(x)$ with $\hat{f}(x)$. To see this, consider, for example, a parameter α_{ML} and its estimator $\hat{\alpha}_{\mathrm{ML}}$,

$$\alpha_{\mathrm{ML}} = \alpha_{\mathrm{ML}}(Q) = A\frac{f_Q^2(1)}{2f_Q(2)}$$

$$\hat{\alpha}_{\mathrm{ML}} = A\frac{\hat{f}^2(1)}{2\hat{f}(2)} = A\frac{n_1^2}{2n_+n_2} \qquad (10)$$

where $A = 1$ (abundance-based) or $A = 1 - 1/T$ (incidence-based).

The unfortunate fact is that the odds α cannot be written as an explicit function of the $f_Q(x)$. There are two general recipes toward estimation of α. The first, termed the approximation recipe, means that one finds some parameter $\alpha^{\#}$ that is an explicit function of the $f_Q(x)$ and is assumed to be close to α under certain situations. The estimator $\hat{\alpha}^{\#}$ will be used as an estimator for α. The parameter α_{ML} in Eq. 10 is such an example. It was shown in Mao and Lindsay (2003) and Mao (2004*a*) that for all assemblages, we have $\alpha_{\mathrm{ML}} \leq \alpha$, with $\alpha_{\mathrm{ML}} = \alpha$ for the homogeneous case. This means that α_{ML} is a universal lower bound for α. Chao (1984) obtained α_{ML} for the abundance-based data and Chao (1989) applied it to the incidence-based data without the factor $A = 1 - 1/T$, which is close to one for a large T. Many estimators for α that are well-known to ecologists, under the name of nonparametric estimators, in the sense that $\hat{f}(x)$ is a nonparametric estimator for $f_Q(x)$, can be thought to be based on the approximation recipe, although the original logic that leads to the development of an estimator might be something else and the development might be done in a different model (Burnham and Overton 1978, Darroch and Ratcliff 1980, Smith and van Belle 1984, Zelterman 1988, Chao and Lee 1992, Lee and Chao 1994).

The second strategy, termed the plug-in recipe, means that one finds an estimator $\hat{Q}$ for the latent distribution Q and then plugs the elements of $\hat{Q}$ (Eq. 1) into Eq. 9 to yield an estimator for α. This is also equivalent to estimating the latent distribution H. Because estimation of the latent distribution Q or H involves complicated iterative algorithms, the plug-in recipe is less known and less used in ecology, although several estimation procedures have been proposed (Burnham 1972, Efron and Thisted 1976, Ord and Whitmore 1986, Mingoti and Meeden 1992, Norris and Pollock 1996, 1998, Pledger 2000, Dorazio and Royle 2003, Mao 2004*a*, *b*, Mao et al., *in press*).

Among various Q/H-estimation procedures, the ones that maximize either the full likelihood $L_1(S, H)$ in Eq. 5 or the conditional likelihood $L_2(Q)$ in Eq. 6 have been addressed thoroughly. The nonparametric maximum likelihood procedures in Norris and Pollock (1996, 1998) used $L_1(S, H)$ in Eq. 5. The nonparametric maximum likelihood procedures in Mao (2004*a*) used $L_2(Q)$ in Eq. 6. The parametric maximum likelihood procedure for incidence-based data in Dorazio and Royle (2003) used $L_1(S, H)$ in Eq. 5. The parametric maximum likelihood procedure for the abundance-based data in Efron and Thisted (1976) used a conditional likelihood similar to $L_2(Q)$ in Eq. 6.

THE ROLE OF RARE SPECIES IN INFERENCE

Rare species are interpreted here as those with small relative abundances ϕ_i, which in turn result in small detection rates/odds λ_i. To see how the existence of rare species will affect statistical inference on the total species richness, we will consider fixing an assemblage, called a reference assemblage, and compare it with a modified assemblage, which differs from the reference assemblage by an additional group of species that have a small detection rate/odds. We will denote the reference assemblage as $C = \{S, H\}$ or $C = \{S, Q\}$ and the modified assemblage by $C^* = \{S^*, H^*\}$ or $C^* = \{S^*, Q^*\}$, where the relative species richness, the relative expected observed richness, and the detection rate/odds of the additional rare species group are denoted by π_0, ϖ_0^*, and ξ_0^*, respectively. For each species group in the reference assemblage, its group relative richness and group relative expected observed richness in the modified assemblage have been changed by the addition of the rare species group, as specified by

$$\xi_k^* = \xi_k \qquad \pi_k^* = \pi_k(1 - \pi_0^*)$$
$$\varpi_k^* = \varpi_k(1 - \varpi_0^*) \qquad k = 1, 2, \ldots, G. \tag{11}$$

Fig. 4 illustrates a reference assemblage and a modified assemblage.

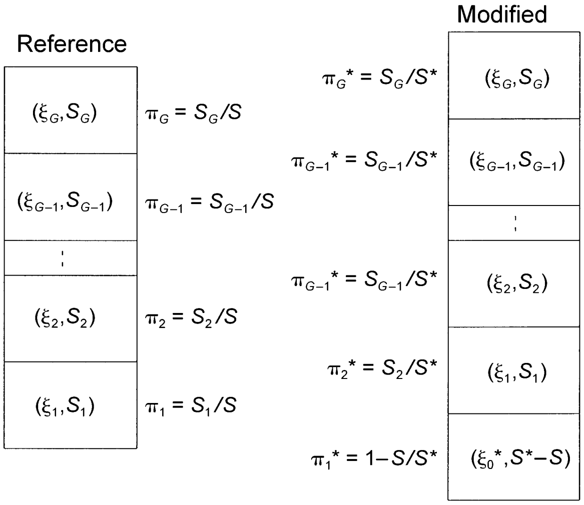

FIG. 4. A reference assemblage and a modified assemblage. Each species group is put in a small box. The modified assemblage has an additional rare species group. The group relative richness changes for a species group in the reference assemblage while the group detection rate/odds is kept the same.

The expected number of species that have frequency x from the kth homogeneous group equals the number of species in the group, $S\pi_k$, times the probability $g(x; \xi_k)$ that a species has frequency x. The expected number of species that have the frequency x from the reference assemblage C is

$$E\{n_x | C\} = \sum_{k=1}^{G} S\pi_k g(x; \xi_k) \qquad x \geqq 0.$$

From Eqs. 2 and 3, and the fact that $S\{1 - g_H(0)\} = E\{n_+ | C\}$, we can write the following:

$$E\{n_x | C\} = E\{n_+ | C\} \sum_{k=1}^{G} \varpi_k f(x; \xi_k) \qquad x \geq 0. \tag{12}$$

Because going from the reference assemblage C to the modified assemblage C^*, we add a group of rare species, we can write the difference as

$$\begin{aligned} E\{n_x | C^*\} - E\{n_x | C\} &= (S^* - S)g(x; \xi_0^*) \\ &= E\{n_x | C^*\}\varpi_0^* f(x; \xi_0^*) \\ &\qquad x \geq 0. \end{aligned}$$

Note that ϖ_0^* is the ratio of the expected number of species being detected from the additional rare species group over the total expected number of species being detected from the modified assemblage, that is,

$$\varpi_0^* = \frac{E\{n_+ | C^*\} - E\{n_+ | C\}}{E\{n_+ | C^*\}} \tag{13}$$

which can be reformulated to

$$E\{n_+|C^*\} - E\{n_+|C\} = E\{n_+|C\}\varpi_0^*/\{1 - \varpi_0^*\}. \tag{14}$$

Due to Eq. 14, we can write Eq. 12 in terms of $E\{n_x | C\}$, a quantity determined by the reference assemblage, and ϖ_0^* and ξ_0^*, the relative expected observed richness and the detection rate/odds of the rare species group, respectively:

$$E\{n_x | C^*\} - E\{n_x | C\} = E\{n_+ | C\}\frac{\varpi_0^*}{1 - \varpi_0^*} f(x; \xi_0^*) \qquad (x \geq 1) \tag{15}$$

$$E\{n_0 | C^*\} - E\{n_0 | C\} = E\{n_+ | C\}\frac{\varpi_0^*}{1 - \varpi_0^*} f(0; \xi_0^*). \tag{16}$$

The right-hand side of each of Eqs. 14, 15, and 16

TABLE 2. Example 1 of modified assemblages.

m	2	10	50	100	1000	2000	5000	Reference
α	0.30	0.47	0.52	0.53	0.53	0.53	0.53	0.07
En_+	84.76	74.91	72.43	72.10	71.81	71.79	71.78	71.77
En_0	25.24	35.09	37.57	37.90	38.19	38.21	38.22	5.23
En_1	21.21	14.19	11.85	11.53	11.24	11.22	11.21	11.20
En_2	14.65	12.30	12.15	12.15	12.15	12.15	12.15	12.15
En_3	9.54	9.13	9.12	9.12	9.12	9.12	9.12	9.12
En_4	5.80	5.75	5.75	5.75	5.75	5.75	5.75	5.75
En_5	3.76	3.76	3.76	3.76	3.76	3.76	3.76	3.76

Notes: The modified assemblage $C^{(m)}$ is obtained by adding to the latent distribution H (the reference assemblage) in Table 1 a support point $\xi_0^{(m)} = 1/m$ with a mixing weight $\pi_0^{(m)} = 0.3$ so that $S^{(m)} = 110$. We present the odds α, the expected observed richness En, and the expected counts En_x for $x = 0, 1, 2, 3, 4, 5$.

stands for the contribution to the expected count or the expected observed species richness from the rare species group. From Eqs. 14 and 15, if the relative expected observed richness of the rare species group ϖ_0^* is small enough, then the contribution from the added rare species group to the expected observed richness or any expected observed count can be as small as possible, that is, for $x \geq 1$, when $\varpi_0^* \approx 0$,

$$0 \leq E\{n_x|C^*\} - E\{n_x|C\}$$
$$\leq E\{n_+|C^*\} - E\{n_+|C\} \approx 0.$$

This means that, with the same sampling effort, empirically one can not tell whether the data are generated from the reference community C or from the modified community C^*. On the other hand, when both ϖ_0^* and the detection rate/odds ξ_0^* of the rare species group are small, using Eq. 4, we can write Eq. 16 as

$$E\{n_0|C^*\} - E\{n_0|C\} = E\{n_+|C\}[\xi_0^* f(0;\, \xi_0^*)](\varpi_0^*/\xi_0^*)$$
$$\approx E\{n_+|C\}B(\varpi_0^*/\xi_0^*).$$

One can easily find ϖ_0^* and ξ_0^* such that the contribution from the rare species group to the expected unobserved count can be as large as possible, for example, using ϖ_0^* and ξ_0^* with $\varpi_0^* = \sqrt{\xi_0^*}$ and $\xi_0^* \approx 0$.

