Journal of Applied Ecology 2007 doi: 10.1111/j.1365-2664.2007.01413.x

Using subsets of species in biodiversity surveys

Mark Vellend[1,2,3]*, Patrick L. Lilley[1,2] and Brian M. Starzomski[2,3,4]

[1]*Department of Botany;* [2]*Biodiversity Research Centre;* [3]*Department of Zoology, University of British Columbia, Vancouver, BC, Canada V6T 1Z4;* [4]*Department of Biology and School for Resource and Environmental Studies, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J1*

Summary

1. In many biodiversity surveys, a small proportion of species require a disproportionate amount of a researcher's time and effort to detect or identify. If we are interested in predicting species diversity or composition, what are the consequences for statistical power of ignoring difficult species – that is, of surveying only a subset of the full suite of species?

2. We analysed 10 data sets on a variety of taxa, at different spatial scales, to assess correlations for species richness and species composition between a full data set and subsets of data with different numbers of species deleted at random, or according to the time investment required for inclusion. Power analyses characterized the trade-off between the number of sites surveyed and the completeness of the survey in each site.

3. For species richness, the majority of information regarding among-site patterns was retained even with large numbers of species removed. With only half the full species pool, the lower 95th percentile of correlations between the full vs. randomly generated reduced data sets was >0·75 in all 10 cases. With 10% of the full species pool removed, correlations were ≥0·95.

4. Subsets of species were not as good at capturing among-site patterns of species composition (ordination scores and pairwise site dissimilarities). With half the full species pool, lower 95th percentile correlations between the full and randomly generated reduced data sets were as low as 0·1. Nonetheless, in most cases the lower 95th percentile correlation with half the number of species was >0·7, and removing 10% of species gave correlations >0·8 across all data sets.

5. For the three data sets in which species were also removed according to the time investment for inclusion, correlations fell within the range of variability observed for random species removals.

6. *Synthesis and applications*. In biological surveys, ignoring a relatively small proportion of species (e.g. <10%), and often a much larger proportion, results in very little loss of information on patterns of biodiversity. As such, statistical power in many biodiversity studies may be maximized by eliminating difficult species from a survey in order to increase the number of sites surveyed.

Key-words: biodiversity survey, sampling design, species composition, species diversity, statistical power, surrogates

Introduction

The goal of biodiversity surveys is to learn as much as we can with limited resources. Developing efficient sampling strategies for capturing spatial variation in measures such as species richness or composition is thus an important goal in ecology and conservation (Magurran 1988; Colwell & Coddington 1994; Lawton *et al*. 1998). In the absence of sufficient resources to survey a particular group of species comprehensively, surrogates of species diversity are often used to set conservation priorities (Caro & O'Doherty 1999; Tognelli 2005; Favreau *et al*. 2006; Pawar *et al*. 2007). Examples of surrogates include particular 'umbrella' species (Sergio *et al*. 2006), species groups that have already been surveyed (Colwell & Coddington 1994), rare and endemic species (Tognelli 2005; Lamoreux *et al*. 2006), or higher taxa (Balmford, Green & Murray 1996; Baldi 2003). Sometimes surrogates work, but often they do not (Andelman & Fagan 2000; Grenyer *et al*. 2006; Magierowski & