DISSERTATION

SIMULATING SPECIES ASSEMBLAGES AND EVALUATING SPECIES RICHNESS ESTIMATORS

Submitted by

Gordon C. Reese

Graduate Degree Program in Ecology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2012

Doctoral Committee:

Advisor: Kenneth R. Wilson

Amy L. Angert
Curtis H. Flather
Thomas J. Stohlgren

ABSTRACT

SIMULATING SPECIES ASSEMBLAGES AND
EVALUATING SPECIES RICHNESS ESTIMATORS

Conservation efforts have long emphasized protecting biologically diverse areas. Species richness, the number of species in a defined area, is the most frequently used biodiversity measure and it can be used for selecting amongst different areas and for studying process effects over time. Despite its intuitive appeal and conceptual simplicity, species richness is often difficult to quantify, even in well-surveyed areas, because of sampling limitations such as survey effort and species detection probability. This has led to the development of numerous species richness estimators.

Nonparametric estimators present the least biased option, but no particular estimator has consistently performed best. Factors such as abundance, behavior, and survey design vary widely between locations, species, and datasets, affecting richness estimates and revealing the limitations of estimators. Increasing our understanding of the relationships between estimator performance and important factors can improve prediction and, ultimately, estimator utility.

My objective was to evaluate the performance of nonparametric species richness estimators, both established and new, across a wide range of species assemblages. Given the difficulties of surveying many different assemblages and of assessing performance when the true state of an assemblage is unknown, I choose to develop a program for estimating the species richness of assemblages simulated with user-specified parameters. I also sought to use the following studies to develop a framework for selecting the best estimator given particular

assemblage attributes and survey design parameters. In the following studies, I assumed that every individual was: 1) independent, i.e., there were no clonal colonies, 2) detectable, 3) correctly identified, and 4) sessile for the duration of a survey.

Simulations were used because they are convenient, possibly the only, means of simultaneously controlling many characteristics on an assemblage. By controlling only those factors that are of interest and excluding others, simulations represent a simplification of the real world, i.e., they trade convenience for realism, which can benefit cause-and-effect assessments because the real world involves many additional factors that can complicate estimation efforts. For example, there are difficult to detect species, e.g., cryptic and extremely small species, as well as limited sampling efforts that can further reduce estimator performance. The real world might therefore not conform to the trends detected in a simulated environment, particularly beyond the range of evaluated factors. For such reasons, I recommend that application of these results to the real world, especially extrapolation, be done with caution. Simulated environments ultimately represent a best case scenario, so if estimators perform poorly there, how can we trust them in the much more complicated real world?

Several factors influence estimator performance including the number of species in the assemblage, total abundance or density, distribution of abundances across species, spatial configuration of individuals, species detection probability, and survey effort. In Chapter 2, I developed a species assemblage simulator for assessing estimator performance across a wide range of conditions. The program, SimAssem, allows a user to specify both assemblage and survey parameters and generates encounter histories as input for various estimators. In addition to nonparametric species richness estimators, SimAssem includes: 1) estimators of the additional

amount of survey effort required to encounter user-specified proportions of the estimates from the Chao estimators and 2) an option to process existing encounter histories.

In Chapter 3, I evaluated the bias, precision, and accuracy of 13 nonparametric estimators across simulated assemblages that are systematically varied for the number of species, distribution of species abundances, total abundance, spatial configuration of individuals, and species detection probability. I also varied sampling effort and survey design.

When averaged across all assemblages, the estimators were less negatively biased than a raw count of species in a sample and there was generally a tradeoff between bias and precision. Two relatively new estimators based on the similarity of repeated subsets of surveys were most accurate and appeared to reach asymptotes more quickly than the other estimators when used with real data. The number of species, distribution of species abundances, and effort had the largest effects on performance, largely by affecting sample coverage, i.e., the proportion of the species pool contained in the sample. Increases in the true number of species and decreases in the evenness of abundances negatively affected bias and accuracy. Increasing the rate of encounters via total abundance, species detection probability, and effort generally improved bias and accuracy. There was a moderate increase in bias when individuals were aggregated and sampled using a non-random survey design. Also, a refined estimator selection framework based on sample coverage showed promising results when applied to real datasets.

Point estimates of species richness are of limited value without some measure of reliability; nevertheless, species richness estimates are often reported without any measure of precision. For many species richness estimators, analytically derived variance estimators exist. For others, approaches such as bootstrap and jackknife resampling can be used.

In Chapter 4, I evaluated variance estimators across levels of the factors with the largest effects on species richness estimators, representing a portion of the data simulated for Chapter 3. Variation in the species richness estimates generally increased with the true number of species. The analytical variance estimates usually exceeded those of the two resampling procedures, but all three methods were negatively biased at most factor levels. Similarly, the analytical estimators often resulted in the largest confidence interval coverage levels, though coverage was less than the nominal 95% in all except one case. Furthermore, there was generally a negative relationship between the achieved coverage level and true number of species. Bootstrap resampling always produced the best coverage for the bootstrap species richness estimator and occasionally performed similarly well with other species richness estimators. Confidence interval coverage was, in general: 1) smallest in assemblages with log-series distributions and largest in assemblages with particulate-niche distributions, 2) positively related to effort and species detection probability, and 3) variable across species richness estimators as a function of total abundance. The abundance-based coverage estimator and its associated analytical variance estimator regularly achieved the largest coverage levels, so I recommended its use when there is little or no information to suggest that another estimator is more appropriate.

ACKNOWLEDGMENTS

I am greatly indebted to my advisor, Kenneth Wilson, for the many hours that he devoted to this project whether it was to answer my questions, edit one of my multiple drafts, find funding, or check on my sanity. I am grateful to Curtis Flather for ideas that stimulated much of this work and to Thomas Stohlgren, Amy Angert, and Ryan Elmore for serving on my committee. I thank Robert Colwell, Robert Gardner, Jim Graham, Tom Hobbs, Jeff Laake, Shirley Pledger, Eric Smith, and Gary White for addressing questions on everything from species abundance and configuration patterns to programming algorithms and estimator formulas.

My work on several outside projects funded some portion of my degree and I am thankful for those interesting and additionally rewarding opportunities. First and foremost was work with the USFS Rocky Mountain Research Station, some of which was completed prior to entering my graduate program, but was nonetheless an instrumental step towards this degree. Other agencies included the USDA Animal and Plant Health Inspection Service, National Park Service, USDA Forest Service, and the Yellowstone Ecological Research Center. I also thank the Departments of Biology and Fish, Wildlife, and Conservation Biology for teaching assistantships that helped make this degree possible.

My family, particularly my mother Elizabeth, father Gordon, sister Allison, and grandmothers Evangeline and Joan, were always there to provide the support, encouragement, and love that was so important to completing this project. I hope that they forever know how much a part of this they were and that my appreciation is immeasurable.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Interests in natural habitat span a wide spectrum, from resource extraction and development to recreation and protection. Human landscape modifications include, but are certainly not limited to, reducing and fragmenting natural habitat (Thiollay 2006). Resulting patch sizes and configurations can be insufficient for the persistence of extant species. Ecology includes the study of species distributions and abundances (Andrewartha 1961, Krebs 1994) as well as the predictable patterns that species exhibit. The ultimate goal of ecology is to understand the processes behind observed patterns. Given sufficient understanding, ecologists can inform policy makers and resource planners on how to manage for environmental change (e.g., climate change, fragmentation).

Efforts aimed at protecting biologically diverse areas often focus on species richness (SR), the number of species in a defined area (McIntosh 1967), sometimes to a degree that neglects consideration of other, potentially parallel, system properties (Rapport et al. 1985). Species richness is a fundamental attribute of a species assemblage, i.e., phylogenetically related species co-occurring in a habitat, and an intuitive measure of biological diversity. Species richness is a useful ecological measure for selecting amongst areas for protection and for studying processes that cause changes over time. For example, one might investigate the influence of factors such as land-use, climate, and degree of habitat heterogeneity on SR (Gotelli et al. 2009). Despite its conceptual simplicity, SR is often difficult to quantify, even in well-surveyed areas, because of sampling limitations such as insufficient survey effort and small species detection probabilities (p).

Survey data are, by definition, incomplete and thus, estimation is often used to reveal the number of species not encountered. The development of estimators that are robust to the wide variations in factors such as abundance, behavior, and survey design that can occur between locations, species, and datasets has been difficult at best. Several factors that affect estimator performance (see Coddington et al. 1996, Chazdon et al. 1998, Brose et al. 2003) are ultimately linked through sample coverage (*sc*), which is the proportion of a species pool represented in a sample (Heltshe and Forrester 1983, Baltanás 1992, Wagner and Wildi 2002). Not surprisingly, SR estimators have generally performed worst at factor levels resulting in the smallest *sc* (Brose et al. 2003). It could be possible to increase the utility of an estimator by improving our understanding of its relationship to influential factors. This chapter provides background information for the remainder of the dissertation with information specifically about: 1) the factors shown or suspected to affect the performance of SR estimators, 2) the available categories of SR estimators as well as the reasons for testing the selected set, 3) the importance of variance estimates, and 4) the objectives, hypotheses, assumptions, usefulness, and structure of the included chapters.

DRIVERS OF SAMPLE COVERAGE

Number of species, species-abundance distribution, and spatial configuration of individuals

Several ecological attributes of a species assemblage can affect *sc* and thereby, the performance of SR estimators. One component of *sc*, the true number of species in an area (S_{true}), has great potential to affect *sc*. The other component of *sc*, the number of species observed (S_{obs}), is affected by *p*, considered here to be a function of both species behavior and survey design. For example, mobile and nocturnal animals can be difficult to detect, especially when combined with an inadequate survey design. Biological factors such as body size and

coloration are important and *p* can also depend on habitat, e.g., some species are more cryptic in certain habitats.

Species abundances can span several orders of magnitude, sometimes within a single assemblage. Methods used to approximate species abundance distributions can be categorized as statistically- or niche-based; the basic distinction being that the former essentially represents a summary statistic that is fit to data and the latter a distribution that results from a systematic division of resources, i.e., individuals, that focuses on process. The most common statistical methods use a log-normal distribution, where a few species are relatively abundant and a few species are relatively rare (Preston 1948, Bulmer 1974, Otis et al. 1978, Sugihara 1980, Magurran 1988, Wilson et al. 1998) or a log-series distribution, where a few species are exceptionally abundant and most species are relatively rare (Fisher et al. 1943, Williams 1964). The log-series and a truncated log-normal distribution are related in that both can successfully fit an incomplete dataset. However, once additional sampling reveals the least abundant species and a symmetrical abundance distribution, only the log-normal distribution can provide a successful fit. Using a statistical distribution to represent ecological abundance patterns is often criticized for not proposing any explanation for observed patterns. Rather, a distribution is simply fit to the data. However, there is at least one proposed biological explanation that is based on species residency status. Magurran and Henderson (2003) concluded that the abundances of permanent (breeding) and non-permanent (occasionally breeding) species tend to follow log-normal and log-series distributions, respectively.

Abundance data being successfully fit by a statistically-based model can be an artifact of the multiplicative product of many positive independent factors, which effectively masks any structure by including so much heterogeneity. For this reason, statistically-based models could

be most appropriate for assemblages with a large number of relatively unrelated species (Tokeshi 1990). In contrast, assemblages with a small number of taxonomically related species are more likely shaped by competition for shared resources. Niche-based models emulate this process by breaking a stick, the length of which represents the available shared resources or, equivalently, the total number of individuals (Tokeshi 1993). Niche-based models of species abundance distributions include the sequential broken-stick, dominance-decay, dominance-preemption, geometric-series, particulate-niche, power fraction, random-fraction, random-assortment, and sequential 75% (Tokeshi 1990, 1993, 1996, also see Chapter 2 for detailed descriptions).

The spatial configuration of individuals and species can exhibit many patterns and aggregation occurs to varying degrees. When surveyed, aggregation can result in spatially autocorrelated data that are more similar as the distance between data decreases. Correlations such as these violate independence assumptions, complicate modeling efforts (Legendre 1993), and could partly explain inconsistent conclusions about the effects of aggregation on SR estimators (see Baltanás 1992, Chazdon et al. 1998, Walther and Morand 1998, Wagner and Wildi 2002, Brose et al. 2003). Abundance or, in another guise, density, is a partial component of spatial configuration. In the same confined space, for example, the average distance between the individuals of a larger population would be less than for a smaller population and this would affect some measures of aggregation. A positive relationship between density and estimator performance could be a result of *p*, or the number of encounters, increasing as a species becomes more abundant (see Baltanás 1992, Walther and Morand 1998).

Survey design and effort

Survey design can also affect *sc* because no design detects all species with equal probability (Boulinier et al. 1998). A random design not biased towards any particular habitat,

taxa, or species is preferable, but often difficult to implement. A design that accesses only a subset of the true species pool, e.g., a non-random survey design that utilizes roads such as the one used by the Breeding Bird Survey (Robbins et al. 1986), results in a comparatively smaller *sc*. Other survey details such as trap and bait type, the mesh size of a net, time of day, and season can also affect S_{obs} and, thereby, *sc*. Increasing effort will often increase *sc*, though the effort required for a census is usually logistically and financially prohibitive.

SPECIES RICHNESS ESTIMATION TECHNIQUES

Species richness estimators can be classified into three categories: 1) extrapolation using a species-area relationship or species accumulation curve, 2) parametric estimation with an abundance distribution or a derived formula, and 3) nonparametric estimation (Bunge and Fitzpatrick 1993, Colwell and Coddington 1994, Palmer 1995, Chazdon et al. 1998). All estimators have limitations, but many of the available SR estimators are less biased than S_{obs} and based on testable assumptions (see Baltanás 1992, Bunge and Fitzpatrick 1993, Walther and Morand 1998, Chiarucci et al. 2003, Walther and Moore 2005). Nonparametric estimators have performed favorably when compared with S_{obs} and estimators from the other two categories and are therefore the focus of this dissertation (Table 1.1; see also tables in Gotelli and Colwell 2001, Walther and Moore 2005).

Methods that address variable detection probabilities are supported by several studies (e.g., Boulinier et al. 1998, MacKenzie et al. 2002). Many of the nonparametric estimators were developed for estimating population size using the capture frequencies from capture-recapture surveys (Burnham and Overton 1978, Chao 1987, Lee and Chao 1994). Originally used to model detection probabilities that can vary across individuals, the nonparametric estimators model detection probabilities that can vary across species when used to estimate SR. Behavior

and abundance can cause greater detectability differences between species than those that occur between the individuals in a population (Burnham and Overton 1979, Brose et al. 2003). Additionally, environmental gradients and spatial aggregation can cause *p*'s to vary over space more than the detection probabilities of individuals vary over time (see Legendre 1993, Brose et al. 2003).

VARIANCE ESTIMATION FOR SPECIES RICHNESS ESTIMATORS

The usefulness of a SR estimate is compromised when it is reported without information regarding its precision, e.g., a variance estimate or confidence interval, as is any estimate. Variance estimates themselves can be biased and the actual coverage levels of associated confidence intervals can be less than or greater than the nominal level.

I assessed the frequency at which variance estimates are reported along with SR estimates by reviewing 21 recent articles (2005–2009) found in the Web of Science database (topic search = non*parametric "species richness" estimat*). A paper had to meet two additional criteria before being reviewed: 1) it had to be published in English and 2) it had to use previously published estimators, i.e., I excluded papers that primarily introduced one or more new estimators. Eleven of the papers reported nonparametric SR estimates without any estimate of variance. Of the 10 papers that reported variance estimates, four used confidence intervals, four used standard errors, one used box-plots, and one included a single line of text noting which estimator had the greatest variance. I found only one reference to the analytical variance estimators that are available for many of the nonparametric estimators, though the frequent use of the program EstimateS (Colwell 2006) would indicate they were occasionally used for a few of the nonparametric estimators. Also, variance estimates were sometimes computed from the

repeated randomizations used to construct species accumulation and rarefaction curves. Thus, the performance of variance estimators is an important, but little studied issue for SR estimators.

OBJECTIVES, HYPOTHESES, AND ASSUMPTIONS

My first objective was to evaluate the performances of nonparametric SR estimators across a wide range of species assemblages. Despite their importance as an indicator of reliability, variance estimators of SR have been little studied and are often not reported. Therefore, a second objective was to evaluate variance estimators. Analytical estimators are unavailable for some of the nonparametric estimators, so I also wanted to compare two general variance estimation procedures, bootstrap and jackknife resampling. The scope of the first two objectives suggested the usefulness of simulations where ecological factors and survey design parameters are controllable and known. A goal that originated with these objectives was the development of a program for simulating and surveying user-defined assemblages and estimating SR from resulting sample data (see Chapter 2). It was also my objective to use these studies to develop a framework for selecting the best estimator given particular assemblage attributes and survey design parameters.

My research included both factor- and estimator-specific hypotheses. I hypothesized that the SR estimators would perform best in assemblages with the most equally abundant species and evenly spaced individuals, i.e., assemblages with relatively homogeneous p's. For a given level of sampling effort, I hypothesized that estimator performance would be negatively related to S_{true} because variance tends to increase with the size of the estimate. Based on previously reported performances, I hypothesized that no estimator would: 1) perform best across all factor combinations and 2) be both least biased and most precise in any particular comparison (see Table 1.1). Furthermore, my hypotheses involving specific estimators were that the first-order

jackknife estimator (Burnham and Overton 1978) would be least biased and that the closed population estimator developed by Chao (1987) would be most precise when averaged across all factor levels. I hypothesized that derived analytical variance estimators would perform better than general procedures such as bootstrap and jackknife resampling. Since the variance estimators are based on the same data as the SR estimators, I hypothesized that they would be similarly affected, e.g., in magnitude and direction, by the same independent factors.

I made the following assumptions in these studies: 1) every individual was independent, i.e., there were no clonal colonies, 2) all individuals in a surveyed area had a positive probability of being detected, 3) every individual was correctly identified, and 4) all individuals were sessile for the duration of a survey.

VALUE OF THE SIMULATION EXPERIMENTS

Conservation efforts could benefit from a better understanding of how SR estimators perform across various datasets, especially if it provides the information needed to select the best estimator for describing and comparing assemblages. The studies presented in this dissertation collectively aim to advance the field by: 1) developing a program with which estimators can be easily evaluated across simulated assemblages, 2) evaluating several relatively untested estimators, and 3) comparing both SR estimators and three different variance estimation methods across systematically varied assemblages. I also updated and expanded a selection framework by Brose et al. (2003) to improve richness estimates based on incomplete survey data and, consequently, aid monitoring efforts and biological reserve design.

DISSERTATION DETAILS

I introduced a computer program for evaluating the performance of SR estimators in Chapter 2, including the algorithms by which assemblages are simulated, the survey design

options, the SR and variance estimators included in the program, and the input and output options. In Chapter 3, I evaluated the performance of several SR estimators across systematically varied assemblages, focusing on nonparametric estimators. Two studies, in particular, have completed similar research on the effects of species abundance distributions, spatial heterogeneity, and survey effort on the performance of SR estimators (Wagner and Wildi 2002, Brose et al. 2003). In comparison to the studies presented in this dissertation, Wagner and Wildi (2002) did not compare their results to a known number of species and neither study included variation in survey design and *p* nor evaluated the precision of the estimates (see Table 1.2 for a comparison of those studies with those conducted in Chapter 3). I considered *p* an umbrella factor that could, theoretically, account for numerous factors including body size, mobility, habitat, and trapping details. In Chapter 4, I evaluated and compared variance estimation techniques including bootstrap and jackknife resampling and, where possible, analytical estimators. A majority of the tests conducted for this dissertation used simulated data, but a few real datasets were also used in Chapter 3. Worked examples of the various species abundance distributions and estimator formulas are given in Appendices I and II, respectively.

Table 1.1. The performance of species richness estimators across studies.

Authors	Year	Evaluated estimators	Best and/or recommended estimators
Palmer	1990	LND, LOGLIN, LOGLOG, MONOD, Boot, Jack1, (incorrect Jack2), S_{obs}	Jack1
Palmer	1991	LND, LOGLIN, LOGLOG, MONOD, Boot, Jack1,2, S_{obs}	Jack2 (for bias), Jack1 (for precision)
Baltanas	1992	SV SAC, LND, Jack1	LND, Jack1
Bunge & Fitzpatrick	1993	3 SAC-based, LND, ACE, Chao1, Jack1,2,3,4,5, Bernoulli, Hypergeometric, Multinomial, Poisson	ACE
Colwell & Coddington	1994	MM, ACE, Boot, Chao1,2, Jack1,2, S_{obs}	Chao2, Jack2
Coddington et al.	1996	LND, MM, Chao1,2, Jack1	None
Chazdon et al.	1998	MM, ACE, Boot, Chao1,2, ICE, Jack1,2	ICE, Chao2, Jack2, MM
Poulin	1998	Boot, Chao2, Jack1, S_{obs}	Boot, S_{obs} (large coverage)
Walther & Morand	1998	2 SAC-based, Boot, Chao1,2, Jack1,2	Chao2, Jack1
Hellmann & Fowler	1999	Boot, Jack1,2	Jack2, Jack1
Schmit et al.	1999	LND, ACE, Boot, Chao1,2, Jack1,2	ACE
Zelmer & Esch	1999	Boot, Jack-int, S_{obs}	Jack-int
Chiarucci et al.	2001	MM-mean, Boot, Chao2, Jack1,2	Jack1, Jack2
Melo & Froehlich	2001	SAC-based, 13 LND, non-parametric	Jack1, Jack2, Chao1, Chao2
Walther & Martin	2001	12 SAC-based, ACE, Boot, Chao1,2, ICE, Jack1,2	Chao2, Chao1, Jack1, Jack2
Brose	2002	Boot, Chao2, Jack1,2	Chao2
Cam et al.	2002		jackknife of Pollock & Otto 1983
Herzog et al.	2002	9 LND, MM-runs, MM-mean, ACE, Boot, Chao1,2, ICE, Jack1,2	MM
Longino et al.	2002	1 LND-based, MM, ICE	None
Wagner & Wildi	2002	Chao2, ICE, Jack1,2, S_{obs}	Jack1, ICE w/ large (80%) coverage
Brose et al.	2003	MM, Expo, Chao2, ICE, Jack1,2,3,4,5,-sel,-int, S_{obs}	Jacks, depends on coverage & evenness
Chiarucci et al.	2003	Boot, Chao2, Jack1,2	Jack2
Foggo et al.	2003	ACE, Boot, Chao1,2, ICE, Jack1,2	Chao1
Petersen & Meier	2003	Presten LND, Poisson LND, ACE, Chao1	Poisson LND (slightly)
Petersen et al.	2003	MM-mean, ACE, Boot, Chao1,2, ICE, Jack1,2	Jack2, Jack1
Brose & Martinez	2004	MM, ACE, Chao1,2, ICE, Jack1,2,3,a1,a2,a3	depends on sample coverage
Cao et al.	2004	Boot, Chao2, CY-1; CY-2, ICE, Jack1,2	CY-2, Jack2
Walther and Moore	2005	MM-mean, ACE, Chao1,2, Jack1,2	Chao2, Jack2, Jack1, Chao1
Ulrich & Ollik	2005	AL, Jack2,5, P5	Elog-series, Elognormal, Jack2
Hortal et al.	2006	3 SAC-based, 1 S-area, MM, ACE, Boot, Chao1,2, F3,5,6, ICE, Jack1,2	ACE, ICE, Chao1, Chao2, Jack1, Jack2

Magnussen et al.	2006	ACE, BBIN, Boot, Chao2, GPOI, JKk, MBIN, MPOI, PET	GPOI, ACE, JKk
Canning-Clode et al.	2008	MM, ACE, Chao1,2, Jack1,2	Jack2, then MM

ACE = abundance-based coverage estimator (Chao and Lee 1992)

AL = aymptotic linear estimators (Ulrich 1999)

BBIN = beta-binomial estimator (Dorazio and Royle 2003)

Bernoulli = (Goodman 1949, Esty 1985); both based on Bernoulli sample

Boot = bootstrap estimator (Smith and van Belle 1984)

Chao1 = abundance-based estimator (Chao 1984)

Chao2 = incidence-based estimator (Chao 1987)

Expo = exponential species accumulation curve (Holdridge et al. 1971)

F3, F5, F6 = extrapolation estimators (Rosenzweig et al. 2003)

GPOI = gamma-mixed Poisson estimator (Chao and Bunge 2002)

Hypergeometric = (Goodman 1949, Shlosser 1981); both based on hypergeometric sample

ICE = incidence-based coverage estimator (Lee and Chao 1994)

Jack# = #-order jackknife estimator (Burnham and Overton 1978)

Jacka# = abundance-based jackknife (Burnham and Overton 1979)

Jack-sel = jacknife select (Burnham and Overton 1978)

Jack-int = jacknife interpolated (Burnham and Overton 1978)

JKk = generalized jackknife estimator (Sharot 1976)

LND = integration of lognormal distribution

LOGLIN = logarithmic-linear regression of species-area curve (non-asymptotic) (Gleason 1922)

LOGLOG = logarithmic-logarithmic regression of species-area curve (non-asymptotic) (Gleason 1922)

MBIN = mixed-binomial estimator (Norris and Pollock 1998)

MM = Michaelis-Menten species accumulation curve (asymptotic) (Michaelis and Menten 1913)

MONOD = Monod function (a hyperbolic model) (Lauga and Joachim 1987)

MPOI = mixed-Poisson estimator

Multinomial = (Darroch 1958, Darroch and Ratcliff 1980, Sichel 1986); all based on multinomial sample

PET = Petersen capture-recapture estimator (Thompson 1992)

Poisson = (Ord and Whitmore 1986; Efron and Thisted 1975); both based on Poisson sample

P5 = parametric estimator (Turner et al. 2003)

SAC = species accumulation/area curve

S_{obs} = number of species observed

SV SAC = species-area curve (Stout & Vandermeer 1975)

Table 1.2. Tested assemblage factors and factor levels.

Factor	**Wagner and Wildi (2002)**	**Brose et al. (2003)**	**Chapter 3 and/or 4**
S_{true}	NA[2]	25, 50, 100, 150, 200, 250, 500	25, 100, 500
Total abundance	NA[2]	Ave. 200/spp.	6250, 12500
Species abundance distributions	Broken-stick Geometric series Log-normal	Broken-stick Random-fraction Random-assortment	Log-normal Log series Particulate-niche
Spatial patterns	Aggregated (species-specific) Combined (aggregation, strong gradient, edge effect) Edge effect Gradient (weak and strong) Homogeneous	Aggregated (3 species-specific types) Gradient (none, weak, and strong)	Aggregated (species-specific) Hyper-dispersed Random
Sample design	Random	NA[2]	Random, linear transect
Sample intensity	20, 50, 100, 500 (sequentially)	25, 200, 500	100 (1%), 500 (5%)
Detection probability	None	None	0.5, 0.9 Decreasing with abundance Increasing with abundance
Species richess estimators[1]	Chao2 ICE Jack1,2 S_{obs}	Chao2 ICE Jack1,2,3,4,5 Jack-selected Jack-interpolated SAC (exponential, Michaelis-Menten) S_{obs}	ACE Bootstrap Chao1,2 CY-1,-2 ICE Jack1,2,3,4,5 Mixture S_{obs}
Variance estimators	None	None	Analytical Bootstrap resampling Jackknife resampling

[1]In addition to the raw count, S_{obs}, the evaluated estimators include the abundance-based coverage (ACE, Chao and Lee 1992), bootstrap (Smith and van Belle 1984), abundance-based (Chao1, Chao 1984), incidence-based (Chao2, Chao 1987), incidence-based coverage (ICE, Lee and Chao 1994), mixture estimator with two groups (Pledger 2000), first- through fifth-order, selected, and interpolated jackknife (Burnham and Overton 1978, Burnham and Overton 1979), similarity of repeated surveys (CY-1, Cao et al. 2001; CY-2, Cao et al. 2004) and two extrapolated species accumulation curves (SAC), the exponential equation (Holdridge et al. 1971) and Michaelis-Menten model (Michaelis and Menten 1913).
[2]The tested factor levels were unknown.

LITERATURE CITED

Andrewartha H.G. 1961. *Introduction to the study of animal populations*. University of Chicago Press, Chicago, IL, USA.

Baltanás A. 1992. On the use of some methods for the estimation of species richness. Oikos 65:484-492.

Boulinier T., Nichols J.D., Sauer J.R., Hines J.E. and Pollock K.H. 1998. Estimating species richness: the importance of heterogeneity in species detectability. Ecology 79:1018-1028.

Brose U. 2002. Estimating species richness of pit catches by non-parametric estimators. Pedobiologia 46:101-107.

Brose U. and Martinez N.D. 2004. Estimating the richness of species with variable mobility. Oikos 2004:292-300.

Brose U., Martinez N.D. and Williams R.J. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. Ecology 84:2364-2377.

Bulmer M.G. 1974. On fitting the Poisson lognormal distribution to species-abundance data. Biometrics 30:101-110.

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

Burnham K.P. and Overton W.S. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika 65:623-633.

Burnham K.P. and Overton W.S. 1979. Robust estimation of population size when capture probabilities vary among animals. Ecology 60:927-936.

Cam E., Nichols J.D., Sauer J.R. and Hines J.E. 2002. On the estimation of species

richness based on the accumulation of previously unrecorded species. Ecography 25:102-108.

Canning-Clode J., Valdivia N., Molis M., Thomason J.C. and Wahl M. 2008. Estimation of regional richness in marine benthic communities: quantifying the error. Limnology and Oceanography: Methods 6:580-590.

Cao Y., Larsen D.P. and Hughes R.M. 2001. Estimating total species richness in fish assemblage surveys: A similarity based approach. Canadian Journal of Fisheries and Aquatic Sciences 58:1782-1793.

Cao Y., Larsen D.P. and White D. 2004. Estimating regional species richness using a limited number of survey units. Ecoscience 11:23-35.

Chao A. 1984. Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11:265-270.

Chao A. 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783-791.

Chao A. and Bunge J. 2002. Estimating the number of species in a stochastic abundance model. Biometrics 58:531-539.

Chao A. and Lee S.M. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association 87:210-217.

Chazdon R.L., Colwell R.K., Denslow J.S. and Guariguata M.R. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rainforests of NE Costa Rica. Pages 185-309 *in* Dallmeier F. and Comiskey J.A. (eds.). *Forest biodiversity research, monitoring and modeling: conceptual background and old world case studies*. Parthenon Publishing Group, Paris, France.

Chiarucci A., Enright N.J., Perry G.L.W. and Miller B.P. 2003. Performance of nonparametric species richness estimators in a high diversity plant community. Diversity and Distributions 9:283-295.

Chiarucci A., Maccherini S. and De Dominicis V. 2001. Evaluation and monitoring of the flora in a nature reserve by estimation methods. Biological Conservation 101:305-314.

Coddington J.A., Young L.H. and Coyle F.A. 1996. Estimating spider species richness in a southern Appalachian cove hardwood forest. The Journal of Arachnology 24:111-128.

Colwell R.K. 2006. *EstimateS* Statistical estimation of species richness and shared species from samples. Version 8. Persistent URL <purl.oclc.org/estimates>. Published at: http://viceroy.eeb.uconn.edu/EstimateS.

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

Darroch J.N. 1958. The multiple-recapture census, I. Estimation of a closed population. Biometrika 45:343-359.

Darroch J.N. and Ratcliff D. 1980. A note on capture-recapture estimation. Biometrics 36:149-153.

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

Efron B. and Thisted R. 1976. Estimating the number of unseen species: how many words did Shakespeare know? Biometrics 63:435-447.

Esty W.W. 1985. Estimation of the number of classes in a population and the coverage of a sample. Mathematical Scientist 10:41-50.

Fisher R.A., Corbet A.S. and Williams C.B. 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.

Foggo A., Attrill M.J., Frost M.T. and Rowden A.A. 2003. Estimating marine species richness: an evaluation of six extrapolative techniques. Marine Ecology Progress Series 248:15-26.

Gleason H.A. 1922. On the relation between species and area. Ecology 3:158-162.

Goodman L.A. 1949. On the estimation of the number of classes in a population. Annals of Mathematical Statistics 20:572-579.

Gotelli N.J., Anderson M.J., Arita H.T., Chao A., Colwell R.K., Connolly S.R., Currie D.J., Dunn R.R., Graves G.R., Green J.L., Grytnes J.A., Jiang Y.H., Jetz W., Lyons S.K., McCain C.M., Magurran A.E., Rahbek C., Rangel T.F.L.V.B., Soberón J., Webb C.O. and Willig M.R. 2009. Patterns and causes of species richness: a general simulation model for macroecology. Ecology Letters 12:873-886.

Gotelli N.J. and Colwell R.K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4:379-391.

Hellmann J.J. and Fowler G.W. 1999. Bias, precision, and accuracy of four measures of species richness. Ecological Applications 9:824-834.

Heltshe J.F. and Forrester N.E. 1983. Estimating species richness using the jackknife procedure. Biometrics 39:1-11.

Herzog S.K., Kessler M. and Cahill T.M. 2002. Estimating species richness of tropical bird communities from rapid assessment data. Auk 119:749-769.

Holdridge L.R., Grenke W.C., Hatheway W.H., Liang T. and Tosi J.A. 1971. *Forest environments in tropical life zones.* Pergamon Press, Oxford, UK.

Hortal J., Borges, P.A.V. and Gaspar C. 2006. Evaluating the performance of species richness estimators: sensitivity to sample grain size. Journal of Animal Ecology 75:274-287.

Krebs C.J. 1994. *Ecology*. 4th ed. Addison-Wesley Educational Publishers, Inc., USA.

Lauga J. and Joachim J. 1987. L'échantillonnage des populations d'oiseaux par la méthode des E. F. P.: interêt d'une etude mathématique de la courbe de richesse cumulée. Acta Oecologica Oecologia Generalis 8:117-124.

Lee S.M. and Chao A. 1994. Estimating population size via sample coverage for closed capture-recapture models. Biometrics 50:88-97.

Legendre P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659-1673.

Longino J., Coddington J. and Colwell R.K. 2002. The ant fauna of a tropical rain forest: estimating species richness three different ways. Ecology 83:689-702.

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

Magnussen S., Pélissier R., He F. and Ramesh B.R. 2006. An assessment of sample-based estimators of tree species richness in two wet tropical forest compartments in Panama and India. International Forestry Review 8:417-431.

Magurran A.E. 1988. *Ecological Diversity and its Measurement*. Chapman & Hall, London, UK.

Magurran A.E. and Henderson P.A. 2003. Explaining the excess of rare species in natural species abundance distributions. Nature 422:714-716.

McIntosh R.P. 1967. An index of diversity and the relation of certain concepts to diversity. Ecology 48:392-404.

Melo A.S. and Froehlich C.G. 2001. Evaluation of methods for estimating macroinvertebrate species richness using individual stones in tropical streams. Freshwater Biology 46:711-721.

Michaelis M. and Menten M.L. 1913. Der kinetic der invertinwirkung. Biochemische Zeitschrift 49:333-369.

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

Ord J.K. and Whitmore G.A. 1986. The Poisson-Inverse Gaussian distribution as a model for species abundance. Communications in Statistics, Part A – Theory and Methods 15:853-871.

Otis D.L., Burnham K.P., White G.C. and Anderson D.R. 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs 62:1-135.

Palmer M.W. 1990. The estimation of species richness by extrapolation. Ecology 71:1195-1198.

Palmer M.W. 1991. Estimating species richness: the second-order jackknife reconsidered. Ecology 72:1512-1513.

Palmer M.W. 1995. How should one count species? Natural Areas Journal 15:124-135.

Petersen F.T. and Meier R. 2003. Testing species-richness estimation methods on single-sample collection data using the Danish Diptera. Biodiversity and Conservation 12:667-686.

Petersen F.T., Meier R. and Larsen M.N. 2003. Testing species richness estimation methods using museum label data on the Danish Asilidae. Biodiversity and Conservation 12:687-701.

Pledger S. 2000. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. Biometrics 56:434-442.

Pollock K.H. and Otto M.C. 1983. Robust estimation of population-size in closed animal populations from capture recapture experiments. Biometrics 39:1035-1049.

Poulin R. 1998. Comparison of three estimators of species richness in parasite component communities. Journal of Parasitology 84:485-490.

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

Rapport D.J., Regier H.A. and Hutchinson T.C. 1985. Ecosystem behavior under stress. The American Naturalist 125:617-640.

Robbins C.S., Bystrak D. and Geissler P.H. 1986. *The Breeding Bird Survey: its first fifteen years, 1965-1979*. U.S. Fish and Wildlife Service Resource Publication 157. Washington D.C., USA.

Rosenzweig M.L., Turner W.R., Cox J.G. and Ricketts T.H. 2003. Estimating diversity in unsampled habitats of a biogeographical province. Conservation Biology 17:864-874.

Schmit J.P., Murphy J.F. and Mueller G.M. 1999. Macrofungal diversity of a temperate oak forest: a test of species richness estimators. Canadian Journal of Botany 77:1014-1027.

Sharot T. 1976. The generalized jackknife: finite samples and sub-sample sizes. Journal American Statistical Association 71:451-454.

Shlosser A. 1981. On estimation of the size of the dictionary of a long text on the basis of

a sample. Engineering Cybernetics 19:97-102.

Sichel H.S. 1986. Parameter estimation for a word frequency distribution based on occupancy theory. Communications in Statistics, Part A – Theory and Methods 15:935-949.

Smith E.P. and van Belle G. 1984. Nonparametric estimation of species richness. Biometrics 40:119-129.

Stout J. and Vandermeer J. 1975. Comparison of species richness for stream-inhabiting insects in tropical and mid-latitude streams. The American Naturalist 109:263-280.

Sugihara G. 1980. Minimal community structure: An explanation of species abundance patterns. The American Naturalist 116:770-787.

Thiollay J.M. 2006. Large bird declines with increasing human pressure in savanna woodlands (Burkina Faso). Biodiversity and Conservation 15:2085-2108.

Thompson S.K. 1992. *Sampling*. Wiley, New York, USA.

Tokeshi M. 1990. Niche apportionment or random assortment: species abundance patterns revisited. Journal of Animal Ecology 59:1129-1146.

Tokeshi M. 1993. Species abundance patterns and community structure. Advances in Ecological Research 24:111-186.

Tokeshi M. 1996. Power fraction: a new explanation of relative abundance patterns in species-rich assemblages. Oikos 75:543-550.

Turner W., Leitner W. and Rosenzweig M. 2003. WS2M. User's manual. Published at: http://eebweb.arizona.edu/diversity/.

Ulrich W. 1999. Estimating species numbers by extrapolation I: comparing the

performance of various estimators using large model assemblages. Polish Journal of Ecology 47:271-291.

Ulrich W. and Ollik, M. 2005. Limits to the estimation of species richness: the use of relative abundance distributions. Diversity and Distributions 11:265-273.

Wagner H.H. and Wildi O. 2002. Realistic simulation of the effects of abundance distribution and spatial heterogeneity on non-parametric estimators of species richness. Ecoscience 9:241-250.

Walther B.A. and Martin J.L. 2001. Species richness estimation of bird communities: how to control for sampling effort? Ibis 143:413-419.

Walther B.A. and Moore J.L. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography 28:815-829.

Walther B.A. and Morand S. 1998. Comparative performance of species richness estimation methods. Parasitology 116:395-405.

Williams C.B. 1964. *Patterns in the balance of nature*. Academic Press, London, UK.

Wilson J.B., Gitay H., Steel J.B. and King W.M. 1998. Relative abundance distributions in plant communities: Effects of species richness and of spatial scale. Journal of Vegetation Science 9:213-220.

Zelmer D.A. and Esch G.W. 1999. Robust estimation of parasite component community richness. Journal of Parasitology 85:592-594.

CHAPTER 2

SIMASSEM: A PROGRAM FOR SIMULATING SPECIES ASSEMBLAGES AND ESTIMATING SPECIES RICHNESS

INTRODUCTION

Species richness, the number of unique species in a defined area, is the most commonly used measure of biological diversity (Gaston 1996, Moreno et al. 2006). Species richness (SR) can be used to delineate protected areas, monitor biological systems, and investigate environmental relationships. Surveys rarely encounter all of the species in an area; therefore, numerous estimators have been proposed to reduce the negative bias of raw counts.

Species richness estimators

Three categories are regularly used to classify SR estimators (Colwell and Coddington 1994). The first category includes methods used to extrapolate from a species accumulation curve, including species-area curves, to an asymptote, i.e., an estimate of SR. The Michaelis-Menten equation (Michaelis and Menten 1913), negative exponential model (Holdridge et al. 1971), and power model (Arrhenius 1921, Tjørve 2009) are the most common. Species accumulation curves are constructed by plotting the number of species encountered against effort, usually expressed in units of time, area, or individuals. Accumulation curves are often smoothed by averaging over repeated randomizations of the survey order at each effort level. This has been found to improve estimates based on extrapolation (Chazdon et al 1998). Similarly, by extrapolating from the richness of a sampled area to either a larger or, though no

longer an accumulation, a spatially independent area, estimates can be generated from the strong positive relationship between area and SR.

Two additional estimator types are categorized by the assumptions upon which they are based. The parametric category is comprised of estimators that make assumptions about the underlying species abundance distribution or species detection probabilities (p). One type of parametric estimator uses the shape of a fitted species abundance distribution. For example, a log-series distribution can be used to predict either how many new species would be found with an additional number of sampled individuals or, if total abundance is known or estimable, to estimate SR (Colwell and Coddington 1994). The area under a fitted log-normal distribution can also be used as an estimate of SR (Magurran 2004). This approach is difficult to implement because of the need to define both a parameter distribution and the discrete abundance classes to which a continuous distribution is fit (Magurran 2004). Parametric estimators also include those based on the assumption that p is constant across species. A third category includes nonparametric estimators, defined as those that are not based on assumptions about the underlying distributions of parameters. Many of the nonparametric estimators used for SR were originally derived to estimate the size, i.e., number of individuals, of a closed population.

Comparisons of specific SR estimators have not found a single best estimator. However, more general comparisons of the three categories listed above have indicated that the nonparametric SR estimators often perform best (see Table 1 in Cao et al. 2004, Table 3 in Walther and Moore 2005, and Table 1.1 in Chapter 1). Nonparametric estimators are therefore the focus of this project and all of the included estimators are further discussed in the following sections.

Factors affecting estimation of species richness

The performance of nonparametric estimators can be affected by species- and assemblage-level attributes as well as by the parameters of a survey design, collectively referred to as factors throughout this dissertation (Keating and Quinn 1998, Brose et al. 2003, Magnussen et al. 2006). Otis et al. (1978) suggested that estimator performance could depend on abundance patterns. Several studies have indicated that the nonparametric estimators are more negatively biased with uneven species abundance distributions than with even distributions (Heltshe and Forrester 1983, Lee and Chao 1994, Wagner and Wildi 2002, O'Dea et al. 2006). One assumption of the closed population nonparametric estimators, translated for species data, holds that species are equally detectable across space. This assumption is often violated and modeling is potentially complicated because species are usually spatially aggregated (see Schmit et al. 1999, Walther and Martin 2001). Other factors found to affect estimator performance include the number of species (Keating and Quinn 1998, Poulin 1998), species abundance, i.e., density (Baltanás 1992, Walther and Morand 1998), and *p* (Boulinear et al. 1998, Ashbridge and Goudie 2000).

Raw sample data and consequently, SR estimates, are also affected by survey design parameters including effort (Burnham and Overton 1979, Brose et al. 2003) and the spatial configuration of surveys, e.g., linear transects versus random quadrats. A random design is unbiased and therefore preferable; however, survey locations are often selected based on accessibility and on the results of previous surveys (see Beck and Kitching 2007). Each of these factors can affect sample coverage (*sc*), which is the proportion of a species pool represented in a sample and the single most important factor with respect to estimator performance (Baltanás 1992, Brose et al. 2003). Unfortunately, one needs to know the true number of species to

calculate *sc* and, if this information were available, estimation would be unnecessary. It is therefore important to understand how the aforementioned factors affect performance.

It is difficult to evaluate SR estimators across a wide range of factors in a field setting because of temporal, financial, and logistical constraints and uncertainty about species- and assemblage-level parameters. Despite making numerous simplifications, simulations are advantageous because they can be systematically varied and randomly surveyed, and most important, the true number of species is known. Most, if not all, of the programs currently available for estimating SR, e.g., EstimateS (Collwell 2006), SPADE (Chao and Shen 2009), SPECRICH (Hines 1996), and Ws2m (Turner et al. 2003), focus on processing existing encounter history data and include little or no simulation functionality. My objective therefore was to develop a program that would allow a user to: 1) simulate species assemblages with specified parameters, 2) survey the assemblage, and 3) evaluate the performance of SR estimators.

PROGRAM SIMASSEM

SimAssem is application software that I developed in Visual Basic 6.0 for estimating SR from a survey history of encountered species. The program requires a 32-bit Microsoft Windows operating system and, possibly, an Intel or Intel-compatible processor. Via a graphical user interface (Fig. 2.1) and an internal dialogue with R software (R Development Core Team 2009), SimAssem can process both existing encounter history data (see Table 2.1 for an example) and encounter data from surveys of assemblages simulated with user-specified parameters. When an assemblage is simulated, a user specifies the SR, total abundance, species abundance distribution, degree of spatial aggregation, and *p*. There are also two survey designs, random and linear transect, by which a specified number of surveys can be conducted. SimAssem includes

several estimators rooted in population estimation and two that were specifically derived for the estimations of SR. Input data for these estimators, as well as for most of the other estimators in SimAssem (see below) are the frequencies of encounter, at the resolution of either individuals (abundance data) or surveys (incidence data). Randomization is performed by the Mersenne twister pseudorandom number generator (Matsumoto and Nishimura 1998). Both the program and source code are available for download at http://warner.colostate.edu/~kenw/program_download/SimAssem.html.

Simulating an assemblage

Species richness, species abundance distribution, and total abundance

Species richness and the distribution of species abundances (see Fig. 2.2) are fundamental attributes of an assemblage. Two established theories about species abundance are: 1) abundances are generally unequal amongst species and 2) most species are relatively rare (Fisher et al. 1943). The statistically-based geometric-series (Motomura 1932), log-normal (Preston 1948), and log-series distributions (Fisher et al. 1943) have been successfully fit to numerous biological datasets. However, the representation of species abundance distributions with purely mathematical models has been criticized for lacking explanations of the patterns. Some of the earliest alternatives that focused instead on process were proposed by MacArthur (1957), including the broken stick and particulate-niche models. More recent work with species abundance distributions has continued to emphasize the methodological steps required to create a distribution and, by way of analogy, the ecological processes that result in real abundance distributions. These niche-based models are assumed to approximate the interactions and subsequent patterns of small groups of taxonomically related species, i.e., species vying for the same resources. A basic premise holds that niche apportionment can be modeled by a stick

being broken, where the units of the stick represent individuals. Many of the niche-based models, presented below, were developed by Tokeshi (1990, 1993, 1996).

In SimAssem, assemblages are simulated by specifying a number of species (*S*), total number of individuals across all species, i.e., total abundance (*N*), and a distribution to which species abundances conform (Fig. 2.2). SimAssem will run only when *N* equals or exceeds *S*. Available species abundance distributions include the geometric-series, log-series, and log-normal models and several niche-based models. The niche-based models include the MacArthur fraction (a sequential representation of MacArthur's broken-stick model), dominance-decay, dominance-preemption, particulate-niche, power-fraction, random-assortment, random-fraction, and sequential 75% models (Tokeshi 1990, 1993, 1996). SimAssem also includes an option to divide individuals based on the zero-sum multinomial model (Hubbell 2001). Worked examples for all models are given in Appendix I. Whether *N* and *S* are fixed or stochastic depends on the selected model (see Table 2.2 for an example of the species abundances from a single iteration of each model and Fig. 2.2 for species rank by relative abundance averaged over repeated replications).

Niche-based models

Each of the niche-based models begins with a hypothetical line or segment with *N* units, which is broken into segments based on rules of the specific model. Unless otherwise specified, each new segment is immediately equated with the abundance of a species (n_i; where abundance represents the degree of niche apportionment). The terms (*number of*) *individuals* and *abundance* are used interchangeably throughout this chapter. Only segments larger than one are, of course, breakable.

MacArthur's (1957) broken-stick model is defined as the simultaneous breakage of a line of length *N* into *S* segments. This is done by splitting the line at *S* - 1 randomly selected points; SimAssem includes the modified MacArthur fraction model that instead uses *S* - 1 sequential breaks (Fig. 2.2; Tokeshi 1990). The probability of a segment being broken is positively related to its length. This occurs because segments retain the original number sequence throughout the procedure and because the random breakage points are always bounded by 1 and *N*. Thus, a segment representing 100 individuals will be selected 10x more often by a random number generator than a segment representing 10 individuals (Appendix I, ex. 1). This model could represent an assemblage comprised of equally competitive species vying for niche space (Tokeshi 1993).

The dominance-preemption and dominance-decay models use inverse procedures by always breaking either the smallest or largest segment, respectively (Fig. 2.2). The dominance-preemption model begins by randomly allocating between 50-100% of the full line to the first species. Then, breakage points are randomly selected such that >50% of the remaining units are allocated to the next new species. Frequently, this algorithm will result in fewer than the specified number of species, *S*. This would occur, for example, if *S* = 10, 75% of *N* is allocated to the first species, all except one individual are allocated to the second species and, therefore, one individual is allocated to the third species. Units that remain following an *S* - 1 break are allocated to the last species (Appendix I, ex. 2). By contrast, the dominance-decay model tends towards more equitable abundances because the largest segment is randomly broken in each step (Appendix I, ex. 3). The dominance-preemption model could be appropriate where colonizing species always vie for the smallest remaining niche, whereas the dominance-decay model could be applied to situations where jostling occurs in the niche of the most abundant species.

The geometric-series model in a niche-based framework (Tokeshi 1990) is similar to the dominance-preemption model in that new species are always generated from a segment not yet allocated to a species. They differ in that the geometric-series model always breaks the same user-specified proportion (k) from the segment not yet allocated to a species whereas the dominance-preemption breaks a randomly selected proportion (Fig. 2.2). Some values of k will result in assemblages with fewer individuals than the specified N or a true number of species (S_{true}) that is less than S (Appendix I, ex. 4 and 5). Such distributions can result when an unsaturated area, possibly species poor or in the early stages of succession as a result of harsh conditions, is settled at regular intervals (Whittaker 1965, 1972, Magurran 2004).

The particulate-niche model randomly allocates each unit to a species, with an equal probability for each species in S (Fig. 2.2; MacArthur 1957, Tokeshi 1993). The probability that $S_{true} < S$ in the resulting assemblage is always greater than zero. Furthermore, the probability that no units will be allocated to a species in S is positively related to the ratio S:N. In other words, the probability that $S_{true} < S$ increases as S approaches N (Appendix I, ex. 6). Similar to the MacArthur fraction model, the particulate-niche model could be appropriate for assemblages with equitable species, where competition ceases to occur in a filled niche (Tokeshi 1993).

The power fraction model uses weighted segments such that a positive relationship exists between segment length and the probability of selection (Tokeshi 1996). As with other models, the first step involves randomly breaking the complete line into two segments. For each subsequent break, the power equation (n_i^k), where n_i is the abundance of species i and k is a user-specified power parameter [0,1], is calculated for each segment and used in $\alpha = 1/\sum n_i^k$. The selection probabilities of segments are weighted by αn_i^k, where α is constant across all species. Then, a random uniform variate (RUV; [0-1]) is compared to the cumulative probabilities of αn_i^k

and the segment within which the variate falls is randomly broken (Fig. 2.2; Appendix I, ex. 7). Unlike other niche-based models, the power fraction can model species-rich assemblages (Magurran 2004).

The underlying assumption of the random-assortment model is that species abundances are independent. In other words, niche apportionment and species abundance are only weakly related (Magurran 2004). SimAssem includes the sequential formulation of Tokeshi (1993). The algorithm begins by allocating a random proportion of N to the first species. The segment not yet allocated to a species is randomly split as in the dominance-preemption model; however, the random-assortment model allocates a random fraction (0-1) of this segment to the next species whereas the dominance-preemption model randomly allocates >0.5. Thus, it is possible for the random-assortment model to result in assemblages with $S_{true} < S$. As with the dominance-preemption model, units that remain following the S - 1 break are allocated to the last species (Fig. 2.2). Tokeshi (1993) noted the similarity of the random-assortment model to a neutral model (Caswell 1976) which makes a similar assumption of species independence (Appendix I, ex. 8). A rapidly changing assemblage, possibly a result of environmental changes that release the bounds of competition, can often be fit by this distribution (Magurran 2004).

The random-fraction model includes two randomization steps. First, one of the segments is randomly selected and second, the selected segment is randomly broken. The model begins by randomly breaking the complete line into two segments. Following the initial break, randomly selected segments are randomly broken until there are S segments (Fig. 2.2; Appendix I, ex. 9). An ecological analogy sees new arrivals securing a random proportion of the niche of a relatively established species. Also, the random-fraction can potentially model speciation events (Tokeshi 1999).

The algorithm for the sequential 75% model is similar to the random-fraction with one major difference. Instead of the randomly selected segment being randomly broken, segments are always split with 75% of the individuals allocated to a new species (Fig. 2.2; Appendix I, ex. 10; Sugihara 1980). There are cases when a 75:25 split is not possible. In the event that the selected segment contains two individuals, one individual is allocated to each species. Selected segments with three individuals are broken such that two individuals are allocated to the new species and one individual is allocated to the selected species.

The zero-sum multinomial model is unique to the set included in SimAssem because S is not specified (Hubbell 2001, pgs. 289-290). Instead, θ is specified for the species generator, $\theta/(\theta + j - 1)$, where j iterates from 1 to N, and thus S_{true} can vary. At each iteration j, a RUV is compared to the value of the species generator. When the RUV exceeds the species generator, the individual is allocated to an existing species; otherwise, the individual begins a new species. In the former case, the fractional abundance of each populated species is calculated as $n_i/(j - 1)$, where n_i is the abundance of species i, and a cumulative abundance distribution is constructed from species abundances in the order in which species were generated. The j^{th} individual is then allocated to the species that corresponds to the interval into which a new RUV falls (Fig. 2.2; Appendix I ex. 11). The zero-sum multinomial model is thus a neutral model that can account for the random fission model of speciation in a set of local communities (Magurran 2004).

Statistically-based models

SimAssem creates log-normal abundance distributions by summing S random log-normal variates and normalizing them to one. The normalized variates are each multiplied by N, giving S abundances. The number of species simulated by this procedure always equals S, however, the number of individuals that are simulated could differ from N by several individuals (Fig. 2.2;

Appendix I, ex. 12). Log-normal variates in SimAssem are generated from the *rlnorm* function ($\mu = 0$, $\sigma = 1$) in the R statistical software. Large datasets comprised of relatively unrelated species, upon which many factors act, have been successfully fit by this distribution (May 1975, Tokeshi 1993).

Log-series abundance distributions are generated with the procedures described by Magurran (2004). SimAssem requires the user-specified parameters S and x, where x is used in $\alpha = N(1 - x)/x$. The number of species with i individuals, where i iterates from 1 to N, is computed with $s_i = \alpha x^i/i$; any fractional portion is added to s_i in the next iteration. Iterations continue until all species are populated with ≥1 individuals and this comprises the *log-series (original)* algorithm in SimAssem (Appendix I, ex. 13). A second log-series algorithm, *log-series (modified)*, assigns to the final species any individuals that remain after populating the penultimate species. This algorithm generates the specified N, but it can slightly alter the relative abundances of the least abundant species (Fig. 2.2; Appendix I, ex. 14). Assemblages receiving newly arriving species at random intervals will eventually result in a log-series distribution (Boswell and Patil 1971, May 1975).

Locating individuals on the landscape: spatial configuration

Individuals are distributed across a square landscape and spatial configuration options include random, hyper-dispersion, assemblage-wide aggregation, and several species-specific aggregation patterns. In the *random* option, one RUV is assigned to the x-coordinate and one to the y-coordinate of each individual (Fig. 2.3, a panels).

Hyper-dispersion is a species-specific option that assumes that all individuals of a species exhibit equal territoriality, which results in individuals being more evenly spaced than expected by chance (Fig. 2.3, b panels). More specifically, a linear species territory size (l) is defined as l_i

= 1 / $\sqrt{n_i}$, where n_i is the abundance of species *i*. A grid of square territories, l_i to a side and snapped to the lower-left corner of the landscape, is then overlaid on the landscape. In each territory, randomly selected horizontal and vertical distances (0-l_i) are added to the coordinates of the lower left corner and an individual is placed at that location. When the summed length of adjacent territories does not exactly equal one, i.e., the linear extent of the landscape, territories on the right and top will overlap landscape boundaries. Using the placement procedures described for other territories, the probability of an individual being placed in these territories equals the proportion of the territory that falls within the landscape. After one individual is randomly distributed in each territory, those falling outside the landscape are randomly redistributed across the entire landscape. This last step increases the probability that an individual will occur in an overlapping territory.

In another option, *clustered (assemblage-wide)*, the same set of clusters is used for all individuals, independent of species identity (Fig. 2.3, c panels). Patterns generated by this algorithm resemble landscapes with biodiversity hotspots, locations with a relatively large number of overlapping species. The algorithm selects a random number of origins, i.e., cluster centers, from one to a user-specified number, and randomly places the origins on the landscape. Each individual is then distributed a random distance and direction from a randomly selected origin, given that two criteria are met. First, the location must fall on the landscape and second, a distance decay formula, $1 - (1 - \omega^{DistanceToOrigin})^{\tau}$, with user-specified parameters ω and τ, must be ≥RUV (Fig. 2.4). A distance decay formula is used here and in algorithms described below to create a negative relationship between the probability of dispersal and the distance from the origin.

The *clustered (species-specific)* option places individuals in much the same way as *clustered (assemblage-wide)*; however, a random number of origins, from one to a user-specified number, is selected for each species instead of for the assemblage as a whole (Fig. 2.3, d panels). Resulting patterns resemble landscapes where most locations are suitable to one or more species. Individuals are placed with the same steps outlined for *clustered (assemblage-wide)*, which involves randomly selecting an origin, a distance, and a direction and comparing the potential location against both the boundaries of the landscape and the distance decay formula. In both options, the probability of some origins never being selected is >0.

There are four additional species-specific aggregation options based on a user-specified Fidelity parameter (F) which influences the number of clusters randomly generated for each species. These algorithms can create patterns ranging from nearly random to patchy and they can be categorized based on the point at which clusters are generated. Three of these options belong to a class where the first cluster begins with a randomly placed individual. Then, one RUV is drawn for each remaining individual and a RUV > F results in the individual being randomly located on the landscape, beginning a new cluster. The expected number of clusters for a species is therefore proportional to its abundance. When a RUV ≤ F, the associated individual is randomly assigned to an existing cluster.

The three options in this first class differ in how individuals are placed within a selected cluster. In the *aggregated (individuals)* option, distances [0-1] are randomly and repeatedly generated and measured from a randomly selected individual until: 1) RUV ≤ $[(1 - D)^{DistanceToIndividual}]$, where D is a user-specified parameter with a positive relationship to fidelity and 2) the resulting location is on the landscape (Fig. 2.3, e panels; see Table 2.3 for dispersal probability examples). The *aggregated (centers)* option differs only in that individuals are

always randomly distributed relative to the cluster origin, which in this class is the first individual allocated to that cluster (Fig. 2.3, f panels). The third option, *aggregated (individuals max distance)*, uses the same algorithm as *aggregated (individuals)* except that a randomly generated distance, i.e., a RUV, is generated until the RUV ≤ (1 – *D*), i.e., less than or equal to the largest possible dispersal distance (Fig. 2.3, g panels). Thus, large *F* values result in most individuals being assigned to the oldest clusters.

In a related fourth algorithm, *aggregated (centers, equal probability)*, a random number of clusters is selected for each species before the distribution of any individuals (Fig. 2.3, h panels). This modification results in more equitable numbers of individuals per cluster by making all clusters available to all individuals. One RUV is generated for each individual of a species and one origin is added for each RUV > *F*. Then, origins are randomly placed on the landscape. Each individual is placed a random distance from a randomly selected origin when RUV ≤ $[(1 - D)^{DistanceToOrigin}]$ and the resulting *x*- and *y*-coordinates are each [0,1].

Creating sample data: species detection probabilities and survey design

Before *p*'s are assigned, species are grouped into thirds based on abundance such that one group is comprised of the least abundant species. A randomly selected group is increased by one for each species that remains when S_{true} is not a factor of three. Within each group, species-specific detection probabilities can be randomly drawn from a beta distribution with specified α and β parameters (R function *rbeta*). Beta distributions are characterized by an expected value (mean),

$$E(X) = \frac{\alpha}{\alpha + \beta},$$

and variance,

$$\mathrm{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2 + (\alpha+\beta+1)},$$

where X is a random beta variate. Additionally, p's can be fixed for each abundance group.

Each simulated landscape is partitioned by a 100 x 100 grid for the purpose of conducting surveys. Parameters include the *number of cells* (1-10000) *to survey* (*t*) and *survey design*, i.e., spatial configuration of surveyed grid cells. Cells are surveyed without replacement.

SimAssem includes two survey designs. Surveyed grid cells can be randomly selected (*random*) or added to randomly oriented, horizontal or vertical linear transects that are each one grid cell wide (*linear transect*). The linear transect option requires a *minimum number of transects* (*m*) across which *t* is divided. Due to landscape dimensions, maximum transect length is the smaller of 100 or t/m. When transect length is truncated to 100 or when t is not a factor of m, additional transects are added until the number of surveyed cells equals t. One RUV is drawn for every individual in a surveyed cell and an individual is encountered when the RUV $\leq p$.

Estimating species richness

SimAssem includes numerous SR and variance estimators (Table 2.4). Some analytically derived variance estimators are yet to be included in SimAssem, so SimAssem will format data for other published programs that provide the variance estimate. Most of the included estimators use data converted to a frequency table, either: 1) the number of species with exactly i encountered individuals (where $i = 1, 2, \ldots, n$, and n is the number of individuals encountered across all surveys) or 2) the number of species that were encountered in exactly j surveys ($j = 1, 2, \ldots, t$).

Most of the estimators assume that the number of species rarely encountered, e.g., once or twice, is the best information from which to estimate the number of species not encountered. Estimators often involve adding some function of the number of rare species to S_{obs} (Table 2.4).

In addition to numerous nonparametric estimators, SimAssem includes S_{obs}. The nonparametric estimator set includes five based on abundance data and 13 based on incidence data. The abundance-based estimators use a function of the number of species encountered by an exact number of individuals, one or two for an abundance-based estimator (Chao1, Chao 1984) and a variable number, usually ≤10, for the abundance-based coverage estimator (ACE, Chao and Lee 1992).

The incidence-based estimators use a function of the number of species encountered in an exact number of surveys, one or two surveys for an incidence-based estimator (Chao2, Chao 1987), usually ≤10 for the incidence-based coverage estimator (ICE, Lee and Chao 1994), and the number equal to the order of each of the jackknife estimators (Jack#, Burnham and Overton 1978). Two versions of a bootstrap estimator are included. One applies the bootstrap estimator to the original set of surveys and is the more common of the two (Boot; Smith and van Belle 1984). A less common procedure averages over a user-specified number of survey randomizations. In each randomization, the formula is applied to *t* surveys randomly drawn with replacement from the original set (Boot-B). The incidence-based estimators could be more appropriate than the abundance-based estimators when the individuals of an identifiable species are difficult to distinguish, as they are for some floral species, e.g., a single individual with numerous stems.

Three of the estimators included in SimAssem are relatively new. The CY-1 and CY-2 estimators are based on the similarity of two replicate subsets as measured by Jaccard's coefficient (Cao et al. 2001, Cao et al. 2004, respectively). As programmed, CY-1 requires encounters in ≥2 surveys and CY-2 in ≥10 surveys; the details of these requirements are given below. Both CY-1 and CY-2 additionally require that individuals from ≥1 species have been

encountered in ≥2 surveys. Some computation errors are avoided by using only the surveys with encounters (Y. Cao, personal communication). Surveys are randomly drawn without replacement from this potentially reduced set until two equally sized subsets are created, SR is estimated, and estimates are averaged over a user-specified number of replications. CY-1 and CY-2 use differently sized subsets, as explained below.

The algorithm for CY-1 randomly splits surveys into the largest possible subsets. When the number of surveys with encounters (q) is odd, the size of each subset is (q - 1)/2. In each iteration, the CY-1 estimate equals the average number of species in each subset ($\overline{SR}$) divided by Jaccard's coefficient (JC), where $JC = c/(a + b + c)$, and a and b are the numbers of species unique to each subset and c is the number of shared species. When $c = 0$, the CY-1 estimate is undefined because $JC = 0$, thus the requirement that ≥1 species must be encountered in ≥2 surveys.

The CY-2 algorithm divides surveys into several equally sized subsets. At each subset size, $\overline{SR}$ and JC are averaged over a user-specified number of iterations. Species richness is not estimated in each iteration; instead, a plot of average $\overline{SR}$ versus average JC, using all subset sizes, is fit with linear regression (R function *lm*), $\overline{SR}$ = Intercept + Slope(JC). The CY-2 estimate equals the intercept plus the slope because that is the point at which $JC = 1$ and, therefore, $\overline{SR} = S_{true}$. When $q \geq 20$, plots include 10 $\overline{SR}$ - JC pairs and subset sizes equal the integer portion of $0.1hq/2$, where $h = 1, 2, \ldots, 10$. When $10 \leq q < 19$, regression is based on five subset sizes with $h = 2, 4, \ldots, 10$.

A closed population estimator that allows heterogeneity in p (Otis et al. 1978) was derived in a maximum likelihood framework by grouping species with similar detection probabilities into "mixtures" (Pledger 2000). SimAssem includes the mixture estimator under

model M_h (heterogeneity in p) with two groups. Both the number of iterations used to maximize the likelihood (y) and the number for the expectation maximization procedure are user-specified (see Pledger 2000 for more details). Mixture estimates should be compared with those from the more powerful optimizer available in program MARK (White and Burnham 1999), particularly when the estimate equals either S_{obs}, $S_{obs} + y$ (i.e., an estimate that failed to converge after the specified number of iterations), or an unreasonably large value. The steps for generating a comparable model are listed in the first line of a file that can be exported for program MARK.

Where possible, I validated estimators against programs EstimateS (Colwell 2006) and SPADE (Chao and Shen 2009). The performance of the Mixture estimator in SimAssem depends on the quantity of encounter data and it can fail to converge with sparse datasets. The CY-1 and CY-2 estimators were validated with a Copper Creek dataset (Angermeier and Smogor 1995) and found to produce estimates that closely approximate (estimates are based on randomizations) those reported in Cao et al. (2001).

Additional Output

SimAssem reports several additional values including both the number of species (S_{true}) and total number of individuals that were simulated as well as the number of surveys with encounters and the total number of individuals encountered. When data are simulated, SimAssem also reports *sc*. Two diversity indices with the potential to further benefit biodiversity investigations are also given, Margalef's diversity index (Appendix II, 19.1; Clifford and Stephenson 1975) and Menhininck's index (Appendix II, 20.1; Whittaker 1977).

Biological surveys are generally expensive and often provide diminishing returns on investment (effort). Thus, an estimate of the number of new species that would be encountered for an additional level of effort could benefit survey design. SimAssem includes a pair of such

estimators derived by Chao et al. (2009). One estimates the number of additional individuals that would need to be encountered before a user-specified proportion of the bias-corrected Chao1 estimate would be detected (Appendix II, 21.1); therefore, this estimator requires abundance data. An incidence-based version estimates the number of additional surveys, e.g., quadrats, that would be needed to encounter a user-specified proportion of the bias-corrected Chao2 estimate (Appendix II, 22.1).

Assemblages can vary widely in patterns of abundance and spatial configuration. SimAssem reports the evenness of sample abundances and, when an assemblage is simulated, the evenness of the true distribution of species abundances (Appendix II, 18.1; Shannon and Weaver 1949).

Importing an encounter history file

SimAssem can import specifically formatted comma-, space-, and tab-delimited encounter history data saved as a plain text file. The first line is useful for documentation, as it is disregarded by SimAssem. Line two must contain two numbers, S_{obs} and t. Encounter history data, in a S_{obs} x t matrix, must begin on line three. So, each row will represent a different species and each column a different survey result, i.e., the number of individuals encountered for a particular species (see Table 2.1 for an example).

Export options

Several options exist for exporting data to a new file or appending data to an existing file. Estimates can be exported to a comma-delimited file, where the first line is a list of estimator names and the second line is a list of estimates. Encounter history data can be formatted for programs EstimateS, MARK, and SPADE with the details described in earlier sections. Individual-level data can be exported to a comma-delimited text file including (in the following

order): a numerical species identifier, *x*-coordinate, *y*-coordinate, grid cell where it was located, *p*, and whether the individual was detected (1), or undetected (0). Data for an accumulation curve are also exportable where, at each possible survey size (1-*t*), surveys are randomly drawn without replacement and estimates at each survey size are averaged over a user-specified number of replications. This is, therefore, a potentially time-consuming procedure.

Utility

Numerous SR estimators can be quickly and easily evaluated across a wide range of assemblages in SimAssem. Such investigations are difficult in the real world because of both sampling limitations and uncertainty in the true assemblage parameters. When one can approximate the parameters of an assemblage from which survey data were collected, similarly structured assemblages can be simulated, providing an initial evaluation of estimator performances. SimAssem also includes a couple of estimators aimed at survey design decisions that can potentially help with the allocation of surveying funds. Thus, SimAssem could benefit studies and applications in the conservation sciences. Additionally, SimAssem includes estimators that performed relatively well (see Chapter 3), but are not known to be available in any other program.

Table 2.1. An example encounter history file formatted for program SimAssem where rows represent species and columns represent surveys.

Encounter history data.[a]							
6[b]	8						
2[c]	0	3	0	2	0	0	1
1	1	0	0	5	3	5	1
0	2	0	0	0	0	0	0
1	0	0	1	0	1	0	0
0	0	0	0	1	0	0	0
3	0	3	0	0	0	0	0

[a]The first row must include some alphanumeric information.

[b]The second row must contain two values separated by a tab character including: 1) the number of species encountered and 2) the number of surveys conducted.

[c]An encounter history data matrix, i.e., one row per species and one column per survey, must begin on the third row and can consist of either abundance data, i.e., number of individuals encountered in each survey, or incidence data, i.e., indicator for whether or not any individuals were encountered.

Table 2.2. Example of species abundances generated by one run of each of the algorithms available in SimAssem, BS for broken-stick, DD for dominance-decay, DP for dominance-preemption, GS for geometric-series, LN for log-normal, LS for log-series (modified version), PF for power-fraction, PN for particulate-niche, RA for random-assortment, RF for random-fraction, S75 for sequential 75%, and ZS for zero-sum multinomial.

	Abundance model											
Species rank	**BS**	**DD**	**DP[a]**	**GS[b]**	**LN[a]**	**LS[b]**	**PF[b]**	**PN**	**RA[a]**	**RF**	**S75**	**ZS[a,c]**
1	2038	847	8120	2900	1304	3197	1876	427	5020	3281	3164	3314
2	926	815	1127	2059	1182	2515	1683	426	3442	2508	1406	2271
3	844	765	741	1462	908	1408	1361	418	623	1833	1406	780
4	606	708	8	1038	637	890	988	418	444	567	1055	746
5	604	659	4	737	573	594	873	416	170	358	1054	600
6	523	655		524	565	407	817	415	111	310	469	491
7	457	653		372	548	284	510	408	60	272	352	474
8	443	640		264	528	201	352	407	46	190	264	468
9	391	615		187	505	143	305	406	38	175	198	444
10	376	592		133	467	102	282	404	26	167	117	156
11	368	409		94	407	73	247	404	9	103	117	125
12	366	362		67	296	53	198	402	7	60	88	65
13	365	359		48	285	38	106	401	2	55	88	21
14	335	320		34	256	27	105	401	1	39	66	17
15	324	276		24	250	20	76	399	1	30	50	10
16	286	270		17	235	14	65	394		21	29	8
17	191	257		12	218	10	41	393		14	22	5
18	149	227		9	205	7	31	392		6	16	2
19	145	152		6	138	5	21	392		3	12	1
20	90	102		4	126	4	19	391		2	10	1
21	86	94		3	119	3	15	390		2	7	1
22	36	80		2	97	2	13	387		1	4	
23	29	67		2	70	1	10	385		1	3	
24	19	48		1	41	1	5	377		1	2	
25	3	28		1	40	1	1	347		1	1	
N	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**

[a]Model includes stochastic steps whereby some iterations can result in numbers of species and individuals that differ from the specified number (see text).

[b]The GS, LS, and PF distributions were parameterized with $k = 0.29$, $k = 0.05$, and $x = 0.9996907406$, respectively.

[c]The ZS algorithm was run with 10,000 individuals and $\theta = 3$ whereas the other algorithms were run with 25 species and 10,000 total individuals (N).

Table 2.3. Probability of dispersing a given distance based on the distance decay formula [(1 - D)^{LinearDistance}], where D is a user-specified parameter.

	Linear distance from a dispersal location (origin)						
D	**0.0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**1.0**
1.00	1.00[a]	0.00	0.00	0.00	0.00	0.00	0.00
0.90	1.00	0.79	0.50	0.32	0.20	0.13	0.10
0.70	1.00	0.89	0.70	0.55	0.43	0.34	0.30
0.50	1.00	0.93	0.81	0.71	0.62	0.54	0.50
0.30	1.00	0.96	0.90	0.84	0.78	0.73	0.70
0.10	1.00	0.99	0.97	0.95	0.93	0.91	0.90
0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

[a]Occasionally undefined, $0^0 = 1$ in this program.

Table 2.4. Species richness estimators and estimator abbreviations (Abbrev.) used in program SimAssem.

Estimator[1]	**Abbrev.**	**Equation no.**[2]	**Citation**
Abundance-based coverage (a)	ACE	3.1-3.5[+]	Chao and Lee 1992
		4.1-4.5[+]	
Bootstrap (i)	Boot	8.1[+]	Smith and van Belle 1984
Bootstrap; iterated (i)	Boot-B	8.3	Smith and van Belle 1984
Chao1 (a)	Chao1	1.1-1.2[+]	Chao 1984
Chao1 (bias-corrected; a)	Chao1BC	2.1[+]	Chao 2005
Chao2 (i)	Chao2	11.1-11.2[+]	Chao 1987
Chao2 (bias-corrected; i)	Chao2BC	12.1[+]	Chao 2005
CY-1 (i)	CY-1	9.1-9.3	Cao et al. 2001
CY-2 (i)	CY-2	10.1	Cao et al. 2004
Darroch-Ratcliff (a)	DR	6.1	Darroch and Ratcliff 1980
Horvitz-Thompson (a)	HT	7.1-7.4	Ashbridge and Goudie 2000
Incidence-based coverage (i)	ICE	13.1-13.4[+]	Lee and Chao 1994
		14.1-14.5[+]	
		15.1-15.5[+]	
1st-order jackknife (i)	Jack1	16.1[+]	Burnham and Overton 1978
2nd-order jackknife (i)	Jack2	16.2[+]	Burnham and Overton 1978
3rd-order jackknife (i)	Jack3	16.3[+]	Burnham and Overton 1978
4th-order jackknife (i)	Jack4	16.4[+]	Burnham and Overton 1978
5th-order jackknife (i)	Jack5	16.5[+]	Burnham and Overton 1978
Mixture-model (i)	Mixture	17.1	Pledger 2000
Observed species count	Sobs	Count	

[1]A description of the estimator where (a) indicates that the estimator uses sample abundance data, i.e., number of individuals, and (i) indicates that the estimator uses sample incidence data, i.e., presence/absence in surveys.
[2]The associated equation numbers in Appendix II.
[+] Indicates that a variance equation is included in Appendix II.

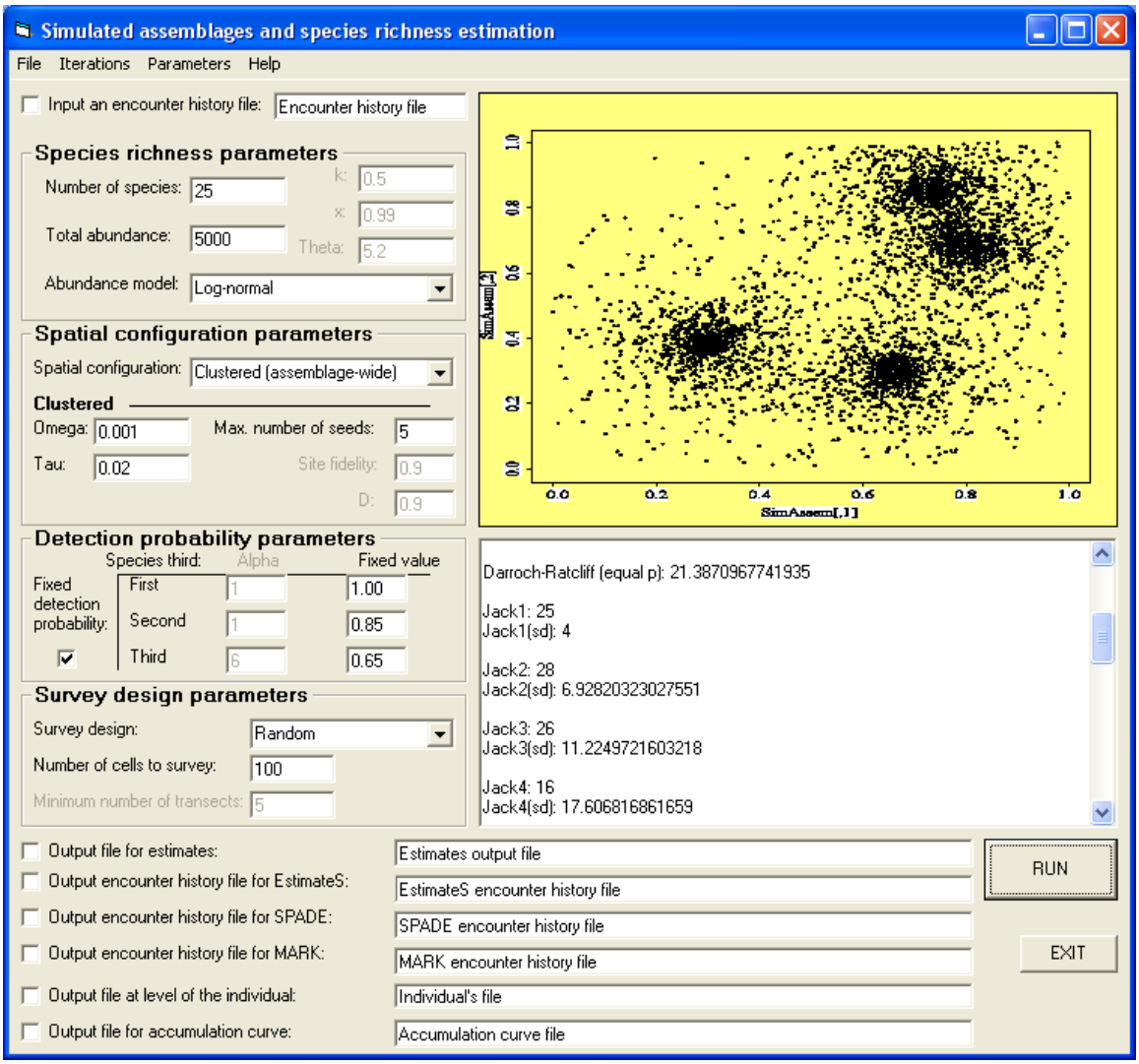


Fig. 2.1. Graphical user interface of program SimAssem.

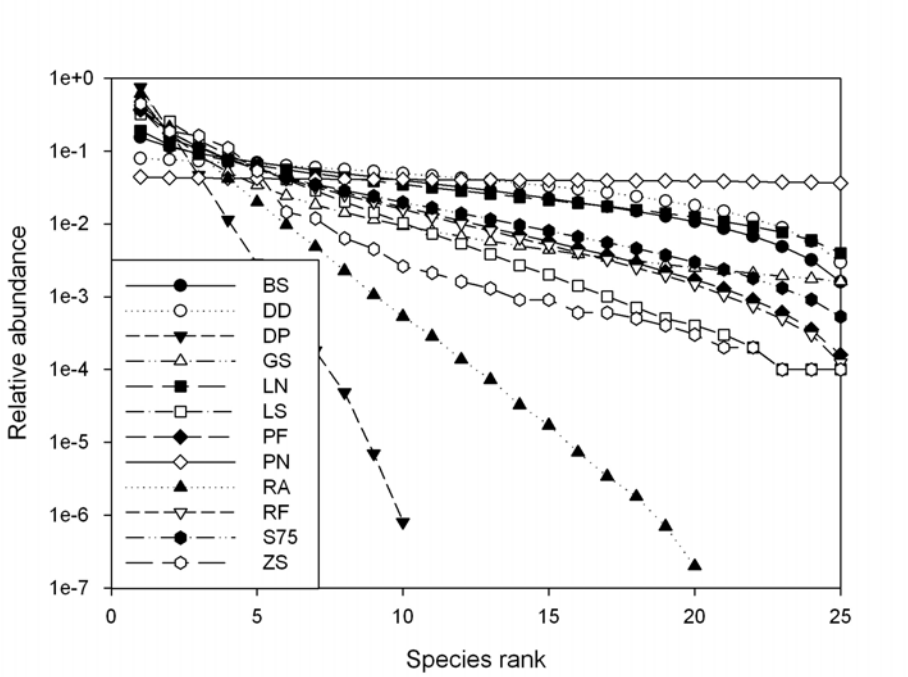
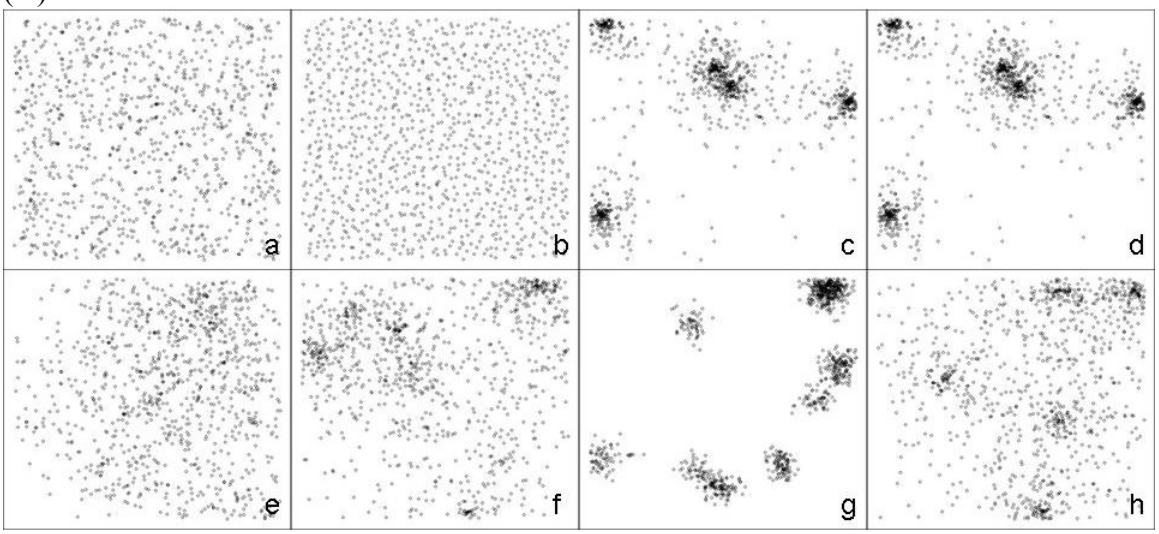


Fig. 2.2. Distribution of species abundances, represented as relative abundances, by species rank for each algorithm available in program SimAssem. Each graph point is the average value over 1,000 iterations with the initial parameters being 25 species and 10,000 total individuals, except for the zero-sum multinomial model as described below. Estimator abbreviations are given in Table 2.2. Due to the stochastic nature of the ZS model, the abundance distribution shown is from the first iteration that resulted in 25 species with $\theta = 4$. The x parameter for the log-series distribution was parameterized at 0.99969074 (see text for parameter description).

(A)



(B)

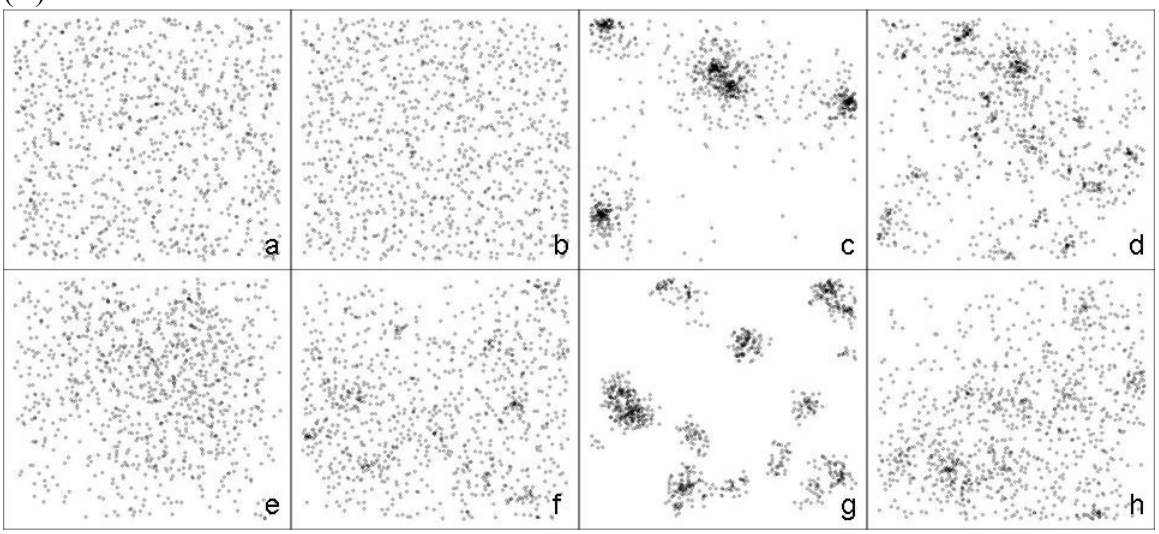


Fig. 2.3. (A) Spatial configuration patterns with one species and 1,000 total individuals. Panel a, random; b, hyper-dispersed; c, clustered (assemblage-wide); d, clustered (species-specific); e, aggregated (distance-decay); f, aggregated (distance-decay with seeds); g, aggregated (fidelity); and h, aggregated (distance-decay with seeds and equal abundances). (B) Spatial configuration patterns with 10 species and 1,000 total individuals. Panels organized as in (A).

(A) (B)

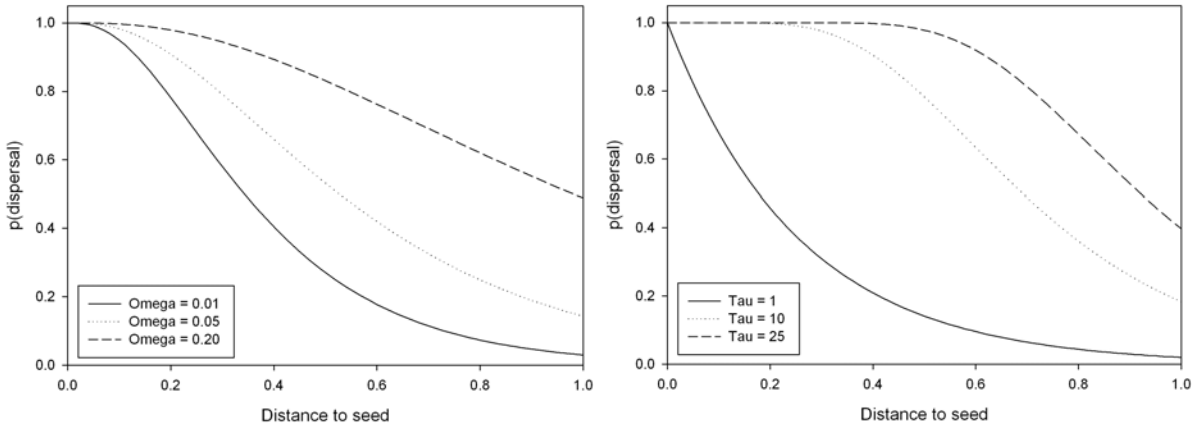


Fig. 2.4. The effect of ω and τ on the shape of the distance decay formula. (A) Omega (with τ = 3) controls the rate at which the probability of dispersal declines. (B) Tau (with ω = 0.02) controls the length of the shoulder, i.e., the distance over which the probability of dispersal equals one.

LITERATURE CITED

Angermeier P.L. and Smogor R.A. 1995. Estimating number of species and relative abundances in stream-fish communities – effects of sampling effort and discontinuous spatial distributions. Canadian Journal of Fisheries and Aquatic Sciences 52:936-949.

Arrhenius O. 1921. Species and area. Journal of Ecology 9:95-99.

Ashbridge J. and Goudie I.B.J. 2000. Coverage-adjusted estimators for mark-recapture in heterogeneous populations. Communications in Statistics. Simulation and Computation 29:1215-1237.

Baltanás A. 1992. On the use of some methods for the estimation of species richness. Oikos 65:484-492.

Beck J. and Kitching I.J. 2007. Estimating regional species richness of tropical insects from museum data: a comparison of a geography-based and sample-based methods. Journal of Applied Ecology 44:672-681.

Boswell M.T. and Patil G.P. 1971. Chance mechanisms generating the logarithmic series distribution used in the analysis of number of species and individuals. Pages 99-130 *in* Patil G.P., Pielou E.C. and Waters W.E. (eds.). *Statistical ecology*, vol. 3. Pennsylvania State University Press, University Park, PA, USA.

Boulinier T., Nichols J.D., Sauer J.R., Hines J.E. and Pollock K.H. 1998. Estimating species richness: the importance of heterogeneity in species detectability. Ecology 79:1018-1028.

Brose U., Martinez N.D. and Williams R.J. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. Ecology 84:2364-2377.

Burnham K.P. and Overton W.S. 1978. Estimation of the size of a closed population

when capture probabilities vary among animals. Biometrika 65:623-633.

Burnham K.P. and Overton W.S. 1979. Robust estimation of population size when capture probabilities vary among animals. Ecology 60:927-936.

Cao Y. Illinois Natural History Survey. 1816 South Oak Street , MC–652, Champaign, IL 61820.

Cao Y., Larsen D.P. and Hughes R.M. 2001. Estimating total species richness in fish assemblage surveys: a similarity based approach. Canadian Journal of Fisheries and Aquatic Sciences 58:1782-1793.

Cao Y., Larsen D.P. and White D. 2004. Estimating regional species richness using a limited number of survey units. Ecoscience 11:23-35.

Caswell H. 1976. Community structure: a neutral model analysis. Ecological Monographs 46:327-354.

Chao A. 1984. Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11:265-270.

Chao A. 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783-791.

Chao A. 2005. Species richness estimation. *In* Balakrishnan N., Read C.B. and Vidakovic B. (eds.). *Encyclopedia of Statistical Sciences*. Wiley, New York, USA.

Chao A., Colwell R.K., Lin C.W. and Gotelli N.J. 2009. Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90:1125-1133.

Chao A. and Lee S.M. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association 87:210-217.

Chao A. and Shen T.-J. 2009. Program SPADE (Species prediction and diversity

estimation). Program and user's guide published at http://chao.stat.nthu.edu.tw.

Chazdon R.L., Colwell R.K., Denslow J.S. and Guariguata M.R. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rainforests of NE Costa Rica. Pages 185-309 *in* Dallmeier F. and Comiskey J.A. (eds.). *Forest biodiversity research, monitoring and modeling: conceptual background and old world case studies*. Parthenon Publishing Group, Paris, France.

Clifford H.T. and Stephenson W. 1975. *An introduction to numerical classification*. Academic Press, London, UK.

Colwell R.K. 2006. *EstimateS* Statistical estimation of species richness and shared species from samples. Version 8. Persistent URL <purl.oclc.org/estimates>. Published at: http://viceroy.eeb.uconn.edu/EstimateS.

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

Darroch J.N. and Ratcliff D. 1980. A note on capture-recapture estimation. Biometrics 36:149-153.

Fisher R.A., Corbet A.S. and Williams C.B. 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.J. 1996. *Biodiversity: a biology of numbers and difference*. Blackwell Science, Oxford, Massachusetts, USA.

Heltshe J.F. and Forrester N.E. 1983. Estimating species richness using the jackknife procedure. Biometrics 39:1-11.

Hines J.E. 1996. *SPECRICH* Software to compute species abundance

from empirical species abundance distribution data. USGS-PWRC. Published at: http://www.mbr-pwrc.usgs.gov/software/specrich.html.

Holdridge L.R., Grenke W.C., Hatheway W.H., Liang T. and Tosi J.A. 1971. *Forest environments in tropical life zones.* Pergamon Press, Oxford, UK.

Hubbell S.P. 2001. *The unified neutral theory of biodiversity and biogeography*. Princeton University Press, Princeton, NJ, USA.

Keating K.A. and Quinn J.F. 1998. Estimating species richness: the Michaelis-Menten model revisited. Oikos 81:411-416.

Lee S.M. and Chao A. 1994. Estimating population size via sample coverage for closed capture-recapture models. Biometrics 50:88-97.

MacArther R.H. 1957. On the relative abundance of bird species. Proceedings of the National Academy of Science USA 43:293-295.

Magnussen S., Pélissier R., He F. and Ramesh B.R. 2006. An assessment of sample-based estimators of tree species richness in two wet tropical forest compartments in Panama and India. International Forestry Review 8:417-431.

Magurran A.E. 2004. *Measuring Biological Diversity*. Blackwell Publishing, MA, USA.

Matsumoto M. and Nishimura T. 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. ACM Transactions on Modeling and Computer Simulation 8:3-30.

Michaelis M. and Menten M.L. 1913. Der kinetic der invertinwirkung. Biochemische Zeitschrift 49:333-369.

Moreno C., Zuria I., García-Zenteno M., Sánchez-Rojas G., Castellanos I., Martínez-Morales M.

and Rojas-Martínez A. 2006. Trends in the measurement of alpha diversity in the last two decades. Interciencia 31:67-71.

Motomura I. 1932. On the statistical treatment of communities. Japanese Journal of Zoology 44:379-383 (in Japanese).

O'Dea N., Whittaker R.J. and Ugland K.I. 2006. Using spatial heterogeneity to extrapolate species richness: a new method test on Ecuadorian cloud forest birds. Journal of Applied Ecology 43:189-198.

Otis D.L., Burnham K.P., White G.C. and Anderson D.R. 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs 62:1-135.

Pledger S. 2000. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. Biometrics 56:434-442.

Poulin R. 1998. Comparison of three estimators of species richness in parasite component communities. Journal of Parasitology 84:485-490.

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

R Development Core Team. 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Schmit J.P., Murphy J.F. and Mueller G.M. 1999. Macrofungal diversity of a temperate oak forest: a test of species richness estimators. Canadian Journal of Botany 77:1014-1027.

Shannon C.E. and Weaver W. 1949. *The mathematical theory of communication*. University of Illinois Press, Urbana, IL, USA.

Smith E.P. and van Belle G. 1984. Nonparametric estimation of species richness.

Biometrics 40:119-129.

Sugihara G. 1980. Minimal community structure: an explanation of species abundance patterns. The American Naturalist 116:770-787.

Tjørve E. 2009. Shapes and functions of species-area curves (II): a review of new models and parameterizations. Journal of Biogeography 36:1435-1445.

Tokeshi M. 1990. Niche apportionment or random assortment: species abundance patterns revisited. Journal of Animal Ecology 59:1129-1146.

Tokeshi M. 1993. Species abundance patterns and community structure. Advances in Ecological Research 24:111-186.

Tokeshi M. 1996. Power fraction: a new explanation of relative abundance patterns in species-rich assemblages. Oikos 75:543-550.

Tokeshi M. 1999. *Species coexistence: ecological and evolutionary perspectives*. Blackwell Science, Oxford, UK.

Turner W., Leitner W. and Rosenzweig M. 2003. *Ws2m* Software for the measurement and analysis of species diversity. University of Arizona. Published at: http://eebweb.arizona.edu/diversity/.

Wagner H.H. and Wildi O. 2002. Realistic simulation of the effects of abundance distribution and spatial heterogeneity on non-parametric estimators of species richness. Ecoscience 9:241-250.

Walther B.A. and Martin J.L. 2001. Species richness estimation of bird communities: how to control for sampling effort? Ibis 143:413-419.

Walther B.A. and Moore J.L. 2005. The concepts of bias, precision and accuracy, and

their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography 28:815-829.

Walther B.A. and Morand S. 1998. Comparative performance of species richness estimation methods. Parasitology 116:395-405.

White G.C. and Burnham K.P. 1999. Program MARK: survival estimation from populations of marked animals. Bird Study 46 Supplement, 120-138.

Whittaker R.H. 1965. Dominance and diversity in land plant communities. Science 147:250-260.

Whittaker R.H. 1972. Evolution and measurement of species diversity. Taxon 21:213-251.

Whittaker R.H. 1977. Evolution of species diversity in land communities. Pages 1–67 *in* Hecht M.K., Steere W.C. and Wallace B. (eds.). *Evolutionary Biology*, Vol. 10, Plenum Publishing Corporation, New York, USA.

CHAPTER 3

PERFORMANCE OF SPECIES RICHNESS ESTIMATORS ACROSS ASSEMBLAGE TYPES AND SURVEY PARAMETERS

INTRODUCTION

Biological diversity encompasses many levels of the living world, from genes to ecosystems, all with the potential to inform decisions in ecological monitoring, conservation management, and reserve design. Despite a long focus on the species level, how best to quantify the diversity of an assemblage or some other subset of species remains a topic of much debate (Brose et al. 2003), though most attempts include: 1) a count or estimate of the unique species in a delineated area (species richness), 2) a measure of the uniformity of abundances amongst species (species evenness), or 3) some measure of species composition (species similarity) (Magurran 2004). Of the three, species richness (SR) is the most conceptually simple and frequently used (Gaston 1996, Brose et al. 2003, Moreno et al. 2006).

The number of species observed (S_{obs}) is the most straightforward measure of SR. S_{obs} can be considered a naïve estimator of SR. In other words, it assumes that all species in the sampling area are detected with a probability equal to one, a rare case especially for organisms that are difficult to see or capture. As an estimator of the true number of species in an assemblage (S_{true}), S_{obs} is negatively biased due to numerous factors, both biological and methodological (Palmer 1990, Chazdon et al. 1998, Nichols et al. 1998). Fortunately, species distribution and abundance patterns can be used to inform estimates of S_{true}. All estimators are based on assumptions that impose limitations; however, estimators of SR are typically less

negatively biased than S_{obs} (Baltanás 1992, Bunge and Fitzpatrick 1993, Walther and Morand 1998, Chiarucci et al. 2003, Walther and Moore 2005).

There are three categories of SR estimators. One category includes approaches for extrapolating a species accumulation curve to an asymptote, often using either a negative exponential model (Holdridge et al. 1971), Michaelis-Menten equation (Michaelis and Menten 1913), or power model (Arrhenius 1921, Tjørve 2009). A second category includes parametric methods that involve either: 1) interpolating under a distribution fit to abundance data, often a log-normal or log-series distribution or 2) applying an estimator that is based on an assumption that all species are equally detectable. A third category of estimators does not involve parametric assumptions, hereafter termed nonparametric estimators. All three categories use one of three types of baseline information comprised of the number of individuals encountered for each species, the number of surveys in which each species was encountered, or a list of species encountered in each survey.

Since no SR estimator has established itself as the clear best choice, there is a debate about which estimator to use. Inconsistencies in performance are a result of whether the underlying assumptions associated with each are met, thus the importance of factors such as the distribution of species abundances, the degree of spatial autocorrelation between individuals, survey effort, and S_{true} (Baltanás 1992, Keating and Quinn 1998, Wagner and Wildi 2002, Brose et al. 2003). The nonparametric estimators, specifically the M_h class of estimators, where h indicates heterogeneity in species detection probability, have generally performed better than other categories (Walther and Morand 1998, Gotelli and Colwell 2001, Brose et al. 2003). Nonparametric estimators are therefore the focus of this study and are further described in the methods (see also Chapter 2).

Nonparametric SR estimators can be grouped based on those that model heterogeneity in detection probability (M_h; e.g., Otis et al. 1978, Burnham and Overton 1978, Chao 1984), use of maximum likelihood methods (Pledger 2000), and use of similarity between replicate subsets of the survey data (Cao et al. 2001, Cao et al. 2004). Many were first developed as population estimators using mark-recapture data under the assumption of geographic and demographic closure; an assumption that must apply to all SR estimators. Brose et al. (2003) detailed some additional challenges that arise when M_h population estimators are used to estimate SR. First, larger differences in detectability between the species of an assemblage can be more difficult to model than those between the individuals of a population of one species. Second, when an estimator of the M_h class is used to estimate population size from the encounter histories of individuals in repeated surveys, where surveys are generally repeated at the same location, an assumption is that detection probabilities vary across individuals, but are constant over time. When those estimators are instead used to estimate SR from the encounter histories of species in replicated surveys, where surveys are generally replicated at different locations, the comparable assumption is that detection probabilities vary amongst species, but are constant across space. This assumption can be violated when distributions are spatially heterogeneous, which is a regular, scale-dependent, occurrence in natural systems (Legendre 1993, Deblauwe et al. 2008).

My objectives included evaluating the performances of nonparametric SR estimators across systematically varied assemblages (number of species, total abundance, distribution of species abundances, spatial configuration of individuals, and species detection probability) and across variation in sampling parameters (effort and survey design). Due to the large number of factor combinations and the benefits of knowing truth in evaluating performance, I mostly used simulated data; however, I also compared estimators with three real datasets that were easily

obtained. Two were from surveys of arboreal spider assemblages in African rainforests and savannahs of Kakum, Ghana and the Luki Biosphere Reserve in the Democratic Republic of the Congo (Fannes et al. 2008). Another dataset was comprised of ant species from five ecosystems in Florida, USA (King and Porter 2005).

Another objective was to expand and evaluate an estimator selection approach proposed by Brose et al. (2003). They used estimator accuracy across simulations to develop their framework and based selection on the ratio of S_{obs} to the mean of all SR estimates. I wanted to additionally incorporate selection criteria based on bias and precision.

METHODS

Simulations

I used program SimAssem (Chapter 2) to evaluate the performances of S_{obs} and 13 nonparametric SR estimators across assemblages that were systematically varied and surveyed. Species assemblages were simulated in a 2x3x3x3x4 factorial design (total abundance, number of species, species abundance distribution, spatial configuration, and species detection probability) or 216 combinations that were each surveyed by the four combinations of a 2x2 factorial design (survey design and effort), for 864 total combinations (Table 3.1). I applied the 13 SR estimators to each of 42 replicates for each factor combination (the term replicate is used hereafter to indicate runs that share certain assemblage properties), for a total of 36,288 realizations (the term realization if used hereafter to indicate any run, i.e., independent of assemblage properties).

Defining the species assemblage

I selected factor levels based on the findings of a preliminary literature review. Assemblages with 25, 100, and 500 species (factor S_{true}) were populated with 6,250 and 12,500

total individuals (factor *N*), so that the range of possible N/S_{true} values was within the range found in several datasets (Williams 1939, Williams 1940, Lewis and Taylor 1967, Dahlberg and Odum 1970, Dallmeier et al. 1991). Individuals were distributed amongst species such that the resulting species abundance distributions (factor *Abund*) approximated the often cited log-normal (Preston 1948) and log-series (Fisher et al. 1943) distributions. In order to contrast performance with a relatively even distribution, I also simulated particulate-niche distributions (MacArthur 1957; see Appendix I for examples of these and other species abundance distributions). SimAssem generates log-normal distributions by drawing a random log-normal variate ($\mu = 0$, $\sigma = 1$) for each species, dividing each variate by the sum of all variates, and rescaling by multiplying each variate by *N* (Appendix I, ex. 12), log-series distributions by calculating the number of species to populate with *z* individuals, where $z = 1, 2, \ldots, N$ (Appendix I, ex. 14; Magurran 2004), and particulate-niche distributions by randomly assigning each individual to a species (Appendix I, ex. 6). In a particulate-niche distribution, $N{:}S_{true}$ is positively related to the expected evenness of abundances.

Spatial configuration

Individuals were spatially distributed with three of the species-specific configuration options available in SimAssem: random, hyper-dispersed, and aggregated (factor *Config*). Each procedure involves assigning *x*- and *y*-coordinates [0, 1] to each individual. A random configuration is created by locating each individual with a pair of random uniform variates [0, 1]. For a hyper-dispersed configuration, SimAssem assigns each individual a square territory with a linear dimension of $1/\sqrt{n_i}$, where n_i is the abundance of species *i*. Territories are adjacent in the horizontal and vertical directions and collectively form a grid across the entire landscape. Individuals are dispersed a random distance and random direction from the bottom left corner of

their territory. Territories occasionally extend beyond the top and right landscape boundaries. When a randomized location in an overlapping territory falls outside the landscape, the individual is randomly located across the full landscape. SimAssem uses a two-step procedure for generating aggregated distributions that involves randomly selecting both a number of clusters and a location for each individual, (option *aggregated [centers, equal probability]*, see Chapter 2). First, for each species i, n_i random uniform variates (RUV) are drawn. Every RUV ≥ 0.98 increases by one the number of randomly placed aggregation centers (seeds), resulting in approximately one cluster for every 50 individuals. Second, each individual of that species is randomly allocated to a seed and distributed when two conditions are met. First, 0.95^d has to equal or exceed a RUV, where d is a randomly selected distance [0, 1]. Second, the individual has to fall on or within landscape boundaries when placed distance d in a random direction (0-359°) from the selected seed.

Detection probability

I used variates randomly drawn from three beta distributions for species-specific detection probabilities (factor p). Beta distributions are parameterized by two positive shape parameters, α and β, which I selected for expected means of 0.5, 0.7, and 0.9 and variances between 0.010 and 0.015. Four different algorithms were used for selecting p's. Two assemblages were created by drawing all p's from the same beta distribution, with expected means of 0.5 and 0.9 (α and β were 10 and 10, and 4.5 and 0.5, respectively). Species detection probability can also vary with abundance (Selmi and Boulinear 2004, Pagano and Arnold 2009). Two additional assemblages were therefore created after species were ranked by abundance and grouped into thirds (see Chapter 2). Since the tested levels of S_{true} are not divisible by three, one or two abundance groups were randomly selected to accommodate an additional species. For

one assemblage, the species in the least, moderate, and most abundant groups were assigned p's randomly drawn from beta distributions with expected means of 0.5, 0.7 ($\alpha = 14$, $\beta = 6$), and 0.9, respectively. The expected means were reversed for another factor level.

Survey design

In SimAssem, every cell of an overlaid 100 x 100 grid is a potential survey site. In addition to two levels of survey effort (factor *Effort*), 100 (1%) and 500 (5%) cells, I compared two survey designs (factor *Design*), random and a linear transect design that can represent, for example, surveys along roads and trails. Each transect is comprised of 50 adjacent cells in a randomly selected horizontal or vertical orientation and transects are added until a specified number of grid cells is surveyed. Previously surveyed grid cells intersected by a new transect are applied to the transect length, but are not double-counted. An encounter occurs when two conditions are met. First, a surveyed cell has to contain at least one individual. Second, a RUV, where one is drawn for each individual, has to be $\leq p$.

Species richness estimators

In addition to S_{obs}, I compared 13 estimators where some are variants of others. Details of the estimators as well as the abbreviations used throughout the text can be found in Table 3.2 and formulas and descriptions are in Appendix II. Those belonging to the M_h class included two that are based on abundance patterns, i.e., the number of species with an exact number of individuals encountered, and nine that are based on incidence patterns, i.e., the number of species encountered in an exact number of surveys. Two additional estimators, CY-1 and CY-2, are based on the similarity of two replicate subsets of surveys. CY-1 is only calculable when individuals are encountered in ≥ 2 surveys because average SR across the replicate sets of surveys ($\overline{SR}$) is divided by Jaccard's coefficient, $JC = c / (a + b + c)$, where a and b are the numbers of

species unique to each subset and *c* is the number of species common to both subsets. The CY-2 estimator equals the slope plus the intercept of a regression line fit to $\overline{SR}$ versus *JC*, where each value, as programmed in SimAssem, is the average of 100 realizations (see Cao et al. 2004). SimAssem uses five regression points ($\overline{SR}$ -*JC* pairs) when there are between 10 and 19 surveys with encounters and 10 regression points when there are ≥20 surveys with encounters (see Chapter 2 for additional details). For the Mixture estimator, I used the Rmark package (Version 1.9.5; Laake and Rextad 2008) to program R (R Development Core Team 2009) to generate estimates based on two groups in program MARK (Version 6.0; White and Burnham 1999).

Performance evaluation

The largest possible estimates of the tested estimators vary widely. The bootstrap estimator can extrapolate to S_{obs} x 2 (Colwell and Coddington 1994), the jackknife estimators to slightly less than S_{obs} x (the order of the estimator + 1), e.g., Jack5 can extrapolate to approximately S_{obs} x 6, Chao2 to S_{obs}^2 / 2, and CY-2 to $\overline{SR}^2$ (Cao et al. 2004). When negative terms dominate the equation, the higher-order jackknife estimators can return estimates $<S_{obs}$ and sometimes even negative estimates (see Kim et al. 2006). I therefore calculated the proportion of every estimate relative to S_{obs} and tracked both the number of estimates that were $<S_{obs}$ as well as the number of realizations in which Mixture failed to converge.

Estimators were evaluated based on bias, precision, and accuracy (Walther and Moore 2005). I evaluated the overall performance of an estimator by combining all realizations, i.e., all factors and all factor levels, and performance at each of the different levels of a specific factor, e.g., *Abund*, by combining the replicates of all remaining factors and factor levels. Interaction effects were not evaluated. Estimator bias was evaluated with scaled mean error (*SME*; Walther and Moore 2005),

$$SME = \frac{1}{X}\sum_{j=1}^{X}\left(\frac{\widehat{S_{est(j)}} - S_{true}}{S_{true}}\right),$$

where X is the number of replicates across combined factors, j is the replicate, $j = 1, 2, \ldots, X$, and S_{est} is an estimate from any SR estimator. Negative and positive values indicate average underestimation and overestimation, respectively, and $SME = 0$ for an unbiased estimator. I evaluated estimator precision with the sample standard deviation of a group of scaled estimates (*SD*),

$$SD = \sqrt{\frac{\sum_{j=1}^{X}\left(\frac{\widehat{S_{est(j)}}}{S_{true}} - \frac{\overline{\widehat{S_{est}}}}{S_{true}}\right)}{X-1}},$$

where $\widehat{S_{est(j)}} / S_{true}$ is the scaled estimate of the j^{th} replicate, $j = 1, 2, \ldots, X$. To estimate accuracy, which accounts for both bias and precision, I used scaled mean square error (*SMSE*; Walther and Moore 2005),

$$SMSE = \frac{1}{X}\sum_{j=1}^{X}\left(\frac{\widehat{S_{est(j)}} - S_{true}}{S_{true}}\right)^2.$$

Values of *SD* and *SMSE* are always positive and, as with *SME*, those closer to zero indicate a better performing estimator.