The different contributions to the expected observed counts and the expected unobserved counts from the additional rare species have profound consequences in statistical inference on the odds α and the estimation of the species richness S. From Eqs. 1, 9, and 11, we can also obtain

$$\sum_{x\geq 1} |f_{Q^*}(x) - f_Q(x)| = \varpi_0^* \sum_{x\geq 1} |f(x;\, \xi_0^*) - f_Q(x)|$$
$$\leq 2\varpi_0^* \qquad (17)$$

$$\alpha(Q^*) - \alpha(Q) = \varpi_0^* f(0;\, \xi_0^*) - \varpi_0^* \alpha(Q)$$
$$\approx B\varpi_0^*/\xi_0^* \qquad (18)$$

where the approximation in Eq. 18 holds when both ϖ_0^* and ξ_0^* are small.

Eqs. 17 and 18 suggest that the estimation procedures based on either the approximation recipe or the plug-in recipe might have trouble producing useful results. For a procedure based on the approximation recipe, for example, for α_{ML} in Eq. 10, we will have $\alpha_{\mathrm{ML}}(Q^*) \approx \alpha_{\mathrm{ML}}(Q)$ as $\varpi_0^* \approx 0$. If $\alpha_{\mathrm{ML}}(Q) \approx \alpha(Q)$, then $\alpha_{\mathrm{ML}}(Q^*) \approx \alpha(Q)$ and $\alpha_{\mathrm{ML}}(Q^*)$ can never be close to $\alpha(Q^*)$. The same logic applies to all estimators based on the approximation recipe. There is no such parameter $\alpha^\#$ for α such that $\alpha^\#$ is close to α for all latent distributions.

Consider any procedure that estimates the latent distribution Q. Even if Q fits a data set well, then, because of Eq. 17, when ϖ_0^* is small enough, Q^* should also fit the same data as well as Q and numerically either Q^* or Q might provide a fit slightly better than the other. However, the estimators for α obtained from Q^* and Q will be dramatically different so that estimation of α by plug-in might produce extremely large values. A theoretic consequence for confidence intervals is that, for a confidence interval, for either the odds α or the species richness S, to achieve its advertised confidence level, the upper confidence limit must often infinite.

As an illustration, the assemblage in Table 1 has $S = 77$ species, taken to be the true, complete reference assemblage. We will consider a sequence of modified assemblages $C^{(m)} = \{S^{(m)}, H^{(m)}\}$ with corresponding derived latent distribution $Q^{(m)}$, that are parameterized by m. Table 2 and Fig. 5 show two sequences of modified assemblages. In Table 2, the last column (reference) shows the expected counts and observed species richness for the reference assemblage. To create the modified assemblages in Table 2, we add a fixed number of rare species (33, so that $S^{(m)} = 110$), with a fixed mixing weight $\pi_0^{(m)}$. We then vary the support point $\xi_0^{(m)}$ for the added species according to $\xi_0^{(m)} = 1/m$ so that $\varpi_0^{(m)}$ also becomes smaller as m gets larger, with m ranging from 2 to 5000, to explore the effect of adding rarer and rarer species. (The function $\xi_0^{(m)} = 1/m$ is just a convenient way to vary $\xi_0^{(m)}$ based on the index m.) When $\xi_0^{(m)}$ is large (e.g., $m = 2$, $\xi_0^{(m)} = 0.5$), the added species are reflected in the expected counts for the observed species (e.g., the expected singletons En_1 goes from 11.20 to 21.21), and thus in the expected number of observed species En_+ (which rises from 71.78 to 84.77). But as $\xi_0^{(m)}$ becomes smaller and smaller

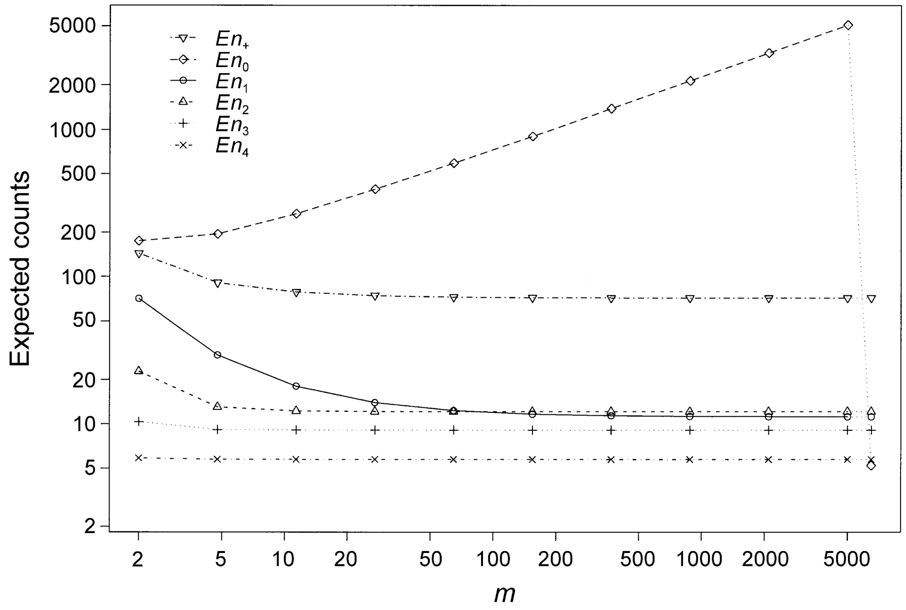

FIG. 5. Example 2 of modified assemblages. Reference (C) and modified ($C^{(m)}$) assemblages with a variable number of added rare species, variable mixing weight, and variable support point for the rare species group. The modified assemblage $C^{(m)}$ is obtained by adding to the derived latent distribution Q in Table 1 a support point $\xi_0^{(m)} = m^{-3/2}$ with a mixing weight $\varpi_0^{(m)} = m^{-1}$. The expected observed richness $E\{n_+|C^{(m)}\}$ and the expected counts $E\{n_x|C^{(m)}\}$ for $x = 0, 1, 2, \ldots, 5$ are plotted. The points not on the curves are for the reference assemblage, but each is linked to a curve by dotted lines. The ordinates are logarithmically scaled.

(index m becomes larger and larger), the expected observed counts (and thus the expected observed richness) for the modified assemblages converge on the expected observed counts for the reference assemblage. Meanwhile, the number of undetected species En_0 rises asymptotically to a constant, as reflected in the odds against detection $\alpha(Q^{(m)})$, which approaches an asymptote of 0.53.

This unsettling result is further illustrated in Fig. 5, where we vary both $\xi_0^{(m)}$ (by the arbitrary function $\xi_0^{(m)} = m^{-3/2}$) and mixing weight $\pi_0^{(m)}$ (indirectly by the arbitrary function $\varpi_{(m)0} = 1/m$), so that the true richness of the modified assemblages $S^{(m)}$ is not constant, but increases with m. The same qualitative pattern seen in Table 2 emerges. The expected observed counts and the expected observed richness for the modified assemblages converge on the expected observed counts and the expected observed richness for the reference assemblage. Meanwhile the number of undetected species and the odds against detection continue to rise (instead of approaching an asymptote), because we have allowed the true richness of the modified assemblages to increase with index m.

Table 3 presents α_{ML} in Eq. 10 calculated in the reference assemblage and modified assemblages that have been considered in Table 2 and Fig. 5. While α_{ML} is close to the true odds α in the reference assemblage, there is a substantial difference between α_{ML} and α in the modified assemblages. If a data set is generated from the modified assemblage $C^{(m)}$, using α_{ML} will yield a large bias. We can conclude that all nonparametric estimators can have a large bias in the presence of rare species because of Eqs. 17 and 18. The confidence intervals based on the asymptotic normality of these

TABLE 3. The approximation parameter α_{ML} in Eq. 10 is compared with α in the reference assemblage in Table 1 and modified assemblages in Table 2 (first pair of rows) and Fig. 5 (second pair of rows).

	m							
Parameter	2	10	50	100	1000	2000	5000	Reference
Tables 1 and 2								
α	0.30	0.47	0.52	0.53	0.53	0.53	0.53	0.07
α_{ML}	0.18	0.11	0.08	0.08	0.07	0.07	0.07	0.07
Fig. 5								
α	1.22	3.18	7.13	10.07	31.70	44.79	70.78	0.07
α_{ML}	0.77	0.19	0.09	0.08	0.07	0.07	0.07	0.07

TABLE 4. The two estimates $\hat{Q}$ and $\hat{Q}_\#$ with $\rho_k = \xi_k/(1 + \xi_k)$ (seed bank data).

	k					
Parameter	1	2	3	4	5	6
$\hat{Q}$						
ρ_k	0	0.0292	0.0680	0.1786	0.3827	0.4885
ω_k	0.0545	0.3179	0.3940	0.0866	0.1138	0.0332
$\hat{Q}_\#$						
ρ_k	$\cdots$	0.0197	0.0636	0.1774	0.3827	0.4885
ω_k	$\cdots$	0.2889	0.4752	0.0889	0.1138	0.0332

nonparametric estimators usually can not achieve their advertised confidence level.

