Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 99–108 doi: 10.1111/2041-210X.12296

Accounting for imperfect detection in Hill numbers for biodiversity studies

Kristin M. Broms[1]*, Mevin B. Hooten[1,2,3] and Ryan M. Fitzpatrick[4]

[1]*Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO 80523, USA;* [2]*U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Unit, Fort Collins, CO 80523, USA;* [3]*Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA; and* [4]*Aquatic Wildlife Research Group, Colorado Parks and Wildlife, Fort Collins, CO 80526, USA*

Summary

1. Hill numbers unify biodiversity metrics by combining several into one expression. For example, species richness, Shannon's diversity index and the Gini–Simpson index are a few of the most used diversity measures, and they can be expressed as Hill numbers. Traditionally, Hill numbers have been calculated from relative abundance data, but the expression has been modified to use incidence data as well. We demonstrate an approach for estimating Hill numbers using an occupancy modelling framework that accounts for imperfect detection.

2. We alter the Hill numbers formula to use occupancy probabilities as opposed to the incidence probabilities that have been used previously and to calculate its summations from the modelled species richness. After introducing the occupancy-based Hill numbers, we demonstrate the differences between them and the incidence-based Hill numbers previously used through a simulation study and two applications.

3. In the simulation study and the two examples using real data, the occupancy-based Hill numbers were larger than the incidence-based Hill numbers, although species richness was estimated similarly using both methods.

4. The occupancy-based Hill number estimators are always at their asymptotic values (i.e. as if an infinite number of samples have been taken for the study region), therefore making it easy to compare biodiversity between different assemblages. In addition, the Hill numbers are computed as derived quantities within a Bayesian hierarchical model, allowing for straightforward inference.

Key-words: Bayesian methods, Gini–Simpson index, incidence matrix, multi-species occupancy model, Shannon entropy, species richness

Introduction

Biodiversity is one of the most important concepts in the study of ecology and is commonly measured by species richness, the Gini–Simpson index and Shannon entropy (Lande 1996; Jost 2006; Mao 2007; Gotelli & Chao 2013; Chao *et al.* 2014). Multiple measurements of biodiversity are valuable because species richness does not account for evenness among species. To adapt the example from Gotelli & Chao (2013), suppose two communities both contain exactly five species. The first community has one species comprising 0·80 of the total number of individuals, with the other species each comprising 0·05 of the assemblage. In the second community, each species comprises 0·20 of the population. Arguably, the second community should be seen as more diverse, but the species richness estimator is not robust enough to distinguish the two communities.

Therefore, other measurements of biodiversity are also used; both the Shannon entropy and the Gini–Simpson index take relative abundances of each species into account. The Shannon entropy 'quantifies the uncertainty in the species identity of a randomly chosen individual in the assemblage' (Gotelli & Chao 2013). It is also called the Shannon's diversity index or the Shannon–Wiener index (Jost 2006). The Gini–Simpson index 'measures the probability that two randomly chosen individuals (selected with replacement) belong to two different species' (Gotelli & Chao 2013). Variations of the Gini–Simpson index include the Simpson concentration, the inverse Simpson concentration, the second-order Renyi entropy or the Hurlbert–Smith–Grassle index (Jost 2006).

Hill numbers conveniently summarize all three types of biodiversity using a single expression, providing a unification (Hill 1973; Chao *et al.* 2014; Chiu, Jost & Chao 2014) and a framework from which to derive alpha and beta diversities (Jost 2007). We describe the Hill number formula and its exact relationship to Shannon entropy and the Gini–Simpson index in the 'Implementation' Section.

Traditionally, biodiversity measurements are functions of the relative abundances of each species in an assemblage, as obtained from the sampling design. However, Hill numbers have also been calculated using presence–absence data (Colwell & Coddington 1994; Colwell, Mao & Chang 2004; Colwell *et al.* 2012; Gotelli & Chao 2013; Chao *et al.* 2014). The presence–absence data are less informative than species counts, but they can be easier to collect, they may allow for comparisons

*Correspondence author. E-mail: kristin.broms@colostate.edu

with historic data, and for some species, including soil microbes and plankton, may be the only data obtainable due to their large counts and life-history traits.

Common attributes of the estimators mentioned above are that they do not take imperfect detection into account and they are sensitive to sample sizes (Lande 1996; Colwell *et al.* 2012; Chao *et al.* 2014). In Hill number estimation, the concepts of rarefaction and extrapolation have been introduced to deal with unequal sample sizes when comparing different assemblages (e.g. Chao & Jost 2012; Colwell *et al.* 2012; Gotelli & Chao 2013). While these methods make up for some shortcomings, the issue of imperfect detection remains. Variation in detectability affects relative abundances and incidence probabilities and therefore affects the interpretation of the Hill numbers. Detectability is one of the reasons that sample size and sample coverage play such an important role in the Hill number estimators.

Recent efforts in occupancy modelling have made it possible to separate detection from the true occurrences of the species. We describe a modification of the incidence-based Hill numbers to explicitly account for imperfect detection of species through a multi-species occupancy model (Dorazio & Royle 2005). Dorazio *et al.* (2006) used the multi-species occupancy model to estimate species richness and a species accumulation curve, but other biodiversity measures can also be obtained using this framework, bridging the divide between theoretical knowledge about biodiversity and hierarchical statistical models (e.g. Iknayan *et al.* 2014).

We demonstrate how to directly account for detectability in estimating Hill numbers and illustrate the discrepancy in inference that results if we ignore this sampling reality. Unlike traditional Hill numbers, the occupancy-based diversity estimators remain invariant to the number of sites and surveys and naturally account for differences in sampling intensities. Occupancy model estimators provide the asymptotic values of interest, stay consistent when the number of surveyed sites (i.e. the number of sampling units) changes and make Hill number estimation accessible in an existing framework. The utilization of Bayesian methods allows for inference concerning the detection-adjusted, occupancy-based Hill numbers and associated uncertainty as derived quantities. Estimating Hill numbers and other biodiversity measurements becomes as easy as adding one line of code to an existing algorithm. This ease of implementation removes the need to rely on asymptotic approximations and takes all sources of variation into account, leading to inference that accurately portrays the uncertainty related to the estimator. Finally, we can use the models to evaluate the sensitivity of the desired inference to different survey designs.

The format of the rest of the paper is as follows. We first reconcile the terminology and notation used in the Hill number theory with that of occupancy models. We describe their similarities and highlight the theoretical differences among the incidence-based Hill numbers and those arising from the multi-species occupancy models. We examine these approaches through a simulation study and apply them to two data sets: the forest ant data from Longino & Colwell (2011) and plains fish data from eastern Colorado. We discuss the implications of the results of our simulation study and provide guidance for the use of these diversity measures in practice.