I evaluated relative factor effects with random-effects models, using proc mixed in SAS (9.2, SAS Institute 1999), by computing variance components and proportionally allocating total variance into the seven factors, S_{true}, N, *Abund*, *Config*, p, *Effort*, and *Design*. The residual captured unexplained experimental error (Ott and Longnecker 2001).

Sample coverage (*sc*) is the proportion of S_{true} represented in a sample, i.e., S_{obs}/S_{true} (Engen 1975), and possibly the single most important factor driving estimator performance (Baltanás 1992, Brose et al. 2003). I therefore investigated estimator performance as a function

of *sc*, using two different methods. First, *sc* was used to group realizations into 10 equally sized bins, e.g., $0.0 < sc \leq 0.1$, $0.1 < sc \leq 0.2$, …, $0.9 < sc \leq 1.0$ and estimator performance was averaged over each of these coverage ranges. Second, a threshold (*th*) was systematically incremented from 0.01-1.0 and, at each *th*, bias and accuracy were averaged across all realizations with $sc \leq th$.

Species survey data from the literature

In addition to the simulated data, I applied estimators to survey data from Fannes et al. (2008) as well as to a potentially less complete dataset from King and Porter (2005). The value of S_{true} is unknown in a real assemblage, however, Fannes et al. (2008) considered the 11 spider species they encountered in 12 surveys at Kakum (Ghana) a near census of the available species and I therefore assumed that $S_{true} = 11$ for that dataset. The accumulation curve resulting from randomizations of the eight species encountered in five surveys on the Luki Biosphere Reserve (Democratic Republic of the Congo) failed to reach an asymptote (see Fannes et al. 2008, their Fig. 4), which is indicative of an incomplete dataset. Despite possibly violating the assumption that $S_{obs} = S_{true}$, I used $S_{true} = 8$ in analyses of the Luki data. In the King and Porter (2005) dataset, absolute estimator bias could not be evaluated because S_{obs} had not yet reached S_{true}, evidenced by the monotonic increase of the accumulation curve (their Fig. 1). However, by assuming that estimates are $<S_{true}$, I was able to compare estimators with relative bias.

I evaluated estimator performance via randomization techniques that allowed for comparisons to my simulation results. At every possible survey size, I randomized surveys without replacement, 200x for the Fannes et al. (2008) datasets and 50x for the King and Porter (2005) dataset, as per the original studies. The SR estimators were applied to each randomized

subset and bias, precision, and accuracy were computed where possible. Some estimators, i.e., CY-1, CY-2, and jackknife estimators, could not be computed at the smallest sample sizes.

I also evaluated estimator performance with randomized accumulation curves, plotting estimates against survey effort. Estimator performance was evaluated by the rate at which the curve reached a reasonable asymptote, i.e., comparable to the asymptotic value of other estimators and/or S_{obs}. Increasing to an asymptote faster than S_{obs} is an important measure of estimator quality (Chazdon et al. 1998, Gotelli and Colwell 2001).

RESULTS

Simulations

The Mixture estimator failed to converge in 0.31% of the 36,288 realizations and several other estimators occasionally returned an estimate $<S_{obs}$ including CY-1 (0.79%), CY-2 (1.10%), Jack2 (5.28%), Jack3 (9.87%), Jack4 (15.28%), and Jack5 (19.13%). Over 69% of the realizations in which the Jack5 estimate was $<S_{obs}$ occurred when $S_{true} = 25$ and only one realization included $S_{true} = 500$. Furthermore, the smaller Jack5 estimates were almost evenly split between the levels of each of the other factors when $S_{true} = 25$. When $S_{true} = 100$, realizations were not so evenly divided between the levels of *Effort* (>83% of the realizations included *Effort* = 500) and *Abund* (>51% included a particulate-niche distribution). Prior to analysis, realizations in which the Mixture estimates failed to converge were removed which affected, at most, five replicates. I rebalanced the factorial design by removing ≤5 replicates from every combination, with preference given to removing replicates that included one or more estimates $<S_{obs}$. This resulted in 37 replicates and 31,968 total realizations (Table 3.3, first rows).

A SR estimate should never be $<S_{obs}$; therefore, such estimates were set equal to S_{obs} and the performances of affected estimators were reevaluated (Table 3.3, second rows). Large outliers, sometimes four orders of magnitude larger than S_{obs}, contributed considerably to the poor performance of Mixture. Relative to S_{obs}, CY-2 returned the next largest estimate (approximately S_{obs} x 76), which I used as a proportionality threshold above which Mixture estimates were set equal to S_{obs} (Table 3.3, third row).

Estimator rank varied by the metric used to evaluate performance (Table 3.3). The modifications just described resulted in all of the estimators being less biased than S_{obs}. CY-1 and CY-2 were the least biased estimators and, after the large Mixture estimates were revalued, the only positively biased estimators. Chao1 and Chao2 were the most precise estimators, slightly greater than S_{obs}, and ACE, ICE, and Jack3 were the most accurate estimators.

Abundance distribution

Estimator performance varied between the species abundance distributions. Estimator bias and accuracy generally improved with increases in evenness, or in moving from the log-series to the log-normal to the particulate-niche abundance distribution (Table 3.4). The CY-1, CY-2, and Jack5 estimators, however, were less biased with log-normal distributions than with particulate-niche distributions where considerable overestimation occurred. Estimators were most precise with log-series distributions and nearly evenly divided on whether they were more precise with log-normal or particulate-niche distributions. Additionally, CY-1 and CY-2 were relatively inaccurate in assemblages with particulate-niche distributions and Jack5 was most accurate with log-series distributions. In assemblages with log-normal, log-series, and particulate-niche distributions, CY-1, Jack5, and ICE were least biased, Chao2, CY-1, and Chao2 were most precise, and CY-1, CY-2, and Chao2 were most accurate, respectively.

The largest number of replicates that included at least one untenable estimate (in as many as 33 of the 37 replicates) occurred in factor combinations that included a particulate-niche distribution. Several of the estimators, particularly CY-1 and CY-2, performed considerably worse in assemblages that included a particulate-niche abundance distribution (Table 3.4). The poor performance of some estimators in assemblages with particulate-niche distributions is not an indication of the general performance of estimators with niche-based models, but rather an assessment of performance as a function of evenness. In other words, performance depends not on whether a niche- or statistically-based model was used, but on the patterns of relative abundance. Given the additional lack of empirical support for particulate-niche distributions, I excluded those assemblages from further analyses.

Additional assemblage factors

After the realizations with a particulate-niche distribution were removed, all estimators were negatively biased (Table 3.5). Jack5 was the least biased estimator and, along with S_{obs}, Chao1, Chao2 were the most precise estimators. Generally, there was a tradeoff between bias and precision such that the most biased estimators were the most precise, and vice versa. CY-1 and CY-2 were the most and Mixture the least accurate estimators. These estimators also typically performed best in the factor-specific analyses and additional relative estimator performances are therefore reported only when there are noteworthy differences. Thus, I focus on the general performance of estimators across factor levels in the remaining comparisons.

Bias was generally greater with larger values of S_{true}; however, Jack5 was least biased in assemblages with 100 species (Table 3.6). Precision generally increased with S_{true}, but there were deviations for individual estimators. Accuracy tended to decline with increasing S_{true} except that Jack3, Jack4, and Jack5 were at their most accurate, and Mixture at its least accurate,

in assemblages with S_{true} = 100. In comparison to the overall results (Table 3.5), Jack5 was the least biased estimator only in assemblages with S_{true} = 100 and, with 25 species, the only estimator that exhibited a positive bias. At all levels of S_{true}, there were estimators other than S_{obs} that exceeded the precision of Chao1 and Chao2 and, when S_{true} = 25 or 100, an estimator other than CY-1 and CY-2 was most accurate.

Estimator bias decreased when *N* was increased from 6,250 to 12,500, with the biases of CY-1 and CY-2 changing by the smallest amounts (Table 3.7). Except for S_{obs}, estimators were more precise in assemblages with more individuals. Also, all estimators were more accurate with the larger *N*.

The patterns of bias as a function of *Config* were the same for all estimators (Table 3.8). Though the differences were minimal, estimators were least and most biased in assemblages with hyper-dispersed and aggregated individuals, respectively. All estimators were also least precise in assemblages with hyper-dispersed individuals. Among the three levels of *Config*, CY-1, CY-2, Jack4, Jack5, and Mixture were most accurate in assemblages with randomly dispersed individuals whereas the remaining estimators were most accurate in assemblages with hyper-dispersed individuals. The Mixture estimator was more accurate than S_{obs} except in assemblages with hyper-dispersed configurations (compare with Table 3.5).

Estimators were less biased and more accurate, though, in general, with slightly less precision when average species detection probability ($\overline{p}$) equaled 0.9 than when $\overline{p}$ = 0.5 (Table 3.9). When species were grouped by ranked abundance all estimators, except for Boot and S_{obs}, were less biased and only Boot, Jack1, Jack2, and S_{obs} were more precise when $\overline{p}$ decreased, on average, with abundance than when $\overline{p}$ increased with abundance (Table 3.9). Furthermore, only ACE, Chao1, Chao2, and ICE were more accurate when $\overline{p}$ decreased with abundance. Compared

to performances averaged across all factors (Table 3.5), CY-2 was less biased than Jack5 when $\overline{p}$ decreased with abundance and Chao1 and Chao2 were not amongst the most precise estimators in any of the assemblages.

Survey factors

Increasing the amount of surveyed landscape from 1% (100 cells) to 5% (500 cells) improved both the bias and accuracy of all estimators and the precision of all estimators, but not S_{obs} (Table 3.10). The precision and accuracy of Mixture improved considerably when effort was increased from 1% to 5%. Notable exceptions to the overall results (Table 3.5) included, with 1% of the landscape surveyed, CY-2 was the least biased estimator and Boot was more precise than Chao1 and Chao2. Also, with 5% of the landscape surveyed, ICE and Mixture were the most precise estimators and ACE, Boot, Jack5, and S_{obs} were all less accurate than Mixture. All estimators were less biased and more accurate, often by a marginal degree, with the random than with the linear transect survey design (Table 3.11). Only Boot, CY-1, CY-2, and S_{obs}, were less precise with a random than with the linear transect survey design and, once again, the differences were small.

Variance component analysis

Averaged across all estimators, a variance component analysis with random-effects models showed that S_{true} had by far the largest effect on estimator bias, precision, and accuracy, with *Effort* ranking 2nd in explaining bias and precision and 3rd in accuracy, and *Abund* ranking 3rd in explaining bias and precision and 2nd in accuracy (Table 3.12, row Mean). The results for individual estimators tended to follow the above patterns with S_{true} always having the largest effect and usually being followed, in either order, by *Effort* and *Abund*. However, *N* ranked slightly above *Abund* for the precision of Boot and *Effort* ranked below 3rd for the precision of

CY-2 and the precision and accuracy of CY-1 and Mixture. The remaining factors, *N*, *p*, *Config*, and *Design* generally had small effects on estimator performance. When averaged across all estimators, the seven tested factors accounted for approximately 95%, 81%, and 81% of the variation in bias, precision, and accuracy, respectively, but explained <50% of the variation in the precision and accuracy of Mixture. Preliminary analyses indicated that interaction effects explain a portion of the variance captured by the residual (data not shown).

Sample coverage

Performance was also evaluated as a function of *sc*, partly for the development of a selection framework (Table 3.13). The least biased estimator was CY-2, CY-1, or Jack5 in each of the five smallest coverage ranges and a different estimator in each of the five largest coverage ranges. S_{obs}, Boot, and Jack1 were, in that order, most precise in all except the largest coverage range where Mixture was second best. CY-1 was the most accurate estimator in the two smallest coverage ranges, progressively smaller-order jackknife estimators were most accurate in the six next larger coverage ranges and Boot was most accurate in the two largest ranges. Bias and accuracy generally improved with larger *sc* whereas there was no general trend in precision.

When bias and accuracy were evaluated across all realizations where $sc \leq th$ (a systematically incremented threshold), CY-2 was the least biased estimator for all $th \leq 0.12$ (*SME* = -0.31), CY-1 from there until $th = 0.38$ (*SME* = -0.30), CY-2 again until $th = 0.67$ (*SME* = -0.23), and then Jack5 until $th = 1.00$ (*SME* = -0.13). CY-2 was the most accurate estimator for all $th \leq 0.05$ (*SMSE* = 0.48) and CY-1 was the most accurate estimator when averaged over all larger thresholds (e.g., *SMSE* = 0.14 for $th = 1.00$). The performance of an estimator is therefore not necessarily best at all values of *sc* within the reported boundaries of the selection frameworks.

Real data

For the real datasets, I report the averages across the randomizations of surveys. The randomization of two surveys from the Kakum dataset (Kakum, Ghana; Fannes et al. 2008) had an $\overline{sc} > 0.60$ (assuming $S_{true} = 11$). The selection framework based on my simulation results (Table 3.13) correctly identified the least biased, CY-1 (*SME* = 0.08), most precise, S_{obs} (*SD* = 0.14), and most accurate estimators, Jack1 and Jack2 (*SMSE* = 0.09). When five surveys were randomized, $\overline{sc} = 0.84$ and Chao2 was the least biased estimator (*SME* = -0.47); however, the selection framework incorrectly predicted ICE which was the fifth least biased estimator, *SME* = 0.11. The selection framework again correctly predicted that S_{obs} would be most precise (*SD* = 0.11), and Boot most accurate (*SMSE* = 0.02). In addition to recreating accumulation curves for the originally compared estimators (ACE, Boot, Chao1, Chao2, Jack1, and Jack2), I created curves for CY-1, CY-2, ICE, and Mixture (Fig. 3.1). The estimates from CY-2, which required at least 10 surveys, appear as a line fragment in the top right corner. When all survey data were used, the Jack2 estimate was $<S_{obs}$ and the Jack3, Jack4, and Jack5 estimates were monotonically decreasing, 7.9, 4.6, and 1.0, respectively (not shown).

The randomization of two surveys from the Luki Biosphere Reserve dataset (Democratic Republic of the Congo; Fannes et al. 2008) included >80% of the species, $\overline{sc} = 0.82$ (assuming $S_{true} = 8$). The selection framework correctly predicted the least biased, ICE (*SME* = -0.01), most precise, S_{obs} (*SD* = 0.12), and most accurate, Boot (*SMSE* = 0.04) estimators (Fig. 3.2). The selection framework did not perform as well when $\overline{sc} = 0.95$ (four surveys), predicting that Chao2 (*SME* = 0.07) would be least biased when it was actually Boot (*SME* = 0.03). Chao1 (*SME* = -0.03) was, however, correctly predicted to perform nearly as well. The most precise estimator was correctly predicted as S_{obs} (*SD* = 0.10). The first and third most accurate

estimators, S_{obs} (*SMSE* = 0.01) and Boot (*SMSE* = 0.02), respectively, were predicted in the reverse order by the selection framework. To minimize clutter, Jack3 (the estimate based on all data equaled 10.7), Jack4 (10.8), Jack5 (10.8), and ICE (9.2) were not displayed and, with only five surveys, there were not enough data for computing CY-2.

When 20, 100, and 400 surveys from King and Porter (2005) were randomized, Jack5 produced the largest estimates, 54.8, 87.3, and 109.1, respectively. If it is assumed that S_{true} = 142, which is the number of species found in historical regional datasets (see Deyrup 2003), then these survey sizes equate to $\overline{sc}$ = 0.15, 0.31, and 0.46, respectively. Based on my selection framework, CY-1 would have been incorrectly selected as the least biased estimator for 20 surveys and Jack5 would have been correctly selected for survey sizes of 100 and 400. Jack3 was the least biased estimator when 1,650 surveys were randomized (*SME* = -0.11), but at the assumed sample coverage, $\overline{sc}$ = 0.65, the framework placed it third behind CY-1 (*SME* = -0.21) and Jack2 (*SME* = -0.14). S_{obs}, Boot, and Jack1 were the first, second, and third most precise estimators, respectively, identical to the selection framework.

DISCUSSION

The use of S_{obs} as an estimate of SR has often been criticized for its strong dependence on sampling effort and the assumption that all species are detected, which frequently leads to large underestimates (Nichols et al. 1998, Brose et al. 2003, Kéry and Plattner 2007). As in other studies, the nonparametric estimators were generally less biased and more accurate (Table 3.5; see also Wagner and Wildi 2002, Brose et al. 2003) in estimating SR. As effort increases at a site, S_{obs} certainly approaches S_{true}, but estimators provide a more reliable approach, especially when effort and thus sample sizes are small (Table 3.10). Furthermore, S_{obs} provides no measure

of precision which is another advantage of using an estimation approach (Nichols et al. 1998, Chapter 4).

Despite the relatively frequent use of Chao1, Chao2, Jack1, and Jack2, my results suggest that CY-1, CY-2, Jack3, Jack4, and to a lesser extent, ACE and ICE, can provide less biased and more accurate estimates. The third- through fifth-order jackknife estimators require at least 3-5 sampling occasions and thus more effort, which certainly contributes to the more frequent use of Jack1 and Jack2 (Otis et al. 1978). In particular, Jack5, CY-1, and CY-2 were often less biased than the ≥20% reported for other estimators (see Canning-Clode 2008, Jobe 2008). Bias reduction involves increased extrapolation, which could partly explain the loss of precision that accompanied the least biased estimators. This tradeoff between bias and precision has been reported elsewhere (see Burnham and Overton 1979, Brose et al. 2003). The higher-order (third- through fifth-order) jackknife estimators and bias-corrected Chao estimators were the least and intermediately biased estimators, respectively, findings similar to Brose et al. (2003). By contrast, two estimators performed better in my study than previously reported, ICE (see Walther and Morand 1998, Walther and Martin 2001) and Jack2 (see Wagner and Wildi 2002).

The results of this study suggest that CY-1 and CY-2 are promising newer estimators, being the most accurate and among the least biased estimators when averaged over many different assemblage types (Table 3.5). CY-1 and CY-2 performed their best and were among the least biased and most accurate estimators with the relatively uneven log-normal and log-series distributions that are regularly considered two of the best approximations to true species abundance patterns (Sugihara 1980, Ulrich et al. 2010).

Cao et al. (2004) stated that CY-2 would be less affected by sample size than other estimators and CY-2 was, in fact, the least biased estimator at the smaller effort level (Table

3.10). CY-1 was more accurate and also more robust, based on accuracy changing by a smaller margin between the two levels of *Effort* than for any of the other estimators. CY-1 and CY-2 performed similarly at both levels of *N*, a factor that affects sample size in a manner similar to *Effort*. The performance of CY-2 could partly depend on the number of regression points. The number used in SimAssem to address occasional data limitations, five or ten, might not be enough with which to expose its full performance potential. CY-1 and CY-2 were, however, relatively imprecise and prone to overestimation for the particulate-niche abundance distributions and larger levels of *N*, *Effort*, and *p*, factor levels generally associated with larger *sc* values. Based on the assemblages simulated in this study, CY-1 and CY-2 begin to overestimate once *sc* exceeds 0.70 and 0.60, respectively.

Another estimator largely untested against established SR estimators, Mixture, often performed relatively poorly, particularly with respect to precision and accuracy. However, there were scenarios that appeared to improve the performance of Mixture. Based on bias, Mixture ranked more favorably in assemblages with more species (Table 3.6) and was among the best estimators in assemblages that conformed to a particulate-niche distribution (Table 3.4). Both precision and accuracy improved dramatically when the number of surveyed cells was increased from 100 to 500 (Table 3.10). The amount of data required to estimate probabilities for two groups, as was done by Mixture, might not have been provided by some of the factor combinations in this study. The conditions under which the performance of Mixture improved would indicate that it should be further evaluated in comparatively data-rich environments.

Similar to Brose et al. (2003), of the seven systematically varied factors, S_{true}, *Effort*, and *Abund* had the largest effects. These factors were also the most strongly correlated with *sc*, i.e., S_{true} ($r^2 = -0.65$), *Effort* ($r^2 = 0.48$), *Abund* ($r^2 = 0.36$), *N* ($r^2 = 0.18$), *p* ($r^2 = 0.04$), *Config* ($r^2 =$

0.02), and *Design* ($r^2 = 0.01$), which supports *sc* being the intermediary through which factors affect estimator performance. The sampling parameters in my study, i.e., *Effort* and *Design*, were identical at each of the three levels of S_{true}, which could partly explain both the decline in *sc*, which was approximately 0.64, 0.46, and 0.22 for S_{true} = 25, 100, and 500 species, respectively, and the negative relationship between S_{true} and performance. However, this explanation does not explain the contradictory positive and relatively weak correlation ($r^2 = 0.14$) found by Brose et al. (2003) who also nested factor levels.

At larger values of S_{true}, bias was generally greater, which supports Baltanás (1992), but not Brose et al. (2003), and accuracy generally declined, which contradicts Walther and Morand (1998) (Table 3.6). Estimator performance generally improved with *Effort*, a finding also reported by Brose et al. (2003) and Wagner and Wildi (2002), and a logical result as more effort often results in more data (see also the results based on *N* and *p*).

The distribution of abundances in an assemblage strongly affected the performance of most estimators, even without considering the effects of particulate-niche distributions (Table 3.12). A broken-stick distribution (MacArthur 1957) was used as the relatively even distribution in other studies (see Wagner and Wildi 2002, Brose et al. 2003). I chose the particulate-niche distribution after not finding a statistical difference between log-normal and broken-stick distributions (unpublished data) using Kolmogorov-Smirnov goodness of fit tests (Magurran 2004, pg. 220). The similarity of these two distributions could have contributed to Wagner and Wildi (2002) finding that estimators are more negatively biased with an even distribution, i.e., a broken-stick, than with an uneven distribution, i.e., a log-normal, which was opposite of my finding. I found that estimators were most negatively biased in assemblages with relatively uneven log-series distributions, supporting both Wagner and Wildi (2002) and Brose et al.

(2003). Accuracy also generally improved with evenness in my study, though Jack5 was most accurate with the relatively uneven log-series distribution.

Estimator performance was better in assemblages with larger *N*, i.e., with 12,500 versus 6,250 individuals (Table 3.7). A positive relationship has been similarly reported with *N* in the form of density (Baltanás 1992, Walther and Morand 1998). Basically, increasing the pool of individuals available to a survey generally increases the amount of sample data.

Species detection probability (*p*) had a larger effect if limited to the factor levels $\overline{p} = 0.5$ and $\overline{p} = 0.9$ ($r^2 = 0.17$), i.e., if the two levels where *p* was a function of abundance were removed. Bias, accuracy, and for all except Boot and S_{obs}, precision, were better in assemblages with the larger average *p*, $\overline{p} = 0.9$ versus $\overline{p} = 0.5$ (Table 3.9). Increasing $\overline{p}$ is another way of increasing the number of encounters and thereby improving estimation.

When $\overline{p}$ was smaller for less abundant species, estimators were less biased (Boot and S_{obs} produced small exceptions), though generally with less precision and accuracy, than when $\overline{p}$ increased with ranked abundance (Table 3.9). Less abundant species, particularly those with a small *p*, are often not encountered in surveys. The smaller estimator bias when $\overline{p}$ decreased with ranked abundance could be a result of a larger proportion of the less abundant species being encountered as rare, e.g., singletons and doubletons, or infrequent, e.g., uniques, species, thereby increasing the estimate. Also, the lack of consistent trends in precision and accuracy could be a result of the differences in $\overline{p}$ being too small. Sample coverage averaged 46.9% when $\overline{p}$ increased with abundance and 39.3% when $\overline{p}$ decreased with abundance.

The effect size of *Config* was small, corroborating the findings of both Wagner and Wildi (2002) and Brose et al. (2003). The three configuration patterns differed considerably for a single species, but at the assemblage level the differences were minimal. This emergent property

could partly explain the relatively weak effects that *Config* had on estimator performance and it is certainly possible that the effect of an assemblage-wide configuration pattern would be greater. Estimators were more negatively biased with increased aggregation of individuals, which supports Baltanás (1992, their Fig. 3), but not Wagner and Wildi (2002). Furthermore, I did not find the positive relationship between accuracy and aggregation reported by Walther and Morand (1998).

Survey design also had a relatively small effect on estimator performance, a result that might not hold in all situations. For example, in a situation where SR varies along one or more gradients, it could be possible to orient linear transects such that they fail to fully represent an area. I am unaware of any other study that has evaluated relationships between survey design and SR estimates, but differences between linear transects and random surveys are important to other estimation issues (see Reese et al. 2005).

For different reasons, some estimates were $<S_{obs}$, even negative, and Mixture occasionally failed to converge or returned unreasonably large estimates. Fortunately, it is intuitively obvious that the final estimate of SR should never be $<S_{obs}$. Formulas for the jackknife estimators include terms that either add to or subtract from an estimate based on the number of species encountered in an exact number of surveys (see Appendix II for formulas). When the count associated with one or more of these negative terms is large, the estimate can be $<S_{obs}$ and such estimates were set equal to S_{obs}. Another approach would be to set these estimates equal to the next lower-order jackknife estimator that produces an estimate $>S_{obs}$. This would reduce bias and would be an approach somewhat similar to the jackknife-select estimator introduced by Burnham and Overton (1979). CY-2 produced negative estimates in approximately 0.15% of the realizations and the number of surveys with encounters was not

adequate for computing CY-1 and CY-2 in some realizations, demonstrating a potential limitation in their applicability. CY-2 can return an estimate $<S_{obs}$ when the relationship between $\overline{SR}$ and JC is negative, an unlikely situation that can occur by chance when data are randomized. In all of the realizations where CY-2 failed to give an estimate, the Mixture estimator either failed to converge or gave an unreasonably large estimate, i.e., $\geq$(1720 x S_{obs}). This appeared to be the result of sparse data. Finding thresholds at which valid estimates become possible could require systematically varying both the number of surveys and the number of encounters.

Estimates from ACE and ICE are entirely derived from the number of rare and infrequent species, typically defined as species with $\leq$10 individuals encountered across all surveys and species encountered in $\leq$10 surveys, respectively. Therefore, ACE and ICE do not produce estimates whenever all species are encountered >10x, a situation that occurred most commonly in assemblages with particulate-niche distributions. I considered a lack of rare or infrequent classes an indication that all species were encountered and therefore set estimates in those realizations equal to S_{obs}. Walther and Morand (1998) attributed such instances to the estimators requiring too many frequency classes. However, it would be incorrect to state that these estimators require an encounter in each of the frequency classes to produce an estimate, i.e., only one rare or infrequent species is needed. It is therefore appropriate to set ACE and ICE equal to S_{obs} or to another estimator such as Chao1 or Chao2, as is done in program EstimateS (Colwell 2006).

Adjusting the untenable estimates improved the performances of several of the estimators. For example, the jackknife estimators were thereafter less negatively biased and more accurate, with the performance of Jack5 then comparable to that of CY-1 and CY-2. The

bias and accuracy of Mixture fell within the range of other estimators after the largest Mixture estimates were revalued, which included all 124 realizations where CY-1 estimates were set equal to S_{obs} because there was no estimate. As was the goal, the selected modifications improved estimator performance.

My results mostly support the guidelines of the selection framework proposed by Brose et al. (2003), but indicate that there are better options at the extremes of *sc*. Specifically, the most accurate estimator was CY-1 for $sc \leq 20\%$ and Boot for $sc > 80\%$ (Table 3.13), estimators that Brose et al. (2003) did not evaluate. I also expanded their framework by including bias and precision selection criteria. The best performing estimators varied by evaluation metric; therefore, estimator selection should be application specific. For example, when the objective is to compare richness across different areas, precision should be heavily weighted so as to maximize the probability that areas will be correctly ranked. When the objective is to estimate the number of species in a single area, an estimator that reduces bias would be preferable.

Many real world factors can complicate estimation efforts. For example, there are many difficult to detect species, e.g., cryptic and extremely small species, as well as limited sampling efforts that can further reduce estimator performance. The real world might therefore not conform to the trends detected in a simulated environment, particularly beyond the range of evaluated factors. I therefore recommend that application of these results to the real world, especially extrapolation, be done with caution. Simulated environments ultimately represent a best case scenario, so if estimators perform poorly there, how can we trust them in the much more complicated real world?

Using the selection framework of either study is difficult because both are based on *sc*. Estimation would, of course, be unnecessary in the event that we knew *sc*, i.e., S_{obs}/S_{true}. The

calculation of CY-1 involves a procedure for estimating $sc, \widehat{sc}$. In my study, the correlation between $\widehat{sc}$ and the true *sc* of simulations exceeded 0.78; therefore, I recommend using $\widehat{sc}$ with my proposed selection frameworks. SimAssem reports $\widehat{sc}$ and, just as importantly, the program can be used to preliminarily test estimator performances in a specified assemblage type (Chapter 2).

I used the Kakum data from Fannes et al. (2008) to test the performance of $\widehat{sc}$. In building the accumulation curve, $\widehat{sc}$ ranged from 0.53 to 0.70 between four and 10 randomized surveys, respectively. If it is assumed that $S_{true} = 11$, then $\widehat{sc}$ underestimated empirical *sc* which was 78.77 and 98.32 with four and 10 surveys, respectively (see Fig. 3.1). Still, my selection framework suggested using either Jack4 or CY-1 to reduce bias and, in both cases, the suggested estimate was either the first or second largest (Jack4 data not shown). Evaluating the selection framework with data from actual surveys is difficult, of course, because there is rarely, if ever, an assurance that selection is based on correct information. For example, several estimators indicated that S_{true} is larger than 11, which could be true given the difficulty in detecting all species.

With the Kakum data, Fannes et al. (2008) considered the performances of the SR estimators unsatisfactory, largely because none reached an asymptote any faster than S_{obs}. While this remained true of the estimators that were also used in their study, the average CY-1 estimate was 12.1 when two surveys were randomized ($S_{obs} = 7$ at that point) which is 110% of an assumed $S_{true} = 11$ and 94% of the CY-1 estimate (12.8) based on all the data (Fig. 3.1). The ICE estimate was even larger with two surveys (15.4), but the continual decline in estimates at larger samples sizes implied that it was an artifact of randomization, e.g., one or more surveys had a

relatively large number of uniques that would cause a large estimate when coupled with only one other survey. The next largest estimate at two surveys was 92% of S_{true} (Chao2 = 10.2).

When the Luki survey data were randomized, S_{obs} never reached an asymptote, but Fannes et al. (2008) noted a possible leveling of the Chao1 estimator. In my randomizations, none of the estimators reached an apparent asymptote as quickly as with the Kakum data and Chao2 and CY-1 changed the least between samples sizes of four and five (Fig. 3.2). Had there been more data with which to evaluate performance, these two estimators possibly would have shown that they had reached their respective asymptotes. As in the simulations with $S_{true} = 25$, these comparisons indicate that CY-1 requires relatively less data to reach an asymptote.

SR estimators appeared not to approach an asymptote faster than S_{obs} with the ant dataset of King and Porter (2005; their Fig. 1). Though not reproduced here, none of the additional estimators performed any better. This could indicate that a considerable number of species were yet to be encountered.

CONCLUSION

My study supports previous findings that the performances of SR estimators depend, to varying degrees, on numerous assemblage characteristics and survey design parameters. Therefore, selecting the best estimator for a particular situation requires information about such relationships. The numerous studies on SR estimators have sometimes occurred in vastly different systems. As my study shows, this can partly explain the reported differences in estimator performance. Additionally, as Walther and Moore (2005) indicated, many different performance metrics have been used and there have even been instances where a metric is used to evaluate bias in one study and accuracy in another. The use of such a wide variety of performance metrics across studies is almost certain to complicate comparisons.

The assemblage factors S_{true}, *Effort*, and *Abund* had the largest effects on estimator performance, the relative strength of which largely depended on a correlation with *sc*. The nonparametric estimators were all less biased and more accurate than a raw count of the number of species and I therefore conclude that S_{obs} is far from the best approach for estimating SR. I included several estimators that have received little previous evaluation. Based on bias and accuracy across a wide range of assemblages, i.e., all tested factor levels except particulate-niche distributions, it appears that CY-1 and CY-2 are among the best available estimators of SR; however, better estimators are often available for particular assemblages. This study further shows the overarching influence of *sc* and, therefore, its value in estimator selection.

A SR estimate without an associated variance estimate is of limited value because one has no measure of its reliability. As this study indicates, the most biased estimates are generally the most precise. This is a particularly dangerous combination because a precise and biased estimator, based on the repeatability of estimates, can easily be considered more correct than an imprecise, but unbiased, estimator. Variance estimators have been derived for many of the SR estimators evaluated in this study and their performance should also be considered when selecting a SR estimator. To the best of my knowledge, the performance of variance estimators has not been evaluated. Such a study should also evaluate the performance of variance estimation procedures such as bootstrapping and jackknifing, as they represent possible substitutes for missing and ineffective derived estimators (see Chapter 4).

Table 3.1. Factor descriptions, abbreviations, and simulated levels.

Description	**Abbreviation**	**Levels**
Total (true) number of species	S_{true}	25
		100
		500
Total abundance across all species	N	6250
		12500
Species abundance distribution	*Abund*	Log-series
		Log-normal
		Particulate-niche[1]
Spatial relationship between individuals of a species	*Config*	Aggregated
		Hyper-dispersed
		Random[2]
Mean detection probability of species abundance groups	p	(0.5, 0.5, 0.5)
		(0.9, 0.9, 0.9)
		(0.5, 0.7, 0.9)
		(0.9, 0.7, 0.5)
Spatial arrangement of surveyed grid cells	*Design*	Linear transect
		Random[3]
Amount of landscape surveyed	*Effort*	1% (100 cells)
		5% (500 cells)

[1]Assemblage abundance patterns ranged from relatively uneven (log-series) to even (particulate-niche).

[2]Individuals were spaced less regularly [Aggregated, *aggregated (centers, equal probability)*] or more evenly (Hyper-dispersed) than expected by chance (Random). See text and Chapter 2 for more details.

[3]Surveys were configured as either random linear transects of 50 grid cells or random grid cells.

Table 3.2. Properties of the tested species richness estimators.

Estimator name	**Abbreviation**	**Citation**
Abundance-based coverage[1,2]	ACE	Chao and Lee 1992
Chao1 (bias-corrected)[1,2]	Chao1	Chao 1984
Bootstrap[1,3]	Boot	Smith and van Belle 1984
Chao2 (bias-corrected)[1,3]	Chao2	Chao 1987
Incidence-based coverage[1,3]	ICE	Lee and Chao 1994
1st-order jackknife[1,3]	Jack1	Burnham and Overton 1978
2nd-order jackknife[1,3]	Jack2	Burnham and Overton 1978
3rd-order jackknife[1,3]	Jack3	Burnham and Overton 1978
4th-order jackknife[1,3]	Jack4	Burnham and Overton 1978
5th-order jackknife[1,3]	Jack5	Burnham and Overton 1978
Mixture-model[1,4]	Mixture	Pledger 2000
CY-1[5]	CY-1	Cao et al. 2001
CY-2[5]	CY-2	Cao et al. 2004

[1]Estimation involves modeling heterogeneity in detection probability.
[2]Estimation involves use of sample abundance patterns, i.e., number of individuals.
[3]Estimation involves use of sample incidence patterns, i.e., number of surveys.
[4]Estimation involves use of maximum likelihood.
[5]Estimation involves use of similarity of replicate surveys of species with Jaccard's coefficient.

Table 3.3. Average performance of species richness estimators across all factors (see Tables 3.1 and 3.2) based on bias, measured as scaled mean error (*SME*), precision, measured as standard deviation of scaled estimates (*SD*), and accuracy, measured as scaled mean square error (*SMSE*). Below each performance measure, estimator rank is given parenthetically for each step. Column *Affected* is the number of realizations, out of 36,288, that were modified in each step.