Consider calculating the nonparametric maximum likelihood estimators (NPMLE) for the latent distribution Q in the seed-bank data and the beetles data (Lindsay 1983*a*, *b*). The NPMLE $\hat{Q}$ for Q that maximizes the conditional likelihood $L_2(Q)$ in Eq. 6 in the seed-bank data is presented in Table 4, where we use the detection probability instead of the detection odds, and the group detection probability is denoted by $\rho_k = \xi_k/(1 + \xi_k)$. Note that $\hat{Q}$ has a support point at zero, which means that an assemblage with infinitely many rare species whose relative abundances are extremely small is the most plausible one to generate such data. Another likelihood-based estimate $\hat{Q}_\#$ is also presented in Table 4, which is obtained by the EM algorithm starting from the distribution that eliminates the zero support point in $\hat{Q}$. Both estimates fit the data very well. While the log-maximized likelihoods $\log L_2(\hat{Q}_\#) = -113.55$ and $\log L_2(\hat{Q}) = -113.24$ are close to one another, the estimates $\alpha(\hat{Q}) = \infty$ and $\alpha(\hat{Q}_\#) = 0.03$ are dramatically different: the former suggests there are many undetected species while the latter suggests there are few. The observed seed-bank dataset has relatively few rare species (e.g., only three singletons and two doubletons), suggesting that few species remain undetected, as indicated by $\hat{Q}_\#$. But the NPMLE $\hat{Q}$ suggests that there might be many rare species, because the likelihood can be increased slightly by allowing many additional rare species, as in $\hat{Q}$. An estimator for the odds α calculated from an NPMLE might be severely biased.

For some data sets, it is possible to have an NPMLE for Q without a zero or tiny support point so that an estimate for the odds α seems acceptable. However, when one tries to construct confidence intervals for α or S by means of bootstrap, there will usually be some bootstrap resamples that yield an NPMLE with a tiny or zero support point, which makes the upper confidence limit extremely large. For example, the NPMLE $\hat{Q}$ for Q in the beetle data is presented in Table 5, yielding an estimate $\alpha(\hat{Q}) = 3.29$. We take 200 bootstrap resamples from the estimated conditional likelihood $L_2(\hat{Q})$. There are 14 resamples that produce the NPMLE for Q with the smallest support point less than 1×10^{-5}, among which one has a zero support point. The 95% conditional confidence interval for α is $(1.71, 1 \times 10^7)$. We remark that the numeric results of bootstrap might vary across different runs if one wishes to run the algorithm again because they are random. Clearly the upper confidence limit is useless. Because the NPMLE for Q based on bootstrap resamples are different in magnitude, a smoothed density of the random variable $\log \alpha(\hat{Q})$ rather than $\alpha(\hat{Q})$ constructed from the 199 resample estimates for α is presented in Fig. 6. Note that the density of $\log \alpha(\hat{Q})$ is still highly skewed and has a long right tail, which means that there is always a small but not negligible probability of obtaining a large value of $\log \alpha(\hat{Q})$ or equivalently a large value of $\alpha(\hat{Q})$. Note that because $\hat{S} = n_+(1 + \alpha(\hat{Q}))$ is simply a linear function in $\alpha(\hat{Q})$, the conclusions also apply if we replace $\alpha(\hat{Q})$ with $\hat{S}$.

TABLE 5. The estimate $\hat{Q}$ (beetle data).

	k		
Parameter	1	2	3
ξ_k	0.2252	3.4105	9.7892
ω_k	0.8288	0.1553	0.0155

CONCLUSION

The problem of estimating the species richness S has long been a challenge to both statisticians and ecologists. We have attempted to provide an overview of the

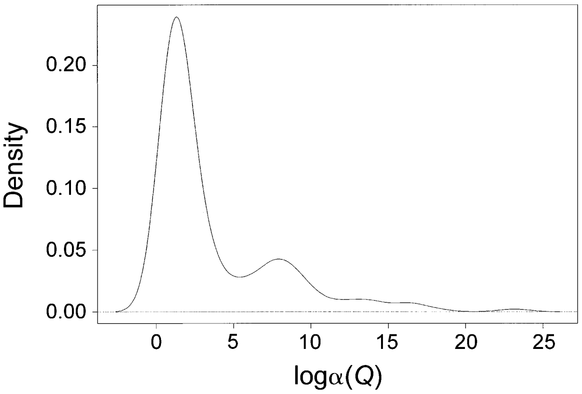

FIG. 6. The estimated density of $\log \alpha(\hat{Q})$ based on the resample.

challenges inherent in estimating richness when many rare species are known or suspected to be present. Our conclusions may appear pessimistic, but we consider it is important to explore the fundamental issues that reflect on claims for estimators and confidence intervals, to caution against overoptimistic conclusions from parametric models, and warn against the blind applications of statistical methods, although such a warning has been issued. For example, statistician I. J. Good, one of the first to tackle this problem (Good 1953), later stated: ''I don't believe it is usually possible to estimate the number of species . . . but only an appropriate lower bound to that number. This is because there is nearly always a good chance that there are a very large number of extremely rare species.'' (Bunge and Fitzpatrick 1993).

Although from typical empirical data, one cannot exclude the existence of many rare species statistically, it is nonetheless possible to infer, with mixture models, how many species should exist at a particular confidence level, that is, to compute with rigor a lower bound for the species richness S, other than the observed richness n_+. Besides mixture models, all existing methods for estimation of S can be treated as improvements upon n_+, the number of observed species, which is a clearly negatively biased estimator for true species richness S.

Using mixture models, it is also possible to estimate rigorously, with confidence intervals, the expected increment in richness that would result from increasing observed samples to two or three times their actual size, without attempting to find a true asymptote (Colwell et al. 2004, Mao et al., *in press*; see also Shen et al. 2003, for a different approach; in fact, it is not at all clear that a true asymptote even exists, for some taxa in some habitats, as discussed in the *Introduction*.) Such an extrapolation, for many purposes in ecology and conservation biology, will often represent a most useful and welcome savings in time and resources. In short, we do not counsel despair, but rather realistic expectations and a cautious approach to inference.

ACKNOWLEDGMENTS

We are grateful to A. Ellison for the invitation to prepare this paper and the constructive comments from the two referees. This work was supported by US-NSF grant DEB-0072702 to R. K. Colwell.

LITERATURE CITED

Bunge, J., and M. Fitzpatrick. 1993. Estimating the number of species: a review. Journal of the American Statistical Association **88**:364–373.

Burnham, K. P. 1972. Estimation of population size in multiple capture–recapture studies when capture probabilities vary among animals. Dissertation. Oregon State University, Corvallis, Oregon, USA.

Burnham, K. P., and W. S. Overton. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika **65**:625–634.

Butler, B. J., and R. L. Chazdon. 1998. Species richness, spatial variation, and abundance of the soil seed bank of a secondary tropical rain forest. Biotropica **30**:214–222.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. Scandinavian Journal of Statistics **11**:265–270.

Chao, A. 1989. Estimating population size for sparse data in capture–recapture experiments. Biometrics **45**:427–438.

Chao, A., and S. M. Lee. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association **87**:210–217.

Chao, A., and T. J. Shen. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environmental and Ecological Statistics **10**: 429–443.

Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B **345**:101–118.

Colwell, R. K., C. X. Mao, and J. Chang. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology **85**:2717–2727.

Darroch, J. N., and D. Ratcliff. 1980. A note on capture-recapture estimation. Biometrics **36**:149–153.

Dorazio, R. M., and J. A. Royle. 2003. Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics **59**:351–364.

Efron, B., and R. Thisted. 1976. Estimating the number of unseen species: how many words did Shakespeare know? Biometrika **63**:435–447.

Falkowski, P. G., and C. de Vargas. 2004. Shotgun sequencing in the sea: a blast from the past? Science **304**(5667): 58–60.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology **12**:42–58.

Gaston, K. E. 1994. Rarity. Chapman and Hall, London, UK.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika **40**: 237–264.

Gotelli, N., and R. K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters **4**:379–391.

Hughes, J. B., J. J. Hellmann, T. H. Ricketts, and B. J. M. Bohannan. 2000. Counting the uncountable: statistical approaches to estimating microbial diversity. Applied and Environmental Microbiology **67**:4399–4406.

Janzen, D. H. 1973. Sweep samples of tropic foliage insects: description of study sites, with data on species abundances and size distributions. Ecology **54**:659–686.

Lee, S. M., and A. Chao. 1994. Estimating population size via sample coverage for closed capture–recapture models. Biometrics **50**:88–97.

Lindsay, B. G. 1983*a*. The geometry of mixture likelihoods: a general theory. The Annals of Statistics **11**:86–94.

Lindsay, B. G. 1983*b*. The geometry of mixture likelihoods. Part II: the exponential family. The Annals of Statistics **11**: 783–792.

Longino, J., R. K. Colwell, and J. A. Coddington. 2002. The ant fauna of a tropical rainforest: estimating species richness three different ways. Ecology **83**:689–702.

Magurran, A. E. 2004. Measuring biological diversity. Blackwell, London, UK.

Magurran, A. E., and P. A. Henderson. 2003. Explaining the excess of rare species in natural species abundance distribution. Nature **422**:714–718.

Mao, C. X. 2004*a*. Estimating the unknown sample size from truncated binomial or geometric mixtures. Technical report. Department of Statistics, University of California, Riverside, California, USA.

Mao, C. X. 2004*b*. Predicting the conditional probability of discovering a new class. Journal of American Statistical Association **99**:1108–1118.

