Diversity and Distributions, (Diversity Distrib.) (2005) 11, 349–359

BIODIVERSITY RESEARCH

Strengths and weaknesses of museum and national survey data sets for predicting regional species richness: comparative and combined approaches

Robert Guralnick* and Jeremy Van Cleve

Department of Ecology and Evolutionary Biology and the University of Colorado Museum, University of Colorado, Boulder, CO 80309--0334, USA

*Correspondence: Robert Guralnick, Department of Ecology and Evolutionary Biology and the University of Colorado Museum, University of Colorado. Boulder, CO 80309–0334, USA.
Tel.: 303-735-0178;
E-mail: Robert.Guralnick@colorado.edu

ABSTRACT

We compiled three independent data sets of bird species occurrences in northeastern Colorado to test how predicted species richness compared to a combined analysis using all the data. The first data set was a georeferenced regional museum data set from two major repositories — the Denver Museum of Nature, and the Science and University of Colorado Museum. The two national survey data sets were the Breeding Bird Survey (summer), and the Great Backyard Bird Count (winter). Resulting analyses show that the museum data sets give richness estimates closest to the combined data set while exhibiting a skewed abundance distribution, whereas survey data sets do not accurately estimate overall richness even though they contain far more records. The combined data set allows the strengths of one data set to augment weaknesses in others. It is likely some museum data sets display skewed abundance distributions due to collectors' potentially self-selecting under-represented species over common ones.

Keywords

Biodiversity informatics, bird species richness, breeding bird survey, Great Backyard Bird Count, museum collections data, rare representation, species richness estimation.

INTRODUCTION

A major argument for the utility of natural history museums in the 21st century is that, taken in total, they contain one of the best sources of information on past and present biodiversity (Krishtalka & Humphrey, 2000; Suarez & Tsutsui, 2004). However, there are still hurdles to using this huge reserve of information. One challenge has been preparing museum data for use in ecological analyses. Over the last 10 years, advances in computing and data sharing over networks have led to solutions to some of the problems of digitizing, georeferencing and distributing data. In the immediate future, we anticipate a flood of museum data to be available to users worldwide over the Internet (<http://www.gbif.org> for an existing portal to millions of records).

As these data become available, two crucial problems need to be addressed. The first is determining if data from natural history collections adequately predict species richness (Meier & Dikow, 2004) for a given region, even if many records are available. Natural history collections are *ad hoc* data sets that have developed from efforts of multiple collectors over long periods of time. Even with potentially billions of specimen records available worldwide, questions remain: are the data resolved and unbiased enough (Soberón *et al.*, 2000) to be an appropriate sample at the spatial and temporal scales of interest? Even if it appears that museum data provide good estimates of species richness, do they match species richness estimates from more systematically collected data sets?

If museum data provide useful estimates of species richness, the second problem is how best to use this information to examine biodiversity and its relationship to environmental change. Species richness can be compared to ecological null models (e.g. the mid-domain effect; Colwell *et al.*, 2004) and correlated with spatiotemporal environmental parameters. Community level measures of biodiversity can be used to infer past biogeographical patterns. A major strength of natural history collections is the potential to determine change in species richness through time, especially at regional and continental scales.

In this study, we compare and combine independent sources of bird species occurrence for a region of the Southern Rockies (Fig. 1). We have chosen to examine bird species richness in this region for three reasons. First, birds are a diverse group (e.g. Cuerto & Casenave, 1999; Rahbek & Graves, 2001) especially in this region, and distributions likely reflect underlying past and present environmental factors. Second, birds are one of the few groups where multiple independent data sets on species richness are already available.

DOI: 10.1111/j.1366-9516.2005.00164.x

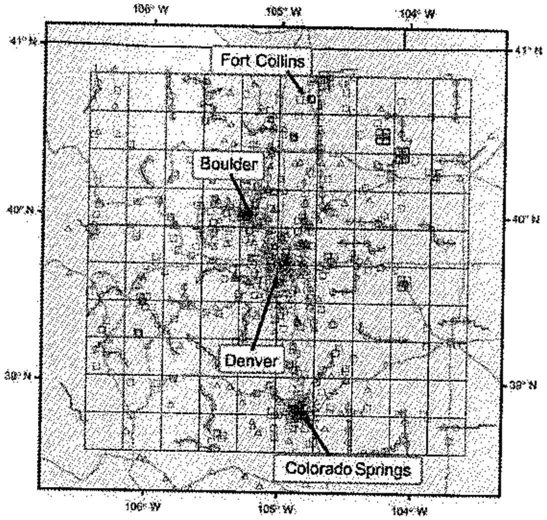

Figure 1 Study region and distribution of records for MAPSTEDI, BBS, and GBBC. The grid cells (fine black lines) in this figure are 25 km wide. Broken grey lines represent major roads, and the solid grey lines are the BBS routes. Each small square marks the location of one MAPSTEDI museum record; circles mark BBS route stops where observations are taken, and triangles mark zip code and city region centroids where GBBC records were grouped together. Both MAPSTEDI and GBBC records are clustered around the Front Range Metropolitan regions, and MAPSTEDI records show some clustering around major roads.

We utilize four data sources to examine species richness in the study area; two from regional museums and two from national observational surveys. The museum-based data set represents the combined species occurrences from the University of Colorado Museum (UCM) and the Denver Museum of Nature and Science (DMNS). The two observational data sets are independently collected sets of observations from the Breeding Bird Survey (BBS) and Greater Backyard Bird Count (GBBC).