Hill numbers and occupancy models

While being similar in most respects, the data collected and analysed with Hill numbers are slightly different from those collected for occupancy studies. In the Hill numbers literature, there are T sampling units, and each is assumed to be sampled randomly and independently. The sampling units are the locations where the species counts and/or recordings of incidences took place (i.e. the trap, net, quadrat, plot or point count). In the occupancy modelling literature, these same sampling units are called the J sites and each of these sites is surveyed K times.

Hill numbers are functions of the number of species in the area of interest and the probabilities of encountering each species at a given site (eqn 5). These calculations are based on the incidence matrix, $\mathbf{W} \equiv \{w_{ij}, \forall i,j\}$ with i representing the $i = 1,\ldots,S_{\text{obs}}$ species, and $j = 1,\ldots,T$ being the sampling units. S_{obs} is the number of species that were ever detected within the study area. In this incidence matrix, $w_{ij} = 1$ if species i was detected at sampling unit j and $w_{ij} = 0$ if it was not detected.

In the multi-species occupancy model literature, there are two matrices related to occurrences: one observed and one latent. The first matrix, $\mathbf{Y} \equiv \{y_{ij}, \forall i,j\}$, is the matrix of detection data. It is similar to $\mathbf{W}$, the incidence matrix from above, but it is augmented with rows of 0s to account for species that occurred in the study area but went undetected on all surveys. Therefore, this matrix has M rows, where $M \geq S_{\text{obs}}$ represents the augmented population. The columns, $j = 1, \ldots, J$, represent the sites at which sampling occurred. If detection probabilities were not affected by survey-specific covariates, then each element of the matrix, y_{ij}, is the number of surveys in which species i was detected at site j. If the model does include survey-specific covariates, the data must be represented by a three-dimensional array, with each element y_{ijk} being a binary variable equalling 1 if species i was detected at site j on survey $k = 1, \ldots, K_j$. Note that the number of surveys per site (K_j) can vary. This detection matrix is very similar to the incidence matrix of the Hill number literature, the main difference being that in the implementation of the multi-species occupancy model, we augment the detection matrix with rows of 0s to account for the species that went completely undetected (Royle & Dorazio 2012).

Inference in occupancy models is typically based on the occurrence matrix, $\mathbf{Z} \equiv \{z_{ij}, \forall i,j\}$, a matrix of latent variables indicating the true occurrences of each species. As above, the indices are $i = 1, \ldots, M$ species and $j = 1, \ldots, J$ sites. If $\sum_k y_{ijk} > 0$, then species i was detected at site j at least once and it is a known occurrence ($z_{ij} = 1$). If the species was never detected at site j, then z_{ij} is an unknown quantity and must be estimated. For further details on the multi-species occupancy model, we refer the reader to Royle & Dorazio (2008) for a comprehensive description.

INCIDENCE AND OCCUPANCY PROBABILITIES

Incidence-based Hill numbers are functions of the incidence probabilities, $\boldsymbol{\pi}$, which are the probabilities that species were detected (Chao *et al*. 2014). They are equivalent to the probability of a species occurring at a site multiplied by the probability of it being detected,

$$\pi_i = \psi_i\left(1 - (1 - p_i)^K\right), \qquad \text{eqn 1}$$

where ψ_i is the probability of occupancy for species i, p_i is the probability of detection given the site is occupied, and K is the number of surveys. In a typical Hill numbers study, only one survey is conducted per site and $K = 1$. The possibility of imperfect detection of a species is acknowledged because it can and does happen for a variety of reasons; for example, the species are mobile and are temporarily absent from an area that they typically occupy; biologists are unable to correctly identify species; the sampling method used is species-selective; and time of day or weather can have a strong impact on detectability.

Because ψ_i and p_i are both unknown quantities in the model, we propose an alternative approach to obtain Hill numbers that removes the detection component from the diversity measures and replaces the incidence probabilities by the occupancy probabilities. Therefore, instead of this Bernoulli representation of the data:

$$\Pr(W = w_{ij}) = \pi_i^{w_{ij}}(1 - \pi_i)^{1-w_{ij}}, \qquad \text{eqn 2}$$

we use a similar expression for the underlying occurrence process:

$$\Pr(Z = z_{ij}) = \psi_i^{z_{ij}}(1 - \psi_i)^{1-z_{ij}}. \qquad \text{eqn 3}$$

IMPLEMENTATION

The theoretical expression for incidence-based Hill numbers is

$${}^q\Delta = \left(\sum_{i=1}^{S}\left(\frac{\pi_i}{\sum_{s=1}^{S}\pi_s}\right)^q\right)^{\frac{1}{1-q}}. \qquad \text{eqn 4}$$

This equation requires $q \neq 1, q \geq 0$ and represents the asymptotic diversity as $T \to \infty$ because it is only when an infinite amount of sampling has been conducted that the species richness, S, will be known. For $q = 1$, the limit is used in place of the direct equation, leading to the following expression:

$${}^1\Delta = \exp\left(-\sum_{i=1}^{S}\frac{\pi_i}{\sum_{s=1}^{S}\pi_s}\log\frac{\pi_i}{\sum_{s=1}^{S}\pi_s}\right). \qquad \text{eqn 5}$$

While all values of q may lead to useful inference concerning the biodiversity of the study area, $q = 0$, 1, or 2 are especially important as ${}^0\Delta$ represents species richness, ${}^1\Delta$ represents Shannon diversity (the Shannon entropy exponentiated), and ${}^2\Delta$ represents Simpson diversity (inverse of the complement of the Gini–Simpson index, Jost 2006).

Because the incidence probabilities and species richness of an assemblage are unknown, a variety of methods and estimators have been proposed to calculate the incidence-based Hill numbers (${}^q\Delta_{\text{incid}}$) while incorporating the fact that many species likely went undetected in all samples. Extrapolated values and bootstrapping techniques are often used to compare assemblages that were sampled at different intensities (iNEXT function, Chao *et al*. 2014), and separate formulas have been created to estimate the asymptotic values of the Hill numbers (Lee & Chao 1994; Colwell *et al*. 2012; Chao, Wang & Jost 2013; Chao *et al*. 2014). The asymptotic estimator for species richness, ${}^0\Delta_{\text{incid}}$, is supplied in Colwell *et al*. (2012), and the asymptotic estimators for ${}^1\Delta_{\text{incid}}$ and ${}^2\Delta_{\text{incid}}$ are supplied in appendix H of Chao *et al*. (2014). For the reader's convenience, we provide these formulas in Appendix A.