Mao, C. X., R. K. Colwell, and J. Chang. *In press*. Estimating the species accumulation curve using mixtures. Biometrics.

Mao, C. X., and B. G. Lindsay. 2003. Estimating the number of classes using Poisson mixtures. Technical report. Department of Statistics, University of California, Riverside, California, USA.

Mingoti, S. A., and G. Meeden. 1992. Estimating the total number of distinct species using presence and absence data. Biometrics **48**:863–875.

Norris, J. L. I., and K. H. Pollock. 1996. Nonparametric MLE under two closed capture–recapture models with heterogeneity. Biometrics **52**:639–649.

Norris, J. L. I., and K. H. Pollock. 1998. Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. Environmental and Ecological Statistics **5**:391–402.

Novotny, V., and Y. Basset. 2000. Rare species in communities of tropical insect herbivores: pondering the mystery of singletons. Oikos **89**:564–572.

Ord, J. K., and G. Whitmore. 1986. The Poisson-inverse Gaussian distribution as a model for species abundance. Communications in Statistics, Theory and Methods **15**: 853–871.

Pledger, S. 2000. Unified maximum likelihood estimates for closed capture–recapture models using mixtures. Biometrics **56**:434–442.

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

Rabinowitz, D. 1981. Seven forms of rarity. Pages 205–217 *in* H. Singe, editor. The biological aspects of rare plant conservation. Wiley, Chichester, UK.

Shen, T. J., A. Chao, and C. F. Lin. 2003. Predicting the number of new species in future taxonomic sampling. Ecology **84**:798–804.

Sichel, H. S. 1997. Modeling species-abundance frequencies and species individual functions with the generalized inverse Gaussian-Poisson distribution. South African Statistical Journal **31**:13–37.

Smith, E. P., and G. van Belle. 1984. Nonparametric estimation of species richness. Biometrics **40**:119–129.

Whittaker, R. H. 1972. Evolution and measurement of species diversity. Taxon **21**:213–251.

Zelterman, D. 1988. Robust estimation in truncated discrete distributions with application to capture–recapture experiments. Journal of Statistical Planning and Inference **18**: 225–237.

Ecology, 86(5), 2005, pp. 1154–1163

UNDERSTANDING AND PREDICTING THE EFFECTS OF SPARSE DATA ON DEMOGRAPHIC ANALYSES

Daniel F. Doak,[1,4] Kevin Gross,[2] and William F. Morris[3]

[1]*Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, California 95064 USA*
[2]*Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695 USA*
[3]*Biology Department, Duke University, Box 90338, Durham, North Carolina 27708-0338 USA*

Abstract. Demographic models are an increasingly important tool in population biology. However, these models, especially stochastic matrix models, are based upon a multitude of parameters that must usually be estimated with only a few years of data and limited sample sizes within each year, calling into question the accuracy of the results of these models. We first discuss how these data limitations create sampling uncertainty and bias in the estimated parameters for a stochastic demography model. Next, we ask whether limited data can favor the construction of deterministic models that ignore variation and correlation of rates. With less than five years of data, the mean squared error of deterministic models will sometimes be smaller than that of stochastic models, favoring the use of simple models, even when their predictions are known to be biased. Finally, we introduce a procedure to estimate the sampling variation around population growth rate estimates made from demographic models that are based on specified sampling durations and intensities.

Key words: *covariance; demography; matrix; rarity; stochasticity; variability; viability.*

Introduction

Demographic models are among the most widely used analytical tools in life history studies, population ecology, and, especially, conservation biology, where they form the backbone of population viability analysis (PVA). However, demographic analyses—including both life tables and matrix models—are also among the most data hungry of methods, requiring estimates of birth, death, and other vital rates for each step of an individual's life. In spite of these needs, there has been relatively little discussion in the demographic literature of what data limitations mean for the construction and interpretation of these models (but see Ludwig 1999, Fieberg and Ellner 2001, Ellner et al. 2002), and in particular, little work that would help a field biologist understand how much demographic data must be collected to have confidence in modeling results, and how the complexity of different demographic models and the scarcity of information with which to parameterize them will interact to influence this confidence.

Two distinct kinds of data scarcity can effect the accuracy of vital rate estimates and hence the accuracy of demographic predictions: the sampling of small numbers of individuals within each year, so that vital rate estimates made for each transition period are uncertain; and the collection of data over only a small number of years, so estimates of the mean, variance, and covariance of vital rates over time are unreliable. We will refer to these as sampling intensity and sampling duration. One or both kinds of data limitation afflict virtually anyone who tries to develop a demographic model. Here, we discuss some of the ramifications of this data scarcity. In particular, we focus on the problems of parameterizing stochastic demographic models, which are increasingly advocated, but which require even more information than do simple deterministic models.

We have three goals in this paper. First, to explain how limited data not only increase the uncertainty in all parameter estimates needed for a demographic analysis, but also create biased estimates of the variance and correlation parameters needed for stochastic models. While this bias is to some extent correctable, few ecologists seem to realize that this is a key complication in the construction of stochastic models. Second, to analyze the trade-off of building more realistic, but complex, stochastic models vs. simpler, but less realistic models that do not include variance and correlation. While variance and correlation in vital rates can strongly affect demography, there is also a great deal of uncertainty around estimates of variance and correlation with limited data. Consequently, we ask: when is the cost of including estimates of variance and correlation in a demographic model worth the benefit? The basic idea behind this question is now well-publicized in ecology: there is a trade-off between model realism (i.e., complexity) and the ''noise'' created by building predictions on the back of poorly estimated parameters (Ludwig et al. 1988, Hilborn and Mangel 1997, Burnham and Anderson 1998). However, this trade-off has not been directly assessed for the incorporation of sto-

---

Manuscript received 2 April 2004; revised 4 August 2004; accepted 5 August 2004; final version received 17 September 2004. Corresponding Editor: A. M. Ellison. For reprints of this Special feature, see footnote 1, p. 1079.

[4] E-mail: doak@biology.ucsc.edu

chasticity into demography models. Third, we address what may be the most widely posed question among empirical ecologists who are initiating demographic studies: how many data are enough to get reasonably accurate predictions? While this is obviously impossible to answer with certainty, we develop a way to get a better answer than a simple guess or a shrug of the shoulders—the usual responses to this question.

Demographic Models in a Nutshell

To make the following discussion of parameter estimation and model construction comprehendible, we briefly review the basic form and analysis of demographic models. Demographic models generally use age, size, or stage (e.g., juveniles and adults) to classify individuals. For the sake of clarity, we will couch our discussion in terms of the rates needed to construct a size-based model, but the issues and methods are the same when using other classifying variables. We consider a simple model with four size classes and a set structure, in which an individual can either survive and remain in the same size class from one year to the next, or survive and grow enough to enter the next largest size class. We can characterize such a life history with three types of vital rates: fecundities (f_a), growth probabilities (g_a), and survival rates (s_a), each of which are indexed by the size class (a) to which they apply. We will assume that, once in the largest size class, there is no senescence, and also that the fecundities, and hence the model as a whole, only apply to females. We also assume that reproduction is synchronous and that the youngest class in the model consists of newborns, with the census of the population occurring just after reproduction. These assumptions yield the following transition matrix **A** for four size classes:

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & e_{13} & e_{14} \\ e_{21} & e_{22} & 0 & 0 \\ 0 & e_{32} & e_{33} & 0 \\ 0 & 0 & e_{43} & e_{44} \end{pmatrix}$$

$$= \begin{bmatrix} 0 & 0 & f_3 s_3 & f_4 s_4 \\ s_1 & s_2(1 - g_2) & 0 & 0 \\ 0 & s_2 g_2 & s_3(1 - g_3) & 0 \\ 0 & 0 & s_3 g_3 & s_4 \end{bmatrix}. \qquad (1)$$

Here, the matrix element e_{ab} is the number of individuals in class a that we expect to arise one year in the future from each individual of size b now. For the top row of the matrix, these elements are functions of survival and fecundity, and for the remaining rows, the elements are functions of survival and growth probabilities. The second matrix in the equation shows how each matrix element is defined in terms of the vital rates.

With an estimated projection matrix in hand, there are several key predictions that are usually made. First is the annual population growth rate, λ, which summarizes the long-term rate of change for a population governed by **A**. If population size is shown by a column vector, $\mathbf{N}_t$, that contains the numbers of individuals in each size class (n_1, n_2, n_3, n_4 for our example), then $\mathbf{N}_{t+1} = \mathbf{A}\mathbf{N}_t = \mathbf{A}^{t+1}\mathbf{N}_0$, and as t becomes large, the annual rate of change of the population will converge to that predicted by λ. At the same time, the population will converge to the stable stage distribution (SSD), in which the fraction of the whole population that is in each stage remains constant. Besides λ and the SSD, the most commonly used outputs of a simple projection matrix are the sets of sensitivity and elasticity values. Sensitivities estimate the rate with which λ will change with changes in each matrix element, and elasticities are rescalings of these values that estimate the proportional change in λ resulting from proportional changes in each matrix element (Caswell 2001).

In reality, a single matrix constructed from a set of mean vital rates will often do a poor job of representing population dynamics, because one or more of these rates will vary substantially between years. One way to include this environmental stochasticity is to estimate a separate matrix for each year of data collected, and then randomly choose among them in simulating the future (Fieberg and Ellner 2001). While this approach is straightforward, it does not allow any ready analysis of, or correction for, the problems created by limited data. Alternatively, we can make explicit estimates of the mean and variance of each vital rate and the covariance, or correlation, between all pairs of rates across years. With these estimates, both analytical approximations or simulations of the model can provide results analogous to those of a deterministic model, as well as estimates of extinction risks through time (Caswell 2001, Morris and Doak 2002). The stochastic analogue to λ, λ_s, represents the average expected rate of population change, accounting for the effects of environmental stochasticity. Because it is more or less normally distributed and hence easy to deal with statistically, the natural logarithm of the of λ_s, log λ_s, forms the basis of most stochastic demography, including our analyses here.