The museum and observational survey data sets incorporate different strategies for sampling diversity. The museum collections represent a set of long-term and heterogeneous approaches to sampling that often focus on capturing total diversity rather than actual abundance distributions. The Breeding Bird Survey (BBS, <http://www.mp2-pwrc.usgs.gov/bbs/>) is more systematic in its approach to surveying bird species occurrence (Boulinier *et al.*, 1998) and has been used elsewhere for estimates of regional and continental species richness (Boulinier *et al.*, 1998; Nichols *et al.*, 1998; Cam *et al.*, 2000; Jones *et al.*, 2000). Each summer for almost 40 years, volunteers have followed well-defined survey routes by car, stopping at 0.5 mile increments to do 3-minute point counts by eye and ear, before travelling to the next interval. The data are then summarized in five segments of 10 stops each across the routes. For more details on sampling strategies and issues with the BBS see Boulinier *et al.* (1998). The GBBC is another observational data set collected from volunteers, but is less systematic than the BBS. Birders record species observations in their backyards and local areas during a 4-day period in February, and summary results are made available to users online (<http://gbbc.birdsource.org>). Taken together, the BBS and GBBC record occurrences for both summer (BBS) and winter (GBBC). Data from the Christmas Bird Count (CBC), another national survey, has a minimum 24-km error, and was considered too imprecise to utilize.

We also compiled a separate mammalian species occurrence data set for the same study region from the UCM and DMNS collections and from the Mammal Networked Information System (MaNIS, <http://elib.cs.berkeley.edu/manis/>, Stein & Wieczorek, 2004). MaNIS is a database portal to multiple United States museum mammal collections providers. We compiled this museum-only data set in order to compare patterns of abundance in a taxonomic group with lower richness against results from the avian data set. Collections of species rich taxa might misrepresent abundances of common species in order to maximize total species coverage. Also, one could expect a difference in collection strategy between the two taxa to result in a more balanced mammal abundance distribution and skewed bird abundance distribution or vice versa.

To summarize, the main questions we will address are:

1 Taken separately, do the museum data sources and the observational data sets give similar estimates of species richness over the whole study region, i.e. do the data sets sample the same set of species? Does the spatial sample size, or grain size, affect richness estimates?

2 Do the accumulation curves for each data set appear to reach their asymptotes? An accumulation curve that plateaus does not imply that the underlying data set is well-sampled and representative of

the whole region; rather, it only implies that the data set well represents a subset of data defined by the limits and biases of the sampling methods used to collect the data set. Based on our analyses, what are these biases in the data sets in terms of rare versus abundant species?

3 If the data sets are combined, are estimates of bird species richness for the region similar to estimates generated from any one data set alone? Do accumulation curves for the combined data set reach an asymptote, indicating a negligible return for more sampling effort using the combined collecting strategies? Which data sets have the most effect on the observed patterns in the combined data set?

4 How likely are biases in other museum data sets? Do patterns seen in the museum data sets for Aves appear consistent with a compiled data set of mammalian records for the region?

METHODS

Spatial extent

The spatial extent was defined as a square region 250 km long on each side encompassing 62 500 km^2 of the southern Rocky Mountains and adjacent foothills and plains regions. The boundaries, major cities and sampling efforts for different data sets are shown in Fig. 1. Using the Environmental Systems Research Institute's (ESRI) ARCMAP desktop version 8.3 (2002), the study region was divided into grid cells of different sizes: 10 × 10 km, 16.67 × 16.67 km, 25 × 25 km, and 50 × 50 km. Occurrence records were loaded into ARCMAP and transformed into point records based on their geographical location.

Data set preparation

For each data set, the following steps were employed to convert the records into a form usable for species richness estimates: (1) geospatially precise raw species occurrence data for the region were accumulated; (2) data were validated by verifying accepted taxonomic names of the species and removing spurious records; (3) spatial annotation and reformatting of the data were performed, and the data exported to species richness calculation programs. Additional information on data set preparation can be found under 'Supplementary material'. Both the UCM and DMNS records were digitized and georeferenced (Neufeld *et al.*, 2003; Murphey *et al.*, 2004) and made available as part of the biodiversity informatics (MAPSTEDI (mountains and plains spatio-temporal database informatics), <http://mapstedi.org>). As records from both of these museum data sets were downloaded from <http://mapstedi.org> and combined into one, this will be referred to as the MAPSTEDI data set.

Species richness estimation

Species richness estimation methods are usually grouped into three categories (Colwell & Coddington, 1994): fitting curves to species accumulation functions, parametric models of relative abundance, and non-parametric estimators based on rarity. A collection of the most common of these was used to assess species richness. We did not subject the mammal data set to such estimates, but did plot the relative abundance distribution.

The curve fitting model used was the Michaelis-Menten equation (Raaijmakers, 1987; Colwell & Coddington, 1994), and the asymptote was calculated using the MMMean estimator in ESTIMATES (R.K. Colwell, versions 5 and 6, <http://viceroy.eeb.uconn.edu/EstimateS>, 2000). The relative abundance distributions used were the continuous log-normal (Preston, 1948), Poisson log-normal (Bulmer, 1974), and log-series (Fisher *et al.*, 1943). Parametric abundance distributions were fit to the data using the statistical computing environment R (version 1.9.0, <http://www.r-project.org>, 2004). The continuous log-normal was fit using the R function, prestondistr in J. Oksanen's Vegan package (version 1.6–3, <http://cc.oulu.fi/~jarioksa/softhelp/vegan.html>, 2004). The Vegan package function fisherfit was used to fit the log-series. Fitting the Poisson log-normal required building a custom R subroutine to evaluate Poisson log-normal probability functions and using the fitdistr function available in the MASS package (Venables & Ripley, 2002). Data were plotted on a log$_2$ scale where each abundance class, or octave, represented 1, 2, 3–4, 5–8, … occurrences (Fig. 2). Predictions from all three parametric distributions were compared to empirical abundance distributions by gathering data in log$_2$ octaves and computing *P* values using a chi-squared test. Small *P* values suggest that the null hypothesis, namely that the empirical and parametric distributions are the same, can be rejected.

Non-parametric estimators are abundance or incidence-based and were developed to estimate the number of species in a random sample from a single population (Colwell & Coddington, 1994). It is important to note that museum data do not constitute a random sample and are gathered from multiple populations. Nonetheless, we use these methods in the hope that they can provide some useful information when applied to collections data. Abundance-based estimators use the relative abundance of each species in the whole sample to estimate species richness. The abundance-based estimators employed were: Chao1 (Chao, 1984) and ACE (Abundance-based Coverage Estimator) (Chao & Lee, 1992; Chao *et al.*, 1993). Incidence-based estimators use the relative rarity and commonness of species in subsamples of the complete sample to estimate richness. Estimators of this variety used in this study were: bootstrap (Smith & van Belle, 1984), first and second order jackknife (Jack1 and Jack 2, Burnham & Overton, 1978, 1979; Heltshe & Forrester, 1983; Smith & van Belle, 1984), Chao2 (Chao, 1987), and ICE (Lee & Chao, 1994). Non-parametric estimation was performed using ESTIMATES with 100 randomizations of sample order.

Species–area curves

Using grid cells of equal area allowed generation of species–area curves from the data. The shape of these curves can reveal characteristics such as habitat heterogeneity (Rosenzweig, 1995) and provide a basis for comparing the data sets used in this study with others found in the literature. Following Rosenzweig's recommendations (1995), sampling was done using contiguous, nested subplots. One thousand random nested subplots were created (details in 'Supplementary material'), and the number of species

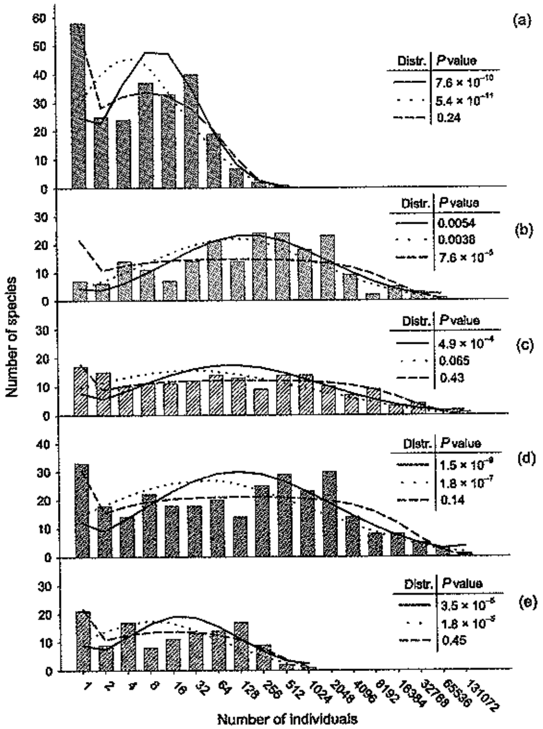

Figure 2 Relative abundance distributions plotted on a log$_2$, or octave, scale. Panel (a) displays MAPSTEDI data (b) BBS data (c) GBBC data (d) combined MAPSTEDI-BBS-GBBC data, and (e) MAPSTEDI-MaNIS mammal data. Fitted parametric abundance distributions are plotted on top of the empirical data. The solid line represents the log-series, the dotted line represents the continuous log-normal and the broken line represents the Poisson log-normal. The *P* values are listed alongside each plot for each distribution. The only data set clearly displaying a mode to the right of the first octave is the BBS. The MAPSTEDI, combined, and MAPSTEDI-MaNIS data sets all have peak abundances in the first octave and smaller peak abundances in later octaves. This pattern is reflected in log-series *P* values that are at least four orders of magnitude larger than *P* values for either the continuous or Poisson log-normal for each of those three data sets.

was calculated for each subplot and averaged over the 1000 orderings to create empirical species-area plots. The curve parameters were estimated in SIGMAPLOT using non-linear regression (Rosenzweig, 1995).

RESULTS

Spatial biases in data sets

Spatially referencing all records revealed a strong bias towards collecting along roads and around major population centres (Fig. 1). The BBS survey design builds in roadside bias, while the tendency of collectors to remain on roads for access explains the bias in MAPSTEDI records (e.g. Soberón *et al.*, 2000). GBBC records show less clustering around roads, probably as a result of the displacement of records towards their zip code or city centroids (see Supplemental materials and Methods). Both GBBC and MAPSTEDI data cluster near the heavily populated Front Range regions. BBS routes, on the other hand, are fairly evenly distributed by design.

Observed species numbers and sampling effort

Observed species richness (S_{obs}) and estimates for species richness from the four different grids and the four data sets are listed in Table 1. S_{obs} is highest in the combined data set (303 species) followed by MAPSTEDI with S_{obs} of 247, BBS at 203, and GBBC at 174. Differences in the number of occurrence records among data sets are evident in accumulation curves (Fig. 3). The curves for the BBS and GBBC are near their asymptotes, which indicate that a moderate increase in sampling will reach the maximum number of species observable under each set of sampling methodologies. In comparison, the MAPSTEDI and combined data sets appear to be far from their methodologically imposed sampling limits as their curves continue to increase as all the grid cells are pooled.

Species estimator trends

Differences in estimated species richness between the different grid sizes were small (Table 1) and species accumulation curves

Table 1 Species richness estimator and species–area curve results. The percent increase is [(estimated value/S_{obs}) — (1) × 100] and in bold to emphasize the overall difference in performance between data sets for each estimator. Percent relative error (% RE) is standard deviation of estimator values for the four grids divided by S_{obs}