We propose that the Hill numbers can also be calculated using the occupancy probabilities such that

$${}^q\Delta_{\text{occu}} = \left(\sum_{i=1}^{N}\left(\frac{\psi_i}{\sum_{s=1}^{N}\psi_s}\right)^q\right)^{\frac{1}{1-q}} \qquad q \neq 1, q \geq 0 \qquad \text{eqn 6}$$

and

$${}^1\Delta_{\text{occu}} = \exp\left(-\sum_{i=1}^{N}\frac{\psi_i}{\sum_{s=1}^{N}\psi_s}\log\frac{\psi_i}{\sum_{s=1}^{N}\psi_s}\right) \qquad q = 1. \qquad \text{eqn 7}$$

In multi-species occupancy models, N is the symbol used to represent species richness. In practice, Hill numbers are calculated as derived quantities within a Markov chain Monte Carlo (MCMC) algorithm. To illustrate the implementation of this model and the calculation of the occupancy-based Hill numbers, we refer the reader to the code in Appendix B.2.

If detection probabilities were equal for every species, site and survey, and an equal number of surveys per site were conducted, then they would cancel and the occupancy-based Hill numbers would be the same as the incidence-based Hill numbers. It is doubtful that detection probabilities will be equal for every species. For example, one can imagine that relative abundances could affect detections; that is, more prevalent species will be detected more often than ones that occur as lower densities (Royle & Nichols 2003). The conspicuousness of species must also be accounted for because more noticeable species will have higher incidence probabilities and hence will appear more often in the incidence-based Hill number calculations than inconspicuous ones, but they are not more valuable to diversity.

OTHER TERMINOLOGY

First, we remind the reader that an important assumption of both the incidence-based and the occupancy-based expressions is that the definition of a 'site' or 'sampling unit' is standardized; incidence probabilities, occupancy probabilities and

detection probabilities will all vary as the areal size of a site enlarges. Assuming that the size of a site has been standardized, different assemblages can be compared. Much work has been done with the incidence-based Hill numbers regarding the different number of samples that may have been taken when comparing one assemblage to another (e.g. Gotelli & Colwell 2001; Colwell *et al.* 2012; Chao *et al.* 2014).

The incidence-based Hill numbers are a function of the number of sites that are sampled; thus, diversity accumulation curves are created to help one compare diversity measures for assemblages that were sampled at different intensities. These curves are traditionally created through rarefaction and extrapolation procedures. Rarefaction is similar to interpolation, in that it is the estimation of Hill numbers for $t < T$ sampling units. Extrapolation is the estimation of Hill numbers for $T + m > T$ sampling units, with the additional sampling units being defined as m. It has been suggested that extrapolation only be used with up to twice the number of sampling units (i.e. when $m = T$) in order for the estimate to remain relatively unbiased (Chao *et al.* 2014).

Sampling coverage is the proportion of the total individuals in an assemblage that are found in a sample (Jost 2010). Rarefaction and extrapolation may be based on sample coverage rather than the number of sampling units because sampling units may have different degrees of completeness, depending on the species–abundance distribution of the community (Chao & Jost 2012). The use of sample coverage is to aid in creating an equal comparison between assemblages.

Diversity accumulation curves are usually functions of the number of sampling units and are plots of the Hill numbers for 0 to $T + m$ units. Alternatively, the diversity accumulation curve can be a function of sample coverage and would be a plot of the Hill numbers for 0–100% coverage (Chao & Jost 2012). The curves approach the asymptotic values of the biodiversity measures if an unlimited number of samples were taken.

In hierarchical occupancy modelling, we estimate the total number of species in the assemblage, $\hat{N}$, which *is* the asymptotic value for S (Dorazio *et al.* 2006). The estimates for the other Hill numbers are also the asymptotic values, thus minimizing the need for rarefaction and extrapolation when comparing assemblages. Additional calculations associated with incidence-based Hill numbers are as follows: $Y_i = \sum_{j=1}^{T} w_{ij}$, the incidence-based frequencies of species i, and Q_k, the incidence frequencies, where Q_k is the number of species that are detected in exactly $k = 1, \ldots, T$ sampling units. In the occupancy framework, the required calculations do not depend on these quantities.

Simulation study

SIMULATING MULTI-SPECIES DATA

We conducted a simulation study to demonstrate how Hill number estimates compare when calculated using incidence probabilities versus the occupancy probabilities. We do not expect the ${}^{q}\Delta_{\text{incid}}$ and ${}^{q}\Delta_{\text{occu}}$ to be equal due to their different interpretations and the different data (one survey per site versus multiple surveys per site) that are used in their calculations. However, the way their differences affect the resulting estimators is of interest.

We simulated data with $N = 100$, 250 or 500 species, $J = 25$ or 100 sites, $K = 2$ or 5 surveys per site, and occupancy probabilities were 'very low' with a median value of 0·10, 'low' with a median value of 0·20, or they were 'moderate' with a median value of 0·50. Our simulated detection probabilities always had a median value of 0·20. Simulation values were chosen to give a range of what might be expected from the study design and from the population. Simulated occupancy and detection probabilities were based on the range of estimated values from the literature (Dorazio *et al.* 2006; Williams 2009; Dorazio, Gotelli & Ellison 2011; Holt *et al.* 2013). A block design of these combinations led to a total of 36 sets of simulations. The scenarios involving $N = 100$ or $N = 250$ species were replicated 50 times; the scenarios involving $N = 500$ species were replicated 20 times due to their much larger matrices and longer computing times.

The occupancy and detection probabilities were simulated from normal distributions on the logit scale, such that

$$\text{logit}(\psi_i) \sim \text{Normal}\left(\alpha, \sigma^2_{\psi}\right), \qquad \text{eqn 8}$$

$$\text{logit}(p_i) \sim \text{Normal}\left(\beta, \sigma^2_{p}\right), \qquad \text{eqn 9}$$

for each species, $i = 1, \ldots, N$. The detection probability parameters were $\beta = -1{\cdot}5$ and $\sigma^2 = 0{\cdot}5$. For the very low occupancy probability scenarios, the parameters were $\alpha = -2$ and $\sigma^2 = 4$; for the low occupancy probability scenarios, the parameters were $\alpha = -1{\cdot}5$ and $\sigma^2 = 0{\cdot}5$; and for the moderate occupancy probability scenarios, the parameters were $\alpha = 0$ and $\sigma^2 = 0{\cdot}5$.