Uncertainty and Bias Created by Limited Data

There are three interrelated issues to consider in using limited data to estimate the parameters of a stochastic matrix model. The first of these is perhaps the most important, and is also the most widely understood. The second two are much less appreciated, but are still critical to understand when building estimates of demographic rates.

First is the simple problem that estimates of vital rate means are highly uncertain when data are limited. While there are various tactics to reduce this problem in the estimation of vital rate parameters (e.g., Easterling et al. 2002, Morris and Doak 2002, Dixon et al. 2005), it will never go away. Quantifying this uncertainty in estimates of means is the grist of most basic statistics classes.

The second issue to appreciate in making parameter estimates for demographic models is that sampling error not only creates uncertainty but also bias in estimates of variances and correlations. The emphasis in standard statistics in estimating and testing effects on *means* is undoubtedly why this basic aspect of sampling theory is so underappreciated by ecologists. Estimated temporal variance in a demographic rate is generally biased high; observed variation across years is due in part to real changes in rates across time (environmental stochasticity) but also in part to variation caused by estimation errors. Conversely, if vital rates that are probabilities, such as survival and growth, have very high variation from year to year, sampling error can bias estimates of this variation low (Kendall 1998). In either case, to arrive at an unbiased estimate of the environmental variation in a demographic rate, one must estimate and remove the portion of observed variation due to sampling errors. This biasing of temporal variance estimates has recently received considerable attention in the technical demographic literature, with at least three suggestions of how to make such a correction (Engen et al. 1998 [whose method is presented for estimation of demographic stochasticity], Kendall 1998, White 2000; Morris and Doak 2002 include programs to implement these corrections). However, these methods are all relatively new, and the vast majority of stochastic demography studies published to date did not use them in estimating the environmental stochasticity in vital rates.

Covariances between two rates are not biased by sampling errors, at least if we can assume that sampling errors for the two rates do not themselves covary. However, correlation coefficients, equal to the covariance of two rates divided by their standard deviations, are biased, with increasing amounts of sampling error leading to estimated correlation coefficients that approach zero even when the true correlation is substantial. This would seem to suggest that we should use the unbiased covariance estimates together with the corrected variance estimates to arrive at unbiased estimates of correlations. However, for any limited sample of data, this procedure will yield many correlation estimates that are greater than one or less than -1, simply due to random (but unbiased) variation in the estimated covariance values. Until some solution is found to this problem, we are left estimating correlations directly, realizing that they will on average be biased toward zero.

The third parameter estimation problem involves the entire set of correlation coefficients between all the vital rates estimated for a population; because of sampling errors or missing data, the full set of correlations you estimate may not be able to exist. For example, three vital rates can't all be highly negatively correlated with each of the others. While this is an extreme example, more subtle forms of this problem commonly arise. This may seem like an esoteric issue, but it becomes important if your goal in estimating demographic rates is to make stochastic predictions; if the set of estimated correlations is impossible, then you can't use them in simulation models to predict population growth or extinction. We describe this issue and how to deal with it in more detail in Appendix A.

THE COSTS AND BENEFITS OF INCLUDING STOCHASTICITY IN DEMOGRAPHIC MODELS

Even when properly accounting for bias in estimates of vital rate variances, the uncertainty in estimated means, variances and correlations arising from low sampling duration and intensity mean that a PVA or other demographic model based on limited data can give highly misleading predictions. The extent to which PVA results can be trusted without truly extensive sampling within and across years has been vigorously debated in the literature, but no clear conclusion exists to date (Ludwig 1999, Brook et al. 2000, Ellner et al. 2002). While it is clear that even moderate amounts of environmental stochasticity in combination with limited data can yield inaccurate results, we would still like to know how to build the best model with the data at hand, and in particular, whether it is worth the hassle to estimate and simulate the variance and correlation of vital rates in order to arrive at stochastic predictions.

Three classes of demographic models are commonly parameterized and used: deterministic models, based only on estimated mean rates; stochastic variance models, which ignore correlations between rates, but do use estimated variances; and full stochastic models, which use estimated means, variances, and correlations for all or most vital rates. The advantage of using a deterministic model is that its predictions rely on a smaller number of parameters, and with sparse data, this means a smaller number of badly estimated parameters. However, this benefit may be more than balanced by the greater realism of stochastic models. Although these use more parameter estimates, they do so to include real features of the population—temporal variability—and thus may give more accurate predictions. If environmental variability in vital rates is moderate or strong, deterministic matrix models will generally give predictions of population growth that are biased high, predicting more rapid growth rates than will actually occur (Tuljapurkar 1990, Caswell 2001). It is this bias that has led to the frequent calls to include variation and correlation in demographic PVA models (e.g., Burgman et al. 1993, Beissinger and Westphal 1998, Morris and Doak 2002). Nonetheless, we usually have only one short stretch of years with which to parameterize a model, and the increased bias of a simplified model may be outweighed by less random variation in its predictions, relative to an unbiased, but potentially more inaccurate model built with many poor estimates of variation and correlation. In statistical terms, there

TABLE 1. Sets of parameter values used in the simulation models.

Parameter	Vital rate							
	s_1	s_2	s_3	s_4	g_2	g_3	f_3	f_4
Mean	0.1	0.3	0.8	0.9	0.5	0.75	5	10
Variances								
High	0.0675	0.1575	0.1200	0.0675	0.1875	0.1400	20	20
Medium	0.0500	0.0500	0.0200	0.0200	0.0100	0.0100	7	10
Low	0.0125	0.0125	0.0050	0.0050	0.0025	0.0025	1.75	2.50

Notes: A total of nine sets of vital rate parameters were used, corresponding to combinations of the three sets of variance values with the three sets of correlation coefficients. In the high-correlation models, all correlation coefficients were +0.80; in the low-correlation models, all correlation coefficients were +0.20. In the mixed correlation models, the correlation between each fecundity (f) and any survival (s) or growth (g) rates was −0.80 (corr(f_i,s_j) = corr(f_i,g_j) = −0.80), while all other correlations were +0.80. Mean values of vital rates were the same in all models.

is a trade-off between the bias and sampling variation of the estimated population growth rate.

To explore this issue, we developed a set of simulation models to test the accuracy of different demographic models parameterized with differing durations and intensities of data sampling. In all, we simulated nine demographic scenarios, each with a different set of vital rate means, variances, and correlations (Table 1), but all based on the life history shown in Eq. 1. These rates generate expected stochastic growth rates ranging from $\lambda_s = 0.81$ to 1.01. Each simulation consisted of three basic steps (see Appendix B for details):

1) Generation of a set of ''real,'' but unobserved, vital rate values across 20 years. For each demographic scenario, we generated 500 sets of these real annual rates.

2) Based on each set of real rates, simulation of an observed data set, taken with limited sampling duration and intensity. We varied sampling intensity from 10 to 40 individuals in each size class, and sampling duration from three to 10 years. For any single simulation, we assumed that the same number of individuals were followed in each size class and each year. Using these simulated data, we estimated means, variances, and covariances of vital rates, employing Kendall's (1998) method of variance correction for survival and growth rates and White's (2000) method for fecundities.

3) For each simulated data set, we predicted log λ_s using the deterministic growth rate predicted from the mean matrix and using the stochastic variance and full stochastic versions of Tuljapurkar's (1990) approximation, each parameterized with the estimated vital rates. For each sampling pattern and model type, we compared the estimated log λ_s values with the growth rate predicted from the real vital rate parameters by calculating mean squared error (MSE) and bias.

As expected, the average predictions of the deterministic models were optimistically biased, with positive differences between the mean deterministic predictions and true log λ_s for almost all model forms and sampling regimes (Fig. 1a–c). Differences between the full and variance model predictions and expected log λ_s values lie closer to zero, although they too are for the most part positively biased for models with moderate to high vital rate variances. This trend is largely due to optimistic estimates of growth for models using the high or mixed correlation structures, which both feature strong and mostly positive correlations between vital rates. In these situations, the tendency of estimated correlations to be biased low is expected to result in just such a bias in the predictions of full models. Predictions of variance models, which implicitly assume that all correlations are zero, are only somewhat more biased than those of full models, but also show a tendency towards more variable results.

The greater average bias in deterministic model predictions would suggest that the stochastic models should provide better estimates of future numbers. However, while the average bias of deterministic models is larger, the variance in the deterministic predictions is substantially lower than that of either stochastic model for some sets of vital rates and sampling regimes (Fig. 1d). In particular, deterministic models showed lower sampling variance for all simulations using high vital rate variances and for some sampling regimes with medium variances. This lower variance in predictions means that even with their biases, deterministic models can sometimes have lower MSE than do full or variance-only models.