Estimator	***SME (Bias)***	***SD (Precision)***	***SMSE (Accuracy)***	***Affected***
ACE	-0.22	0.31	0.15	0
	(7, 7, 7)	(5, 5, 5)	(2, 2, 2)	
Boot	-0.43	0.31	0.28	0
	(12, 12, 13)	(4, 4, 4)	(11, 12, 12)	
Chao1	-0.28	0.28	0.16	0
	(10, 10, 11)	(1, 1, 1)	(4, 5, 5)	
Chao2	-0.28	0.28	0.16	0
	(9, 9, 10)	(2, 2, 2)	(3, 4, 4)	
CY-1	0.02, 0.02	0.51, 0.51	0.26, 0.26	0, 288
	(1, 1, 1)	(11, 11, 11)	(9, 9, 9)	
CY-2	0.03, 0.03	0.51, 0.51	0.27, 0.26	0, 398
	(2, 2, 2)	(12, 12, 12)	(10, 10, 10)	
ICE	-0.22	0.32	0.15	0
	(6, 6, 6)	(6, 6, 6)	(1, 1, 1)	
Jack1	-0.34	0.32	0.22	0
	(11, 11, 12)	(7, 7, 7)	(8, 8, 8)	
Jack2	-0.26, -0.26	0.32, 0.32	0.17, 0.17	0, 1917
	(8, 8, 9)	(8, 8, 8)	(6, 6, 6)	
Jack3	-0.21, -0.19	0.35, 0.34	0.16, 0.15	0, 3583
	(5, 5, 5)	(9, 9, 9)	(5, 3, 3)	
Jack4	-0.17, -0.12	0.43, 0.39	0.21, 0.17	0, 5545
	(4, 4, 4)	(10, 10, 10)	(7, 7, 7)	
Jack5	-0.13, -0.04	0.62, 0.52	0.40, 0.27	0, 6942
	(3, 3, 3)	(13, 13, 13)	(13, 11, 11)	
Mixture	1.71 1.71 -0.23	23.53, 23.53, 0.57	556.57, 556.57, 0.38	113[a], 0[b], 403[c]
	(14, 14, 8)	(14, 14, 14)	(14, 14, 14)	
S_{obs}	-0.50	0.30	0.34	0
	(13, 13, 14)	(3, 3, 3)	(12, 13, 13)	

[a]The first value for each estimator is performance after realizations in which Mixture failed to converge were removed and the factorial was rebalanced by removing 5 replicates from each combination (resulting in 37 replicates and 31,968 total realizations).

[b]A second performance measure indicates that the estimator returned estimates $<S_{obs}$ that were set equal to S_{obs}.

[c]The third measure for Mixture is performance after estimates $>(S_{obs} \times 76)$ were set equal to S_{obs}.

Table 3.4. Average performance of species richness estimators as a function of species abundance distribution, log-normal (LN), log-series (LS), and particulate-niche (PN). Specific performance metrics are given in Table 3.3.

	Bias			Precision			Accuracy		
Estimator	**LN**	**LS**	**PN**	**LN**	**LS**	**PN**	**LN**	**LS**	**PN**
ACE	-0.19	-0.47	-0.01	0.26	0.19	0.28	0.10	0.26	0.08
Boot	-0.39	-0.60	-0.29	0.31	0.18	0.33	0.25	0.40	0.19
Chao1	-0.24	-0.50	-0.08	0.25	0.19	0.22	0.12	0.29	0.06
Chao2	-0.24	-0.50	-0.08	0.25	0.19	0.22	0.12	0.29	0.06
CY-1	0.02	-0.40	0.44	0.28	0.16	0.57	0.08	0.19	0.51
CY-2	0.03	-0.35	0.41	0.32	0.25	0.58	0.10	0.18	0.51
ICE	-0.18	-0.47	0.00	0.26	0.19	0.29	0.10	0.26	0.08
Jack1	-0.30	-0.53	-0.19	0.31	0.20	0.32	0.18	0.32	0.14
Jack2	-0.21	-0.46	-0.11	0.31	0.22	0.31	0.14	0.26	0.11
Jack3	-0.14	-0.40	-0.04	0.34	0.25	0.32	0.13	0.22	0.10
Jack4	-0.06	-0.33	0.02	0.40	0.31	0.37	0.16	0.21	0.14
Jack5	0.03	-0.26	0.11	0.54	0.43	0.51	0.30	0.25	0.27
Mixture	-0.20	-0.51	0.01	0.61	0.39	0.55	0.41	0.41	0.31
S_{obs}	-0.47	-0.66	-0.37	0.30	0.17	0.33	0.31	0.46	0.25

Table 3.5. Average performance of species richness estimators across all factors except particulate-niche distributions (see Tables 3.1 and 3.2). Specific performance metrics are given in Table 3.3.

Estimator	Bias	Precision	Accuracy
ACE	-0.33	0.27	0.18
Boot	-0.50	0.27	0.32
Chao1	-0.37	0.26	0.21
Chao2	-0.37	0.26	0.21
CY-1	-0.19	0.31	0.13
CY-2	-0.16	0.34	0.14
ICE	-0.33	0.27	0.18
Jack1	-0.41	0.29	0.25
Jack2	-0.33	0.30	0.20
Jack3	-0.27	0.33	0.18
Jack4	-0.20	0.38	0.19
Jack5	-0.11	0.51	0.27
Mixture	-0.35	0.54	0.41
S_{obs}	-0.56	0.26	0.38

Table 3.6. Average performance of species richness estimators as a function of the true number of species, 25, 100, and 500 species, across all factors except particulate-niche distributions (see Tables 3.1 and 3.2). Specific performance metrics are given in Table 3.3.

	Bias			Precision			Accuracy		
Estimator	**25**	**100**	**500**	**25**	**100**	**500**	**25**	**100**	**500**
ACE	-0.23	-0.31	-0.44	0.26	0.25	0.25	0.12	0.16	0.26
Boot	-0.30	-0.46	-0.73	0.24	0.22	0.16	0.15	0.26	0.55
Chao1	-0.27	-0.34	-0.51	0.27	0.24	0.20	0.14	0.18	0.30
Chao2	-0.27	-0.34	-0.51	0.27	0.24	0.20	0.14	0.17	0.30
CY-1	-0.18	-0.18	-0.22	0.29	0.32	0.32	0.12	0.14	0.15
CY-2	-0.11	-0.15	-0.21	0.33	0.32	0.37	0.12	0.13	0.18
ICE	-0.23	-0.31	-0.44	0.26	0.25	0.25	0.12	0.16	0.25
Jack1	-0.23	-0.37	-0.65	0.25	0.23	0.19	0.12	0.19	0.45
Jack2	-0.17	-0.27	-0.55	0.28	0.24	0.22	0.11	0.13	0.35
Jack3	-0.11	-0.21	-0.48	0.34	0.26	0.24	0.12	0.11	0.29
Jack4	-0.01	-0.15	-0.43	0.43	0.31	0.26	0.19	0.12	0.25
Jack5	0.14	-0.09	-0.39	0.63	0.40	0.28	0.42	0.17	0.23
Mixture	-0.31	-0.32	-0.43	0.47	0.66	0.45	0.32	0.54	0.39
S_{obs}	-0.36	-0.54	-0.78	0.23	0.20	0.13	0.18	0.33	0.63

Table 3.7. Average performance of species richness estimators as a function of the total number of individuals, 6,250 and 12,500 individuals, across all factors except particulate-niche distributions (see Tables 3.1 and 3.2). Specific performance metrics are given in Table 3.3.

	Bias		Precision		Accuracy	
Estimator	**6250**	**12500**	**6250**	**12500**	**6250**	**12500**
ACE	-0.35	-0.31	0.28	0.25	0.20	0.16
Boot	-0.54	-0.45	0.27	0.27	0.37	0.27
Chao1	-0.41	-0.34	0.26	0.25	0.23	0.18
Chao2	-0.41	-0.34	0.26	0.25	0.23	0.18
CY-1	-0.19	-0.19	0.33	0.29	0.14	0.12
CY-2	-0.16	-0.16	0.37	0.31	0.16	0.12
ICE	-0.34	-0.31	0.28	0.25	0.20	0.16
Jack1	-0.46	-0.37	0.29	0.28	0.29	0.21
Jack2	-0.37	-0.29	0.31	0.28	0.23	0.16
Jack3	-0.31	-0.23	0.34	0.31	0.21	0.14
Jack4	-0.24	-0.16	0.40	0.36	0.22	0.16
Jack5	-0.15	-0.07	0.53	0.49	0.30	0.24
Mixture	-0.37	-0.34	0.56	0.51	0.45	0.37
S_{obs}	-0.61	-0.52	0.25	0.26	0.43	0.33

Table 3.8. Average performance of species richness estimators as a function of the spatial configuration of individuals of a species, aggregated (Agg), hyper-dispersed (Hyp), and random (Rnd), across all factors except particulate-niche distributions (see Tables 3.1 and 3.2). Specific performance metrics are given in Table 3.3.

	Bias			Precision			Accuracy		
Estimator	**Agg**	**Hyp**	**Rnd**	**Agg**	**Hyp**	**Rnd**	**Agg**	**Hyp**	**Rnd**
ACE	-0.35	-0.32	-0.32	0.26	0.27	0.27	0.19	0.17	0.18
Boot	-0.50	-0.49	-0.49	0.27	0.28	0.27	0.33	0.32	0.32
Chao1	-0.39	-0.36	-0.37	0.26	0.26	0.25	0.22	0.20	0.20
Chao2	-0.39	-0.36	-0.37	0.26	0.26	0.25	0.22	0.20	0.20
CY-1	-0.22	-0.18	-0.18	0.30	0.32	0.31	0.14	0.13	0.13
CY-2	-0.18	-0.14	-0.15	0.33	0.35	0.34	0.14	0.15	0.14
ICE	-0.34	-0.32	-0.32	0.26	0.27	0.27	0.19	0.17	0.17
Jack1	-0.42	-0.41	-0.41	0.29	0.29	0.28	0.26	0.25	0.25
Jack2	-0.34	-0.32	-0.33	0.30	0.30	0.29	0.21	0.19	0.20
Jack3	-0.27	-0.26	-0.27	0.33	0.33	0.32	0.18	0.17	0.17
Jack4	-0.21	-0.19	-0.20	0.39	0.39	0.37	0.19	0.19	0.18
Jack5	-0.12	-0.11	-0.12	0.51	0.51	0.50	0.28	0.27	0.26
Mixture	-0.37	-0.34	-0.36	0.50	0.63	0.47	0.38	0.51	0.35
S_{obs}	-0.57	-0.56	-0.56	0.26	0.26	0.26	0.39	0.38	0.38

Table 3.9. Average performance of species richness estimators as a function of species detection probability, across all factors except particulate-niche distributions (see Tables 3.1 and 3.2). Detection probabilities were randomly drawn from either: 1) beta distributions with expected values of 0.5 or 0.9 or 2) beta distributions with expected values that either increased (I) or decreased (D) with abundance after first grouping species into thirds based on ranked abundance. Specific performance metrics are given in Table 3.3.

	Bias				**Precision**				**Accuracy**			
Estimator	**0.5**	**0.9**	**I**	**D**	**0.5**	**0.9**	**I**	**D**	**0.5**	**0.9**	**I**	**D**
ACE	-0.37	-0.30	-0.36	-0.29	0.49	0.49	0.48	0.51	0.21	0.16	0.19	0.17
Boot	-0.54	-0.45	-0.49	-0.50	0.38	0.41	0.40	0.39	0.37	0.28	0.31	0.33
Chao1	-0.42	-0.34	-0.39	-0.35	0.48	0.48	0.47	0.49	0.25	0.18	0.21	0.20
Chao2	-0.42	-0.34	-0.39	-0.35	0.48	0.49	0.47	0.49	0.24	0.18	0.21	0.20
CY-1	-0.22	-0.18	-0.25	-0.12	0.54	0.54	0.53	0.56	0.15	0.11	0.13	0.15
CY-2	-0.19	-0.14	-0.22	-0.08	0.53	0.53	0.51	0.55	0.16	0.12	0.13	0.17
ICE	-0.36	-0.30	-0.36	-0.29	0.49	0.49	0.48	0.51	0.21	0.15	0.18	0.17
Jack1	-0.47	-0.37	-0.41	-0.41	0.40	0.43	0.42	0.41	0.30	0.21	0.24	0.26
Jack2	-0.38	-0.29	-0.33	-0.32	0.43	0.46	0.45	0.44	0.24	0.16	0.19	0.20
Jack3	-0.32	-0.22	-0.27	-0.25	0.45	0.47	0.46	0.46	0.21	0.14	0.17	0.18
Jack4	-0.25	-0.16	-0.20	-0.18	0.46	0.48	0.47	0.47	0.22	0.15	0.18	0.19
Jack5	-0.17	-0.08	-0.12	-0.09	0.47	0.49	0.47	0.48	0.30	0.23	0.26	0.29
Mixture	-0.39	-0.33	-0.39	-0.31	0.52	0.53	0.51	0.53	0.51	0.30	0.33	0.52
S_{obs}	-0.61	-0.52	-0.55	-0.57	0.37	0.39	0.39	0.37	0.43	0.34	0.37	0.40

Table 3.10. Average performance of species richness estimators as a function of survey effort, 100 (1%) and 500 (5%) cells, across all factors except particulate-niche distributions (see Tables 3.1 and 3.2). Specific performance metrics are given in Table 3.3.

	Bias		**Precision**		**Accuracy**	
Estimator	**1%**	**5%**	**1%**	**5%**	**1%**	**5%**
ACE	-0.41	-0.25	0.30	0.20	0.26	0.10
Boot	-0.62	-0.37	0.25	0.24	0.45	0.19
Chao1	-0.48	-0.27	0.27	0.21	0.30	0.12
Chao2	-0.48	-0.27	0.26	0.21	0.30	0.12
CY-1	-0.23	-0.16	0.37	0.22	0.19	0.07
CY-2	-0.22	-0.10	0.41	0.24	0.22	0.07
ICE	-0.40	-0.25	0.30	0.20	0.25	0.10
Jack1	-0.55	-0.28	0.28	0.23	0.37	0.13
Jack2	-0.46	-0.20	0.31	0.22	0.31	0.09
Jack3	-0.39	-0.14	0.34	0.24	0.27	0.08
Jack4	-0.33	-0.07	0.41	0.31	0.27	0.10
Jack5	-0.25	0.02	0.53	0.45	0.34	0.20
Mixture	-0.37	-0.34	0.73	0.20	0.67	0.16
S_{obs}	-0.69	-0.44	0.22	0.24	0.52	0.25

Table 3.11. Average performance of species richness estimators as a function of survey design, random (Random) and linear transect (Transect), across all factors except particulate-niche distributions (see Tables 3.1 and 3.2). Specific performance metrics are given in Table 3.3.

	Bias		Precision		Accuracy	
Estimator	**Random**	**Transect**	**Random**	**Transect**	**Random**	**Transect**
ACE	-0.33	-0.33	0.26	0.27	0.18	0.18
Boot	-0.49	-0.50	0.27	0.27	0.32	0.32
Chao1	-0.37	-0.38	0.26	0.26	0.20	0.21
Chao2	-0.37	-0.38	0.26	0.26	0.20	0.21
CY-1	-0.18	-0.20	0.31	0.30	0.13	0.13
CY-2	-0.15	-0.17	0.35	0.34	0.14	0.14
ICE	-0.32	-0.33	0.27	0.27	0.17	0.18
Jack1	-0.41	-0.42	0.29	0.29	0.25	0.26
Jack2	-0.33	-0.33	0.30	0.30	0.20	0.20
Jack3	-0.26	-0.27	0.32	0.33	0.17	0.18
Jack4	-0.20	-0.20	0.38	0.39	0.18	0.19
Jack5	-0.11	-0.11	0.50	0.51	0.27	0.28
Mixture	-0.35	-0.36	0.52	0.55	0.39	0.43
S_{obs}	-0.56	-0.56	0.26	0.26	0.38	0.39

Table 3.12. The percent variance attributable to each of the independent factors (see Tables 3.1 and 3.2). Specific performance metrics are given in Table 3.3.

Bias

Estimator	S_{true}	*N*	*Abund*	*Config*	*Design*	*Effort*	*p*	Residual
ACE	87.32	0.17	5.54	0.04	0.01	2.83	0.21	3.87
Boot	67.45	2.65	5.61	0.02	0.00	18.51	0.93	4.83
Chao1	84.94	0.50	5.44	0.04	0.01	4.96	0.28	3.83
Chao2	85.02	0.49	5.47	0.02	0.01	4.92	0.28	3.79
CY-1	89.60	0.01	6.60	0.03	0.01	0.59	0.20	2.96
CY-2	88.86	0.03	5.62	0.03	0.01	1.23	0.22	3.99
ICE	87.46	0.16	5.58	0.03	0.01	2.68	0.21	3.87
Jack1	74.41	1.78	5.21	0.02	0.01	13.26	0.65	4.67
Jack2	80.27	1.09	4.50	0.01	0.00	8.97	0.44	4.70
Jack3	82.43	0.79	4.02	0.01	0.00	6.99	0.35	5.41
Jack4	82.18	0.66	3.73	0.01	0.00	6.02	0.30	7.10
Jack5	79.60	0.60	3.56	0.01	0.00	5.53	0.29	10.42
Mixture	84.28	0.23	5.96	0.02	0.01	1.32	0.26	7.92
S_{obs}	52.49	3.25	13.75	0.03	0.00	23.77	1.05	5.67
Mean	80.45	0.89	5.76	0.02	0.01	7.26	0.40	5.22

Precision

Estimator	S_{true}	*N*	*Abund*	*Config*	*Design*	*Effort*	*p*	Residual
ACE	70.97	0.12	5.78	0.10	0.01	3.12	0.50	19.39
Boot	48.43	3.62	3.34	0.02	0.00	25.50	1.20	17.90
Chao1	71.53	0.69	5.17	0.08	0.01	7.16	0.49	14.86
Chao2	71.96	0.67	5.23	0.05	0.02	7.11	0.49	14.48
CY-1	75.08	0.00	7.45	0.08	0.04	0.22	0.57	16.57
CY-2	70.25	0.00	6.63	0.08	0.03	0.51	0.65	21.86
ICE	71.33	0.11	5.87	0.06	0.02	2.86	0.51	19.25
Jack1	53.46	2.86	3.41	0.02	0.00	21.84	0.96	17.46
Jack2	59.89	2.06	3.36	0.02	0.00	17.62	0.71	16.34
Jack3	63.69	1.61	3.31	0.02	0.00	15.04	0.59	15.75
Jack4	65.69	1.33	3.25	0.02	0.00	13.27	0.52	15.92
Jack5	65.36	1.14	3.19	0.02	0.00	12.02	0.48	17.79
Mixture	45.96	0.22	3.25	0.01	0.00	0.00	0.44	50.11
S_{obs}	52.49	3.25	13.75	0.03	0.00	23.77	1.05	5.67
Mean	63.29	1.26	5.21	0.04	0.01	10.72	0.65	18.81

Accuracy								
Estimator	S_{true}	***N***	***Abund***	***Config***	***Design***	***Effort***	***p***	**Residual**
ACE	52.82	0.40	22.06	0.15	0.02	6.84	0.80	16.91
Boot	53.52	2.81	14.92	0.03	0.01	21.76	0.96	5.98
Chao1	52.12	1.24	19.66	0.11	0.01	12.54	0.76	13.56
Chao2	52.28	1.21	19.82	0.07	0.02	12.60	0.77	13.24
CY-1	49.13	0.00	32.46	0.14	0.07	0.90	1.23	16.06
CY-2	49.06	0.01	24.50	0.15	0.06	2.38	1.24	22.60
ICE	52.80	0.36	22.41	0.09	0.04	6.43	0.81	17.05
Jack1	54.22	2.29	15.82	0.04	0.01	19.36	0.86	7.40
Jack2	55.06	1.68	15.28	0.04	0.01	16.27	0.75	10.91
Jack3	54.62	1.28	13.70	0.03	0.00	13.61	0.65	16.10
Jack4	51.82	0.98	11.63	0.02	0.00	11.00	0.53	24.02
Jack5	45.22	0.69	9.00	0.01	0.00	8.19	0.39	36.51
Mixture	21.87	0.09	12.37	0.04	0.00	0.10	0.49	65.04
S_{obs}	61.96	3.02	7.27	0.02	0.00	22.76	1.06	3.92
Mean	50.47	1.15	17.21	0.07	0.02	11.05	0.81	19.23

Table 3.13. The three best performing species richness estimators, averaged over all realizations, in the specified sample coverage (*sc*) range. Specific performance metrics are given in Table 3.3.

	Bias			**Precision**			**Accuracy**		
Coverage range	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
$0.0 < sc \leq 0.1$	-0.35	-0.37	-0.56	0.02	0.02	0.03	0.26	0.34	0.43
	CY-2	CY-1	Mixture	S_{obs}	Boot	Jack1	CY-1	CY-2	ICE
$0.1 < sc \leq 0.2$	-0.25	-0.29	-0.38	0.03	0.04	0.05	0.24	0.28	0.28
	CY-1	CY-2	Mixture	S_{obs}	Boot	Jack1	CY-1	Jack5	CY-2
$0.2 < sc \leq 0.3$	-0.30	-0.30	-0.31	0.03	0.04	0.05	0.15	0.16	0.18
	CY-1	CY-2	Jack5	S_{obs}	Boot	Jack1	Jack5	Jack4	CY-2
$0.3 < sc \leq 0.4$	-0.18	-0.25	-0.265	0.03	0.04	0.07	0.11	0.12	0.14
	Jack5	Jack4	CY-2	S_{obs}	Boot	Jack1	Jack4	Jack3	Jack5
$0.4 < sc \leq 0.5$	-0.03	-0.11	-0.16	0.03	0.04	0.07	0.06	0.07	0.08
	Jack5	Jack4	CY-2	S_{obs}	Boot	Jack1	Jack3	Jack4	Jack2
$0.5 < sc \leq 0.6$	-0.04	0.08	-0.09	0.03	0.04	0.08	0.06	0.06	0.06
	Jack4	Jack5	CY-2	S_{obs}	Boot	Jack1	Jack2	CY-2	Jack3
$0.6 < sc \leq 0.7$	-0.03	-0.03	0.04	0.02	0.04	0.08	0.02	0.02	0.04
	CY-1	Jack2	Jack3	S_{obs}	Boot	Jack1	Jack1	Jack2	CY-1
$0.7 < sc \leq 0.8$	-0.03	0.05	-0.06	0.03	0.04	0.08	0.01	0.02	0.02
	Jack1	Jack2	ICE	S_{obs}	Boot	Jack1	Jack1	ICE	Boot
$0.8 < sc \leq 0.9$	-0.02	-0.03	0.03	0.02	0.04	0.08	0.00	0.01	0.01
	ICE	ACE	Jack1	S_{obs}	Boot	Jack1	Boot	Jack1	ACE
$0.9 < sc \leq 1.0$	0.00	0.00	0.00	0.03	0.03	0.04	0.00	0.00	0.00
	Chao2	Chao1	ACE	S_{obs}	Mixture	Boot	Boot	Mixture	S_{obs}

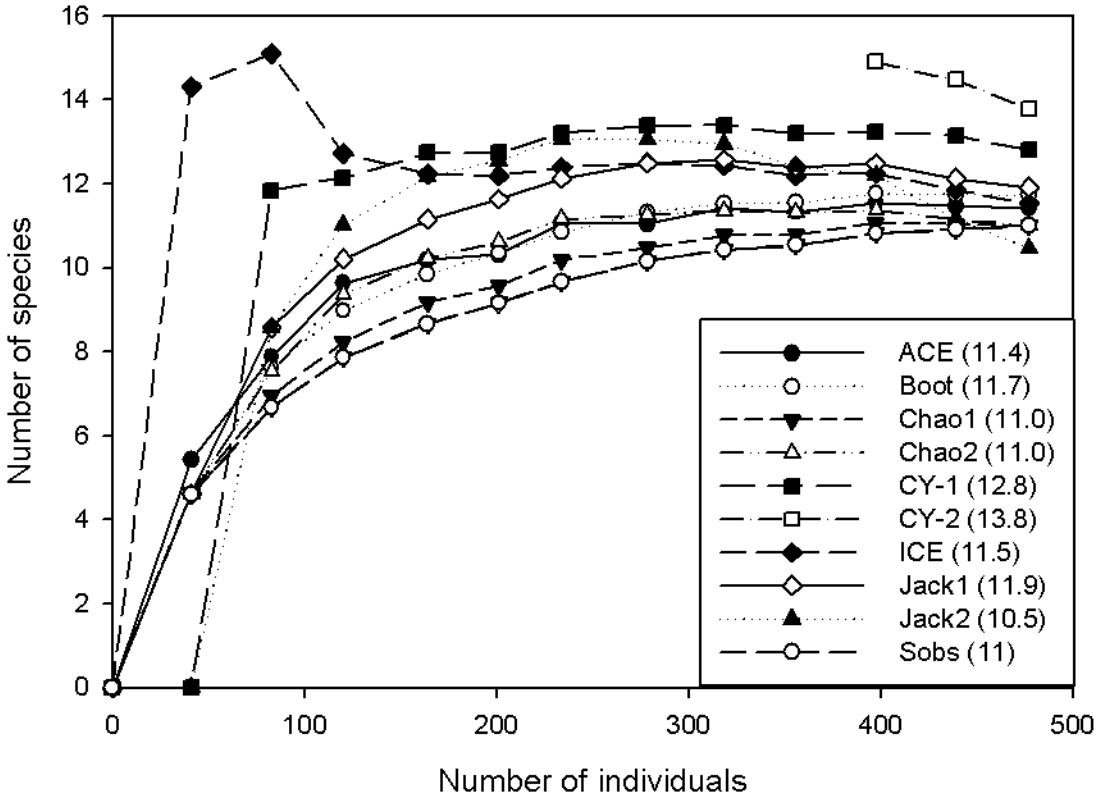


Fig. 3.1. Species accumulation curves for 10 species richness estimators based on arboreal spider data from Kakum National Park, Ghana. At each possible survey size (12 total), 200 surveys were randomly drawn without replacement, species richness was estimated, and estimates were averaged across the 200 randomizations. The x-axis is the average number of individuals contained in randomized surveys (each successive symbol on a line represents an increment of one in the number of randomized surveys).

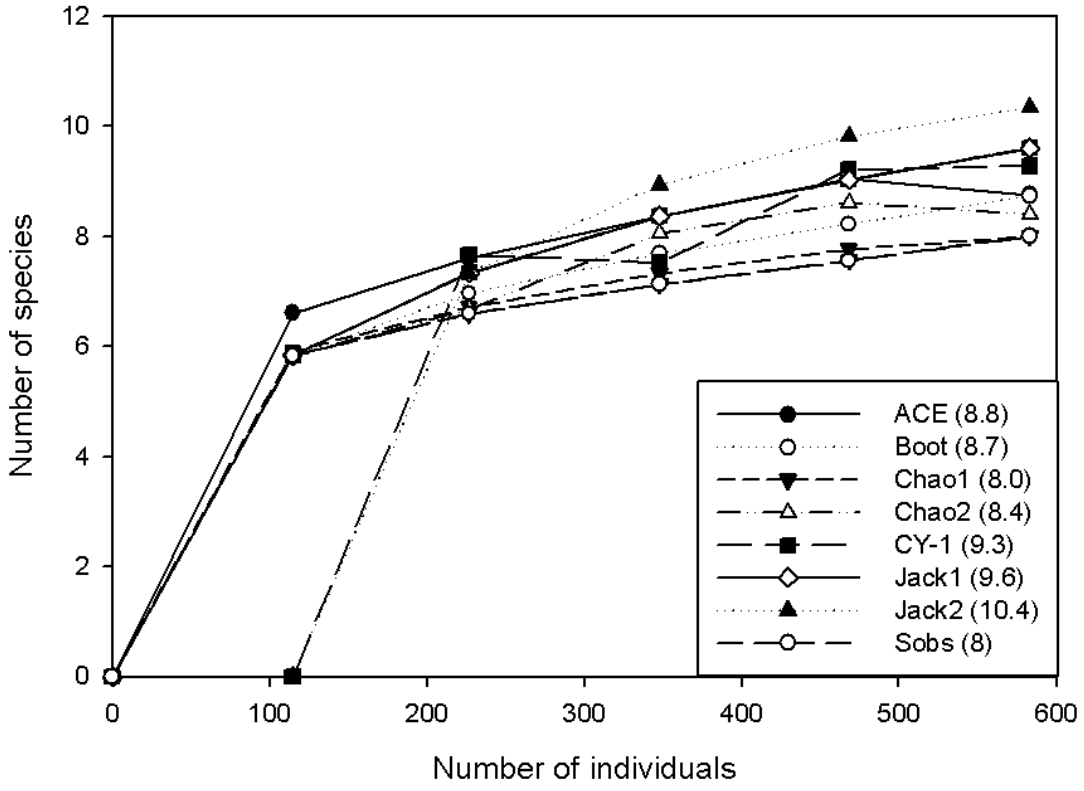


Fig. 3.2. Species accumulation curves for eight species richness estimators based on arboreal spider data from Luki Biosphere Reserve, DR Congo. At each possible survey size (five total), 200 surveys were randomly drawn without replacement, species richness was estimated, and estimates were averaged across the 200 randomizations. The x-axis is the average number of individuals contained in randomized surveys (each successive symbol on a line represents an increment of one in the number of randomized surveys).

LITERATURE CITED

Arrhenius O. 1921. Species and area. Journal of Ecology 9:95-99.

Baltanás A. 1992. On the use of some methods for the estimation of species richness. Oikos 65:484-492.

Brose U., Martinez N.D. and Williams R.J. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. Ecology 84:2364-2377.

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

Burnham K.P. and Overton W.S. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika 65:623-633.

Burnham K.P. and Overton W.S. 1979. Robust estimation of population size when capture probabilities vary among animals. Ecology 60:927-936.

Canning-Clode J., Valdivia N., Molis M., Thomason J.C. and Wahl M. 2008. Estimation of regional richness in marine benthic communities: quantifying the error. Limnology and Oceanography: Methods 6:580-590.

Cao Y., Larsen D.P. and Hughes R.M. 2001. Estimating total species richness in fish assemblage surveys: A similarity based approach. Canadian Journal of Fisheries and Aquatic Sciences 58:1782-1793.

Cao Y., Larsen D.P. and White D. 2004. Estimating regional species richness using a limited number of survey units. Ecoscience 11:23-35.

Chao A. 1984. Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11:265-270.

Chao A. 1987. Estimating the population size for capture-recapture data with unequal

catchability. Biometrics 43:783-791.

Chao A. and Lee S.M. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association 87:210-217.

Chazdon R.L., Colwell R.K., Denslow J.S. and Guariguata M.R. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rainforests of NE Costa Rica. Pages 185-309 *in* Dallmeier F. and Comiskey J.A. (eds.). *Forest biodiversity research, monitoring and modeling: conceptual background and old world case studies*. Parthenon Publishing Group, Paris, France.

Chiarucci A., Enright N.J., Perry G.L.W. and Miller B.P. 2003. Performance of nonparametric species richness estimators in a high diversity plant community. Diversity and Distributions 9:283-295.

Colwell R.K. 2006. *EstimateS*: Statistical estimation of species richness and shared species from samples. Version 8. Persistent URL <purl.oclc.org/estimates>. Published at: http://viceroy.eeb.uconn.edu/EstimateS.

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

Dahlberg M.D. and Odum E.P. 1970. Annual cycles of species occurrence, abundance, and diversity in Georgia estuarine fish populations. The American Midland Naturalist 83:382-392.

Dallmeier F., Foster R. B., Romano C., Rice R. and Kabel M. 1991. *A user's guide to the Beni Biosphere Reserve biodiversity plots.* Smithsonian Institution, Washington, DC., USA. 250 pp.

Deblauwe V., Barbier N., Couteron P., Lejeune O. and Bogaert J. 2008. The global

biogeography of semi-arid periodic vegetation patterns. Global Ecology and Biogeography 17:715-723.

Deyrup M. 2003. An updated list of Florida ants (Hymenoptera: Formicidae). Florida Entomologist 86:43-48.

Engen S. 1975. The coverage of a random sample from a biological community. Biometrics 31:201-208.

Fannes W., Bakker D.D., Loosveldt K. and Jocqué R. 2008. Estimating the diversity of arboreal oonopid spider assemblages (Araneae, Oonopidae) at Afrotropical sites. The Journal of Arachnology 36:322-330.

Fisher R.A., Corbet A.S. and Williams C.B. 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.J. 1996. *Biodiversity: a biology of numbers and difference*. Blackwell Science, Oxford, Massachusetts, USA.

Gotelli N.J. and Colwell R.K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4:379-391.

Holdridge L.R., Grenke W.C., Hatheway W.H., Liang T. and Tosi J.A. 1971. *Forest environments in tropical life zones*. Pergamon Press, Oxford, UK.

Jobe R.T. 2008. Estimating landscape-scale species richness: reconciling frequency- and turnover-based approaches. Ecology 89:174-182.

Keating K.A. and Quinn J.F. 1998. Estimating species richness: the Michaelis-Menton model revisited. Oikos 81:411-416.

Kéry M. and Plattner M. 2007. Species richness estimation and determinants of species

detectability in butterfly monitoring programmes. Ecological Entomology 32:53-61.

Kim J.K., Brick J.M, Fuller W.A. and Kalton G. 2006. On the bias of the multiple-imputation variance estimator in survey sampling. Journal of the Royal Statistical Society Series B, Series B (Statistical Methodology) 68:509-521.

King J.R. and Porter S.D. 2005. Evaluation of sampling methods and species richness estimators for ants in upland ecosystems in Florida. Environmental Entomology 34:1566-1578.

Laake J. and Rexstad E. 2008. RMark – an alternative approach to building linear models in MARK. Pages C-1-C-115. *in* Cooch E. and White G.C. (eds.). *Program MARK: 'A Gentle Introduction'*, 7th ed. Published at: http://www.phidot.org/software/mark/docs/book.

Lee S.M. and Chao A. 1994. Estimating population size via sample coverage for closed capture-recapture models. Biometrics 50:88-97.

Legendre P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659-1673.

Lewis T. and Taylor L.R. 1967. *Introduction to experimental ecology*. Academic Press, London, UK.

MacArther R.H. 1957. On the relative abundance of bird species. Proceedings of the National Academy of Science USA 43:293-295.

Magurran A.E. 2004. *Measuring Biological Diversity*. Blackwell Publishing, MA, USA.

Michaelis M. and Menten M.L. 1913. Der kinetic der invertinwirkung. Biochemische Zeitschrift 49:333-369.

Moreno C., Zuria I., García-Zenteno M., Sánchez-Rojas G., Castellanos I., Martínez-Morales M.

and Rojas-Martínez A. 2006. Trends in the measurement of alpha diversity in the last two decades. Interciencia 31:67-71.

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

Otis D.L., Burnham K.P., White G.C. and Anderson D.R. 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs 62:1-135.

Ott R.L. and Longnecker M.T. 2001. *An introduction to statistical methods and data analysis*. 5th edition. Duxbury Press, Pacific Grove, CA, USA.

Pagano A.M. and Arnold T.W. 2009. Estimating detection probabilities of waterfowl broods from ground-based surveys. Journal of Wildlife Management 73:686-694.

Palmer M.W. 1990. The estimation of species richness by extrapolation. Ecology 71:1195-1198.

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.

R Development Core Team. 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reese G.C., Wilson K.R., Hoeting J.A. and Flather C.H. 2005. Factors affecting species distribution predictions: a simulation modeling experiment. Ecological Applications 15:554-564.

SAS Institute. 1999. SAS/STAT user's guide, version 8. SAS Institute, Cary, North

Carolina, USA.

Selmi S. and Boulinier T. 2004. Distribution-abundance relationship for passerines breeding in Tunisian oases: test of the sampling hypothesis. Oecologia 139:440-445.

Smith E.P. and van Belle G. 1984. Nonparametric estimation of species richness. Biometrics 40:119-129.

Sugihara G. 1980. Minimal community structure: an explanation of species abundance patterns. The American Naturalist 116:770-787.

Tjørve E. 2009. Shapes and functions of species-area curves (II): a review of new models and parameterizations. Journal of Biogeography 36:1435-1445.

Ulrich W., Ollik M. and Ugland K.I. 2010. A meta-analysis of species-abundance distributions. Oikos 119:1149-1155.

Wagner H.H. and Wildi O. 2002. Realistic simulation of the effects of abundance distribution and spatial heterogeneity on non-parametric estimators of species richness. Ecoscience 9:241-250.

Walther B.A. and Martin J.L. 2001. Species richness estimation of bird communities: how to control for sampling effort? Ibis 143:413-419.

Walther B.A. and Moore J.L. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography 28:815-829.

Walther B.A. and Morand S. 1998. Comparative performance of species richness estimation methods. Parasitology 116:395-405.

White G.C. and Burnham K.P. 1999. Program MARK: Survival estimation from populations of marked animals. Bird Study 46 Supplement, pages 120-138.

Williams C.B. 1939. An analysis of four years' captures of insects in a light trap. Part I. Transactions of the Royal Entomological Society of London 89:79-132.

Williams C.B. 1940. An analysis of four years' captures of insects in a light trap. Part II. Transactions of the Royal Entomological Society of London 90:227-306.