Under these parameters, for the most undersampled situation, when $K = 2$ surveys per site and the population has a very low occupancy probability distribution, then the interquartile range of incidence probabilities will run from 0·009 to 0·1 with a median near 0·03. If only one survey had been conducted at each site, then the interquartile range for the incidence probabilities would drop to (0·005, 0·06) with a median of 0·02.

The occurrences and detections for each species and each site ($j = 1,\ldots,J$) were simulated from Bernoulli and Binomial distributions, respectively,

$$z_{ij} \sim \text{Bernoulli}(\psi_i) \qquad \text{eqn 10}$$

$$y_{ij} \sim \text{Binomial}(K, p_i z_{ij}). \qquad \text{eqn 11}$$

To estimate the occupancy-based Hill numbers, multi-species occupancy models were fit to the resulting data (model formulation given in Appendix B). A scale prior was used for the probability of an element of the augmented population being a species in the assemblage, ϕ, following the recommendation of Link (2013). Normal priors were specified for the mean parameters, α and β, and gamma priors were specified for the precisions, $1/\sigma^2_{\psi}$ and $1/\sigma^2_{p}$ (Appendix B). We obtained three chains of length 50 000 with a burn-in of 25 000, all thinned by 5, leaving a total of 5000 saved

iterations per chain. If parameters had not converged, as assessed by the Gelman–Rubin statistic ($\hat{R}$, Gelman & Rubin 1992), an additional 20 000 iterations, thinned by 5, were obtained.

To estimate the incidence-based Hill numbers, we aggregated the simulated data across the number of surveys because the study design associated with these expressions requires only one survey per site. If species i was detected on at least one of the K surveys of site j, then $w_{ij} = 1$; otherwise $w_{ij} = 0$. Under this scheme, $T = J$.

We compared four values: (i) the true, occupancy-based Hill numbers (${}^{q}\Delta_{\text{occu}}$) resulting from the true occupancy probabilities and using eqns (6) and (7); (ii) the estimated, occupancy-based Hill numbers (${}^{q}\hat{\Delta}_{\text{occu}}$) derived from the parameters fit in the multi-species occupancy model and eqns (6) and (7); (iii) the true, incidence-based Hill numbers (${}^{q}\Delta_{\text{incid}}$) calculated from the incidence probabilities using eqns (1), (4) and (5), and N, the known species richness; and (iv) the estimated, asymptotic incidence-based Hill numbers (${}^{q}\hat{\Delta}_{\text{asymp}}$) calculated with the incidence matrix **W** from the simulated data, as described above, and the formulas from Appendix A. In Appendix C, we also show some results from estimating the extrapolated, incidence-based Hill numbers, based on a doubling of the number of sites as recommended in Chao *et al.* (2014) and determined using their iNEXT function. We calculated the extrapolated values for a subset of the simulations to gain an idea of the estimated standard errors associated with the estimates.

All comparisons in the Results section are made against the true, occupancy-based Hill numbers, ${}^{q}\Delta_{\text{occu}}$. Comparison are reported in terms of relative differences, D, which we alternatively call the relative biases. Averaged across all simulations, S, the relative differences are calculated as $D = \sum_{i=1}^{S}\left({}^{q}\hat{\Delta}_i - {}^{q}\Delta_{\text{occu,i}}\right)/{}^{q}\Delta_{\text{occu,i}}$.

SIMULATION RESULTS

The occupancy models led to unbiased estimates of species richness, ${}^{0}\Delta \equiv N$, for most sets of simulations, albeit with higher standard errors when fewer sites were surveyed, when fewer surveys per sites were conducted and/or when the median occupancy probabilities were lower (Table 1, standard errors are reported in Appendix C). The occupancy model estimates had a positive bias when only 25 sites were sampled, 2 surveys per site were used, and the species overall had low occupancy probabilities. When compared to the true, occupancy-based Shannon diversity values, the ${}^{1}\Delta$ numbers, the occupancy model estimates slightly underestimated the diversity measure (Table 2). This bias was greater when the species richness increased but was not affected by the number of sites sampled or the number of surveys per site. The Simpson diversity values, the ${}^{2}\Delta$ numbers, were slightly overestimated for the worst-case scenario (i.e. when the number of sites and surveys were low and the occupancy probabilities were very low or low), but the estimates were otherwise unbiased (Table 3).

The true, incidence-based ${}^{0}\Delta$ numbers were identical to N, the species richness. However, the incidence-based ${}^{1}\Delta$ numbers were smaller than the occupancy-based numbers by 4–10% (Table 2), and the incidence-based ${}^{2}\Delta$ numbers were smaller than the occupancy-based numbers by 7–18% (Table 3). Both ${}^{1}\Delta_{incid}$ *and* ${}^{2}\Delta_{incid}$ were consistently lower when two surveys per site were conducted than when five surveys were conducted.

The estimated, asymptotic incidence-based Hill numbers, ${}^{0}\hat{\Delta}_{\text{asymp}}$, underestimated N by as much as 25% under the worst-case scenarios. The relative biases were often more than twice as high when 25 sites were sampled compared to 100 sites, with all else kept constant. The biases were severe when occupancy probabilities were very low, prevalent when occupancy probabilities were low and 25 sites were surveyed, and mostly disappeared for the other scenarios. As noted in Colwell *et al.* (2012), these results are not surprising as it is known that other estimators, such as the incidence-based coverage estimator (ICE, Lee & Chao 1994), are more appropriate for assemblages with many rare and elusive species.

The ${}^{q}\hat{\Delta}_{\text{asymp}}$ performed much better when estimating the higher indexed Hill numbers. All relative differences were within 3% of the true ${}^{q}\Delta_{incid}$, with most percentages being equal, indicating that they are proper estimators of the incidence-based Hill numbers for $q = 1$ and $q = 2$.

Example 1: Forest ants in Costa Rica

For a direct comparison with incidence-based Hill numbers, we fit the multi-species occupancy model to the ant data that were analysed in Chao *et al.* (2014) (originally analysed and collected by Longino & Colwell 2011) and related the resulting occupancy-based Hill numbers to the incidence-based Hill numbers. We focused on the data associated with the 50 m elevation only, which consisted of 15 sampling periods. Within the collected data, each sampling period was broken into four transects with 10 samples taken along each transect (although along one transect only nine samples were taken). Because the occupancy model framework requires replicate sampling of a given site, we set each of these 10 samples to be spatial replicates (i.e. the K surveys) and each transect was treated as a separate site, giving $15\times4 = 60$ sites. While the surveys should be repeated temporally, the treatment of separate samples along a transect as different sampling occasions is possible as the study design (MacKenzie *et al.* 2006; Royle & Kéry 2007). These spatial replicates are valid as long as their coverage is small compared to the size of the site and the study area, that is, as long as the choice of survey sites is equivalent to sampling with replacement (Kendall & White 2009).