To show how model form and sampling regimes influence the predictive power of these models, we separately plot MSE against sampling duration and intensity for results from different vital rate scenarios. Here and below, we only present results for the full and deterministic models, since the full and variance models show nearly identical results for virtually all parameter combinations. Unsurprisingly, with higher sampling intensity and higher sampling duration, the MSE of all model predictions declines (Figs. 2 and 3; results for low-variance models are qualitatively identical). For the most part, there is little interaction between intensity and duration of sampling. Duration has by far the larger effect on MSE, and if duration is low, intensive within-year sampling usually has little effect

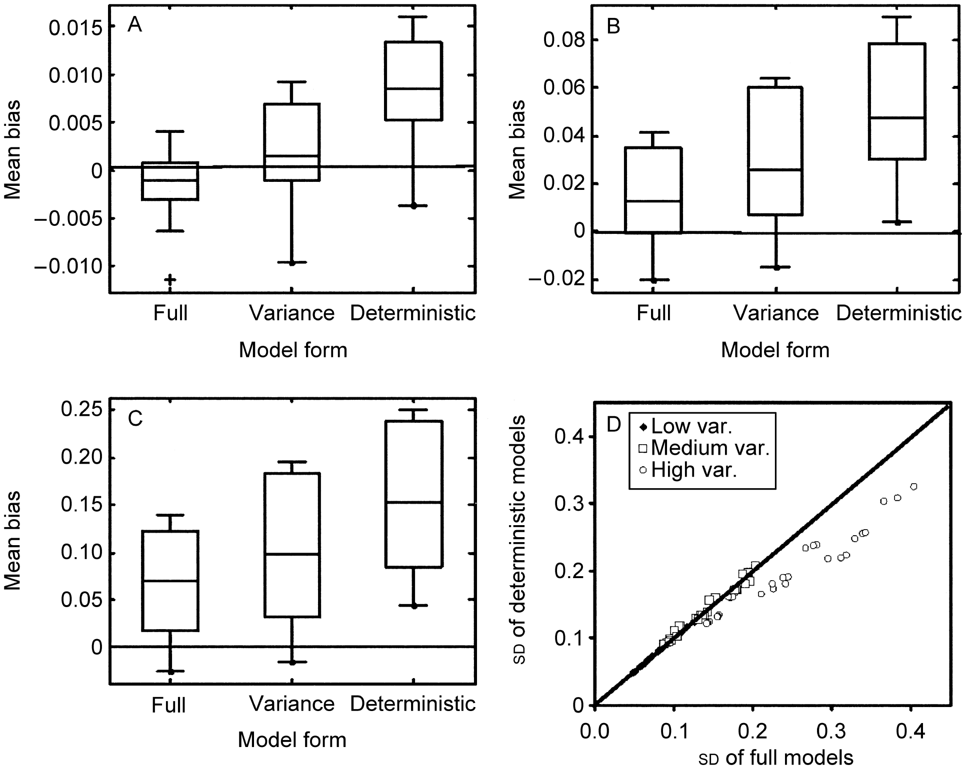

FIG. 1. Bias and variance in predictions of deterministic, variance-only, and full stochastic models. (A–C) Box plots showing the distribution of mean differences between predicted and expected log λ_s values for 27 means, one for each combination of correlation parameters, sampling intensity, and sampling duration: (A) low-variance models, (B) medium-variance models, and (C) high-variance models. (D) The standard deviation in predicted log λ_s (stochastic lambda) values for each combination of vital rate parameters and sampling regimes for full and deterministic models. Symbols indicate results from models with high, medium, and low variance in vital rates (Table 1). Points falling below the diagonal line indicate situations in which the deterministic predictions were less variable than those of the full models for the same vital rate parameters and sampling regimes.

on MSE. The exception to this comes at the lowest sampling durations, when higher intensity can substantially reduce MSE for some vital rate combinations (e.g., Fig. 2b).

For parameter sets with low vital rate variances, the difference in quality of different model predictions are quite small, amounting to <10% of MSE of the full models in all cases, and usually <5% (results not shown). For medium and high variance models, differences between the predictions of deterministic and full models are more substantial and more complex. With high correlations in vital rates (Figs. 2a and 3a), Full models have substantially higher accuracy at all but the lowest sampling durations. In contrast, with low correlation and Medium variance in vital rates, the two model forms show very similar MSEs, and with low correlation and high variances, the deterministic models have much lower MSEs for all but the longest sampling durations (Figs. 2b and 3b). Finally, with mixed correlations in vital rates, full models are more accurate with high duration and less so for short sampling durations (Figs. 2c and 3c). Within these details, the general pattern is that full model MSE values are quite sensitive to data scarcity, rising sharply with decreasing sampling duration and also increasing with lower within-year sample sizes. In contrast, deterministic model MSEs are generally less responsive, such that there is some sampling duration at which the deterministic MSE becomes less than the full MSE, with the exact value depending critically on the underlying vital rate parameters.

Altogether, these results provide a rather mixed picture of the problems of predictive power of stochastic models. For some life histories and sampling regimes deterministic models and full stochastic models essentially tie one another in accuracy, as judged by their MSE values. In these cases, the lower bias of stochastic models will make them preferable to the overly optimistic deterministic predictions. However, when variances in vital rates are high, full stochastic models are sometimes more accurate than simple deterministic predictions, but are also sometimes spectacularly less so. With real data, we can't know the real underlying rates, so we could also ask, averaging across all the vital rate combinations we ran, which model form has the lower MSE values? The answer is deterministic

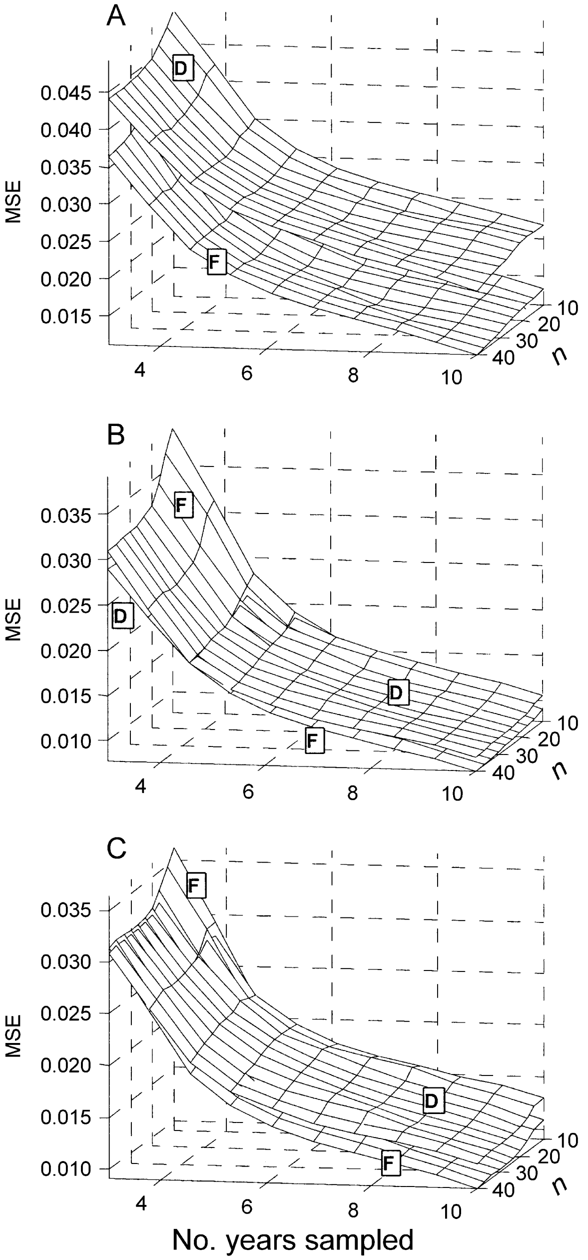

FIG. 2. Mean-squared error (MSE) of predicted population sizes for medium-variance models for sample sizes $n = 10$–40 and 4–10 sampling years. Two surfaces are shown (drawn with cubic interpolation), one for deterministic and one for full stochastic models. The labels "D" and "F" indicate which model corresponds to each surface (deterministic and full, respectively). The model with the lower MSE provides more accurate predictions of population size. Results are from life histories with (A) high positive correlations, (B) low positive correlations, and (C) mixed correlations.

models for sampling durations of 3–5 yr and full models for durations of 5–10 yr (Fig. 3d), implying that if we have few data, simpler models may provide better predictions than will stochastic model forms. To put this in perspective, only nine of the 29 studies listed in Fieberg and Ellner's (2001) review of stochastic demography analyses were based on five or more years of vital rate estimates. Depending on your love of stochastic analysis, this may seem a disheartening or a comforting conclusion. However, we reiterate that we have only explored results for one life history pattern, and it is conceivable that other, very different life histories will show quite different patterns.

HOW MANY DATA ARE ENOUGH?

The final question we consider is how high sampling intensity and duration must be to achieve reasonable accuracy in demographic predictions. In particular, how do the duration and intensity of sampling translate into uncertainty of population growth predictions? Obviously, you can't really answer this question without perfect knowledge of all the rates and processes you are trying to estimate. However, by making a few guesses and assumptions, you can use the derivation we give below to arrive at a good estimate of the accuracy you can hope to achieve (see Caswell [2001] for a review of other approaches, mostly for assessing uncertainty in deterministic λ after data collection is completed).

To begin with, you must guess the form of the matrix model and the values of the vital rate means, variances and correlations governing your species. While you can't know the real values, usually natural history knowledge or data from related species can be used to make rough estimates of the form of the life history of your species and the values of its vital rates. As Dixon et al. (2005) emphasize, data on related species are a great untapped source of information for such guesses. We will call the guessed mean and environmental variance for a vital rate i, ν_i and $\sigma^2_{i,E}$, and its correlation with another rate j, $\rho_{i,j}$.

Second, you must have some guess as to how the within-year sampling intensity for a vital rate, the M_i number of individuals measured, influences the sampling variance of the estimated annual values of the rate. For survival and growth probabilities of individuals with identical mean values, this *within*-year sampling variance is $\sigma^2_{i,W} = \nu_{i,t}(1 - \nu_{i,t})/M_i$, where $\nu_{i,t}$ is the mean probability in year t. For fecundities, in which each individual must give birth to an integer number of offspring, we will assume that the variance between individuals is Poisson, so that the variance of a mean annual estimate is $\sigma^2_{i,W} = \nu_{i,t}/M_i$. There are many other forms of individual variation in fecundities that will give more or less variation in annual mean values and which can be used in place of the Poisson-distributed variation we assume.