CHAPTER 4

ESTIMATING THE VARIANCE OF SPECIES RICHNESS ESTIMATORS: VARIATIONS DUE TO ASSEMBLAGE CHARACTERISTICS AND SURVEY DESIGN

INTRODUCTION

Biodiversity is most often measured at the species level, being based on some combination of the number of unique species, (species richness), the distribution of abundances among species (evenness), and the identities of species (species composition) (Magurran 2004). Species richness (SR) is conceptually simple and the most frequently used biodiversity measure (Gaston 1996, Brose et al. 2003, Moreno et al. 2006). It has been used, for example, to compare habitats (Sweeney et al. 2010), design reserves (Woolhouse 1987, Chiarucci et al. 2005), investigate the effects of environmental and evolutionary factors (Zobel 1997, Mirkin et al. 2010), rapidly assess biodiversity (Obrist and Duelli 2010), and investigate the effects of various habitat uses and modifications (Proulx and Mazumder 1998, Suzart de Albuquerque and Rueda 2010). Unfortunately, a simple raw count of species typically underestimates the true number of species due to survey and sampling error, so SR is often estimated.

A point estimate such as SR is of limited value without some indication of its reliability. Analytical variance estimators (*Analyts*) have been derived for several SR estimators including most of those in program SimAssem (see Chapter 2 for program details and Appendix II for formulas). The *Analyts* are based on either the number of species with specific numbers of individuals encountered or the number of species encountered in specific numbers of surveys, as are the associated SR estimators. Despite being an important component of SR estimation

(Magurran 2004, p. 95), variance estimates are seldom evaluated and often not used. Based on a sample of 21 peer-reviewed articles, approximately one-half of studies using nonparametric estimators do not estimate or otherwise discuss variance (see Chapter 1).

Through repeated randomizations of a dataset, bootstrap and jackknife resampling procedures can be used to estimate the variance of an unknown parameter, θ. In bootstrap resampling (B_{Resamp}), sampling units are randomly drawn with replacement until a new sample equals the size of the original dataset (Efron 1979). Recommendations vary on the number of bootstrap samples to generate, but there should probably be ≥100 (Krebs 1999). Jackknife resampling (J_{Resamp}), as originally conceived, involved randomly resampling (without replacement) a dataset into halves (Quenouille 1949); it was later generalized to include resampling into groups of any size (Quenouille 1956). J_{Resamp} is probably most widely applicable when each sampling unit is successively removed, resulting in jackknife samples with one fewer unit than the original dataset (Miller 1974). For both B_{Resamp} and J_{Resamp}, resampled data can be used to estimate the average and variance of θ across replications (the term replicate is used hereafter to indicate runs that share certain assemblage properties; Krebs 1999, see also Miller 1964). A few of the *Analyts* as well as B_{Resamp} are commonly used to estimate variance, partly a result of the popularity of program EstimateS (see Colwell 2006).

Variance changes in proportion to the size of an estimate and, conceivably, so too could the performance of variance estimators (see Mao and Colwell 2005). If true, assemblage characteristics that affect SR estimates could also affect the performance of variance estimators. The number of species in an assemblage, species abundance distribution, survey effort, and to a lesser degree, density, have the largest effects on the performance of SR estimators (Wagner and

Wildi 2002, Brose et al. 2003, Chapter 3). Density affects encounter rate and similar results can be obtained by changing species detection probabilities.

In this study, I evaluated and compared the performances of three procedures for estimating the variance of SR estimators: analytical derivation, bootstrap resampling, and jackknife resampling. Since variance is partly a function of estimate size, a related objective was to test the performances of the three variance estimation procedures across several levels of the factors with the largest reported effects on SR estimators.

METHODS

I evaluated the performances of the *Analyts*, B_{Resamp}, and J_{Resamp} against datasets simulated with SimAssem. Data were collected across assemblages that varied in the true number of species (S_{true}), species abundance distribution (*Abund*), total abundance across all species (*N*), spatial configuration of individuals (*Config*), species detection probability (*p*), survey effort (*Effort*), and survey design (*Design*) (see Chapter 3). I used the data from the three levels of S_{true} (25, 100, 500), three types of *Abund* (log-normal, log-series, and particulate-niche), two levels of *Effort* (1%, 5% of the landscape), and two levels of *N* (6250, 12500). To reduce computer time, I used a subset of the data for the three other factors that included two levels of *p* (0.5 or 0.9 for all species) and the random levels of *Config* and *Design*. The variance estimation procedures were evaluated against the first 24 replications of each factor combination, totaling 1,728 realizations (the term realization is used hereafter to indicate any run, i.e., independent of assemblage properties).

For each realization, SR estimates were produced by 12 estimators including two based on abundance patterns, eight based on incidence patterns, and two based on the similarity of repeated subsets of surveys (see Table 4.1 for estimator descriptions and abbreviations and

Appendix II for formulas). I also applied the *Analyts* for the abundance-based coverage estimator (ACE; Chao and Lee 1992), Chao1 (Chao 1984), Chao2 (Chao 1987), and the first- through fifth-order jackknife estimators, Jack1-5 (Burnham and Overton 1978). There are three *Analyts* each for both of the Chao estimators where selection depends on characteristics of the survey data, e.g., numbers of species with exactly one and two individuals encountered. The *Analyt* for ACE was calculated by Species Prediction and Diversity Estimation (Chao and Shen 2009); the analogous estimator for ICE was not evaluated. Presently, there are no *Analyts* for CY-1 (Cao et al. 2001) and CY-2 (Cao et al. 2004), estimators that are based on the similarity of repeated subsets of surveys.

I also applied B_{Resamp} and J_{Resamp} to each dataset. A bootstrap sample was created by randomly selecting, with replacement, either 100 or 500 surveys, equal to the number in the original dataset. Each survey involved determining which individuals, if any, were encountered from one particular cell of an overlaid 100 x 100 grid. This was repeated until the number of bootstrap samples with ≥1 encounter equaled the number of surveys in the original dataset, again either 100 or 500. Jackknife samples were created by sequentially removing each survey from the original dataset, resulting in samples with either 99 or 499 surveys. The 12 SR estimators were applied to each bootstrap and jackknife sample and, for each individual SR estimator, variance was estimated by the standard deviation of all positive estimates (some estimators occasionally returned a negative estimate; see Chapter 3 for details). Therefore, there was a variable number of bootstrap or jackknife samples over which variance was computed.

Depending on the availability of an *Analyt*, each realization had either three or four estimates associated with each SR estimator including: 1) SR based on the full original dataset, 2) analytical standard deviation based on the full original dataset, 3) the standard deviation of all

positive estimates from bootstrap samples, and 4) the standard deviation of all positive estimates from jackknife samples. I removed all realizations in which any SR estimate was ≤0 and rebalanced the factorial design, resulting in 13 replicates per factorial combination. All comparisons were made within a single level of S_{true}, so as not to average standard deviation estimates across replicates with different means. Each set of replicates was summarized by the percent coefficient of variation of SR estimates (CV), the standard deviation of SR estimates, i.e., the empirical standard deviation (*Emp*), and the average of each of the three standard deviation estimation procedures (Fig. 4.1).

A confidence interval (CI) was constructed from every standard deviation estimate to assess reliability. I computed 95% CI's using a log-transformation of variance estimates that limits the lower bound to the number of species observed (S_{obs}; Burnham et al. 1987, Part 3) (see Appendix II for the formula). Confidence interval performance was summarized by the proportion containing S_{true}, i.e., the coverage level. Since a CI was based on both the SR and standard deviation estimates, a coverage level measured the combined performance of the two estimators. Average standard deviation estimates were used to isolate the performance of the standard deviation estimators. I also computed the proportion of estimates that were $<S_{obs}$ for each SR estimator.

RESULTS

Except for the CV of estimates from the third- through fifth-order jackknife estimators, variance increased with every increase in S_{true} (Table 4.2). The *Analyts*, B_{Resamp}, and J_{Resamp} almost always underestimated *Emp*, the lone exception being the *Analyt* for Jack5 in assemblages with $S_{true} = 100$. Of the three methods, the *Analyts* usually produced estimates closest to the *Emps* as well as 95% CI's with the largest coverage levels. However, B_{Resamp} always resulted in

the largest estimates of *Emp* and generally the largest coverage levels with Boot, this result is, therefore, only briefly mentioned in later comparisons. The reported estimates of J_{Resamp} and B_{Resamp} were averaged across replicates and thus, they too have associated variances (data not shown; the term variance will hereafter be used to refer to the variance across the samples of a resampling procedure). The variance of J_{Resamp} increased much more rapidly with S_{true} than the variance of B_{Resamp}, becoming more than two orders of magnitude larger for many of the SR estimators when $S_{true} = 500$. This contributed to J_{Resamp} occasionally resulting in larger average standard deviation estimates than B_{Resamp}, i.e., ACE, CY-1, CY-2, and ICE when $S_{true} = 500$, without translating into larger coverage levels.

Coverage levels were <0.80 for all estimator combinations and relationships with S_{true} were mostly negative (Table 4.2). In addition to providing the largest coverage levels at $S_{true} = 25$ and $S_{true} = 500$, the *Analyt* for ACE was the only standard deviation estimator that resulted in >50% coverage levels at all three levels of S_{true}; however, larger coverage levels were achieved by several other *Analyts* and B_{Resamp} when $S_{true} = 100$. Several estimators occasionally produced estimates $< S_{obs}$ in the 312 replicates with $S_{true} = 25$ including CY-2 (2 replicates), Jack2 (24), Jack3 (40), Jack4 (71), and Jack5 (84). The Jack2 (9), Jack3 (16), Jack4 (36), and Jack5 (57) estimators also produced estimates that were $< S_{obs}$ in the 312 replicates with $S_{true} = 100$.

Standard deviation estimates and coverage levels tended to be smallest with log-series distributions and largest with particulate-niche distributions (Appendix 4.A, Table 4.A.1). The *Analyts* generally produced larger estimates than B_{Resamp} and J_{Resamp} at each of the three levels of *Abund*. In assemblages with particulate-niche distributions and $S_{true} = 25$, B_{Resamp} resulted in the largest average standard deviation estimates for Chao1, Chao2, and Jack2. J_{Resamp} estimates were larger than those from B_{Resamp} only in assemblages with $S_{true} = 500$ and log-normal (ACE, CY-2,

and ICE) or particulate-niche (ACE, Chao1, Chao2, CY-1, CY-2, and ICE) distributions. The only standard deviation estimates that exceeded *Emp* were the *Analyts* for Jack3, Jack4, and Jack5 in assemblages with S_{true} = 25 and either log-normal or log-series distributions and the *Analyts* again for Jack4 and Jack5 in assemblages with S_{true} = 100 and log-normal or particulate-niche distributions. A larger standard deviation estimate, as before, did not necessarily achieve larger coverage levels and, even when the B_{Resamp} estimate was larger, the *Analyts* generally resulted in CI's with the largest coverage level (but see B_{Resamp} with Jack5 in assemblages with S_{true} = 500 and log-normal or particulate-niche distributions). The 95% CI's based on the *Analyt* for ACE performed relatively well in all factor combinations and resulted in >50% in all assemblages except those with log-series abundance distributions and either 100 or 500 species. Only the *Analyt* for Jack5 resulted in >40% coverage levels at all levels of S_{true} in assemblages with log-series distributions.

Coverage levels were generally positively related to *Effort*, but there were numerous exceptions (Appendix 4.A, Table 4.A.2). Other than with Boot and Jack4 in assemblages with 100 surveys and 500 species, the *Analyts* produced the largest standard deviation estimates and coverage levels. The only instances where J_{Resamp} resulted in larger estimates and coverage levels than B_{Resamp} were in assemblages with S_{true} = 500 and *Effort* = 100. The *Analyt* for ACE resulted in the largest coverage at both effort levels when S_{true} = 25 and with 100 surveys when S_{true} = 500; otherwise, the *Analyt* for one of the higher-order jackknifes resulted in the largest coverage level. Coverage levels for both resampling procedures were relatively small for all SR estimators when *Effort* = 100 and S_{true} = 500.

Relationships between performance and *N* were inconsistent across SR estimators, even varying across levels of S_{true} for a single SR estimator (Appendix 4.A, Table 4.A.3). The *Analyts*

produced the largest estimates and coverage levels for all except Boot and the coverage levels of J_{Resamp} were never greater than those of B_{Resamp}, despite several larger standard deviation estimates when $S_{true} = 500$. The instances where a coverage level exceeded that of the *Analyt* for ACE (which was always >50%) were few, i.e., the *Analyts* for Chao1 and Chao2 when $N = 12500$ and $S_{true} = 100$, as well as the *Analyts* and occasionally B_{Resamp} for Jack3, Jack4, and/or Jack5 either when $N = 6250$ or 12500 and $S_{true} = 100$ or when $N = 12500$ and $S_{true} = 500$.

Increasing the average p often resulted in smaller standard deviation estimates and larger coverage levels (Appendix 4.A, Table 4.A.4). As in previous comparisons, the *Analyts* resulted in the largest standard deviation estimates and coverage levels except for the *Analyt* for Boot. A coverage level of J_{Resamp} exceeded that of B_{Resamp} only once, when $p = 0.5$ and $S_{true} = 500$ for CY-2. The *Analyt* for ACE resulted in coverage levels >50% in all factor combinations except those with $S_{true} = 500$ and $p = 0.9$, where less biased levels were achieved by the *Analyt* and B_{Resamp} for Jack5. The effects of p and N were similar, such that the coverage levels of the *Analyt* for ACE were exceeded in the same comparisons if the small and large values of p are replaced with those of N.

DISCUSSION

Standard deviation estimators for species richness estimates have received little attention despite the importance that a reliability measure would have to disciplines such as conservation biology and conservation management. I used survey data from systematically varied simulations to test three standard deviation estimation methods including bootstrap and jackknife resampling and, when available, analytically derived estimators. All three methods generally underestimated *Emp*. On average, the *Analyts* produced the least biased estimates of *Emp*. Only average estimates from an *Analyt* of the higher-order jackknife estimators ever exceeded *Emp*

and only in assemblages with fewer than 500 species. The standard deviation of the SR estimators (*Emp*) and the variance of the standard deviation estimators both increased with S_{true}, a combination that could reduce correct estimation. Of the two resampling procedures, B_{Resamp} usually exhibited a smaller negative bias, but there were exceptions in assemblages with 500 species.

The larger of two average standard deviation estimates did not always achieve a less biased coverage level. When a larger estimate involves more variance, it can result in a greater number of CI's that either completely underestimate or overestimate S_{true}. Most often, the *Analyts* resulted in coverage levels greater than those from the resampling procedures. Only the *Analyt* for ACE exceeded the nominal 95% confidence level and only in assemblages with S_{true} = 500 and particulate-niche distributions. Other than with Boot, the only instance where a resampling procedure achieved a larger coverage level than an *Analyt* was B_{Resamp} with Jack4 or Jack5 (see Appendix 4.A, Tables 4.A.1, 4.A.2, and 4.A.4). A smaller coverage level can also result from a larger standard deviation estimate involving less variance. This could occur when variation in the standard deviation estimator is not large enough to regularly overcome a biased estimate. Thus, variance was found to affect performance in both directions. I am unaware of other studies that have assessed the performance of standard deviation estimators with species richness, but in the context of population estimation it has been reported that CI coverage was poor without some modification of the standard deviation estimates (Stanley and Burnham 1998, Walsh et al. 2009).

There was often a negative relationship between coverage levels and S_{true}. Similar relationships have occurred elsewhere (see Otis et al. 1978, Wilson and Anderson 1985) and could result from several factors. First, the SR estimators exhibited more variation with larger

values of S_{true}, even when based on the CV (Table 4.2). The variance of the standard deviation estimators similarly increased with S_{true}. Furthermore, the SR estimators were more negatively biased with larger values of S_{true} (Chapter 3). The increased negative bias of SR estimators is, by itself, enough to cause a negative relationship between coverage levels and S_{true}, but the relationship could be further heightened by increases in variance (both the SR and standard deviation estimators). Increasing the amount of encounter data via *Effort*, *N*, and *p* did not consistently decrease the variance of the standard deviation estimators (data not shown) or the achieved coverage of the CI's, again indicating the complexity of the effects of variance (see Appendix 4.A, Tables 4.A.2, 4.A.3, and 4.A.4).

In contrast to simulation studies, researchers in a real environment are usually restricted to single estimates of SR and the associated standard deviation. A negatively biased standard deviation estimator presents a troubling situation because, with no estimate of bias, there is a tendency to assign more weight to an apparently precise estimate. Estimates of the standard deviation of SR estimators appear quite variable and thus, the reliability of a single CI estimate will be questionable. To illustrate, suppose that an estimated 95% confidence interval is 60-80 species for an area which, if calculated with the true larger standard deviation, becomes 40-100 species. If one also considers the possibility that the CI was centered on an underestimate of S_{true}, then it becomes apparent that the true number of species could be much larger than 100.

This study shows that the standard deviation estimators are, similar to the SR estimators, affected by assemblage factors, which is understandable given that both are based on the same data summaries, e.g., number of species encountered once and twice. Both positive and negative relationships were regularly found between the coverage level of 95% CI's and each of the factors *Effort*, *N*, and *p*. Inconsistent estimator performance across assemblages can pose

problems when conducting a comparative analysis, particularly if an estimator is used to rank compare assemblages that differ in the very properties that greatly affect estimator performance. There is obviously room to improve both the SR and standard deviation estimators. Some of the *Analyts* produced estimates at a relatively regular proportion of *Emp*. Thus, it might be possible to reduce bias by applying a multiplier, possibly by using performance as a function of measured or estimated assemblage properties. Programs such as SimAssem that allow exploration of estimator performance under various scenarios can help address these issues and test new estimators.

Recent studies have shown that there are SR estimators that are less biased and more accurate, though less precise, than the more frequently used estimators Chao1, Chao2, Jack1, and Jack2 (Brose et al. 2003, Chapter 3). With knowledge of assemblage properties, the selection frameworks of Brose et al. (2003) and Chapter 3 can still increase the chance of reliable estimation. In addition to selecting and using a SR estimator based on its likely bias or precision, it is important to include a measure of reliability such as a CI. This study indicates that the *Analyts* are the least negatively biased option over a wide range of conditions. Without knowledge of assemblage properties, I recommend using ACE (and presumably ICE) with the associated *Analyt* which were found to regularly achieve the largest, albeit negatively biased, CI coverage levels. B_{Resamp} could be useful for estimating the standard deviation of a SR estimate when no *Analyt* exists, e.g., for CY-1 and CY-2, at least in assemblages with a small number of species because that is where the procedure appears to perform relatively well (see Table 4.2). However, this study indicates that resampling procedures are less effective than most *Analyts*, resulting in smaller coverage levels. It could be worth testing what effect increasing the number

of samples has on B_{Resamp}, but evidence suggests that no additional amount of resampling would bring it on par with the *Analyts* (see Efron 1979).

Table 4.1. Abbreviations and categorizations of the tested species richness estimators.

Species richness estimator	**Abbreviation**	**Category**	**Citation**
Abundance-based coverage	ACE	M_h[1]	Chao and Lee 1992
Bootstrap	Boot	M_h[1]	Smith and van Belle 1984
Chao1 (bias-corrected)	Chao1	M_h[1]	Chao 1984; Colwell 2006
Chao2 (bias-corrected)	Chao2	M_h[1]	Chao 1987; Colwell 2006
CY-1	CY-1	Similarity[2]	Cao et al. 2001
CY-2	CY-2	Similarity[2]	Cao et al. 2004
Incidence-based coverage	ICE	M_h[1]	Lee and Chao 1994
1st-order jackknife	Jack1	M_h[1]	Burnham and Overton 1978
2nd-order jackknife	Jack2	M_h[1]	Burnham and Overton 1978
3rd-order jackknife	Jack3	M_h[1]	Burnham and Overton 1978
4th-order jackknife	Jack4	M_h[1]	Burnham and Overton 1978
5th-order jackknife	Jack5	M_h[1]	Burnham and Overton 1978

[1]Estimation involves modeling heterogeneity in detection probability of species.
[2]Estimation involves use of similarity of replicate surveys of species with Jaccard's coefficient.

Table 4.2. Empirical and estimated variance of 12 species richness (SR) estimators. Each level of the true number of species (S_{true}) is comprised of 312 realizations or 13 replications of all combinations of a 2x2x2x3 factorial design based on effort, total abundance, detection probability, and abundance distribution, respectively. Results include the standard deviation of SR estimates across the 312 replicates (*Emp*) and the mean variance estimates from analytical estimators (*Analyt*) and bootstrap (B_{Resamp}) and jackknife (J_{Resamp}) resampling procedures. The proportion of 95% confidence intervals that included S_{true} for *Analyt*, B_{Resamp}, and J_{Resamp} are listed as CovA, CovB, and CovJ, respectively. The proportion of SR estimates less than the observed number of species is listed under $<S_{obs}$.

S_{true}	Estimator	CV	*Emp*	*Analyt*	B_{Resamp}	J_{Resamp}	CovA	CovB	CovJ	$<S_{obs}$
25	**ACE**	25.91	5.69	2.99	1.99	0.43	0.78	0.67	0.28	0.00
	Boot	27.57	5.67	0.97	1.23	0.19	0.28	0.34	0.06	0.00
	Chao1	27.70	5.90	2.77	2.09	0.55	0.73	0.63	0.34	0.00
	Chao2	27.85	5.93	2.74	2.09	0.56	0.72	0.63	0.34	0.00
	Jack1	25.69	5.72	2.20	1.79	0.32	0.66	0.53	0.27	0
	Jack2	26.18	6.16	3.81	3.26	0.66	0.72	0.64	0.25	0.08
	Jack3	31.41	7.79	6.69	5.81	1.34	0.67	0.62	0.30	0.13
	Jack4	44.55	11.98	11.67	9.36	2.62	0.53	0.51	0.32	0.23
	Jack5	68.87	20.94	20.86	15.16	4.80	0.43	0.39	0.27	0.27
	CY-1	32.10	7.76		2.02	0.44		0.33	0.12	0.00
	CY-2	29.26	7.36		2.62	0.73		0.41	0.13	0.01
	ICE	25.97	5.70		2.00	0.43		0.66	0.28	0.00
100	**ACE**	32.82	26.32	13.90	6.19	2.25	0.65	0.43	0.12	0.00
	Boot	42.07	26.35	2.57	3.26	0.46	0.13	0.19	0.01	0.00
	Chao1	34.09	26.24	13.78	6.92	2.55	0.65	0.43	0.13	0.00
	Chao2	33.91	26.05	13.70	6.92	2.53	0.66	0.44	0.12	0.00
	Jack1	36.42	26.69	6.12	4.70	0.75	0.29	0.25	0.04	0.00
	Jack2	31.16	25.69	10.60	8.29	1.43	0.50	0.39	0.08	0.03
	Jack3	30.41	26.71	17.37	14.47	2.62	0.75	0.67	0.14	0.05
	Jack4	35.30	32.41	28.48	23.83	4.78	0.79	0.71	0.20	0.12
	Jack5	49.27	46.91	47.27	36.64	8.52	0.71	0.68	0.27	0.18
	CY-1	45.37	47.98		7.35	3.03		0.19	0.05	0.00
	CY-2	42.32	45.00		7.87	3.93		0.21	0.08	0.00
	ICE	32.64	26.22		6.24	2.25		0.44	0.11	0.00
500	**ACE**	54.17	199.70	120.55	15.03	18.10	0.53	0.14	0.11	0.00
	Boot	65.23	100.96	4.90	6.35	0.78	0.00	0.00	0.00	0.00
	Chao1	45.40	143.14	75.16	15.59	14.06	0.44	0.12	0.07	0.00
	Chao2	45.42	142.98	74.95	15.55	14.06	0.44	0.13	0.08	0.00
	Jack1	61.50	124.84	12.32	9.12	1.22	0.03	0.02	0.00	0.00
	Jack2	56.26	147.12	21.34	15.52	2.15	0.12	0.10	0.02	0.00
	Jack3	51.79	157.90	32.28	26.07	3.53	0.18	0.13	0.02	0.00
	Jack4	48.46	163.78	47.19	42.23	5.63	0.31	0.28	0.03	0.00
	Jack5	46.89	170.67	68.73	64.65	8.91	0.47	0.46	0.05	0.00
	CY-1	55.18	309.31		20.02	20.36		0.12	0.08	0.00
	CY-2	63.59	363.30		19.88	34.44		0.13	0.12	0.00
	ICE	54.24	200.83		15.10	18.54		0.14	0.11	0.00

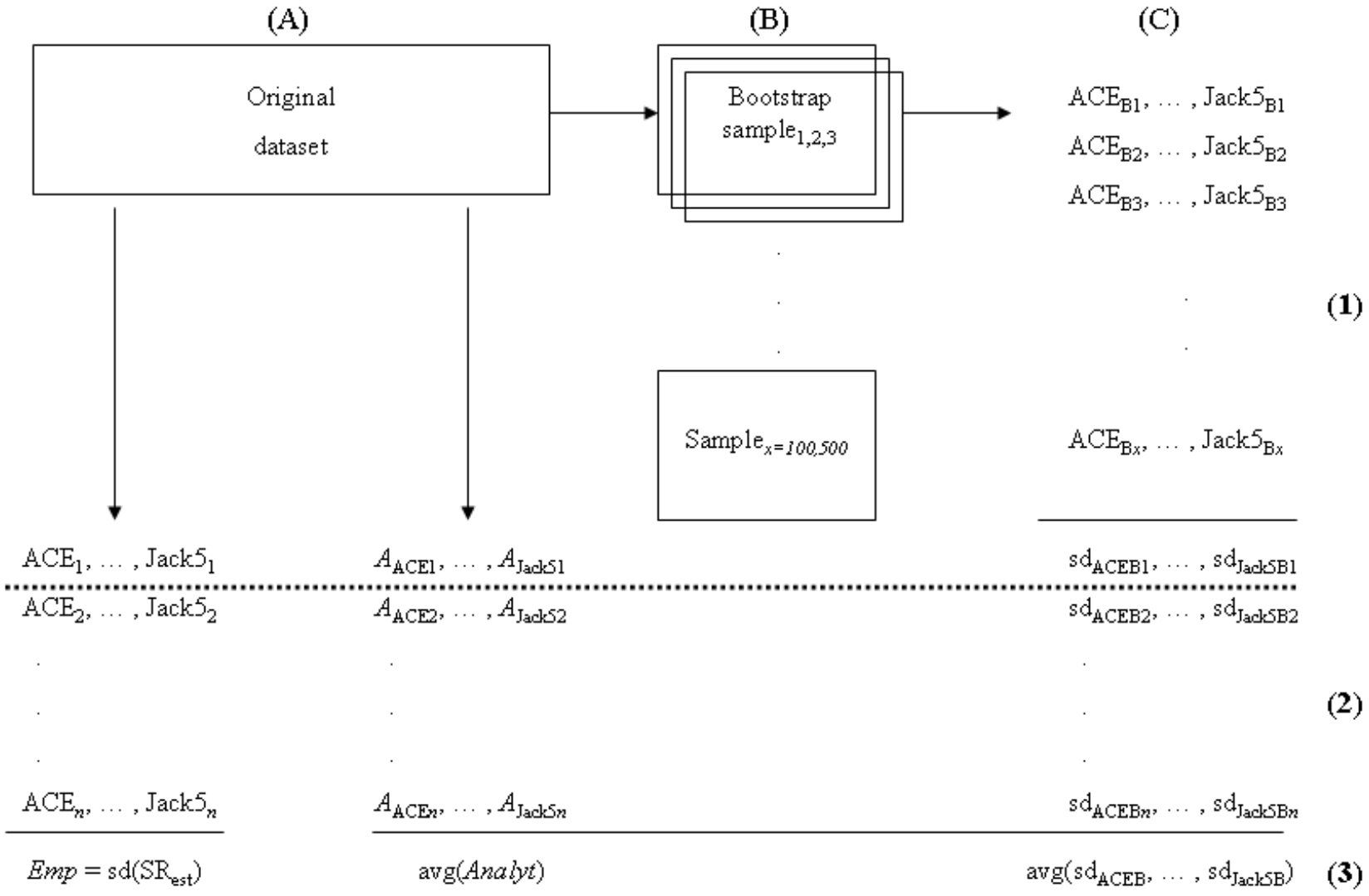


Fig. 4.1. The steps involved in estimating species richness (SR) and variance both for a single replicate (above dashed line) and all replicates in a set. A box indicates a sample dataset, i.e., a collection of surveys. In each replicate, the SR and variance estimators are applied directly to each original dataset (column A), indicated by ACE_1, …, $Jack5_1$ and A_{ACE1}, …, A_{Jack51}, respectively (the subscript,

here 1, indicates the replicate). For bootstrap resampling, each original dataset is resampled with replacement (column B) a number of times equal to the number of surveys in the original dataset (x = 100 or 500). SR estimators are applied to each bootstrap sample (column C; ACE_{B1}, …, $Jack5_{B1}$) and the standard deviation across the x samples represents the variance estimate (sd_{ACEB1}, …, $sd_{Jack5B1}$). Section 2 (below dashed line) indicates that the steps are repeated for each replicate in a set. Final variance estimates (below solid line) are based on the n replicates, including the standard deviation of each SR estimator [Emp = $sd(SR_{est})$], i.e., empirical variance, and the average of each variance estimator [avg($Analyt$) and avg(sd_{ACEB}, …, sd_{Jack5B})]. Jackknife resampling would involve steps similar to those in columns B and C, with jackknife samples created by the successive removal of each survey in column B.

LITERATURE CITED

Brose U., Martinez N.D. and Williams R.J. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. Ecology 84:2364-2377.

Burnham K.P., Anderson D.R., White G.C., Brownie C. and Pollock K.H. 1987. Design and analysis methods for fish survival experiments based on release-recapture. American Fisheries Society Monograph No. 5. Bethesda, MD, USA. 437pp.

Burnham K.P. and Overton W.S. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika 65:623-633.

Cao Y., Larsen D.P. and Hughes R.M. 2001. Estimating total species richness in fish assemblage surveys: a similarity based approach. Canadian Journal of Fisheries and Aquatic Sciences 58:1782-1793.

Cao Y., Larsen D.P. and White D. 2004. Estimating regional species richness using a limited number of survey units. Ecoscience 11:23-35.

Chao A. 1984. Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11:265-270.

Chao A. 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783-791.

Chao A. and Lee S.M. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association 87:210-217.

Chao A. and Shen T.-J. 2009. Program SPADE (Species prediction and diversity estimation). Program and user's guide. Published at: http://chao.stat.nthu.edu.tw.

Chiarucci A., D'Auria F., De Dominicis V., Laganá A., Perini C. and Salerni E. 2005.

Using vascular plants as a surrogate taxon to maximize fungal species richness in reserve design. Conservation Biology 19:1644-1652.

Colwell R.K. 2006. *EstimateS* Statistical estimation of species richness and shared species from samples. Version 8. Persistent URL <purl.oclc.org/estimates>. Published at: http://viceroy.eeb.uconn.edu/EstimateS.

Efron B. 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7:1-26.

Gaston K.J. 1996. *Biodiversity: a biology of numbers and difference*. Blackwell Science, Oxford, Massachusetts, USA.

Krebs C.J. 1999. *Ecological Methodology*. Benjamin/Cummings, Menlo Park, CA, USA.

Lee S.M. and Chao A. 1994. Estimating population size via sample coverage for closed capture-recapture models. Biometrics 50:88-97.

Magurran A.E. 2004. *Measuring Biological Diversity*. Blackwell Publishing, MA, USA.

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

Miller R.G. 1964. A trustworthy jackknife. Annals of Mathematical Statistics 35:1594-1605.

Miller R.G. 1974. The jackknife - a review. Biometrika 61:1-15.

Mirkin B.M., Shirokikh P.S., Martynenko V.B. and Naumova L.G. 2010. Analysis of trends in the formation of species richness of plant communities using syntaxonomy and ecological scales. Russian Journal of Ecology 41:279-283.

Moreno C., Zuria I., García-Zenteno M., Sánchez-Rojas G., Castellanos I., Martínez-Morales M.

and Rojas-Martínez A. 2006. Trends in the measurement of alpha diversity in the last two decades. Interciencia 31:67-71.

Obrist M.K. and Duelli P. 2010. Rapid biodiversity assessment of arthropods for monitoring average local species richness and related ecosystem services. Biodiversity and Conservation 19:2201-2220.

Otis D.L., Burnham K.P., White G.C. and Anderson D.R. 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs 62:1-135.

Proulx M. and Mazumder A. Reversal of grazing impact on plant species richness in nutrient-poor vs. nutrient-rich ecosystems. Ecology 79:2581-2592.

Quenouille M.H. 1949. Approximate tests of correlation in time series. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 11:68-84.

Quenouille M.H. 1956. Notes on bias in estimation. Biometrika 43:353-360.

Smith E.P. and van Belle G. 1984. Nonparametric estimation of species richness. Biometrics 40:119-129.

Stanley T.R. and Burnham K.P. 1998. Information-theoretic model selection and model averaging for closed-population capture-recapture studies. Biometrical Journal 40:475-494.

Suzart de Albuquerque F. and Rueda M. 2010. Forest loss and fragmentation effects on woody plant species richness in Great Britain. Forest Ecology and Management 260:472-479.

Sweeney O.F.McD., Wilson M.W., Irwin S., Kelly T.C. and O'Halloran J. 2010. Are bird

density, species richness and community structure similar between native woodlands and non-native plantations in an area with a generalist bird fauna? Biodiversity Conservation 19:2329-2342.

Wagner H.H. and Wildi O. 2002. Realistic simulation of the effects of abundance distribution and spatial heterogeneity on non-parametric estimators of species richness. Ecoscience 9:241-250.

Walsh D.P., Page C.F., Campa III H., Winterstein S.R. and Beyer Jr. D.E. 2009. Incorporating estimates of group size in sightability models for wildlife. The Journal of Wildlife Management 73:136-143.

Wilson K.R. and Anderson D.R. 1985. Evaluation of a density estimator based on a trapping web and distance sampling theory. Ecology 66:1185-1194.

Woolhouse M.E.J. 1987. On species richness and nature reserve design: an empirical study of UK woodland avifauna. Biological Conservation 40:167-178.

Zobel M. 1997. The relative role of species pools in determining plant species richness: an alternative explanation of species coexistence? Trends in Ecology and Evolution 12:266-269.

APPENDIX 4.A

Table 4.A.1. Empirical and estimated variance of 12 species richness (SR) estimators for three levels of species abundance distribution (*Abund*), log-normal (LN), log-series (LS), and particulate-niche (PN). Each level of *Abund* is comprised of 104 realizations or 13 replications of all combinations of a 2x2x2 factorial design based on effort, total abundance, and detection probability. Results include the standard deviation of SR estimates across the 104 replicates (*Emp*) and the mean variance estimates from analytical estimators (*Analyt*) and bootstrap (B_{Resamp}) and jackknife (J_{Resamp}) resampling procedures. The proportion of 95% confidence intervals that included the true number of species (S_{true}) for *Analyt*, B_{Resamp}, and J_{Resamp} are listed as CovA, CovB, and CovJ, respectively. The proportion of SR estimates less than the observed number of species is listed as $<S_{obs}$.

Abund	S_{true}	Estimator	%CV	*Emp*	*Analyt*	B_{Resamp}	J_{Resamp}	CovA	CovB	CovJ	$<S_{obs}$
LN	**25**	**ACE**	15.48	3.69	2.69	2.07	0.44	0.78	0.76	0.19	0.00
		Boot	15.76	3.58	1.09	1.37	0.20	0.54	0.64	0.09	0.00
		Chao1	16.53	3.84	2.75	2.41	0.60	0.74	0.70	0.29	0.00
		Chao2	16.19	3.76	2.75	2.40	0.61	0.73	0.71	0.30	0.00
		Jack1	14.51	3.57	2.44	1.99	0.34	0.74	0.67	0.30	0.00
		Jack2	17.44	4.50	4.22	3.66	0.71	0.61	0.58	0.22	0.10
		Jack3	24.54	6.61	7.51	6.56	1.45	0.52	0.46	0.26	0.17
		Jack4	40.08	11.48	13.34	10.86	2.91	0.36	0.31	0.23	0.32
		Jack5	67.51	21.23	24.13	17.57	5.36	0.29	0.27	0.17	0.36
		CY-1	17.09	4.53		2.19	0.45		0.63	0.20	0.00
		CY-2	16.23	4.35		2.51	0.60		0.62	0.18	0.00
		ICE	15.52	3.70		2.08	0.44		0.75	0.19	0.00
LS	**25**	**ACE**	35.10	5.83	4.42	2.27	0.53	0.76	0.43	0.05	0.00
		Boot	24.53	3.43	0.98	1.25	0.19	0.01	0.05	0.00	0.00
		Chao1	36.24	5.61	3.72	1.95	0.63	0.62	0.37	0.06	0.00
		Chao2	36.18	5.57	3.64	1.94	0.65	0.61	0.37	0.06	0.00
		Jack1	25.12	3.99	2.51	1.78	0.31	0.52	0.22	0.00	0.00
		Jack2	28.90	5.19	4.34	3.11	0.58	0.92	0.74	0.07	0.02
		Jack3	35.99	7.14	6.93	5.29	1.08	0.88	0.84	0.16	0.07
		Jack4	48.63	10.63	11.00	7.95	1.96	0.75	0.73	0.34	0.15
		Jack5	70.25	17.15	17.99	12.45	3.45	0.59	0.53	0.32	0.23
		CY-1	26.32	4.29		1.79	0.34		0.17	0.03	0.00
		CY-2	39.17	7.56		3.29	1.12		0.54	0.19	0.02
		ICE	34.95	5.79		2.25	0.52		0.42	0.06	0.00
PN	**25**	**ACE**	9.16	2.33	1.79	1.64	0.33	0.81	0.81	0.60	0.00
		Boot	8.24	2.06	0.85	1.08	0.17	0.30	0.34	0.09	0.00
		Chao1	9.46	2.38	1.83	1.92	0.43	0.84	0.83	0.67	0.00
		Chao2	10.15	2.56	1.84	1.93	0.43	0.84	0.83	0.67	0.00
		Jack1	9.84	2.59	1.66	1.60	0.30	0.71	0.69	0.52	0.00
		Jack2	16.74	4.49	2.88	3.02	0.68	0.63	0.61	0.47	0.12
		Jack3	25.83	7.14	5.63	5.59	1.48	0.60	0.56	0.47	0.14
		Jack4	40.49	12.21	10.67	9.28	3.00	0.49	0.48	0.38	0.21
		Jack5	64.29	22.74	20.45	15.46	5.60	0.40	0.37	0.33	0.22
		CY-1	22.30	6.63		2.08	0.52		0.18	0.13	0.00
		CY-2	19.22	5.65		2.06	0.48		0.09	0.03	0.00
		ICE	9.24	2.35		1.66	0.33		0.82	0.60	0.00

LN	**100**	**ACE**	22.74	19.62	15.37	6.70	2.39	0.84	0.56	0.12	0.00
		Boot	35.67	23.50	2.73	3.46	0.49	0.15	0.20	0.01	0.00
		Chao1	27.02	22.03	14.80	7.32	2.90	0.78	0.57	0.12	0.00
		Chao2	26.99	21.97	14.69	7.31	2.87	0.79	0.57	0.11	0.00
		Jack1	29.64	23.07	6.57	4.97	0.79	0.49	0.43	0.06	0.00
		Jack2	23.92	21.15	11.38	8.73	1.49	0.71	0.60	0.12	0.00
		Jack3	23.70	22.34	18.45	15.14	2.72	0.88	0.81	0.18	0.01
		Jack4	30.55	29.83	29.92	24.83	4.93	0.83	0.77	0.27	0.11
		Jack5	47.35	46.98	49.05	37.89	8.83	0.73	0.72	0.39	0.16
		CY-1	23.09	25.03		7.59	2.98		0.53	0.13	0.00
		CY-2	27.96	30.91		8.25	4.15		0.49	0.18	0.00
		ICE	22.32	19.25		6.75	2.39		0.57	0.10	0.00
LS	**100**	**ACE**	26.58	14.23	10.10	4.81	1.24	0.24	0.01	0.00	0.00
		Boot	31.70	13.45	2.07	2.62	0.38	0.00	0.00	0.00	0.00
		Chao1	40.70	22.20	12.57	5.83	2.06	0.33	0.03	0.00	0.00
		Chao2	40.43	22.08	12.60	5.86	2.08	0.34	0.03	0.00	0.00
		Jack1	28.76	14.45	5.24	3.76	0.60	0.00	0.00	0.00	0.00
		Jack2	27.42	16.05	9.07	6.59	1.12	0.18	0.08	0.01	0.00
		Jack3	29.70	19.35	14.37	11.40	2.04	0.71	0.56	0.03	0.00
		Jack4	36.87	26.38	22.79	18.32	3.74	0.92	0.79	0.09	0.04
		Jack5	50.89	39.92	37.12	27.76	6.78	0.87	0.85	0.22	0.12
		CY-1	19.79	11.58		4.79	1.18		0.00	0.00	0.00
		CY-2	26.46	16.80		5.96	1.87		0.10	0.02	0.00
		ICE	26.24	14.09		4.85	1.25		0.01	0.00	0.00
PN	**100**	**ACE**	17.68	17.83	16.19	7.07	3.12	0.88	0.73	0.23	0.00
		Boot	32.05	25.50	2.90	3.71	0.51	0.25	0.36	0.03	0.00
		Chao1	16.66	15.80	13.98	7.62	2.67	0.85	0.70	0.27	0.00
		Chao2	16.53	15.63	13.81	7.60	2.63	0.85	0.71	0.25	0.00
		Jack1	24.57	22.55	6.55	5.38	0.85	0.37	0.31	0.07	0.00
		Jack2	18.71	18.77	11.35	9.56	1.66	0.61	0.50	0.13	0.09
		Jack3	20.44	21.28	19.30	16.87	3.11	0.66	0.63	0.20	0.14
		Jack4	28.65	30.44	32.74	28.34	5.69	0.61	0.57	0.24	0.20
		Jack5	45.16	48.77	55.65	44.27	9.96	0.52	0.47	0.19	0.27
		CY-1	29.19	43.88		9.67	4.94		0.05	0.02	0.00
		CY-2	26.72	38.73		9.39	5.79		0.05	0.03	0.00
		ICE	17.80	17.98		7.11	3.13		0.73	0.23	0.00
LN	**500**	**ACE**	46.47	162.81	109.91	14.88	15.70	0.50	0.04	0.09	0.00
		Boot	60.71	92.63	4.87	6.28	0.79	0.00	0.00	0.00	0.00
		Chao1	36.76	109.03	68.41	15.32	12.90	0.36	0.00	0.01	0.00
		Chao2	36.74	108.69	68.07	15.28	12.85	0.37	0.00	0.01	0.00
		Jack1	56.24	112.12	12.24	9.00	1.23	0.00	0.00	0.00	0.00
		Jack2	50.50	129.19	21.20	15.33	2.14	0.11	0.09	0.00	0.00
		Jack3	46.25	137.53	32.12	25.85	3.48	0.26	0.21	0.04	0.00
		Jack4	43.61	143.37	47.09	42.08	5.56	0.41	0.38	0.03	0.00
		Jack5	43.25	152.44	68.85	64.47	8.92	0.48	0.49	0.07	0.00
		CY-1	35.48	173.18		19.05	17.03		0.29	0.18	0.00
		CY-2	52.56	266.05		19.09	24.97		0.32	0.18	0.00
		ICE	46.48	163.29		14.94	15.82		0.04	0.09	0.00

LS	**500**	**ACE**	32.33	73.44	52.12	12.21	8.07	0.10	0.00	0.00	0.00
		Boot	52.83	64.92	4.22	5.45	0.71	0.00	0.00	0.00	0.00
		Chao1	37.04	77.59	44.28	12.74	7.34	0.06	0.01	0.00	0.00
		Chao2	37.05	77.42	44.05	12.73	7.30	0.06	0.01	0.00	0.00
		Jack1	48.78	77.05	10.73	7.80	1.13	0.00	0.00	0.00	0.00
		Jack2	44.35	88.82	18.58	13.32	2.03	0.00	0.00	0.00	0.00
		Jack3	41.84	97.48	28.29	22.52	3.45	0.01	0.01	0.01	0.00
		Jack4	41.14	106.92	42.01	36.62	5.71	0.16	0.10	0.00	0.00
		Jack5	43.02	121.75	62.87	55.96	9.42	0.44	0.40	0.06	0.00
		CY-1	26.71	82.18		15.01	10.36		0.01	0.04	0.00
		CY-2	27.37	83.18		15.24	13.65		0.03	0.07	0.00
		ICE	32.03	73.04		12.31	8.23		0.00	0.00	0.00
PN	**500**	**ACE**	38.87	205.42	199.62	18.00	30.51	0.98	0.39	0.23	0.00
		Boot	66.51	125.60	5.59	7.33	0.84	0.00	0.00	0.00	0.00
		Chao1	29.31	128.94	112.80	18.72	21.96	0.89	0.35	0.21	0.00
		Chao2	29.24	128.55	112.71	18.65	22.03	0.89	0.37	0.23	0.00
		Jack1	61.71	155.25	14.00	10.57	1.31	0.09	0.07	0.00	0.00
		Jack2	54.83	180.08	24.25	17.91	2.28	0.24	0.20	0.07	0.00
		Jack3	48.86	187.72	36.43	29.84	3.66	0.26	0.17	0.01	0.00
		Jack4	44.04	187.28	52.47	47.99	5.61	0.37	0.35	0.05	0.00
		Jack5	40.70	185.77	74.48	73.53	8.40	0.49	0.50	0.04	0.00
		CY-1	31.03	274.87		26.00	33.70		0.05	0.03	0.00
		CY-2	40.32	364.37		25.31	64.71		0.06	0.10	0.00
		ICE	38.91	206.79		18.06	31.57		0.38	0.23	0.00

Table 4.A.2. Empirical and estimated variance of 12 species richness (SR) estimators for two levels of survey effort (*Effort*). Each level of *Effort* is comprised of 156 realizations or 13 replications of all combinations of a 3x2x2 factorial design based on abundance distribution, total abundance, and detection probability, respectively. Results include the standard deviation of SR estimates across the 156 replicates (*Emp*) and the mean variance estimates from analytical estimators (*Analyt*) and bootstrap (B_{Resamp}) and jackknife (J_{Resamp}) resampling procedures. The proportion of 95% confidence intervals that included the true number of species (S_{true}) for *Analyt*, B_{Resamp}, and J_{Resamp} are listed as CovA, CovB, and CovJ, respectively. The proportion of SR estimates that were less than the observed number of species is listed as $<S_{obs}$.

Effort	S_{true}	Estimator	%CV	*Emp*	*Analyt*	B_{Resamp}	J_{Resamp}	CovA	CovB	CovJ	$<S_{obs}$
100	**25**	**ACE**	32.34	6.85	4.08	2.75	0.72	0.78	0.62	0.21	0.00
		Boot	33.48	6.38	1.27	1.60	0.31	0.33	0.43	0.08	0.00
		Chao1	33.76	6.78	3.72	2.89	0.92	0.73	0.58	0.25	0.00
		Chao2	34.01	6.82	3.68	2.89	0.94	0.72	0.59	0.26	0.00
		Jack1	32.31	6.96	2.95	2.31	0.51	0.56	0.45	0.13	0.00
		Jack2	33.03	7.69	5.11	4.16	1.06	0.72	0.61	0.16	0.06
		Jack3	38.98	9.68	8.82	7.22	2.15	0.69	0.60	0.24	0.12
		Jack4	52.96	14.62	15.29	11.06	4.14	0.56	0.53	0.31	0.20
		Jack5	76.52	24.89	26.86	17.47	7.41	0.45	0.40	0.31	0.25
		CY-1	40.02	10.09		2.90	0.76		0.30	0.12	0.00
		CY-2	36.39	9.26		3.34	0.99		0.40	0.17	0.00
		ICE	32.44	6.87		2.75	0.73		0.61	0.22	0.00
500	**25**	**ACE**	18.04	4.10	1.87	1.24	0.14	0.79	0.72	0.35	0.00
		Boot	19.89	4.38	0.68	0.87	0.07	0.23	0.26	0.03	0.00
		Chao1	20.32	4.57	1.81	1.30	0.18	0.73	0.68	0.43	0.00
		Chao2	20.31	4.57	1.81	1.30	0.18	0.73	0.68	0.43	0.00
		Jack1	17.50	4.02	1.45	1.27	0.12	0.76	0.61	0.41	0.00
		Jack2	17.31	4.12	2.52	2.36	0.25	0.72	0.67	0.35	0.10
		Jack3	21.39	5.30	4.56	4.41	0.53	0.64	0.64	0.35	0.13
		Jack4	32.68	8.55	8.05	7.66	1.10	0.51	0.49	0.32	0.26
		Jack5	56.04	15.85	14.85	12.86	2.20	0.40	0.38	0.24	0.29
		CY-1	17.84	4.13		1.14	0.12		0.35	0.12	0.00
		CY-2	19.25	4.79		1.91	0.48		0.42	0.10	0.01
		ICE	18.04	4.10		1.24	0.14		0.72	0.35	0.00
100	**100**	**ACE**	42.31	31.71	21.19	7.67	3.96	0.64	0.33	0.17	0.00
		Boot	38.57	17.25	2.61	3.35	0.66	0.01	0.03	0.00	0.00
		Chao1	42.69	29.43	19.33	8.00	4.23	0.62	0.31	0.18	0.00
		Chao2	42.19	28.96	19.16	8.00	4.19	0.63	0.31	0.17	0.00
		Jack1	37.75	21.66	6.48	4.83	1.04	0.13	0.10	0.02	0.00
		Jack2	36.79	26.36	11.22	8.30	1.90	0.43	0.31	0.06	0.00
		Jack3	36.92	30.06	17.36	13.90	3.30	0.73	0.60	0.13	0.00
		Jack4	39.95	35.32	26.36	21.03	5.66	0.89	0.77	0.25	0.02
		Jack5	48.94	45.84	40.21	30.43	9.51	0.82	0.78	0.35	0.09
		CY-1	54.68	61.40		9.56	5.44		0.19	0.09	0.00
		CY-2	53.27	58.43		9.90	6.86		0.21	0.12	0.00
		ICE	41.99	31.53		7.75	3.96		0.34	0.16	0.00
500	**100**	**ACE**	21.21	18.13	6.66	4.71	0.54	0.66	0.53	0.06	0.00
		Boot	26.33	21.21	2.53	3.18	0.26	0.26	0.34	0.03	0.00

		Chao1	23.09	19.63	8.23	5.85	0.86	0.68	0.56	0.08	0.00
		Chao2	23.16	19.70	8.24	5.85	0.86	0.68	0.56	0.07	0.00
		Jack1	23.83	21.25	5.76	4.58	0.45	0.44	0.40	0.06	0.00
		Jack2	21.33	19.88	9.98	8.28	0.96	0.57	0.47	0.10	0.06
		Jack3	22.38	21.10	17.38	15.03	1.95	0.77	0.74	0.14	0.10
		Jack4	30.38	28.93	30.60	26.62	3.91	0.68	0.65	0.15	0.21
		Jack5	49.66	48.06	54.33	42.85	7.54	0.59	0.58	0.19	0.28
		CY-1	27.84	27.62		5.14	0.63		0.19	0.01	0.00
		CY-2	24.30	25.02		5.83	1.01		0.21	0.03	0.00
		ICE	21.23	18.16		4.72	0.55		0.53	0.06	0.00
100	**500**	**ACE**	73.98	258.42	192.31	13.98	32.24	0.58	0.08	0.17	0.00
		Boot	40.02	29.66	3.76	4.92	0.88	0.00	0.00	0.00	0.00
		Chao1	58.02	151.33	101.76	14.24	23.75	0.45	0.04	0.11	0.00
		Chao2	57.98	150.83	101.46	14.17	23.76	0.46	0.04	0.12	0.00
		Jack1	39.35	40.74	9.60	7.14	1.36	0.00	0.00	0.00	0.00
		Jack2	38.75	56.22	16.62	12.12	2.30	0.00	0.00	0.00	0.00
		Jack3	38.58	70.10	24.30	20.07	3.56	0.00	0.00	0.00	0.00
		Jack4	38.87	83.31	33.46	30.55	5.32	0.04	0.05	0.00	0.00
		Jack5	39.76	96.87	45.06	43.23	7.82	0.12	0.13	0.03	0.00
		CY-1	65.06	363.78		20.67	34.12		0.06	0.12	0.00
		CY-2	79.08	460.82		19.77	61.06		0.11	0.18	0.00
		ICE	73.89	260.33		14.10	33.12		0.07	0.17	0.00
500	**500**	**ACE**	28.80	111.74	48.78	16.08	3.95	0.47	0.21	0.04	0.00
		Boot	34.20	80.51	6.03	7.78	0.68	0.00	0.00	0.00	0.00
		Chao1	29.91	110.60	48.57	16.95	4.38	0.42	0.20	0.04	0.00
		Chao2	29.96	110.70	48.44	16.94	4.36	0.42	0.21	0.04	0.00
		Jack1	32.57	98.49	15.05	11.11	1.09	0.06	0.04	0.00	0.00
		Jack2	30.15	113.93	26.06	18.92	2.01	0.23	0.19	0.04	0.00
		Jack3	28.20	120.73	40.26	32.07	3.49	0.35	0.26	0.04	0.00
		Jack4	27.50	126.92	60.92	53.91	5.94	0.58	0.50	0.05	0.00
		Jack5	29.13	141.07	92.41	86.07	10.00	0.82	0.80	0.08	0.00
		CY-1	43.46	244.17		19.36	6.60		0.17	0.05	0.00
		CY-2	40.80	228.47		19.99	7.83		0.16	0.05	0.00
		ICE	28.81	111.84		16.10	3.96		0.21	0.04	0.00

Table 4.A.3. Empirical and estimated variance of 12 species richness (SR) estimators for two levels of total abundance (*N*). Each level of *N* is comprised of 156 realizations or 13 replications of all combinations of a 3x2x2 factorial design based on abundance distribution, effort, and detection probability, respectively. Results include the standard deviation of SR estimates across the 156 replicates (*Emp*) as well as the mean variance estimates from analytical estimators (*Analyt*) and bootstrap (B_{Resamp}) and jackknife (J_{Resamp}) resampling procedures. The proportion of 95% confidence intervals that included the true number of species (S_{true}) for *Analyt*, B_{Resamp}, and J_{Resamp} are listed as CovA, CovB, and CovJ, respectively. The proportion of SR estimates that were less than the observed number of species is listed under $<S_{obs}$.

N	S_{true}	Estimator	%CV	*Emp*	*Analyt*	B_{Resamp}	J_{Resamp}	CovA	CovB	CovJ	$<S_{obs}$
6250	**25**	**ACE**	28.64	6.24	3.58	2.33	0.54	0.81	0.69	0.24	0.00
		Boot	29.43	5.84	1.08	1.36	0.21	0.29	0.38	0.06	0.00
		Chao1	29.93	6.28	3.38	2.41	0.69	0.76	0.67	0.30	0.00
		Chao2	30.12	6.33	3.40	2.43	0.70	0.75	0.68	0.31	0.00
		Jack1	27.83	6.12	2.53	1.98	0.35	0.65	0.51	0.22	0.00
		Jack2	28.73	6.85	4.39	3.56	0.71	0.75	0.67	0.22	0.04
		Jack3	33.76	8.63	7.46	6.23	1.39	0.71	0.63	0.26	0.09
		Jack4	46.12	12.92	12.73	9.93	2.69	0.51	0.48	0.24	0.21
		Jack5	68.92	22.08	22.33	15.93	4.86	0.42	0.36	0.24	0.25
		CY-1	36.18	8.98		2.42	0.58		0.32	0.12	0.00
		CY-2	31.39	8.03		2.99	0.84		0.46	0.15	0.00
		ICE	28.66	6.26		2.34	0.54		0.69	0.24	0.00
12500	**25**	**ACE**	23.02	5.09	2.38	1.66	0.32	0.75	0.65	0.32	0.00
		Boot	25.46	5.41	0.87	1.10	0.16	0.27	0.31	0.06	0.00
		Chao1	25.43	5.50	2.15	1.77	0.41	0.70	0.60	0.38	0.00
		Chao2	25.53	5.50	2.09	1.76	0.43	0.70	0.59	0.38	0.00
		Jack1	23.51	5.29	1.87	1.61	0.28	0.67	0.54	0.32	0.00
		Jack2	23.21	5.39	3.24	2.97	0.61	0.69	0.61	0.29	0.11
		Jack3	28.28	6.81	5.92	5.39	1.28	0.62	0.60	0.33	0.17
		Jack4	42.23	10.87	10.60	8.80	2.56	0.55	0.53	0.40	0.24
		Jack5	68.36	19.68	19.38	14.39	4.75	0.44	0.42	0.30	0.29
		CY-1	26.66	6.27		1.62	0.30		0.33	0.12	0.00
		CY-2	26.79	6.64		2.25	0.63		0.37	0.12	0.01
		ICE	23.13	5.11		1.65	0.32		0.64	0.32	0.00
6250	**100**	**ACE**	35.47	28.28	17.71	6.71	2.93	0.70	0.42	0.14	0.00
		Boot	46.68	26.51	2.58	3.27	0.45	0.09	0.12	0.00	0.00
		Chao1	37.04	28.20	17.16	7.24	3.20	0.69	0.40	0.15	0.00
		Chao2	36.95	28.07	17.06	7.21	3.17	0.69	0.40	0.14	0.00
		Jack1	40.54	27.67	6.36	4.69	0.72	0.22	0.17	0.03	0.00
		Jack2	33.28	26.24	11.02	8.21	1.36	0.47	0.36	0.08	0.01
		Jack3	30.82	26.37	17.72	14.16	2.44	0.77	0.67	0.13	0.04
		Jack4	35.03	31.90	28.37	22.89	4.36	0.83	0.72	0.21	0.09
		Jack5	48.13	46.37	45.97	34.62	7.59	0.74	0.72	0.24	0.15
		CY-1	49.31	54.84		8.20	3.96		0.17	0.04	0.00
		CY-2	46.99	51.74		8.65	5.33		0.21	0.08	0.00
		ICE	35.50	28.35		6.74	2.93		0.43	0.13	0.00
12500	**100**	**ACE**	30.10	24.30	10.06	5.67	1.58	0.60	0.44	0.09	0.00
		Boot	36.41	24.93	2.55	3.26	0.47	0.18	0.25	0.03	0.00

		Chao1	31.07	24.17	10.40	6.61	1.89	0.62	0.47	0.11	0.00
		Chao2	30.80	23.93	10.34	6.64	1.88	0.62	0.47	0.10	0.00
		Jack1	31.60	24.75	5.88	4.72	0.77	0.35	0.32	0.06	0.00
		Jack2	28.71	24.69	10.18	8.38	1.49	0.53	0.42	0.09	0.04
		Jack3	29.90	26.94	17.03	14.77	2.80	0.73	0.67	0.15	0.06
		Jack4	35.65	33.00	28.59	24.77	5.21	0.74	0.70	0.19	0.14
		Jack5	50.56	47.57	48.58	38.66	9.46	0.67	0.63	0.29	0.21
		CY-1	39.29	39.40		6.50	2.10		0.21	0.06	0.00
		CY-2	35.92	36.83		7.08	2.54		0.21	0.07	0.00
		ICE	29.69	23.98		5.73	1.58		0.44	0.09	0.00
6250	**500**	**ACE**	55.14	197.22	144.69	14.62	19.69	0.54	0.12	0.12	0.00
		Boot	63.36	77.52	4.37	5.67	0.71	0.00	0.00	0.00	0.00
		Chao1	48.86	141.20	80.75	14.67	15.17	0.43	0.10	0.07	0.00
		Chao2	49.15	141.64	80.40	14.62	15.15	0.43	0.11	0.08	0.00
		Jack1	61.22	100.73	11.38	8.12	1.11	0.00	0.00	0.00	0.00
		Jack2	58.20	127.43	19.71	13.76	1.91	0.08	0.06	0.01	0.00
		Jack3	55.43	144.96	29.41	22.96	3.03	0.17	0.10	0.00	0.00
		Jack4	52.99	156.26	41.98	36.54	4.66	0.28	0.22	0.03	0.00
		Jack5	51.16	164.05	59.25	54.98	7.05	0.41	0.40	0.03	0.00
		CY-1	56.08	309.26		19.95	20.72		0.13	0.11	0.00
		CY-2	62.35	358.70		19.65	38.44		0.15	0.15	0.00
		ICE	55.44	199.03		14.67	20.31		0.12	0.12	0.00
12500	**500**	**ACE**	53.26	202.19	96.41	15.44	16.51	0.51	0.17	0.10	0.00
		Boot	59.31	111.02	5.43	7.04	0.84	0.00	0.00	0.00	0.00
		Chao1	41.16	140.63	69.57	16.51	12.96	0.44	0.14	0.08	0.00
		Chao2	40.93	139.75	69.49	16.49	12.97	0.45	0.14	0.08	0.00
		Jack1	55.80	134.72	13.27	10.13	1.33	0.06	0.04	0.00	0.00
		Jack2	50.46	153.43	22.98	17.28	2.40	0.15	0.13	0.04	0.00
		Jack3	45.62	158.85	35.15	29.18	4.02	0.19	0.16	0.04	0.00
		Jack4	42.05	160.21	52.39	47.92	6.60	0.35	0.33	0.02	0.00
		Jack5	40.90	166.58	78.22	74.33	10.77	0.53	0.53	0.08	0.00
		CY-1	54.45	310.09		20.08	20.00		0.10	0.06	0.00
		CY-2	65.03	368.94		20.11	30.44		0.12	0.08	0.00
		ICE	53.11	202.62		15.53	16.77		0.16	0.10	0.00

Table 4.A.4. Empirical and estimated variance of 12 species richness (SR) estimators for two levels of species detection probability (p). Each level of p is comprised of 156 realizations or 13 replications of all combinations of a 3x2x2 factorial design based on abundance distribution, effort, and total abundance, respectively. Results include the standard deviation of SR estimates across the 156 replicates (*Emp*) as well as the mean variance estimates from analytical estimators (*Analyt*) and bootstrap (B_{Resamp}) and jackknife (J_{Resamp}) resampling procedures. The proportion of 95% confidence intervals that included the true number of species (S_{true}) for *Analyt*, B_{Resamp}, and J_{Resamp} are listed as CovA, CovB, and CovJ, respectively. The proportion of SR estimates that were less than the observed number is listed as $<S_{obs}$.

p	S_{true}	Estimator	%CV	*Emp*	*Analyt*	B_{Resamp}	J_{Resamp}	CovA	CovB	CovJ	$<S_{obs}$
0.5	**25**	**ACE**	29.80	6.42	3.43	2.20	0.50	0.74	0.62	0.22	0.00
		Boot	30.32	5.97	1.04	1.33	0.20	0.31	0.36	0.06	0.00
		Chao1	30.72	6.37	3.16	2.26	0.63	0.69	0.58	0.30	0.00
		Chao2	30.68	6.36	3.14	2.25	0.65	0.69	0.59	0.31	0.00
		Jack1	28.45	6.16	2.38	1.92	0.33	0.62	0.49	0.24	0.00
		Jack2	29.23	6.78	4.12	3.46	0.67	0.69	0.60	0.19	0.07
		Jack3	34.11	8.47	7.15	6.08	1.35	0.64	0.57	0.22	0.13
		Jack4	46.03	12.62	12.39	9.69	2.63	0.50	0.47	0.23	0.23
		Jack5	67.24	21.37	22.01	15.65	4.79	0.40	0.36	0.24	0.25
		CY-1	37.02	9.00		2.32	0.54		0.32	0.09	0.00
		CY-2	33.89	8.40		2.85	0.77		0.43	0.12	0.01
		ICE	29.79	6.42		2.20	0.51		0.61	0.23	0.00
0.9	**25**	**ACE**	21.62	4.84	2.55	1.79	0.36	0.83	0.71	0.34	0.00
		Boot	24.39	5.22	0.91	1.14	0.18	0.26	0.33	0.05	0.00
		Chao1	24.49	5.35	2.37	1.93	0.48	0.77	0.68	0.38	0.00
		Chao2	24.83	5.42	2.35	1.94	0.47	0.76	0.68	0.38	0.00
		Jack1	22.69	5.19	2.02	1.66	0.30	0.69	0.56	0.31	0.00
		Jack2	22.94	5.47	3.51	3.07	0.65	0.76	0.69	0.31	0.08
		Jack3	28.57	7.08	6.23	5.54	1.33	0.69	0.67	0.37	0.13
		Jack4	42.94	11.32	10.94	9.03	2.62	0.56	0.54	0.40	0.22
		Jack5	70.53	20.48	19.71	14.68	4.82	0.46	0.42	0.30	0.29
		CY-1	26.24	6.31		1.72	0.34		0.33	0.15	0.00
		CY-2	24.12	6.17		2.39	0.70		0.40	0.15	0.00
		ICE	21.77	4.87		1.79	0.35		0.72	0.33	0.00
0.5	**100**	**ACE**	36.49	28.34	16.42	6.54	2.71	0.67	0.40	0.13	0.00
		Boot	47.04	26.69	2.56	3.24	0.46	0.13	0.19	0.03	0.00
		Chao1	37.85	27.46	14.92	7.04	2.75	0.64	0.42	0.12	0.00
		Chao2	37.55	27.24	14.93	7.05	2.76	0.65	0.42	0.12	0.00
		Jack1	41.28	27.96	6.23	4.65	0.73	0.24	0.21	0.02	0.00
		Jack2	34.78	26.94	10.79	8.16	1.38	0.47	0.34	0.08	0.01
		Jack3	32.26	26.72	17.46	14.07	2.48	0.72	0.64	0.12	0.04
		Jack4	35.61	30.68	28.12	22.72	4.40	0.86	0.75	0.19	0.09
		Jack5	48.59	42.90	45.66	34.43	7.59	0.77	0.76	0.27	0.16
		CY-1	48.47	51.84		7.91	3.66		0.19	0.07	0.00
		CY-2	45.25	47.99		8.40	4.64		0.22	0.09	0.00
		ICE	36.23	28.16		6.58	2.72		0.40	0.13	0.00
0.9	**100**	**ACE**	28.95	23.96	11.36	5.85	1.79	0.63	0.46	0.10	0.00
		Boot	36.07	24.72	2.58	3.29	0.47	0.13	0.19	0.00	0.00

		Chao1	29.78	24.24	12.64	6.81	2.34	0.66	0.45	0.13	0.00
		Chao2	29.76	24.14	12.47	6.80	2.29	0.66	0.45	0.12	0.00
		Jack1	30.69	24.20	6.01	4.76	0.76	0.33	0.29	0.06	0.00
		Jack2	26.79	23.42	10.41	8.43	1.47	0.53	0.44	0.09	0.04
		Jack3	27.81	25.82	17.29	14.87	2.76	0.78	0.69	0.16	0.06
		Jack4	34.05	33.20	28.84	24.94	5.16	0.71	0.67	0.21	0.14
		Jack5	48.74	49.79	48.88	38.85	9.45	0.64	0.60	0.27	0.21
		CY-1	42.01	43.92		6.79	2.40		0.19	0.03	0.00
		CY-2	39.34	41.94		7.33	3.23		0.21	0.06	0.00
		ICE	28.86	23.93		5.89	1.79		0.47	0.09	0.00
0.5	**500**	**ACE**	60.54	220.38	148.85	14.59	20.92	0.56	0.12	0.09	0.00
		Boot	64.90	81.62	4.43	5.73	0.71	0.00	0.00	0.00	0.00
		Chao1	50.79	150.78	82.64	14.76	16.06	0.44	0.09	0.08	0.00
		Chao2	50.73	150.18	82.25	14.72	16.00	0.44	0.09	0.08	0.00
		Jack1	62.66	105.72	11.47	8.21	1.10	0.00	0.00	0.00	0.00
		Jack2	59.51	133.46	19.87	13.92	1.88	0.09	0.07	0.04	0.00
		Jack3	56.74	152.42	29.65	23.23	3.00	0.15	0.09	0.01	0.00
		Jack4	54.61	166.58	42.39	37.13	4.68	0.24	0.22	0.04	0.00
		Jack5	53.61	179.85	60.14	55.89	7.30	0.37	0.35	0.02	0.00
		CY-1	58.87	316.26		19.57	19.80		0.11	0.08	0.00
		CY-2	70.04	397.82		19.51	39.54		0.14	0.17	0.00
		ICE	60.49	220.99		14.67	21.16		0.12	0.09	0.00
0.9	**500**	**ACE**	47.48	177.22	92.25	15.47	15.27	0.49	0.17	0.12	0.00
		Boot	59.85	109.98	5.36	6.97	0.85	0.00	0.00	0.00	0.00
		Chao1	39.85	133.01	67.68	16.43	12.07	0.44	0.15	0.07	0.00
		Chao2	39.96	133.26	67.64	16.39	12.13	0.44	0.16	0.08	0.00
		Jack1	56.12	133.16	13.18	10.04	1.35	0.06	0.04	0.00	0.00
		Jack2	50.56	151.07	22.82	17.12	2.42	0.14	0.12	0.01	0.00
		Jack3	45.57	155.42	34.92	28.91	4.05	0.21	0.17	0.03	0.00
		Jack4	41.68	154.57	51.98	47.33	6.57	0.38	0.33	0.01	0.00
		Jack5	39.85	156.39	77.33	73.42	10.53	0.57	0.58	0.09	0.00
		CY-1	51.63	301.40		20.46	20.93		0.12	0.08	0.00
		CY-2	56.80	326.40		20.25	29.35		0.13	0.06	0.00
		ICE	47.71	178.99		15.53	15.92		0.16	0.12	0.00

APPENDIX I:

EXAMPLES OF THE SPECIES ABUNDANCE DISTRIBUTIONS

GENERATED IN PROGRAM SIMASSEM

This appendix includes step-by-step examples of the abundance distributions available in program SimAssem. The results of individual steps are given as separate rows and columns. In many of the models, total abundance is represented by a 'line' that is 'broken' into a number of segments by model-specific rules, with the length of each segment representing the abundance of one species. The first section gives examples of nine niche-based abundance distributions, the broken-stick, dominance-decay, dominance pre-emption, geometric-series, power fraction, particulate-niche, random-assortment, random-fraction, sequential 75%, and zero-sum multinomial. Examples of two statistically-based abundance distributions, the log-normal and log-series, are also shown.

NICHE-BASED DISTRIBUTION MODELS:

Broken-stick			
Breaks	**Sorted**	**Differ**	**Abun**
9216	1371	1371	2193
6302	1851	480	1910
7306	1926	75	1371
4109	3113	1187	1187
1371	4109	996	996
1851	6302	2193	784
3113	6567	265	739
6567	7306	739	480
1926	9216	1910	265
		784	75
Total		**10000**	**10000**

Example 1. Allocation of 10,000 'individuals' to 'species' using a broken-stick model for the species abundance distribution. The steps were: 1) selecting nine breakage points (column Breaks), 2) sorting the breakage points from smallest to largest (column Sorted), and 3) calculating the differences between breakage points (column Differ). Sorted species abundances are listed in column Abun. The last row is total abundance across all species.

		Dominance pre-emption						
		Break 1	**Break 2**	**Break 3**	**Break 4**	**Break 5**	**Break 6**	**Abun**
		7948	7948	7948	7948	7948	7948	7948
			1341	1341	1341	1341	1341	1341
				591	591	591	591	591
					102	102	102	102
						14	14	14
							3	3
								1
RUV	**0.795**	**0.653**	**0.830**	**0.842**	**0.776**	**0.611**	**0.703**	
Remain	**10000**	**2052**	**711**	**120**	**18**	**4**	**1**	**0**
Total	**0**	**7948**	**9289**	**9880**	**9982**	**9996**	**9999**	**10000**

Example 2. Allocation of 10,000 'individuals' to 'species' using a dominance pre-emption model for the species abundance distribution. The steps were: 1) drawing a random uniform variate greater than 0.5 (row RUV) and 2) multiplying the remaining abundance (row Remain) by the random variate. Each successive species is listed as the last value in the Break columns. Sorted species abundances are listed in column Abun. The last row is total abundance across all species.

				Dominance-decay					
Break 1	**Break 2**	**Break 3**	**Break 4**	**Break 5**	**Break 6**	**Break 7**	**Break 8**	**Abun**
7944	**5828**	**3191**	**3028**	1922	1922	1922	1922	1922
2056	2056	2056	2056	2056	2056	2056	**2056**	1693
	2116	2116	2116	2116	2116	**2116**	1238	1256
		2637	2637	**2637**	**2140**	884	884	1238
			163	163	163	163	163	1106
				1106	1106	1106	1106	884
					497	497	497	878
						1256	1256	497
							878	363
								163
RUV	**0.734**	**0.548**	**0.949**	**0.871**	**0.812**	**0.989**	**0.972**		
Total	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**

Example 3. Allocation of 10,000 'individuals' to 'species' using a dominance-decay model for the species abundance distribution. The first step involved randomly breaking the complete line (random uniform variate ≈ 0.794) into two segments. Subsequent steps included: 1) drawing a random uniform variate (row RUV) and 2) multiplying the largest remaining segment (bold values) by the RUV. Sorted species abundances are listed in column Abun. The last row is total abundance across all species.

	Geometric-series (k = 0.6)									
	Break 1	**Break 2**	**Break 3**	**Break 4**	**Break 5**	**Break 6**	**Break 7**	**Break 8**	**Break 9**	**Abun**
	6000	6000	6000	6000	6000	6000	6000	6000	6000	6000
		2400	2400	2400	2400	2400	2400	2400	2400	2400
			960	960	960	960	960	960	960	960
				384	384	384	384	384	384	384
					153	153	153	153	153	153
						62	62	62	62	62
							24	24	24	24
								10	10	10
									4	4
										2
Remain	**4000**	**1600**	**640**	**256**	**103**	**41**	**17**	**7**	**3**	**1**
Total	**6000**	**8400**	**9360**	**9744**	**9897**	**9959**	**9983**	**9993**	**9997**	**9999**

Example 4. Allocation of 10,000 'individuals' to 'species' using a geometric-series model, k = 0.6, for the species abundance distribution. The first step involved randomly breaking the complete line into two segments. Subsequent steps involved multiplying the remaining abundance (row Remain) by k. Each successive species is listed as the last value in the Break columns. Sorted species abundances are listed in column Abun. The last row is total abundance across all species.