	625 cells		225 cells		100 cells		25 cells		Average		
MAPSTEDI estimator	Value	% increase	Value	% increase	Value	% increase	Value	% increase	Value	% increase	% RE
S_{obs}	246.00		246.00		246.00		246.00		246.00		
ACE	298.17	21.21	298.17	21.21	298.17	21.21	298.17	21.21	298.17	**21.21**	
Chao 1	309.62	25.86	309.62	25.86	309.62	25.86	309.62	25.86	309.62	**25.86**	
Bootstrap	278.35	13.15	279.85	13.76	279.20	13.50	280.70	14.11	279.53	**13.63**	0.35
Jack 1	317.88	29.22	319.67	29.95	319.26	29.78	319.92	30.05	319.18	**29.75**	0.32
Jack 2	359.80	46.26	357.49	45.32	360.74	46.64	357.26	45.23	358.82	**45.86**	0.60
ICE	297.69	21.01	298.94	21.52	299.50	21.75	305.97	24.38	300.53	**22.16**	1.31
Chao 2	328.49	33.53	319.03	29.69	327.88	33.28	321.05	30.51	324.11	**31.75**	1.68
MMMean	293.27	19.22	302.71	23.05	300.97	22.35	333.53	35.58	307.62	**25.05**	6.25
Poisson log-normal	267.82	8.87	267.82	8.87	267.82	8.87	267.82	8.87	267.82	**8.87**	
Continuous log-normal	269.17	9.42	269.17	9.42	269.17	9.42	269.17	9.42	269.17	**9.42**	
Log-series	246.00	0.00	246.00	0.00	246.00	0.00	246.00	0.00	246.00	**0.00**	
Species–area curve	264.65	7.58	263.95	7.30	265.72	8.02	263.60	7.15	264.48	**7.51**	0.33
BBS											
S_{obs}	203.00		203.00		203.00		203.00		203.00		
ACE	205.94	1.45	205.94	1.45	205.94	1.45	205.94	1.45	205.94	**1.45**	
Chao 1	206.07	1.51	206.07	1.51	206.07	1.51	206.07	1.51	206.07	**1.51**	
Bootstrap	212.84	4.85	213.03	4.94	214.97	5.90	215.21	6.01	214.01	**5.42**	0.53
Jack 1	221.97	9.34	221.92	9.32	227.75	12.19	227.96	12.30	224.90	**10.79**	1.46
Jack 2	223.99	10.34	222.00	9.36	234.79	15.66	236.01	16.26	229.20	**12.91**	3.08
ICE	214.92	5.87	214.37	5.60	219.68	8.22	218.05	7.41	216.76	**6.78**	1.08
Chao 2	212.53	4.69	211.57	4.22	218.82	7.79	220.14	8.44	215.77	**6.29**	1.85
MMMean	204.09	0.54	206.97	1.96	209.42	3.16	215.77	6.29	209.06	**2.99**	2.12
Poisson log-normal	207.69	2.31	207.69	2.31	207.69	2.31	207.69	2.31	207.69	**2.31**	
Continuous log-normal	206.90	1.92	206.90	1.92	206.90	1.92	206.90	1.92	206.90	**1.92**	
Log-series	203.00	0.00	203.00	0.00	203.00	0.00	203.00	0.00	203.00	**0.00**	
Species–area curve	212.03	4.45	211.83	4.35	211.94	4.40	210.54	3.72	211.59	**4.23**	0.30
GBBC											
S_{obs}	174.00		174.00		174.00		174.00		174.00		
ACE	187.84	7.95	187.84	7.95	187.84	7.95	187.84	7.95	187.84	**7.95**	
Chao 1	183.63	5.53	182.53	4.90	182.53	4.90	182.53	4.90	182.81	**5.06**	
Bootstrap	187.57	7.80	188.46	8.31	188.47	8.32	189.61	8.97	188.53	**8.35**	0.42
Jack 1	202.95	16.64	205.86	18.31	205.68	18.21	207.60	19.31	205.52	**18.12**	0.96
Jack 2	214.94	23.53	222.77	28.03	222.49	27.87	224.81	29.20	221.25	**27.16**	2.16
ICE	200.96	15.49	203.15	16.75	202.45	16.35	202.18	16.20	202.19	**16.20**	0.45
Chao 2	198.74	14.22	205.06	17.85	205.06	17.85	207.11	19.03	203.99	**17.24**	1.81
MMMean	162.99	−6.33	166.94	−4.06	171.37	−1.51	183.19	5.28	171.12	**−1.65**	4.35
Poisson log-normal	186.03	6.91	186.03	6.91	186.03	6.91	186.03	6.91	186.03	**6.91**	
Continuous log-normal	198.09	13.84	198.09	13.84	198.09	13.84	198.09	13.84	198.09	**13.84**	
Log-series	174.00	0.00	174.00	0.00	174.00	0.00	174.00	0.00	174.00	**0.00**	
Species–area curve	186.02	6.91	185.61	6.67	185.49	6.61	185.06	6.36	185.55	**6.64**	0.20
MAPSTEDI-BBS-GBBC											
S_{obs}	303.00		303.00		303.00		303.00		303.00		
ACE	332.44	9.72	332.44	9.72	332.44	9.72	332.44	9.72	332.44	**9.72**	
Chao 1	333.25	9.98	330.84	9.19	330.84	9.19	330.84	9.19	331.44	**9.39**	
Bootstrap	323.33	6.71	323.65	6.82	324.02	6.94	324.62	7.14	323.91	**6.90**	0.16
Jack 1	346.93	14.50	346.80	14.46	347.55	14.70	349.08	15.21	347.59	**14.72**	0.30
Jack 2	367.90	21.42	364.76	20.38	365.46	20.61	370.32	22.22	367.11	**21.16**	0.72
ICE	341.15	12.59	339.67	12.10	339.08	11.91	338.40	11.68	339.58	**12.07**	0.33
Chao 2	345.09	13.89	338.07	11.57	338.39	11.68	344.81	13.80	341.59	**12.74**	1.11
MMMean	287.93	−4.97	292.68	−3.41	300.56	−0.81	313.79	3.56	298.74	**−1.41**	3.23
Poisson log-normal	321.54	6.12	321.54	6.12	321.54	6.12	321.54	6.12	321.54	**6.12**	
Continuous log-normal	336.73	11.13	336.73	11.13	336.73	11.13	336.73	11.13	336.73	**11.13**	
Log-series	303.00	0.00	303.00	0.00	303.00	0.00	303.00	0.00	303.00	**0.00**	
Species–area curve	318.07	4.97	317.27	4.71	317.52	4.79	315.51	4.13	317.09	**4.65**	0.32