In calculating the incidence-based Hill numbers, we used the asymptotic estimators (Appendix A) to make a more equal comparison with the occupancy-based Hill numbers than if we had used the extrapolated values found in Chao *et al.* (2014). We assumed $T = 60$ sites and collapsed the 10 surveys from

Table 1. Comparison of the species richness estimators from the simulation study

						Relative differences, D		
N	J	ψ	K	S_{obs}	${}^0\Delta_{occu}$	${}^0D_{occu}$	${}^0D_{incid}$	${}^0D_{asymp}$
100	25	Very Low	2	57	100	−0·02	0	−0·22
250	25	Very Low	2	142	250	−0·06	0	−0·26
500	25	Very Low	2	287	500	0·02	0	−0·26
100	100	Very Low	2	79	100	0·01	0	−0·10
250	100	Very Low	2	198	250	0	0	−0·11
500	100	Very Low	2	395	500	0	0	−0·12
100	25	Low	2	73	100	0·12	0	−0·06
250	25	Low	2	181	250	0·03	0	−0·09
500	25	Low	2	360	500	0·03	0	−0·08
100	100	Low	2	96	100	0·01	0	0
250	100	Low	2	240	250	0	0	−0·01
500	100	Low	2	478	500	0	0	−0·01
100	25	Moderate	2	94	100	0·04	0	0·01
250	25	Moderate	2	235	250	0·01	0	−0·01
500	25	Moderate	2	469	500	0	0	−0·01
100	100	Moderate	2	100	100	0	0	0
250	100	Moderate	2	249	250	0	0	0
500	100	Moderate	2	499	500	0	0	0
100	25	Very Low	5	70	100	−0·01	0	−0·16
250	25	Very Low	5	173	250	0·01	0	−0·18
500	25	Very Low	5	342	500	0	0	−0·19
100	100	Very Low	5	87	100	0·02	0	−0·05
250	100	Very Low	5	216	250	0	0	−0·06
500	100	Very Low	5	434	500	0	0	−0·05
100	25	Low	5	88	100	0·02	0	−0·02
250	25	Low	5	220	250	0	0	−0·03
500	25	Low	5	440	500	0·01	0	−0·04
100	100	Low	5	99	100	0	0	0·01
250	100	Low	5	248	250	0	0	0
500	100	Low	5	497	500	0	0	0
100	25	Moderate	5	99	100	0·01	0	0·01
250	25	Moderate	5	247	250	0	0	0
500	25	Moderate	5	495	500	0	0	0
100	100	Moderate	5	100	100	0	0	0
250	100	Moderate	5	250	250	0	0	0
500	100	Moderate	5	500	500	0	0	0

The relative differences, D, are calculated for the estimated, occupancy-based species richness (${}^0D_{occu}$); the theoretical, incidence-based species richness (${}^0D_{incid}$); and the estimated, asymptotic, incidence-based species richness (${}^0D_{asymp}$), when compared to the true occupancy-based species richness from eqn (6) (${}^0\Delta_{occu}$). N is the species richness, J is the number of sites, ψ describes the median occupancy probability, K is the number of surveys, and S_{obs} is the mean number of species that were detected in a simulation.

each transect as we did in the simulation. In calculating the occupancy-based Hill numbers, we used the same formulation and priors as in the simulation study (Appendix B.1) and obtained 50 000 MCMC samples, with a thinning rate of 5 and a burn-in of 25 000. Under this set-up, the model converged for all parameters based on the Gelman–Rubin statistic.

The resulting occupancy-based Hill numbers were much larger than the asymptotic, incidence-based Hill numbers for all three diversity measures. The estimated species richness rose from 287 to 329 species (95% CI: 282–404); the estimated Shannon diversity, ${}^1\Delta$, rose from 123·2 to 218·5 (95% CI: 189·2–257·6); and the estimated Simpson's diversity, ${}^2\Delta$, rose from 89·3 to 198·8 (95% CI: 167·1–239·6). Because the occupancy-based Hill numbers are derived quantities in the Bayesian models, they have associated posterior distributions and we report their 95% credible intervals to provide a sense of the uncertainty surrounding these estimates.

Example 2: Plains fish of eastern Colorado

For our second example, we considered incidence records of fish in the South Platte River basin in eastern Colorado. The distributions and abundances of Colorado's eastern plains native fishes have declined since 1900 such that many are now state-listed and in need of conservation activities (Fausch & Bestgen 1997; Nesler *et al.* 1997). Anthropogenic changes including stream barriers,

Table 2. Comparison of the Shannon diversity estimators from the simulation study

						Relative differences, D		
N	J	ψ	K	S_{obs}	${}^1\Delta_{occu}$	${}^1D_{occu}$	${}^1D_{incid}$	${}^1D_{asymp}$
100	25	Very Low	2	57	57·3	0·06	−0·09	−0·12
250	25	Very Low	2	142	142·5	0	−0·10	−0·11
500	25	Very Low	2	287	285·6	−0·02	−0·10	−0·13
100	100	Very Low	2	79	57·8	0·02	−0·10	−0·09
250	100	Very Low	2	198	143·9	−0·01	−0·10	−0·10
500	100	Very Low	2	395	284·6	−0·04	−0·10	−0·10
100	25	Low	2	73	87·3	0·05	−0·10	−0·10
250	25	Low	2	181	218·2	−0·02	−0·10	−0·12
500	25	Low	2	360	434·8	−0·04	−0·10	−0·12
100	100	Low	2	96	87·1	0	−0·10	−0·10
250	100	Low	2	240	217·4	−0·02	−0·10	−0·11
500	100	Low	2	478	436·1	−0·05	−0·10	−0·10
100	25	Moderate	2	94	94·6	−0·01	−0·10	−0·10
250	25	Moderate	2	235	237·2	−0·02	−0·10	−0·10
500	25	Moderate	2	469	474·3	−0·05	−0·10	−0·10
100	100	Moderate	2	100	95	−0·01	−0·10	−0·10*
250	100	Moderate	2	249	237·2	−0·02	−0·10	−0·10*
500	100	Moderate	2	499	474·1	−0·05	−0·10	−0·10
100	25	Very Low	5	70	57·9	0·02	−0·05	−0·04
250	25	Very Low	5	173	143·8	−0·01	−0·05	−0·06
500	25	Very Low	5	342	285·1	−0·04	−0·05	−0·06
100	100	Very Low	5	87	57·4	−0·01	−0·04	−0·04
250	100	Very Low	5	216	142·8	−0·02	−0·05	−0·06
500	100	Very Low	5	434	286·9	−0·04	−0·05	−0·05
100	25	Low	5	88	87·4	0	−0·05	−0·05
250	25	Low	5	220	218·4	−0·02	−0·05	−0·06
500	25	Low	5	440	435·8	−0·05	−0·05	−0·05
100	100	Low	5	99	87·2	−0·01	−0·05	−0·05 *
250	100	Low	5	248	217·9	−0·02	−0·05	−0·05*
500	100	Low	5	497	437·6	−0·05	−0·05	−0·05
100	25	Moderate	5	99	94·8	−0·01	−0·05	−0·05*
250	25	Moderate	5	247	237·1	−0·02	−0·05	−0·05
500	25	Moderate	5	495	474·1	−0·05	−0·05	−0·05
100	100	Moderate	5	100	94·9	−0·01	−0·05	NA*
250	100	Moderate	5	250	237·1	−0·02	−0·05	−0·06*
500	100	Moderate	5	500	474·1	−0·05	−0·05	NA*