	Geometric-series (k = 0.3)									
	Break 1	**Break 2**	**Break 3**	**Break 4**	**Break 5**	**Break 6**	**Break 7**	**Break 8**	**Break 9**	**Abun**
	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000
		2100	2100	2100	2100	2100	2100	2100	2100	2100
			1470	1470	1470	1470	1470	1470	1470	1470
				1029	1029	1029	1029	1029	1029	1029
					720	720	720	720	720	720
						504	504	504	504	504
							353	353	353	353
								247	247	247
									173	173
										121
Remain	**7000**	**4900**	**3430**	**2401**	**1681**	**1177**	**824**	**577**	**404**	**283**
Total	**3000**	**5100**	**6570**	**7599**	**8319**	**8823**	**9176**	**9423**	**9596**	**9717**

Example 5. Allocation of 10,000 'individuals' to 'species' using a geometric-series model, k = 0.3, for the species abundance distribution. The first step involved randomly breaking the complete line into two segments. Subsequent steps involved multiplying the remaining abundance (row Remain) by k. Each successive species is listed as the last value in the Break columns. Sorted species abundances are listed in column Abun. The last row is total abundance across all species.

Particulate-niche									
Indiv. 1	**Indiv. 2**	**Indiv. 3**	**Indiv. 4**	**Indiv. 5**	**Indiv. 6**	**Indiv. 7**	**Indiv. 8**	**Indiv. N**	**Abun**
1	1	1	1	1	1	1	1	…	1036
								…	1030
		1	**2**	2	2	2	2	…	1028
								…	1003
	1	1	1	1	1	1	1	…	998
								…	987
					1	1	1	…	986
				1	1	1	1	…	982
						1	**2**	…	981
								…	969
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**…**	**10000**

Example 6. Allocation of 10,000 'individuals' to 'species' using a particulate-niche model for the species abundance distribution. The steps included: 1) randomly selecting a species (bold values) and 2) incrementing the abundance of the species by one. This example shows allocation of the first eight individuals. Steps are repeated once for each remaining individual in total abundance, represented by column Indiv. N. Sorted abundances are listed in column Abun. The last row is total abundance across all species.

	Power fraction (k = 0.1)							
	Break 1	**Cume prob**	**Break 2**	**Cume prob**	**Break 3**	**Cume prob**	**Break 4**	**Abun**
10000	**5959**	0.510	2099	0.319	2099	0.248	2099 …	2912
	4041	1	**4041**	0.660	2876	0.503	2876 …	2876
			3860	1	3860	0.766	3860 …	2099
					1165	1	199 …	948
							966 …	657
							…	218
							…	91
							…	79
							…	65
							…	55
Alpha		**0.214**		**0.149**		**0.115**		
RUV		**0.040**		**0.549**		**0.934**		
Total	**10000**		**10000**		**10000**		**10000**	**10000**

Example 7. Allocation of 10,000 'individuals' to 'species' using a power fraction model, k = 0.1, for the species abundance distribution. The first step involved randomly breaking the complete line into two segments (column Break 1). Subsequent steps included: 1) calculating $\alpha = 1 / \sum n_i^k$ where n_i is the abundance of species i and i = 1, 2, … (row Alpha), 2) drawing a random uniform variate (row RUV), 3) calculating a cumulative probability distribution as αn_i^k (columns Cume prob), 4) determining the species in the cumulative probability distribution to which the RUV coincides (bold values), and 5) multiplying the abundance of the species by another RUV. Each successive species is listed as the last value in the Break columns. Three replications of these steps are shown. Sorted species abundances are listed in column Abun. The last row is total abundance across all species.

		Random-assortment							
		Break 1	**Break 2**	**Break 3**	**Break 4**	**Break 5**	**Break 6**	**Break 7**	**Abun**
		8384	8384	8384	8384	8384	8384	8384	8384
			760	760	760	760	760	760	760
				150	150	150	150	150	150
					517	517	517	517	517
						126	126	126	126
							51	51	51
								11	11
									1
RUV	**0.838**	**0.47**	**0.175**	**0.731**	**0.663**	**0.805**	**0.889**	**0.771**	
Remain	**10000**	**1616**	**856**	**706**	**189**	**63**	**12**	**1**	**0**
Total	**0**	**8384**	**9144**	**9294**	**9811**	**9937**	**9988**	**9999**	**10000**

Example 8. Allocation of 10,000 'individuals' to 'species' using a random-assortment model for the species abundance distribution. The steps included: 1) drawing a random uniform variate (row RUV) and 2) multiplying the remaining abundance (row Remain) by the RUV. Each successive species is listed as the last value in the Break columns. Sorted species abundances are listed in column Abun. The last row is total abundance across all species.

	Random-fraction									
	Break 1	**Break 2**	**Break 3**	**Break 4**	**Break 5**	**Break 6**	**Break 7**	**Break 8**	**Break 9**	**Abun**
10000	4104	4104	4104	4104	4104	4104	4104	4104	4104	4758
	5896	**609**	556	556	**556**	523	523	523	523	4104
		5287	**5287**	4758	4758	4758	4758	4758	4758	529
			53	**53**	27	27	27	27	27	523
				529	529	529	529	529	529	27
					26	26	26	26	26	26
						33	**28**	4	4	17
							5	5	5	7
								24	17	5
									7	4
RUV	**0.103**	**0.913**	**0.900**	**0.512**	**0.942**	**0.859**	**0.143**	**0.737**		
Total	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**

Example 9. Allocation of 10,000 'individuals' to 'species' using a random-fraction model for the species abundance distribution. The first step involved randomly breaking (random uniform variate ≈ 0.410) the complete line into two segments (column Break 1). Subsequent steps included: 1) randomly selecting a line segment (bold values), 2) drawing a random uniform variate (row RUV), and 3) multiplying the selected segment length by the RUV. Each successive species is listed as the last value in the Break columns. Sorted species abundances are listed in column Abun. The last row is total abundance across all species.

	Sequential 75%									
	Break 1	**Break 2**	**Break 3**	**Break 4**	**Break 5**	**Break 6**	**Break 7**	**Break 8**	**Break 9**	**Abun**
10000	7500	7500	**7500**	1875	1875	1875	1875	1875	1875	4219
	2500	625	625	625	**625**	156	156	156	156	1875
		1875	469	**469**	117	117	117	117	117	1406
			1406	1406	1406	1406	**1406**	352	352	1054
				5625	5625	**5625**	1406	1406	1406	352
					352	352	352	352	352	352
						469	469	**469**	117	352
							4219	4219	4219	156
								1054	1054	117
									352	117
Total	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**	**10000**

Example 10. Allocation of 10,000 'individuals' to 'species' using a sequential 75% model for the species abundance distribution. The first step involved breaking the complete line into two segments, one with 75% of the individuals and the other with 25% (column Break 1). Subsequent steps included: 1) randomly selecting a line segment (bold values) and 2) allocating 75% of the selected segment abundance to a new segment, i.e., last value in the Break columns. Sorted species abundances are listed in column Abun. The last row is total abundance across all species.

	Zero-sum multinomial									
	Indiv. 1	**Indiv. 2**	**Indiv. 3**	**Indiv. 4**	**Indiv. 5**	**Indiv. 6**	**Indiv. 7**	**Indiv. 8**	**Indiv. N**	**Abun**
				1		1	**1**			
	1	1	1	(0.25)	1	(0.16)	**(0.14)**	2	…	6930
				1		**1**	2			
		1	1	(0.50)	1	**(0.33)**	(0.43)	2	…	1915
				1		1	1			
			1	(0.75)	1	(0.50)	(0.57)	1	…	676
				1		2	2			
				(1.00)	2	(0.83)	(0.86)	2	…	374
						1	1			
						(1.00)	(1.00)	1	…	40
									…	24
									…	14
									…	13
									…	5
									…	4
									…	3
									…	2
Generator	**1.000**	**0.839**	**0.722**	**0.634**	**0.565**	**0.510**	**0.464**	**0.426**		
RUV 1	**0.309**	**0.745**	**0.217**	**0.130**	**0.963**	**0.468**	**0.807**	**0.838**		
RUV 2				**0.953**		**0.315**	**0.141**			
Total	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**…**	**10000**

Example 11. Allocation of 10,000 'individuals' to 'species' using a zero-sum multinomial model for the species abundance distribution with $\theta = 1.5$. The steps included: 1) calculating [θ / ($\theta + j - 1$)] (row Generator), where j iterates from 1 to total abundance, 2) drawing a random uniform variate (row RUV 1), 3a) creating a new segment of length one when RUV 1 ≤ Generator (new segments are shown as the last value in the Indiv. columns), *or* 3b) incrementing by one the value of the segment to which a RUV (row RUV 2) coincides with the cumulative abundance distribution (bold values; cumulative abundance distribution shown inside parentheses, Indiv. columns). This example shows allocation of the first eight individuals. Steps are repeated once for each remaining individual in total abundance, represented by column Indiv. N. Sorted species abundances are listed in column Abun. The last row is total abundance across all species.

STATISTICALLY-BASED DISTRIBUTION MODELS:

	Log-normal			
	Variates	**Normalized**	**Abun**	**Rounded**
	0.147	0.017	169.550	4411
	0.921	0.106	1062.284	1062
	0.314	0.036	362.168	1029
	3.824	0.441	4410.611	885
	0.751	0.087	866.205	866
	0.767	0.088	884.660	595
	0.258	0.030	297.578	362
	0.516	0.060	595.156	323
	0.892	0.103	1028.835	298
	0.280	0.032	322.953	170
Sum	**8.670**	**1.000**		
Total			**10000**	**10001**

Example 12. Allocation of 10,000 'individuals' to 'species' with a log-normal model for the species abundance distribution. The steps included: 1) drawing 10 random log-normal variates ($\mu = 0$, $\sigma = 1$, column Variates), 2) normalizing the variates (column Normalized), and 3) multiplying each normalized variate by total abundance, i.e., 10,000. Species abundances are listed in column Abun and rounded and sorted species abundances are listed in column Rounded. The last row is total abundance across all species.

Log-series				
Iteration	**Number of species**	**Number (+ remain)**	**Cumulative species**	**Cumulative abundance**
1	1.097	1.097	1	1
3	0.365	1.010	2	4
9	0.122	1.102	3	13
22	0.050	1.045	4	35
54	0.020	1.011	5	89
135	0.008	1.001	6	224
345	0.003	1.002	7	569
911	0.001	1.000	8	1480
2713	0.000	1.000	9	4193
74678	4.08E-09	1.000	10	78871
Total				**78871**

Example 13. Allocation of 10,000 'individuals' to 'species' using a log-series model for the species abundance distribution, $x = 0.99989034$. The first step involved calculating $\alpha = N(1 - x) / x$ where N is total abundance (10,000 in this example). Subsequent steps included: 1) calculating the number of species with z individuals, where z iterates from 1, $\alpha x^z / z$ (column Number of species), 2) adding any fractional portion of step 1 calculations from previous iterations (column Number (+ remain)), and 3) creating new species for each integer portion of column Number (+ remain). These steps are repeated until cumulative abundance (column Cumulative abundance) equals or exceeds N. Only those steps resulting in the creation of a species are displayed. Column Cumulative species tracks the total number of species created. The last row is total abundance across all species.

Log-series				
Iteration	**Number of species**	**Number (+ remain)**	**Cumulative species**	**Cumulative abundance**
1	1.097	1.097	1	1
2	0.548	0.645	1	1
3	0.365	1.010	2	4
4	0.274	0.284	2	4
5	0.219	0.504	2	4
6	0.183	0.686	2	4
7	0.157	0.843	2	4
8	0.137	0.980	2	4
9	0.122	1.102	3	13
…	…	…	…	…
22	0.050	1.045	4	35
54	0.020	1.011	5	89
135	0.008	1.001	6	224
345	0.003	1.002	7	569
911	0.001	1.000	8	1480
2713	0.000	1.000	9	4193
			10	10000
Total				**10000**

Example 14. Allocation of 10,000 'individuals' to 'species' using a modified log-series model for the species abundance distribution, $x = 0.99989034$. The first step involved calculating $\alpha = N(1 - x) / x$ where N is total abundance (10,000 in this example). Subsequent steps included: 1) calculating the number of species with i individuals, where i iterates from 1, $\alpha x^i / i$ (column Number of species), 2) adding any fractional portion of step 1 calculations from previous iterations (column Number (+ remain)), and 3) creating new species for each integer portion of column Number (+ remain). These steps are repeated until individuals are allocated to the penultimate species, then remaining individuals are allocated to the last species. The first nine iterations and those steps resulting in the creation of a later species are displayed. Column Cumulative species tracks the total number of species created. The last row is total abundance across all species.

APPENDIX II:

FORMULAS FOR ESTIMATORS

INCLUDED IN PROGRAM SIMASSEM

This appendix includes formulas for the estimators included in program SimAssem. The analytical variance estimator formulas for the abundance-based (ACE) and incidence-based (ICE) coverage estimators are also given. These two variances estimators are yet to be included in SimAssem, so SimAssem will format data for a program that provides the variance estimates, program SPADE (Chao and Shen 2003). Estimator equations are from the original papers, the user's guides for programs EstimateS (Colwell 2006) and SPADE (Chao and Shen 2003), and from communications with the original authors.

The first section defines notation that is shared between one or more formulas. The remainder of the document divides estimators into five classes. The first two sections present formulas for the species richness estimators that require either abundance data, i.e., the number of individuals of each species that were encountered, or incidence data, i.e., the number of surveys in which each species was encountered. Three additional sections give formulas for an estimator of evenness, species richness indices, and estimators of the additional effort, i.e., individuals or surveys, required to encounter a specified fraction of two nonparametric estimators.

COMMON NOTATION FOR SPECIES RICHNESS ESTIMATORS:

α_{jK} Coefficient for the K^{th}-order jackknife estimator associated with survey j.

a Number of species occurring only in replicate 1.

b Number of species occurring only in replicate 2.

c Number of species occurring in both replicate 1 and replicate 2.

d Number of individuals encountered that divides abundant species from rare

species (ACE); also, number of surveys in which the species was encountered that divides frequent species from infrequent species (ICE).

$\check{D}_{index}$ Estimate from species richness index 'index'.

f The fraction of $\widehat{S_{est}}$ one wants to encounter through additional surveys.

f_i, f_j Number of species with exactly i or j number of individual encounters, $i, j = 0, 1, \ldots, n$. (Species with only one individual encountered are singletons, those with two are doubletons, etc.)

g Group identifier, $g = 1, 2, \ldots, G$.

G Number of groups in which species detection probabilities are homogeneous.

h Species identifier, $h = 1, 2, \ldots, S_{obs}$.

i Species identifier, $i = 1, 2, \ldots, S_{obs}$.

j Survey identifier, $j = 1, 2, \ldots, t$.

K Order of jackknife estimator.

k Species identifier, $k = 1, 2, \ldots, S_{obs}$.

l Species identifier, $l = 1, 2, \ldots, S_{obs}$.

m Number of incidences across all species and surveys. For example, if 10 individuals of the same species were encountered in the same survey, this counts as one incidence.

M_{infr} Number of surveys in which at least one infrequent species, i.e., species encountered in less than or equal to c number of surveys, was encountered.

n Number of individuals encountered across all species and surveys.

N Number of survey randomizations over which the Bootstrap is averaged.

p_h Proportion of surveys in which species h was encountered.

Q_j Number of species encountered in exactly j number of surveys, $j = 0, 1, \ldots, t$.

q_{kl} Proportion of surveys in which neither species k nor species l was encountered.

S_{obs} Number of species encountered at least once.

$\widehat{S_{est}}$ Estimated number of species using estimator 'est'.

S_{true} True number of species in the surveyed area.

t Number of surveys, e.g., trapping occasions, quadrats, etc.

X Number of iterations over which the calculation is averaged.

y Iteration number.

Z Encounter history matrix.

ABUNDANCE-BASED SPECIES RICHNESS ESTIMATORS:

Chao estimators:

- Chao1 estimator, if $f_2 > 0$ (Chao 1984, Chao and Shen 2003):

$$\hat{S}_{Chao1} = S_{obs} + \left(\frac{f_1^2}{2f_2} \right) \tag{1.1}$$

- Chao1 estimator, if $f_2 = 0$ (Chao 1984, Chao and Shen 2003):

$$\hat{S}_{Chao1} = S_{obs} + \left(\frac{f_1(f_1 - 1)}{2} \right) \tag{1.2}$$

- Variance for Chao1 estimator, (Chao 1987):

$$\hat{\text{var}}\left(\hat{S}_{Chao1} \right) = f_2 \left[\frac{1}{2}\left(\frac{f_1}{f_2} \right)^2 + \left(\frac{f_1}{f_2} \right)^3 + \frac{1}{4}\left(\frac{f_1}{f_2} \right)^4 \right] \tag{1.3}$$

- Chao1 estimator, bias-corrected (Chao 2005):

$$\hat{S}_{Chao1BC} = S_{obs} + \left(\frac{f_1(f_1 - 1)}{2(f_2 + 1)} \right) \tag{2.1}$$

- Variance for Chao1, bias-corrected estimator (Colwell 2006):

$$\hat{\text{var}}(\hat{S}_{Chao1BC}) = \frac{f_1(f_1-1)}{2(f_2+1)} + \frac{f_1(2f_1-1)^2}{4(f_2+1)^2} + \frac{f_1^2 f_2(f_1-1)^2}{4(f_2+1)^4} \tag{2.2}$$

- Variance for Chao1, bias-corrected estimator, if $f_1 > 0$ & $f_2 = 0$ (Colwell 2006):

$$\hat{\text{var}}(\hat{S}_{Chao1BC}) = \frac{f_1(f_1-1)}{2} + \frac{f_1(2f_1-1)^2}{4} - \frac{f_1^4}{4\hat{S}_{Chao1BC}} \tag{2.3}$$

- Variance for Chao1, bias-corrected estimator, if $f_1 = 0$ & $f_2 \geq 0$ (Colwell 2006):

$$\hat{\text{var}}(\hat{S}_{Chao1BC}) = S_{obs}\left(e^{-(\frac{n}{Sobs})}\right)\left(1-e^{-(\frac{n}{Sobs})}\right) \tag{2.4}$$

Coverage estimator:

- ACE estimator, use for small heterogeneity values (Chao and Lee 1992, Chao and Shen 2003):

$$\hat{S}_{ACE} = S_{abund} + \frac{S_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}}\hat{\gamma}^2_{rare}, \tag{3.1}$$

$$\text{where } S_{abund} = \sum_{i>d}^{n} f_i, \tag{3.2}$$

$$S_{rare} = \sum_{i=1}^{d} f_i, \tag{3.3}$$

$$\hat{C}_{rare} = 1 - \frac{f_1}{\sum_{i=1}^{d} i f_i}, \tag{3.4}$$

$$\text{and } \hat{\gamma}^2_{rare} = \max\left\{\frac{S_{rare}}{\hat{C}_{rare}}\frac{\sum_{i=1}^{d} i(i-1)f_i}{\left(\sum_{i=1}^{d} i f_i\right)\left(\sum_{i=1}^{d} i f_i - 1\right)} - 1, 0\right\}. \tag{3.5}$$

- ACE estimator, use for large heterogeneity values (Chao and Lee 1992, Chao and Shen 2003):

$$\hat{S}_{ACE} = S_{abund} + \frac{S_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}}\tilde{\gamma}^2_{rare}, \tag{4.1}$$

$$\text{where } S_{abund} = \sum_{i>d}^{n} f_i, \tag{4.2}$$

$$S_{rare} = \sum_{i=1}^{d} f_i, \tag{4.3}$$

$$\hat{C}_{rare} = 1 - \frac{f_1}{\sum_{i=1}^{d} i f_i}, \tag{4.4}$$

$$\text{and } \tilde{\gamma}^2_{rare} = \max\left\{\hat{\gamma}^2_{rare}\left[1 + \frac{\left(1-\hat{C}_{rare}\right)\sum_{i=1}^{d} i(i-1) f_i}{\hat{C}_{rare}\left(\sum_{i=1}^{d} i f_i - 1\right)}\right], 0\right\}. \tag{4.5}$$

- Variance for ACE estimator, (Chao and Lee 1992):

$$\hat{\text{var}}(\hat{S}_{ACE}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial \hat{S}_{ACE}}{\partial f_i}\frac{\partial \hat{S}_{ACE}}{\partial f_j}\text{cov}\left(f_i, f_j\right), \tag{5.1}$$

$$\text{where } \text{cov}\left(f_i, f_j\right) = f_i\left(1 - \frac{f_i}{\hat{S}_{ACE}}\right), \qquad \text{if } i = j \tag{5.2}$$

$$= -\frac{f_i f_j}{\hat{S}_{ACE}}, \qquad \text{if } i \neq j.$$

- *Darroch and Ratcliff estimator* (Darroch and Ratcliff 1980):

$$\hat{S}_{Darroch-Ratcliff} = \frac{S_{obs}}{\left(1 - \frac{f_1}{n}\right)} \tag{6.1}$$

- Modified *Horvitz-Thompson estimator* (Ashbridge and Goudie 2000):

$$\hat{S}_{Ashbridge-Goudie} = \sum_{j=1}^{t} f_j \left\{ 1 - \left(1 - \frac{j\hat{C}_u}{t} \right)^t \right\}^{-1} \qquad \text{(u = 1, 2, 3),} \tag{7.1}$$

where $\hat{C}_1 = 1 - \frac{f_1}{n}$, (7.2)

$$\hat{C}_2 = \min\left\{ 1, 1 - \frac{f_1}{n} + \frac{2}{(t-1)} \frac{f_2}{n} \right\}, \tag{7.3}$$

and $$\hat{C}_3 = \min\left\{ 1, 1 - \frac{f_1}{n} + \frac{2}{(t-1)} \frac{f_2}{n} - \frac{6}{(t-1)(t-2)} \frac{f_3}{n} \right\}. \tag{7.4}$$

INCIDENCE-BASED SPECIES RICHNESS ESTIMATORS:

Bootstrap estimator:

- Bootstrap estimator (Smith and van Belle 1984):

$$\hat{S}_{Boot} = S_{obs} + \sum_{h=1}^{S_{obs}} (1 - p_h)^t \tag{8.1}$$

- Variance for Bootstrap estimator (Smith and van Belle 1984):

$$\hat{\text{var}}\left(\hat{S}_{Boot} \right) = \sum_{h=1}^{S_{obs}} (1 - p_h)^t \left(1 - (1 - p_h)^t \right) + \sum_{i=1}^{S_{obs}} \sum_{h \neq i} \left(q_{hi}^t - (1 - p_h)^t (1 - p_i)^t \right) \tag{8.2}$$

- Bootstrap estimator (bootstrap samples) (Smith and van Belle 1984):

$$\hat{S}_{Boot-B} = \frac{1}{N} \sum_{y=1}^{N} \hat{S}_{Boot(y)} \tag{8.3}$$

Cao estimators (require splitting surveys into two equally sized replicates):

- CY-1 estimator (Cao et al. 2001): CY-1 uses the largest possible replicates.

$$\hat{S}_{CY-1} = \frac{1}{X} \sum_{y=1}^{X} \frac{\overline{SR}_y}{JC_y}, \tag{9.1}$$

where $\overline{SR}_y = \frac{1}{2}(a+b)+c$ for iteration y, (9.2)

$JC_y = \frac{c}{a+b+c}$ for iteration y. (9.3)

- CY-2 estimator (Cao et al. 2004): CY-2

$\hat{S}_{CY-2} = Slope + Intercept$ of the best fit linear regression line (10.1)

of $\overline{SR} - JC$ plot.

Chao estimators:

- Chao2 estimator, if $Q_2 > 0$ (Chao 1987, Chao and Shen 2003):

$$\hat{S}_{Chao2} = S_{obs} + \left(\frac{t-1}{2t}\right)\left(\frac{Q_1^2}{Q_2}\right) \quad (11.1)$$

- Chao2 estimator, if $Q_2 = 0$ (Chao 1987, Chao and Shen 2003):

$$\hat{S}_{Chao2} = S_{obs} + \left(\frac{t-1}{2t}\right)(Q_1-1)Q_1 \quad (11.2)$$

- Variance for Chao2 estimator, (Chao 1987):

$$\hat{\text{var}}\left(\hat{S}_{Chao2}\right) = Q_2\left[\frac{1}{2}\left(\frac{Q_1}{Q_2}\right)^2 + \left(\frac{Q_1}{Q_2}\right)^3 + \frac{1}{4}\left(\frac{Q_1}{Q_2}\right)^4\right] \quad (11.3)$$

- Chao2 estimator, bias-corrected (Chao 2005):

$$\hat{S}_{Chao2BC} = S_{obs} + \left(\frac{t-1}{t}\right)\left(\frac{Q_1(Q_1-1)}{2(Q_2+1)}\right) \quad (12.1)$$

- Variance for Chao2, bias-corrected estimator (Colwell 2006): (12.2)

$$\hat{\text{var}}(\hat{S}_{Chao2BC)} = \left(\frac{t-1}{t}\right)\frac{Q_1(Q_1-1)}{2(Q_2+1)} + \left(\frac{t-1}{t}\right)^2\frac{Q_1(2Q_1-1)^2}{4(Q_2+1)^2} + \left(\frac{t-1}{t}\right)^2\frac{Q_1^2Q_2(Q_1-1)^2}{4(Q_2+1)^4}$$

- Variance for Chao2, bias-corrected estimator, if $Q_1 > 0$ & $Q_2 = 0$ (Colwell 2006):

$$\hat{\text{var}}(\hat{S}_{Chao2BC}) = \left(\frac{t-1}{t}\right)\frac{Q_1(Q_1-1)}{2} + \left(\frac{t-1}{t}\right)^2 \frac{Q_1(2Q_1-1)^2}{4} - \left(\frac{t-1}{t}\right)^2 \frac{Q_1^4}{4\hat{S}_{Chao2BC}} \quad (12.3)$$

- Variance for Chao2, bias-corrected estimator, if $Q_1 = 0$ & $Q_2 \geq 0$ (Colwell 2006):

$$\hat{\text{var}}(\hat{S}_{Chao1BC}) = S_{obs}\left(e^{-(\frac{m}{Sobs})}\right)\left(1 - e^{-(\frac{m}{Sobs})}\right) \quad (12.4)$$

Coverage estimator:

- ICE estimator, (Lee and Chao 1994):
  - as in Magurran (2004) and EstimateS (Colwell 2006):

$$\hat{S}_{ICE} = S_{freq} + \frac{S_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}}\hat{\gamma}^2_{infreq}, \quad (13.1)$$

$$\text{where } S_{freq} = \sum_{j>d}^{n} Q_j, \quad (13.2)$$

$$S_{infreq} = \sum_{j=1}^{d} Q_j, \quad (13.3)$$

$$\hat{C}_{infreq} = 1 - \frac{Q_1}{\sum_{j=1}^{d} jQ_j}, \quad (13.4)$$

$$\hat{\gamma}^2_{infreq} = \max\left\{\frac{S_{infreq}}{\hat{C}_{infreq}}\frac{M_{infr}}{(M_{infr}-1)}\frac{\sum_{j=1}^{d} j(j-1)Q_j}{\left(\sum_{j=1}^{d} jQ_j\right)^2} - 1, 0\right\}. \quad (13.5)$$

  - as in SPADE user's manual, use for small heterogeneity (Chao and Shen 2003):

$$\hat{S}_{ICE} = S_{freq} + \frac{S_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}}\hat{\gamma}^2_{infreq}, \quad (14.1)$$

where $S_{freq} = \sum_{j>d}^{n} Q_j$, (14.2)

$$S_{infreq} = \sum_{j=1}^{d} Q_j \,, \tag{14.3}$$

$$\hat{C}_{infreq} = 1 - \frac{Q_1}{\sum_{j=1}^{d} jQ_j}\left[\frac{(t-1)Q_1}{(t-1)Q_1 + 2Q_2}\right], \tag{14.4}$$

$$\hat{\gamma}^2_{infreq} = \max\left\{\frac{S_{infreq}}{\hat{C}_{infreq}}\frac{t}{(t-1)}\frac{\sum_{j=1}^{d} j(j-1)Q_j}{\left(\sum_{j=1}^{d} jQ_j\right)\left(\sum_{j=1}^{d} jQ_j - 1\right)} - 1, 0\right\}. \tag{14.5}$$

- as in SPADE user's manual, use for large heterogeneity (Chao and Shen 2003):

$$\hat{S}_{ICE} = S_{freq} + \frac{S_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}}\tilde{\gamma}^2_{infreq} \,, \tag{15.1}$$

where $S_{freq} = \sum_{j>d}^{n} Q_j$, (15.2)

$$S_{infreq} = \sum_{j=1}^{d} Q_j \,, \tag{15.3}$$

$$\hat{C}_{infreq} = 1 - \frac{Q_1}{\sum_{j=1}^{d} jQ_j}\left[\frac{(t-1)Q_1}{(t-1)Q_1 + 2Q_2}\right], \tag{15.4}$$

$$\tilde{\gamma}^2_{infreq} = \max\left\{\left(\hat{S}_{ICE}\right)\frac{t}{(t-1)}\frac{\sum_{j=1}^{d} j(j-1)Q_j}{\left(\sum_{j=1}^{d} jQ_j\right)\left(\sum_{j=1}^{d} jQ_j - 1\right)} - 1, 0\right\}, \tag{15.5}$$

where $\hat{S}_{ICE}$ is the estimate from the small heterogeneity estimator (15.1).

Jackknife estimators:

- 1st-order jackknife estimator (Burnham and Overton 1978):

$$\hat{S}_{Jack1} = S_{obs} + \left(\frac{t-1}{t}\right)Q_1 \tag{16.1}$$

- 2nd-order jackknife estimator (Burnham and Overton 1978):

$$\hat{S}_{Jack2} = S_{obs} + \left(\frac{2t-3}{t}\right)Q_1 - \left(\frac{(t-2)^2}{t(t-1)}\right)Q_2 \tag{16.2}$$

- 3rd-order jackknife estimator (Burnham and Overton 1978):

$$\hat{S}_{Jack3} = S_{obs} + \left(\frac{3t-6}{t}\right)Q_1 - \left(\frac{3t^2-15t+19}{t(t-1)}\right)Q_2 + \left(\frac{(t-3)^3}{t(t-1)(t-2)}\right)Q_3 \tag{16.3}$$

- 4th-order jackknife estimator (Burnham and Overton 1978): (16.4)

$$\hat{S}_{Jack4} = S_{obs} + \left(\frac{4t-10}{t}\right)Q_1 - \left(\frac{6t^2-36t+55}{t(t-1)}\right)Q_2 + \left(\frac{4t^3-42t^2+148t-175}{t(t-1)(t-2)}\right)Q_3$$

$$- \left(\frac{(t-4)^4}{t(t-1)(t-2)(t-3)}\right)Q_4$$

- 5th-order jackknife estimator (Burnham and Overton 1978): (16.5)

$$\hat{S}_{Jack5} = S_{obs} + \left(\frac{5t-15}{t}\right)Q_1 - \left(\frac{10t^2-70t+125}{t(t-1)}\right)Q_2 + \left(\frac{10t^3-120t^2+485t-660}{t(t-1)(t-2)}\right)Q_3$$

$$- \left(\frac{(t-4)^5-(t-5)^5}{t(t-1)(t-2)(t-3)}\right)Q_4 + \left(\frac{(t-5)^5}{t(t-1)(t-2)(t-3)(t-4)}\right)Q_5$$

- Variance of the incidence-based jackknife estimators (Rexstad and Burnham 1992):

$$\hat{\text{var}}(\hat{S}_{JackK}) = \sum_{j=1}^{t} a_{jK}Q_j \tag{16.6}$$

- *Mixture estimator* (Pledger 2000):

$$\hat{S}_{Mix} = L\left(S_{true}, \{\pi\}, \{\theta\} \mid Z\right) = \frac{S_{true}!}{\left(S_{true} - S_{obs}\right)!} \prod_{j=0}^{S_{obs}} \left[\left\{ \sum_{g=1}^{G} \pi_g \theta_g^j \left(1-\theta_g\right)^{t-j} \right\}^{Q_j} \right] \tag{17.1}$$

EVENNESS ESTIMATOR:

- Shannon information index (Shannon and Weaver 1949):

$$Evenness = \left(\frac{-\sum_{k=1}^{Sobs} p_k (\ln p_k)}{\ln S_{obs}} \right) \tag{18.1}$$

SPECIES RICHNESS INDEX:

- Margalef's diversity index (Clifford and Stephenson 1975):

$$\hat{\mathrm{D}}_{\mathrm{Margalef}} = \frac{S_{obs} - 1}{\ln n} \tag{19.1}$$

- Menhininck's index (Whittaker 1977):

$$\hat{\mathrm{D}}_{\mathrm{Menhinick}} = \frac{S_{obs}}{\sqrt{n}} \tag{20.2}$$

ADDITIONAL SURVEY EFFORT FORMULAS:

- Number of individuals needed to reach fraction f, n_f, of $\hat{S}_{Chao1}$ (Chao et al. 2009):

$$n_f = \frac{n f_1}{2 f_2} \log\left[\frac{\hat{f}_0}{(1-f)\hat{S}_{Chao1}} \right], \tag{21.1}$$

where $\hat{f}_0 = f\hat{S}_{Chao1} - S_{obs}$.

- Number of surveys needed to reach fraction f, t_f, of $\hat{S}_{Chao2}$ (Chao et al. 2009):

$$t_f = \frac{\log\left[1 - \frac{t}{(t-1)} \frac{2Q_2}{Q_1^2} \left(f\hat{S}_{Chao2} - S_{obs} \right) \right]}{\log\left[1 - \frac{2Q_2}{(t-1)Q_1 + 2Q_2} \right]} \tag{22.1}$$

CONFIDENCE INTERVAL FORMULAS:

- Lower and upper bounds for a 95% confidence interval (Burnham et al. 1987, Chao 1987):

$$\left[\left(S_{obs}+\left(\frac{S_{est}-S_{obs}}{C}\right), S_{obs}+C\left(S_{est}-S_{obs}\right)\right)\right], \tag{23.1}$$

where $C=\exp\left\{1.96\left[\sqrt{\ln\left(1+\frac{\mathrm{var}(S_{est})}{\left(S_{est}-S_{obs}\right)^2}\right)}\right]\right\}$.

LITERATURE CITED

Ashbridge J. and Goudie I.B.J. 2000. Coverage-adjusted estimators for mark-recapture in heterogeneous populations. Communications in statistics. Simulation and computation 29:1215-1237.

Burnham K.P. and Overton W.S. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika 65:623-633.

Burnham K.P., Anderson D.R., White G.C., Brownie C. and Pollock K.H. 1987. Design and analysis methods for fish survival experiments based on release-recapture. American Fisheries Society Monograph No. 5. Bethesda, MD, USA. 437pp.

Cao Y., Larsen D.P. and Hughes R.M. 2001. Estimating total species richness in fish assemblage surveys: a similarity based approach. Canadian Journal of Fisheries and Aquatic Sciences 58:1782-1793.

Cao Y., Larsen D.P. and White D. 2004. Estimating regional species richness using a limited number of survey units. Ecoscience 11:23-35.

Chao A. 1984. Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11:265-270.

Chao A. 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783-791.

Chao A. 2005. Species richness estimation. *In* Balakrishnan N., Read C.B. and Vidakovic B. (eds.). *Encyclopedia of Statistical Sciences*. Wiley, New York, USA.

Chao A., Colwell R.K., Lin C.W. and Gotelli N.J. 2009. Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90:1125-1133.

Chao A. and Lee S.M. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association 87:210-217.

Chao A. and Shen T.-J. 2003. Program SPADE (Species prediction and diversity estimation). Program and user's guide. Published at: http://chao.stat.nthu.edu.tw.

Clifford H.T. and Stephenson W. 1975. *An introduction to numerical classification*. Academic Press, London, UK.

Colwell R.K. 2006. *EstimateS*: Statistical estimation of species richness and shared species from samples. Version 8. Persistent URL <purl.oclc.org/estimates>. Published at: http://viceroy.eeb.uconn.edu/EstimateS.

Darroch J.N. and Ratcliff D. 1980. A note on capture-recapture estimation. Biometrics 36:149-153.

Lee S.M. and Chao A. 1994. Estimating population size via sample coverage for closed capture-recapture models. Biometrics 50:88-97.

Magurran A.E. 2004. *Measuring Biological Diversity*. Blackwell Publishing, MA, USA.

Pledger S. 2000. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. Biometrics 56:434-442.

Rexstad E. and Burnham K. 1991. Users guide for interactive program CAPTURE. Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, CO, USA. 29pp.

Smith E.P. and van Belle G. 1984. Nonparametric estimation of species richness. Biometrics 40:119-129.

Shannon C.E. and Weaver W. 1949. *The mathematical theory of communication*. University of Illinois Press, Urbana, IL, USA.

Whittaker R.H. 1977. Evolution of species diversity in land communities. Evolutionary Biology 10:1-66.