With guesses as to the true demography of your species and the form of sampling variance, there are two ways to proceed. First, simulations like those described in the last section can directly yield estimates of the sampling variance in estimated growth rates. However, constructing such a simulation is not a task that most

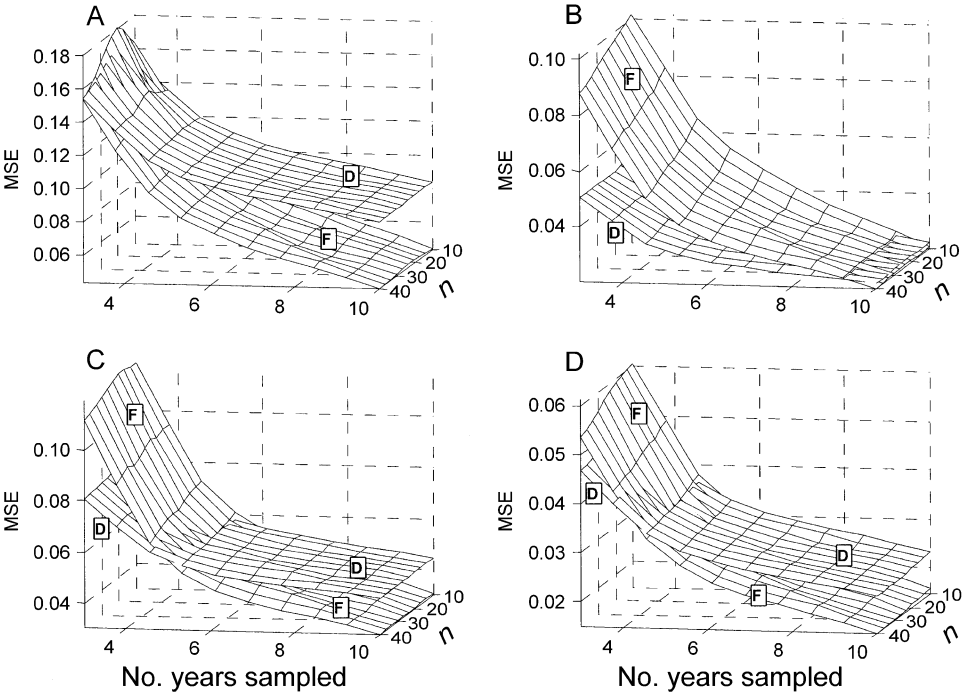

FIG. 3. Mean-squared error (MSE) of predicted population sizes for high-variance models and all models combined for sample sizes $n = 10$–40 and 4–10 sampling years. Two surfaces are shown (drawn with cubic interpolation), one for deterministic models (labeled "D") and one for full stochastic models (labeled "F"). The model with the lower MSE provides more accurate predictions of population size. Results are from life histories with (A) high positive correlations, (B) low positive correlations, and (C) mixed correlations. Results for MSE over all vital rate parameter sets are shown in (D).

empirical ecologists are likely to tackle at the onset of their own field study. Therefore, we derived an analytical approximation for the sampling variance of $\log\lambda_s$, based on Tuljapurkars's (1990) "small noise" approximation for $\log\lambda_s$. Like the derivation of this approximation, our further use of it relies on assumptions of small, normally-distributed variation in vital rates (see Appendix C for details).

The basic result is an equation for the sampling variance in estimated population growth rate as a function of the vital rate parameters, the number of years sampled, and the number of individuals sampled in each year to estimate each rate:

$$\begin{aligned}\mathrm{Var}(\log\hat{\lambda}_s) \approx{}& \sum_i \mathrm{Var}(\hat{\nu}_i)\left(\frac{\partial \log\lambda_s}{\partial \nu_i}\right)^2\\ &+ \frac{1}{\lambda_1^4}\sum_i \mathrm{Var}(\hat{\sigma}_{i,E})\left(\sum_j S_i S_j \rho_{i,j}\sigma_{j,E}\right)^2\\ &+ \frac{1}{\lambda_1^4}\sum_{i,j}\mathrm{Var}(\hat{\rho}_{i,j})(S_i S_j \sigma_{i,E}\sigma_{j,E})^2. \qquad (2)\end{aligned}$$

Here, S_i is the sensitivity of the growth rate predicted by the mean matrix, λ_1, to changes in ν_i, and $\partial \log \lambda_s/\partial\nu_i$ is the sensitivity of the log-stochastic growth rate to the mean vital rate ν_i (see Appendix C for the formula for this sensitivity). $\mathrm{Var}(\hat{\nu}_i)$, $\mathrm{Var}(\hat{\sigma}_{i,E})$, and $\mathrm{Var}(\hat{\rho}_{i,j})$ are approximations for the sampling variation in estimated means, variances and correlations of the vital rates, respectively, and are where limited sampling has its influence. Assuming normally distributed variation and uncorrelated errors among rates, these quantities are:

$$\begin{aligned}\mathrm{Var}(\hat{\nu}_i) &\approx \frac{1}{N}\left(\sigma_{i,E}^2 + \frac{\sigma_{i,W}^2}{M_i}\right)\\ \mathrm{Var}(\hat{\sigma}_{i,E}) &\approx \frac{N}{2(N-1)^2}\left(\sigma_{i,E}^2 + \frac{2\sigma_{i,W}^2}{M_i}\right) \qquad (3)\\ \mathrm{Var}(\hat{\rho}_{i,j}) &\approx \frac{N(\rho_{i,j}^2-1)^2}{(N-1)^2}.\end{aligned}$$

In addition to these basic contributions to $\mathrm{Var}(\log\hat{\lambda}_s)$, a further source of variability comes from the correlations between the estimated means and variances of vital rates that are probabilities. Because the mean of a set of probabilities sets a cap on the possible values of their variance (Morris and Doak 2004), there can be substantial correlations between estimated means and variances across sets of samples, with the sign and the magnitude of the correlation depending on the true mean and variance of the vital rate in question. While Eq. 2 does a good job of approximating $\mathrm{Var}(\log\hat{\lambda}_s)$, these additional terms have nontrivial effects on the predictions, and we discuss how to estimate them in Appendix C. With the addition of these terms, the approximation does an excellent job of predicting the sampling variation we could expect for a given life history and sampling regime, with the caveat that it will tend to give optimistic estimates of accuracy for populations with very high real environmental variability in vital rates (Appendix C). A MATLAB program that automates all the calculations needed to arrive at Var(log $\hat{\lambda}_s$) estimates with a range of sampling regimes is available in a Supplement to this paper.

To illustrate the use of this approximation, we first ask how sampling intensity and duration will affect estimates of growth rates for a winter annual with highly variable demographic rates, *Collinsia verna*. We use the single stochastic matrix model formulated by Doak et al. (2002) from the much more extensive results of Kalisz (1991). For this species, we first plotted Var(log $\hat{\lambda}_s$) for sampling intensities ranging from 10 to 200 plants followed to estimate each of the four vital rates for this matrix, and for sampling durations ranging from three to 20 years (Fig. 4a). The results indicate that sampling of 50 individuals or less is sufficient to achieve good accuracy, but that uncertainty of estimates will only gradually decline with increased sampling duration, with no clear cut-off beyond which further gains in accuracy are minimal. To further examine these results, and to present them in terms of annual growth rates, which can be easier to interpret, we estimated the 95% confidence limits on $\hat{\lambda}_s$ as $\exp(\log \lambda_s \pm 2\sqrt{\text{Var}(\log \hat{\lambda}_s)})$ and plotted these bounds against sampling duration for intensities of 20, 100, and 200 individuals (Fig. 4b). This plot emphasizes the minimal gains that result even from the ten-fold increase in sampling intensity from 20 to 200 individuals. At the same time, environmental variation is high enough for this species that even 200 years of data would not be enough to clearly predict whether λ_s is above or below one. Luckily, this species is not endangered, so the exact value of λ_s is not of critical importance.

To illustrate a contrasting situation, and one where estimating population growth rate is of real importance, we analyzed the duration of sampling needed to reach narrow confidence limits on $\hat{\lambda}_s$ for desert tortoise populations in the western Mojave desert. For this threatened and sparse population, we assumed that sample sizes of 30 animals for each rate in each year were the best that are ever likely to be achieved, and used this sampling intensity in conjunction with vital rates presented in Morris and Doak (2002) from data in Doak et al. (1994). For the tortoise, only 10 years of data would be sufficient to reach fairly narrow confidence limits on population growth estimates (Fig. 4c), and even less would be needed to have confidence that decline was occurring, given the vital rates we use. Given that considerably more years of data than this formed the basis of these models (Doak et al. 1994), we can thus have considerable confidence that the population truly was declining.