The relative differences, D, are calculated for the estimated, occupancy-based Shannon diversity (${}^1D_{occu}$); the theoretical, incidence-based Shannon diversity (${}^1D_{incid}$); and the estimated, asymptotic, incidence-based Shannon diversity (${}^1D_{asymp}$), when compared to the true occupancy-based Shannon diversity from eqn (7) (${}^1\Delta_{occu}$). N is the species richness, J is the number of sites, ψ describes the median occupancy probability, K is the number of surveys, and S_{obs} is the mean number of species that were detected in a simulation.

*For several simulations, the asymptotic Shannon diversity estimators did not give a value because no species were detected at exactly 1 or 2 sites. In particular, none of the simulations for $N = 100$, one of the simulations for $N = 250$, and none of the simulations for $N = 500$, all with $J = 100$, $K = 5$ and moderate occupancy probabilities, yielded an asymptotic value. Going down the column of stars, the relative differences are based on 18, 40, 41, 49, 48, 0, 1 and 0 simulations.

altered flow regime, siltation, channelization, changes in water quality and introduced species have been implicated in the decline of native fishes (Fausch & Bestgen 1997; Falke, Bestgen & Fausch 2010; Perkin & Gido 2011). Understanding the distribution of plains fish species is essential to promoting conservation and potential expansion of remaining populations.

We estimated Hill numbers for the fish community in the main stem of the South Platte River downstream of the confluence with the St. Vrain Creek in north-eastern Colorado using data from recent sampling (Fig. 1). This stretch of river is a homogeneous, high plains landscape with centre-pivot irrigation for agriculture occurring along the river. This homogeneity allowed us to make direct comparisons of the incidence-based and occupancy-based Hill numbers.

With surveys taking place from 2009 to 2013, a total of 36 species of fish were detected at 60 sites within this region. At each site, one to five surveys were conducted. These surveys were either electro-fishing or seining passes where fish were identified and counted. For this analysis, we consider the data from the electro-fishing passes. Because counts can be extremely variable throughout the season and from year to year, we collapsed them into presence–absence data as has been

Table 3. Comparison of the Simpson diversity estimators from the simulation study

						Relative differences, D		
N	J	ψ	K	S_{obs}	${}^2\Delta_{occu}$	${}^2D_{occu}$	${}^2D_{incid}$	${}^2D_{asymp}$
100	25	Very Low	2	57	45·1	0·13	−0·15	−0·16
250	25	Very Low	2	142	111·7	0·06	−0·16	−0·15
500	25	Very Low	2	287	225·1	0·04	−0·18	−0·18
100	100	Very Low	2	79	45·6	0·06	−0·16	−0·15
250	100	Very Low	2	198	113	0·02	−0·18	−0·17
500	100	Very Low	2	395	223·6	0·01	−0·17	−0·17
100	25	Low	2	73	77·7	0·06	−0·17	−0·18
250	25	Low	2	181	194·4	0	−0·18	−0·18
500	25	Low	2	360	386·2	0	−0·18	−0·19
100	100	Low	2	96	77·4	0·01	−0·17	−0·17
250	100	Low	2	240	193	0·01	−0·17	−0·18
500	100	Low	2	478	388·7	−0·01	−0·17	−0·17
100	25	Moderate	2	94	90·5	−0·01	−0·17	−0·18
250	25	Moderate	2	235	227·2	0	−0·17	−0·18
500	25	Moderate	2	469	454·2	0	−0·17	−0·17
100	100	Moderate	2	100	91	0	−0·17	−0·16
250	100	Moderate	2	249	227·2	0	−0·18	−0·18
500	100	Moderate	2	499	454	0	−0·17	−0·17
100	25	Very Low	5	70	45·9	0·05	−0·09	−0·07
250	25	Very Low	5	173	113·3	0·01	−0·09	−0·09
500	25	Very Low	5	342	224	0·01	−0·09	−0·09
100	100	Very Low	5	87	44·9	0	−0·07	−0·07
250	100	Very Low	5	216	112·3	0	−0·09	−0·09
500	100	Very Low	5	434	225·8	0·01	−0·09	−0·09
100	25	Low	5	88	77·9	0·02	−0·09	−0·08
250	25	Low	5	220	194·4	0·01	−0·09	−0·09
500	25	Low	5	440	388·2	0	−0·10	−0·09
100	100	Low	5	99	77·6	0·01	−0·09	−0·09
250	100	Low	5	248	193·8	0	−0·09	−0·09
500	100	Low	5	497	390·7	0	−0·09	−0·09
100	25	Moderate	5	99	90·8	0	−0·09	−0·09
250	25	Moderate	5	247	227	0	−0·09	−0·09
500	25	Moderate	5	495	453·9	0	−0·09	−0·09
100	100	Moderate	5	100	91	0	−0·09	−0·09
250	100	Moderate	5	250	227·1	0	−0·09	−0·09
500	100	Moderate	5	500	454	0	−0·09	−0·09

The relative differences, D, are calculated for the estimated, occupancy-based Simpson diversity (${}^2D_{occu}$); the theoretical, incidence-based Simpson diversity (${}^2D_{incid}$); and the estimated, asymptotic, incidence-based Simpson diversity (${}^2D_{asymp}$), when compared to the true occupancy-based Simpson diversity from eqn (6) (${}^2\Delta_{occu}$). N is the species richness, J is the number of sites, ψ describes the median occupancy probability, K is the number of surveys, and S_{obs} is the mean number of species that were detected in a simulation.

done previously for other count data (e.g. Royle & Nichols 2003; Dorazio & Royle 2005). As with the other model fits, we obtained 50 000 MCMC samples, with a thinning of 5 and a burn-in of 25 000. To calculate the incidence-based Hill numbers, we again collapsed the detections from all surveys so that $w_{ij} = 1$ if species i was ever detected at site j, and $w_{ij} = 0$ otherwise. A list of all species detected in the study is provided in Appendix D.