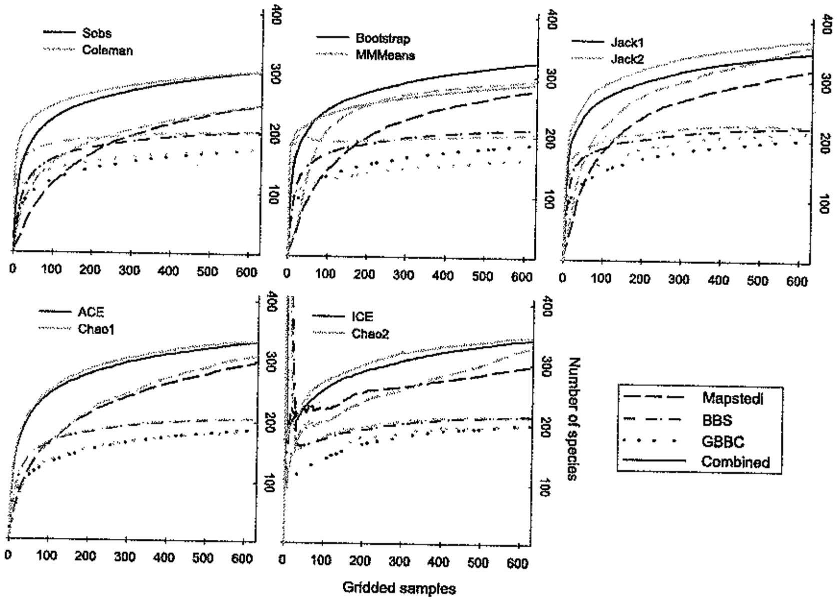

Figure 3 Species accumulation curves and estimated richness curves for the 10 × 10 km cell grid (625 total cells). Coleman is the individual based rarefaction curve and Sobs is the sample based rarefaction curve (see Gotelli & Colwell, 2001 for a detailed explanation of both types of rarefaction curve). The BBS and GBBC data set curves nearly reach asymptote while the MAPSTEDI and combined data set curves are still increasing when all samples are added.

for each grid were nearly indistinguishable (see Fig. 3 for curves from the 10 × 10 km cell grid and 'Supplementary material' for curves from the other grids). The non-parametric incidence based estimators generally provided the highest richness estimates across the data sets. The estimator with the highest richness values was Jack2 averaging between 45% and 13% above S_{obs} (bold values in Table 1). Jack1 provided the second highest estimates, 30% to 11% above S_{obs}, except for the MAPSTEDI data set, where it was third. Chao2 and ICE usually provided the third and fourth highest estimates. The Chao2 and ICE estimates ranged from 32% to 6% and 22% to 7% above S_{obs}, respectively. On all the data sets, the bootstrap produced low estimates. The results for the abundance based estimators, ACE and Chao1, were mixed. Chao1 provided low estimates for the BBS, GBBC, and combined data sets, 9% to 1% increase over S_{obs}, while it gave a moderate estimate of a 26% increase for the MAPSTEDI data set. The ACE estimator followed a similar pattern.

The species accumulation curve extrapolation estimator MMMean yielded moderate estimates with an average increase of 25% over S_{obs} on the MAPSTEDI data set but low estimates on the others. MMMean estimated a 3% increase over S_{obs} on the BBS and actually averaged to a decrease below S_{obs} for the GBBC and combined data sets. This decrease below S_{obs} is likely due to a poor fit of the Michaelis-Menten function to the observed accumulation curve (R. Colwell, personal communication 2004).

The two parametric distributions that provided richness estimates, Poisson and continuous log-normal, generally produced low estimates compared to non-parametric estimators, 14% to 9% over S_{obs} and 7% to 2% over S_{obs}, respectively. The Poisson log-normal almost always provided lower estimates than continuous log-normal, except for the BBS where its estimate is higher by a single species.

Species–area curves

Across the four different grid sizes, the fitted species–area curves were very similar for each bird data set. The *z*-values averaged across the four different grids for MAPSTEDI, BBS, GBBC, and MAPSTEDI-BBS-GBBC were, respectively: 0.513 ± 0.005, 0.275 ± 0.006, 0.304 ± 0.006, and 0.268 ± 0.004. All fitted curves also had high r^2 values greater than 0.95. The shapes of the empirical curves were slightly different from the fitted curves (see Fig. 4). Because

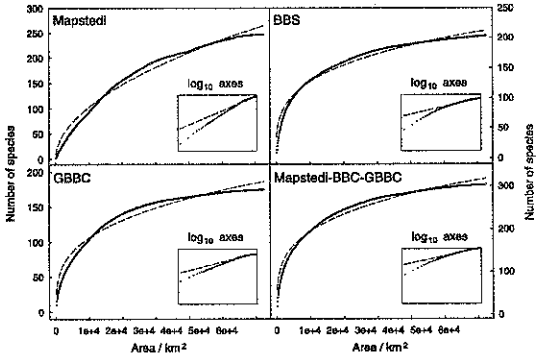

Figure 4 Empirical and fitted species area curves for 10 × 10 km cell grid. Solid lines denote the empirical curves and dashed lines denote the fitted ones. Each inset plot represents the same data plotted on a log-log scale. Apparent in all the datasets, especially in Mapstedi, is a tendency for the empirical curve to have a steeper initial slope that tapers off more gradually than the fitted curve. This results in a hump in the empirical curve near midpoint of the area-axis where the empirical curve overshoots the fitted curve.

the fitted species richness values at 62 500 km^2 were higher than the empirical values, these were also included in Table 1 to compare with other estimators. The species–area curve estimates were low (4–8% increase over S_{obs}) compared to most other estimators.

Relative abundance and richness estimates

The MAPSTEDI data set saw the largest increases in species richness measured as a fraction of S_{obs} (see bold values in Table 1). This pattern is likely as a result of the high fraction of rare species in the MAPSTEDI data set (Table 2). When compared with the two field data sets and even the combined data set, MAPSTEDI contains more uniques and nearly twice as many singletons as a fraction of S_{obs}. Because the non-parametric richness estimators are primarily a function of rare and uncommon species in the data, the prevalence of rare records in MAPSTEDI results in high richness estimates. Overall, the GBBC data set has the second highest overall richness estimates as a fraction of S_{obs}. The GBBC fraction of uniques was higher than in the BBS or combined data sets but its fraction of singletons was lower than the combined, likely explaining why abundance-based estimators like ACE and Chao1 are higher in the combined data set. The BBS had the lowest fraction of singletons and uniques, and thus, the lowest richness estimates as a fraction of S_{obs}.