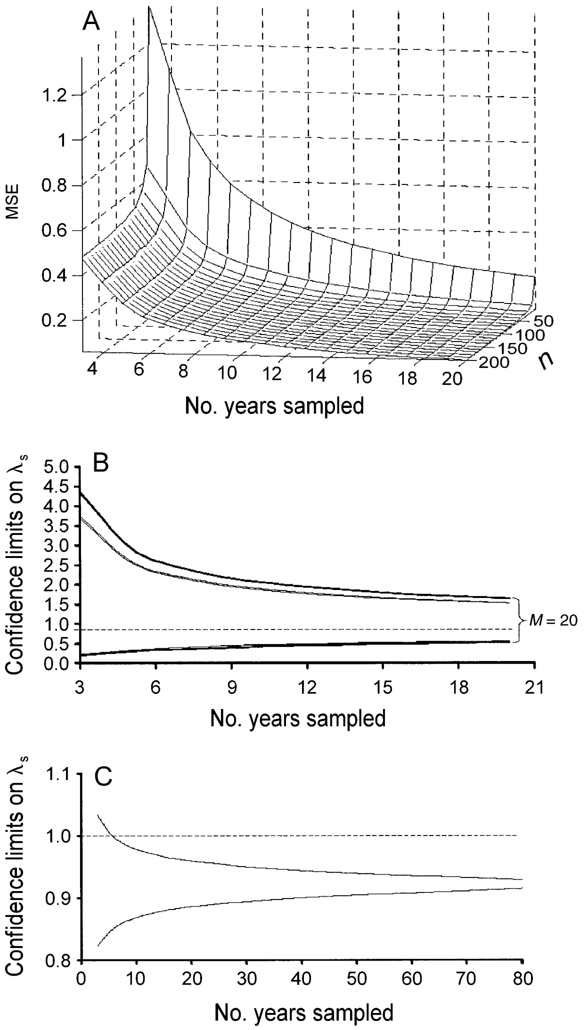

FIG. 4. Effects of sampling regimes on uncertainty in growth rates for *Collinsia verna* and the desert tortoise. (A) Predicted sampling variance in estimates of log λ_s (stochastic lambda) for different sampling intensities and durations for *C. verna*. (B) Upper and lower 95% confidence limits around estimates of λ_s for sampling intensities of M = 200, 100, and 20 individuals for *C. verna*. The outermost two lines are for 20 individuals sampled for each vital rate in each year, while the limits for intensities of 100 and 200 individuals are inward and nearly identical. (C) 95% confidence limits around estimates of λ_s for the desert tortoise. In (B) and (C), dashed lines indicate $\lambda_s = 1$, a stable population.

As these examples show, clear and simple results are easily generated using this approximation. Three more points about its use should be mentioned. First, while our examples all use the same sampling intensity for all rates, there is no difficulty in analyzing more realistic patterns of sampling intensity. Given that larger, older, and less common individuals are frequently the most important members of populations and are also less intensively sampled, these inequities in sampling intensity can have important effects on Var(log $\hat{\lambda}_s$). Second, we usually really will be guessing about the ‘‘true’’ vital rates used to make these approximations, and it therefore behooves us to make several different guesses, especially about patterns of environmental variability, about which we are likely to be least certain. Finally, simply by leaving out the terms in Eq. 2 for the effects of sampling variation in vital rate variances and correlations, the approximation can also be used to assess the variance in predictions of deterministic models (see also Alvarez-Buylla and Slatkin 1994), allowing a contrast between the best model form to use in generating predictions with limited data.

CONCLUSIONS

In the ecological literature, a great deal of effort has gone into arguing the merits of different demographic outputs and developing elaborate forms of population analysis. Ironically, the problems of how to use these models with real data have received far less attention. Our simulations of limited sampling show that inaccuracies in estimated growth rates can be large with even reasonable amount of sampling effort (see also Fieberg and Ellner 2001). Still, we find that stochastic models will generally perform better than will deterministic ones with moderate to long sampling durations. However, with less than five years of data stochastic models may give uncomfortably variable answers. If sampling duration is low, and the variation in vital rates is high, improved estimates of population behavior may come from deterministic models—a counterintuitive result. In this case, the inaccuracies involved in estimating the effects of environmental stochasticity are so large that they outweigh their own strong effects on population dynamics. More encouragingly, our results suggest that unfeasibly large sampling duration or intensity is not necessary in order for stochastic models to perform reasonably well.

While these simulations provide some insights into the interplay of model complexity and data scarcity, this is only a first stab at the problem of when and how to judge the right degree of complexity to build into a demographic model, whether for use in a PVA or more basic ecological work. Furthermore, we have addressed only one of the many decisions that must be made in setting up a demographic study or analyzing its results (Caswell 2001, Easterling et al. 2000, Gross 2002, Morris and Doak 2002, Kaye and Pyke 2003). In particular, we have emphasized only a single measure of model performance, the variance in predicted population growth rate. Many other outputs of demographic models are of keen interest, including elasticity values and extinction risk, which are likely to be even more sensitive to the problems of data scarcity. In addition, any use of demographic models to forecast growth or extinction relies on measured rates remaining relatively stable into the future, an increasingly worrisome assumption.

It could be assumed that it is always better to construct less biased predictions of population growth, regardless of accuracy. However, we would argue against this view. A more precise estimate of population growth, even if biased, is far more useful than a less reliable one. This is especially true if we know the direction of this bias, as is the case with deterministic growth rate predictions. Especially for PVA models, it can be much more useful to have a prediction of the most optimistic likely outcome than it is to have such an uncertain, but unbiased, prediction that it can support few or no conclusions at all. This said, anyone who is actually doing a demographic analysis should probably perform both stochastic and deterministic analyses, and carefully draw inferences using information from both.

While there have been numerous calls for the presentation of demographic results to include analyzes of their uncertainty, the practical difficulties of doing so mean that most demographic studies still do not include any such quality assessment. The approximation for the sampling variance of population growth rate that we provide should allow such an assessment to be made far more easily. This estimator can also be used to plan future demographic studies to better fulfill the goals of investigators. A great deal of heartache can be expended in the decisions about how long and how intensively to pursue a demographic study. To date, these decisions have been almost always made from intuitions about the trade-off between crippled knees (at least for plant demographers) and data quantity, but with almost no analysis of the power of the study. By providing a formula for sampling uncertainty in growth rates, and software to easily implement it, our hope is that these decisions can be made more carefully, improving the collection and interpretation of demographic information.

ACKNOWLEDGMENTS

A. P. Rose and M. Kauffman provided valuable comments on several drafts of this work. A. Ellison and two reviewers also provided useful comments. This work was partially supported by USDA 2002-00610 to D. F. Doak and DEB-0087078 to W. F. Morris and D. F. Doak.

LITERATURE CITED

Alvarez-Buylla, E. R., and M. Slatkin. 1994. Finding confidence limits on population growth rates: three real examples revised. Ecology **75**:255–260.

Beissinger, S., and M. I. Westphal. 1998. On the use of demographic models of population viability in endangered species management. Journal of Wildlife Management **62**: 821–841.

Brook, B. W., J. J. O'Grady, A. P. Chapman, M. A. Burgman, H. R. Akcakaya, and R. Frankham. 2000. Predictive accuracy of population viability analysis in conservation biology. Nature **404**:385–387.

Burgman, M. A., S. Ferson, and H. R. Akçakaya. 1993. Risk assessment in conservation biology. Chapman and Hall, London, UK.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information-theoretic approach. Springer, New York, New York, USA.

Caswell, H. 2001. Matrix population models: construction, analysis and interpretation. Second edition. Sinauer, Sunderland, Massachusetts, USA.

Dixon, P. M., A. M. Ellison, and N. J. Gotelli. 2005. Improving the precision of estimates of the frequency of rare events. Ecology **86**:1114–1123.

Doak, D. F., P. Kareiva, and B. Klepetka. 1994. Modeling population viability for the desert tortoise in the Mojave desert. Ecological Applications **4**:446–460.

Doak, D. F., D. M. Thomson, and E. S. Jules. 2002. PVA for plants: understanding the demographic consequences of seed banks for population health. Pages 312–337 *in* S. R. Beissinger and D. R. McCullough, editors. Population viability analysis. University of Chicago Press, Chicago, Illinois, USA.

Easterling, M. R., S. P. Ellner, and P. M. Dixon. 2000. Size-specific sensitivity: applying a new structured population model. Ecology **81**:694–708.

Ellner, S. P., J. Fieberg, D. Ludwig, and C. Wilcox. 2002. Precision of population viability analysis. Conservation Biology **16**:1–5.

Engen, S., O. Bakke, and A. Islam. 1998. Demographic and environmental stochasticity: concepts and definitions. Biometrics **54**:840–846.

Fieberg, J., and S. P. Ellner. 2001. Stochastic matrix models for conservation and management: a comparative review of methods. Ecology Letters **4**:244–266.

Gross, K. 2002. Efficient data collection for estimating growth rates of structured populations. Ecology **83**:1762–1767.

Hilborn, R., and M. Mangel. 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, New Jersey, USA.

Kalisz, S. 1991. Experimental determination of seed bank age structure in the winter annual, *Collinsia verna*. Ecology **72**:575–585.

Kaye, T. N., and D. A. Pyke. 2003. The effect of stochastic technique on estimates of population viability from transition matrix models. Ecology **84**:1464–1476.

Kendall, B. E. 1998. Estimating the magnitude of environmental stochasticity in survivorship data. Ecological Applications **8**:184–193.

Ludwig, D. 1999. Is it meaningful to estimate a probability of extinction? Ecology **80**:298–310.

Ludwig, D., C. J. Walters, and J. Cooke. 1988. Comparison of two models and two estimation methods for catch and effort data. Natural Resource Modeling **2**:457–498.

Morris, W. F., and D. F. Doak. 2002. Quantitative conservation biology: the theory and practice of population viability analysis. Sinauer, Sunderland, Massachusetts, USA.

Morris, W. F., and D. F. Doak. 2004. Buffering of life histories against environmental stochasticity: accounting for a spurious correlation between the variabilities of vital rates. American Naturalist **163**:579–590.

Tuljapurkar, S. 1990. Population dynamics in variable environments. Spring-Verlag, New York, New York, USA.

White, G. C. 2000. Population viability analysis: data requirements and essential analyses. Pages 288–331 *in* L. Boitani and T. K. Fuller, editors. Research techniques in animal ecology: controversies and consequences. Columbia University Press, New York, New York, USA.

APPENDIX A

A description of problems with estimating and using correlations in vital rates is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-062-A1.

APPENDIX B

A description of the simulation models used is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-062-A2.

APPENDIX C

An approximation for sampling variance in stochastic growth rate estimates is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-062-A3.

SUPPLEMENT

A MATLAB program to calculate the sampling variance in estimated log(stochastic lambda) values given different sampling intensities and durations is available in ESA's Electronic Data Archive: *Ecological Archives* E086-062-S1.