As with the simulation study, the occupancy-based Hill numbers for the plains fish were higher than the incidence-based Hill numbers (Table 4). For this example, the occupancy-based estimators were 14–29% higher than their incidence-based counterparts. Using the occupancy model, we estimated richness to be 46 species even though only 36 species were detected. This estimate is in line with sampling from the 1990s that identified 41 species in the South Platte River basin (Nesler *et al.* 1997). The posterior mean detection probability associated with these data was 0·54, and the posterior mean occupancy probability was 0·31.

Discussion

The estimated, occupancy-based Hill numbers had minimal biases and were comparable with each other when either 25 or 100 sites were surveyed and when either 2 or 5 surveys per site took place in simulation. In contrast, the true and estimated, incidence-based Hill numbers were consistently lower when fewer surveys per site were conducted; their values were 85–90% of the occupancy-based Hill numbers. The disagreement between the incidence-based and the occupancy-based Hill number estimators existed even with five surveys per site. This incongruity is important because increasing the number of surveys increases overall detection. For example, with only two

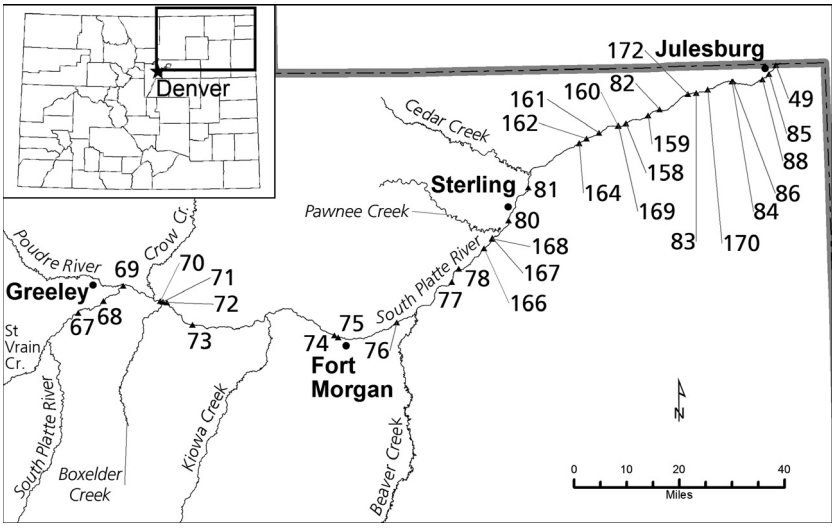

Fig. 1. Sampling sites for plains fish on the main stem of the South Platte River.

Table 4. Diversity measures for plains fish in the main stem of the South Platte River in Colorado, with 95% credible and confidence intervals

Parameter	Model	Estimate	Lower bound	Upper bound
${}^0\hat{\Delta}_{occu}$	Occupancy	45·88	37·00	64·00
${}^1\hat{\Delta}_{occu}$	Occupancy	28·84	24·58	39·78
${}^2\hat{\Delta}_{occu}$	Occupancy	25·06	20·88	35·45
${}^0\hat{\Delta}_{asymp}$	Incidence	40·07	–	–
${}^1\hat{\Delta}_{asymp}$	Incidence	23·50	–	–
${}^2\hat{\Delta}_{asymp}$	Incidence	19·49	–	–
${}^0\hat{\Delta}_{extrap}$	Incidence	38·89	33·39	44·40
${}^1\hat{\Delta}_{extrap}$	Incidence	23·21	21·93	24·48
${}^2\hat{\Delta}_{extrap}$	Incidence	19·38	18·39	20·38

${}^0\Delta$ estimates species richness, ${}^1\Delta$ estimates the Shannon diversity, and ${}^2\Delta$ estimates the Simpson diversity. The ${}^q\hat{\Delta}_{asymp}$ are the asymptotic, incidence-based Hill numbers (Appendix A), and the ${}^q\hat{\Delta}_{extrap}$ are the extrapolated, incidence-based Hill numbers, calculated using the iNEXT function in R and extrapolated to $T = 2J = 120$ sites.

surveys per site, a species that has a 50% chance of being detected during one sampling occasion given that it is present has a 75% chance of being detected on at least one of the surveys $(1-(1-0{\cdot}5)^2 = 0{\cdot}75)$. With five surveys per site, that probability jumps to 97%. Even at such high detection rates, the occupancy framework consistently provided larger diversity measure values. These discrepancies highlight the influence of detectability on the incidence probabilities even if detection rates are high.

Diversity measures should account for imperfect detection because detectability will not be equal across all surveys, sites or species, affecting the proportions in which individuals are seen versus their true proportions in the community. Hill numbers can be readily estimated as derived quantities in multi-species occupancy models. The use of a Bayesian hierarchical model simplified estimation and incorporated the uncertainty surrounding the total species richness and the uncertainty related to the occupancy probabilities in the posterior distributions of Hill numbers. We demonstrated that these occupancy-based Hill number estimators are accurate and stable.

We recognize that other diversity measures exist: alpha, beta and gamma diversities, the Sørensen coefficient and the Jaccard coefficient, to name a few. The estimation of these other indices, as well as the untransformed Shannon entropy or Gini–Simpson index, can be incorporated into an occupancy model framework in a similar fashion to what we presented here. Indeed, beta diversities have already been incorporated into the framework (Dorazio, Gotelli & Ellison 2011), and the species richness estimates are a fundamental component of the multi-species occupancy model (Dorazio & Royle 2005).

While study design changes may be necessary to obtain occupancy model estimates (e.g. multiple surveys per site), such changes are becoming standard in contemporary ecological data collection efforts. In some cases, as we illustrated, design changes may not even be necessary if spatial replicates can be used in place of temporal replicates (but see Kendall & White 2009).