Each of the parametric distributions was fit to the data sets (Fig. 2). The log-series is a much better fit to the MAPSTEDI data ($P = 0.24$) than either the Poisson or the continuous log-normal distribution ($P = 8 \times 10^{-10}$ and $P = 5 \times 10^{-11}$, respectively). The MAPSTEDI-BBS-GBBC and GBBC data sets also showed this pattern, although the contrasts between P values for the log-series and the two log-normal distributions were smaller. The reverse pattern was found for the BBS data set where both log-normal distributions fit better than the log-series ($P = 0.004, 0.005$ for the continuous and Poisson log-normal and $P = 0.0001$ for the log-series). The highest P values were found with log-series fits, although none was greater than 0.5.

Log$_2$ octave relative abundance distributions and associated fits were also generated for the mammal data set. The plot (Fig. 2) reveals a pattern similar to that in the MAPSTEDI and combined data sets, with the main peak in the first octave and at least one smaller peak in the larger octaves. This indicates that a significant proportion of the 122 species in the MAPSTEDI-MaNIS mammal data set were rare. The P values for the mammal data are also

Table 2 Rare species (i.e. under-represented in the sample). The number of species in each rarity class is listed along with the fraction of S_{obs} in parentheses. Singleton and doubleton species are defined as having only one and two recorded individuals in the whole data set whereas unique and duplicate species occur in only one and two of the samples. All non-parametric estimators used in this study are a function of one or more of these rarity classes

Grid size (# cells)	Singletons	Doubletons	Uniques	Duplicates
MAPSTEDI				
625	58 (0.24)	25 (0.10)	72 (0.29)	30 (0.12)
225	58 (0.24)	25 (0.10)	74 (0.30)	36 (0.15)
100	58 (0.24)	25 (0.10)	74 (0.30)	32 (0.13)
25	58 (0.24)	25 (0.10)	77 (0.31)	38 (0.15)
BBS				
625	7 (0.03)	6 (0.03)	19 (0.09)	17 (0.08)
225	7 (0.03)	6 (0.03)	19 (0.09)	19 (0.09)
100	7 (0.03)	6 (0.03)	25 (0.12)	18 (0.09)
25	7 (0.03)	6 (0.03)	26 (0.13)	18 (0.09)
GBBC				
625	17 (0.10)	15 (0.09)	29 (0.17)	17 (0.10)
225	17 (0.10)	15 (0.09)	32 (0.18)	15 (0.09)
100	17 (0.10)	15 (0.09)	32 (0.18)	15 (0.09)
25	17 (0.10)	15 (0.09)	35 (0.20)	17 (0.10)
MAPSTEDI-BBS-GBBC				
625	33 (0.11)	18 (0.06)	44 (0.15)	23 (0.08)
225	33 (0.11)	18 (0.06)	44 (0.15)	26 (0.09)
100	33 (0.11)	18 (0.06)	45 (0.15)	27 (0.09)
25	33 (0.11)	18 (0.06)	48 (0.16)	26 (0.09)

similar to the Aves MAPSTEDI data set. The continuous and Poisson log-normal have poor fits ($P = 1.8 \times 10^{-5}$ and 3.5×10^{-5}, respectively) while the log-series fits much better ($P = 0.45$). Compared against the MAPSTEDI Aves data, the MAPSTEDI-MaNIS mammal log$_2$ abundance distribution is broader and flatter. This difference likely reflects that the mammal data have 50% more records distributed among less than half the species of the bird data.

DISCUSSION

Estimated species richness from the combined data set likely best approximates the true regional species richness for two reasons. First, each individual data set had unique species records not found in any of the other data sets. MAPSTEDI has the most unique species (49), and both the BBS and the GBBC had less than half that number (21). All of these species have been confirmed to potentially occur in Colorado via reference to species lists available from online sources (Colorado Division of Wildlife, <http://ndis.nrel.colostate.edu/wildlife.asp>). Second, the overall estimates appear consistent with known number of bird species in the region. The state of Colorado, a larger region than examined here, has approximately 465 species (<http://www.camacdonald.com/birding/uscolorado.htm>) whereas Rocky Mountain National Park, a smaller region within our sampling area has 280 (<http://www.nps.gov/romo/resources/plantsandanimals/names/birds.html>). Our estimates for the mid-sized region studied here is approximately 350 species. Although this number has not been directly independently verified, it does appear to fall between the values above for smaller and larger regions above.

The museum data set has the fewest species occurrence records of the data sets used by almost two orders of magnitude. Despite the limited number of occurrences, the number of observed species (S_{obs}) is much higher (Table 1) than in either survey data set. We considered two alternatives to explain the results. One is that BBS and GBBC data are collected at limited times during the year, whereas the Museum data are not and therefore each survey data set misses some seasonally occurring birds. To examine this further, we accumulated life history information for regional species from the Colorado Division of Wildlife (<http://ndis.nrel.colostate.edu/wildlife.asp>) and then determined seasonal trends in collecting. MAPSTEDI data show a weak summer species skew (47% summer residents or migrants and 35% winter migrants or residents). The BBS data show a slightly stronger skew towards summer residents or migrants (55% summer vs. 36% winter). Winter migrants or residents are slightly more common in the GBBC (50% winter, 49% summer). Although there is a trend for data sets to preferentially cover one season or another, it is not strong enough to explain why MAPSTEDI museum data has the highest S_{obs}.

A more likely explanation is that the museum records are collected with an emphasis on rare specimens, what we call the 'rare representation' effect. Both professional and amateur collectors are more likely to focus on novel or rare finds and collect those while leaving behind the more common species that would take up valuable space in already cramped collections cabinets. Also, museum collections often represent a longer record of collecting than surveying data sets, and it is possible that they might accumulate more 'odd' or 'rare' years where individuals are found outside their normal ranges. As a result, collections include fewer samples of common individuals and species across their normal range in order to capture as many species as possible. If true, museum collections may have excellent coverage of species from a region, but display more flat or distorted abundance distributions.