In general, one could incorporate survey-specific and site-specific covariates into the occupancy and detection components of the model (as demonstrated in Kéry & Royle 2009; Zipkin *et al.* 2010). Such heterogeneity may allow for greater insight about Hill numbers, provide more intricate comparisons between assemblages and account for different sampling protocols. Traditional Hill numbers do not explicitly allow for diversity estimates in heterogeneous landscapes. With occupancy-based Hill numbers, we can use the site-specific covariates to compare biodiversity measures within a landscape and among landscapes in one model. For example, elevation could be used as a covariate in the Costa Rican forest ant data instead of analysing each elevation separately (e.g. Longino & Colwell 2011; Chao *et al.* 2014). This model would lead to inference on how elevation affects the occurrences and detections of each species in addition to providing Hill number estimates for each level of elevation.

The model-based approach described herein can easily be adapted to accommodate count data. One can use derived quantities in a multi-species N-mixture model, or similar statistical model, to calculate detection-adjusted, abundance-based Hill numbers. We believe that such an approach will similarly

lead to estimated Hill numbers that are consistent across different sampling schemes.

Acknowledgements

Funding for this research was provided by Colorado Parks and Wildlife (1401). The authors would like to thank Harry Crockett, Boyd Wright, Viviana Ruiz-Gutiérrez and John Tipton. Two anonymous reviewers provided thoughtful feedback that improved the manuscript. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Data accessibility

The forest ant data used in the first case study in this paper are already published and available in the supplementary material for Longino & Colwell (2011). The Colorado fish data used in the second case study are available in Appendix D.

References

Chao, A. & Jost, L. (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, **93**, 2533–2547.

Chao, A., Wang, Y.T. & Jost, L. (2013) Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution*, **4**, 1091–1100.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K. & Ellison, A.M. (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, **84**, 45–67.

Chiu, C.H., Jost, L. & Chao, A. (2014) Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. *Ecological Monographs*, **84**, 21–44.

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

Colwell, R.K., Mao, C.X. & Chang, J. (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, **85**, 2717–2727.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.Y., Mao, C.X., Chazdon, R.L. & Longino, J.T. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, **5**, 3–21.

Dorazio, R.M. & Royle, J.A. (2005) Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, **100**, 389–398.

Dorazio, R.M., Royle, J.A., Söderström, B. & Glimskär, A. (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, **87**, 842–854.

Dorazio, R.M., Gotelli, N.J. & Ellison, A.M. (2011) Modern methods of estimating biodiversity from presence-absence surveys. *Biodiversity Loss in a Changing Planet* (ed. P.O. Grillo), pp. 277–302. InTech, Rijeka, Croatia.

Falke, J.A., Bestgen, K.R. & Fausch, K.D. (2010) Streamflow reductions and habitat drying affect growth, survival and recruitment of brassy minnow across a Great Plains riverscape. *Transactions of the American Fisheries Society*, **135**, 1566–1583.

Fausch, K.D. & Bestgen, K.R. (1997) Ecology of fishes indigenous to the central and southwestern Great Plains. *Ecology and Conservation of Great Plains Vertebrates, volume 125 of Ecological Studies* (eds F.L. Knopf & F.B. Samson), pp. 131–166. Springer-Verlag, New York.

Gelman, A. & Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Gotelli, N.J. & Chao, A. (2013) Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. *Encyclopedia of Biodiversity*, volume 5, 2nd edn (ed. S.A. Levin), pp. 195–211. Academic Press, Waltham, Massachusetts, USA.

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.

Hill, M.O. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**, 427–432.

Holt, B.G., Rioja-Nieto, R., MacNeil, M.A., Lupton, J. & Rahbek, C. (2013) Comparing diversity data collected using a protocol designed for volunteers with results from a professional alternative. *Methods in Ecology and Evolution*, **4**, 383–392.

Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014) Detecting diversity: emerging methods to estimate species diversity. *Trends in Ecology & Evolution*, **29**, 97–106.

Jost, L. (2006) Entropy and diversity. *Oikos*, **113**, 363–375.

Jost, L. (2007) Partitioning diversity into independent alpha and beta components. *Ecology*, **88**, 2427–2439.

Jost, L. (2010) The relation between evenness and diversity. *Diversity*, **2**, 207–232.

Kendall, W.L. & White, G.C. (2009) A cautionary note on substituting spatial subunits for repeated temporal sampling in studies of site occupancy. *Journal of Applied Ecology*, **46**, 1182–1188.

Kéry, M. & Royle, J.A. (2009) Inference about species richness and community structure using species-specific occupancy models in the national Swiss Breeding Bird Survey MHB. *Modeling Demographic Processes in Marked Populations, volume 3 of Environmental and Ecological Statistics* (eds D.L. Thomson, E.G. Cooch & M.J. Conroy), pp. 639–656. Springer, New York, New York, USA.

Lande, R. (1996) Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos*, **76**, 5–13.

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

Link, W.A. (2013) A cautionary note on the discrete uniform prior for the binomial N. *Ecology*, **94**, 2173–2179.

Longino, J.T. & Colwell, R.K. (2011) Density compensation, species composition, and richness of ants on a neotropical elevational gradient. *Ecosphere*, **2**, art29.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Elsevier Academic Press, San Diego, California, USA.

Mao, C.X. (2007) Estimating species accumulation curves and diversity indices. *Statistica Sinica*, **17**, 761–774.

Nesler, T.P., VanBuren, R., Stafford, J.A. & Jones, M. (1997) *Inventory and status of South Platte River native fishes in Colorado*. Aquatic Wildlife Section, Colorado Division of Wildlife.

Perkin, J.S. & Gido, K.B. (2011) Stream fragmentation thresholds for a reproductive guild of Great Plains fishes. *Fisheries*, **36**, 371–383.

Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in Ecology*, 1st edn. Academic Press, San Diego, California, USA.

Royle, J.A. & Dorazio, R.M. (2012) Parameter-expanded data augmentation for Bayesian analysis of capture-recapture models. *Journal of Ornithology*, **152**, S521–S537.

Royle, J.A. & Kéry, M. (2007) A Bayesian state-space formulation of dynamic occupancy models. *Ecology*, **88**, 1813–1823.

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

Williams, M.R. (2009) *Diversity of butterflies and day-flying moths in Urban habitat fragments, south-western Australia*. Ph.D. thesis, Curtin University of Technology.

Zipkin, E.F., Royle, J.A., Dawon, D.K. & Bates, S. (2010) Multi-species occurrence models to evaluate the effects of conservation and management actions. *Biological Conservation* **143**, 479–484.

Received 20 July 2014; accepted 19 October 2014
Handling Editor: Robert B. O'Hara

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix A. Asymptotic incidence-based Hill number.

Appendix B. Multi-species occupancy model.

Appendix C. Extended simulation study results.

Appendix D. Colorado plains fish data.