Performance of estimators for species richness

For each data set, the estimators in our study present a wide range of improvements over S_{obs}. Studies using field data (e.g. Palmer, 1990, 1991; Colwell & Coddington, 1994; Walther & Morand, 1998; Chiarucci *et al.*, 2003) and simulated data (e.g. Walther & Morand, 1998; Brose *et al.*, 2003) suggest that some of these estimators may be less biased than others when sampling is less than optimal. Some work (Palmer, 1990, 1991; Colwell & Coddington, 1994; Hellmann & Fowler, 1999; Chiarucci *et al.*, 2003) concludes that of the estimators used in this study, the second order jackknife is the least biased. There is less consensus about bias in the remaining estimators, although the first order jackknife sometimes comes in second (Walther & Morand, 1998; Chiarucci *et al.*, 2003). The bootstrap showed a larger bias than both jackknife estimators and the Chao2 (Palmer, 1990, 1991; Colwell & Coddington, 1994; Walther & Morand, 1998; Hellmann & Fowler, 1999; Chiarucci *et al.*, 2003).

Estimators appear to show higher estimates of richness with decreasing bias and that pattern is seen here: the least biased estimator, Jack2, has the largest increase over S_{obs} followed by Jack1, Chao2, and Bootstrap. Other studies using museum collections data reported similar results; Petersen *et al.* (2003) used Diptera collections data with Jack2 yielding the highest estimates followed by Jack1, Chao2, ICE, MMMean, Chao1, and bootstrap. Meier and Dikow (2004) looked at a smaller set of estimators and saw high estimates from Jack2 followed by ICE, Bootstrap, and Chao1. Although so far consistent for museum data sets, this pattern does not hold for field data sets (e.g. Chazdon *et al.*, 1998). The fact that the least biased estimators provided the highest estimates of species richness lends support to the high jackknife estimates in MAPSTEDI and the combined data set. Notably, the Jack2 estimates from MAPSTEDI and the combined data sets are remarkably close, 358.82 and 367.11 species, respectively. The Jack2 estimator, like other incidence-based non-parametric estimators, is a function of the number of unique and duplicate species (i.e. species present in only one and two samples, respectively); only the Jack2 estimator is a monotonically increasing function of unique species and a monotonically decreasing function of duplicate species (Colwell & Coddington, 1994). Thus, the high Jack2 estimate for the MAPSTEDI data set is probably explained by the fact that the latter had nearly twice the fraction of unique species as any other data set (see Table 2) and the highest S_{obs} of any single data set.

Advantages of combined approaches

Both the survey and museum data sets reviewed here have their strengths and weaknesses. The survey data sets have many records and appear close to their maximum observable number of species given their sampling methodologies even though they have fewer observed species. In this case study, museum data have better representation of species but have fewer records and show an accumulation curve far from its asymptote. The combined data set appears to have a mixture of these qualities; it shows an accumulation curve closer to its asymptote than the museum data and represents the most number of species. Although improved compared to the museum data set, curves for those combined still do not reach an asymptote. Given the two order of magnitude increase in the number of records in the combined data set versus the museum data set, the improvement in the combined estimation was less than we anticipated. Nonetheless, the improved shape of the accumulation curve and highest S_{obs} strongly suggests that combined approaches using museum collection information along with survey information may give the overall best estimates of species richness. The GBBC and BBS do not well represent regional species richness given the observed and estimated richness in the combined analysis. In this sense, the museum data set appears to more precisely reflect true species richness although it shows a distorted abundance distribution and increasing accumulation curve.

Generality of results presented here

To understand whether our results are unique to this study group and area, we examined patterns of rarity and commonness in a mammal data set from the same region. Mammals are less diverse than birds in Colorado, and there are slightly more museum records available (approximately 5000 instead of 4000) for the study region. When we plotted occurrences versus the number of species on our $\log_2$ octave plots, mammal data appear to show patterns similar to those seen in the avian museum data set. The proportion of singletons and doubletons is lower in the mammal than in the bird data set, but rare species are still the most common category. The combined mammal museum data set has a $\log_2$ octave plot shape that appears most similar to the overall combined museum-BBS-GBBC data for birds.

If museum collectors do take a 'rare representation' perspective to collecting biodiversity information, we argue that results presented here should be applicable to regionally diverse groups. The more diverse the group, the more likely it is that many species will appear rare in the museum collections. For example, Longino *et al.* (2002) show that rare specimens are more dominant in a spider sample collected by a specialist than in mass sampling methods. While we believe this argument to be generally true, there are some groups that may be idiosyncratically well sampled because of individual collector's area of research interest or expertise. This claim also depends in part on the effort required to collect specimens, which is higher for some taxa than others. A lower effort per specimen could result in a larger specimen collection, a less skewed abundance distribution, and the appearance of better sampling. It is also possible that differing collecting strategies play a role in sampling adequacy. For example, snap-traps are the most common method for accumulating mammal specimens, whereas with birds and other taxa, the collector selects samples. Samples accumulated by traps and not selected by the collector could exhibit less potential bias. Additionally, an abundance of rare records could depend on the number of species occurrence records available for the region examined. In this study, we were able to accumulate several thousand museum records, and it is unknown how the patterns scale with orders of magnitude more records. We doubt that much more data are available for this area at present. We performed an online database check of mammal and bird records for Colorado available from institutions not included here (e.g. the National Museum of Natural History) and found few records, if any, that could be used. However, many potentially usable specimen records have yet to be assigned explicit computer-readable geospatial coordinates (i.e. retrospectively georeferenced), a necessary first step for inclusion in this study. As more georeferenced specimens become available, it may be possible to test how much more data are needed before sampling from museum data sets appears satisfactory. Our work is a case study of the use of multiple data sets to understand regional species richness. Similar approaches in different regions, and across broader geographical scales, will provide further tests of the generality of the results presented here. Despite the daunting tasks ahead of documenting and understanding patterns of biodiversity, we believe the time is right to begin leveraging multiple data sets to approach these crucial questions and evaluate the strengths and limitations of these data sets.

ACKNOWLEDGEMENTS

We would like to thank the data administrators at the Breeding Bird Survey and Greater Backyard Bird Count for their help accumulating species occurrences from their databases. Dave Neufeld and Gregory Hill provided us useful feedback at the inception of this project. The MAPSTEDI team, especially Dave Neufeld, Ryan Allen, Robbie Glaubitz, and Paul Murphey, were instrumental in making this project feasible through georeferencing museum data and converting collections databases into an online geodatabase. David M. Armstrong, Sharon Collinge, Robert K. Colwell, Alex Cruz, and John Wieczorek provided comments on a previous version of the manuscript. NSF grant DBI-0110133 to PI R. P. Guralnick helped support this work.

SUPPLEMENTARY MATERIAL

The following supplementary material is available at <http://www.mapstedi.org/supp_mat.html>

Supplementary material and methods describes some data formatting, georeferencing, and data analysis methods in more detail.

Figure S1: Species accumulation curves and estimates richness curves for all Aves data sets on the 16.67 × 16.67 km cell grid.

Figure S2: Species accumulation curves and estimates richness curves for all Aves data sets on the 25 × 25 km cell grid.

Figure S3: Species accumulation curves and estimates richness curves for all Aves data sets on the 50 × 50 km cell grid.

Figure S4: Pearson correlation coefficient values between each of three data set pairs. Correlation values are shown for each grid cell in the study area and for each of the four different grids.

Table S1: Species lists and number of occurrences for the MAPSTEDI, BBS, and GBBC data sets.

REFERENCES

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

Brose, U., Martinez, N.D. & 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 log-normal distribution to species-abundance data. *Biometrics*, **30**, 101–110.

Burnham, K.P. & 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. & 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., Hines, J.E. & Flather, C.H. (2000) Relative species richness and community completeness: Birds and urbanization in the Mid-Atlantic states. *Ecological Applications*, **10**, 1196–1210.

Chao, A. (1984) Non-parametric estimation of the number of classes in a population. *Scandanavian 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. & Lee, S.-M. (1992) Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, **87**, 210–217.

Chao, A., Ma, M.-C. & Yang, M.C.K. (1993) Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, **80**, 193–201.

Chazdon, R.L., Colwell, R.K., Denslow, J.S. & Guariguata, M.R. (1998) Statistical methods for estimating species richness of woody regeneration in primary and secondary rainforests of northeastern Costa Rica. *Forest biodiversity research, monitoring and modeling: conceptual background and old world case studies* (ed. by F. Dallmeier and J.A. Comiskey), pp. 285–309. Parthenon Publishing Group, Paris.

Chiarucci, A., Enright, N.J., Perry, G.L.W. & 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. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London, Series B*, **345**, 101–118.

Colwell, R.K., Rahbek, C. & Gotelli, N. (2004) The mid-domain effect and species richness patterns: what have we learned so far? *American Naturalist*, **163**, E1–E23.

Cuerto, V.R. & de Casenave, J.L. (1999) Determinants of bird species richness: role of climate and vegetation structure at a regional scale. *Journal of Biogeography*, **26**, 487–492.

Fisher, R.A., Corbet, A.S. & Williams, C.B. (1943) The relation between the number of species and the number of individuals in a random sample of animal population. *Journal of Animal Ecology*, **12**, 42–58.

Gotelli, N.J. & 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. & Fowler, G.W. (1999) Bias, precision, and accuracy of four measures of species richness. *Ecological Applications*, **9**, 824–834.

Heltshe, J. & Forrester, N.E. (1983) Estimating species richness using the jackknife procedure. *Biometrics*, **39**, 1–11.

Jones, K.B., Neale, A.C., Nash, M.S., Riitters, K.H., Wickham, J.D., O'Neill, R.V. & Van Remortel, R.D. (2000) Landscape correlates of breeding bird richness across the United States mid-Atlantic region. *Environmental Monitoring and Assessment*, **63**, 159–174.

Krishtalka, L. & Humphrey, P.S. (2000) Can natural history museums capture the future? *Bioscience*, **50**, 611–617.

Lee, S.-M. & Chao, A. (1994) Estimating population size via sample coverage for closed capture–recapture models. *Biometrics*, **50**, 88–97.

Longino, J.T., Coddington, J.A. & Colwell, R.K. (2002) The ant fauna of a tropical rain forest: estimating species richness three different ways. *Ecology*, **83**, 689–702.

Meier, R. & Dikow, T. (2004) Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. *Conservation Biology*, **18**, 478–488.

Murphey, P.C., Guralnick, R.P., Glaubitz, R., Neufeld, D. & Allen, J.R. (2004) Georeferencing of museum collections: a review of the problems and automated tools, and the methodology developed by the mountain and plains spatial-temporal database-informatics initiative (MAPSTEDI). *Phyloinformatics*, **1** (3), 1–29.

Neufeld, D., Guralnick, R., Glaubitz, R. & Allen, J.R. (2003) Museum collections data and on-line mapping applications: a new resource for land managers. *Mountain Research and Development*, 23(4), 334–337.

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

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.

Petersen, T.P., Meier, R. & Larsen, M.N. (2003) Testing species richness estimation methods using the museum label data on the Danish Asilidae. *Biodiversity and Conservation*, **12**, 687–701.

Preston, F.W. (1948) The commonness, and rarity, or species. *Ecology*, **29**, 254–283.

Raaijmakers, J.G.W. (1987) Statistical analysis of the Michaelis-Menten equation. *Biometrics*, **43**, 793–803.

Rahbek, C. & Graves, G.R. (2001) Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4534–4539.

Rosenzweig, M.L. (1995) *Species diversity in space and time*. Cambridge University Press, Cambridge, UK.

Smith, E.P. & van Belle, G. (1984) Nonparametric estimation of species richness. *Biometrics*, 40, 119–129.

Soberón, J.M., Llorente, J.B. & Oñate, L. (2000) The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies. *Biodiversity and Conservation*, 9, 1441–1466.

Stein, B. & Wieczorek, J. (2004) Mammals of the World: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics [Online]:4. Available: http://jbi.nhm.ku.edu/viewarticle.php?id=11.

Suarez, A.V. & Tsutsui, N.D. (2004) The value of museum collections for research and society. *Bioscience*, 54, 66–74.

Venables, W.N. & Ripley, B.D. (2002) *Modern applied statistics with S*, 4th edn. Springer-Verlag, New York, New York, USA.

Walther, B.A. & Morand, S. (1998) Comparative performance of species richness estimation methods. *Parasitology*, 116, 395–405.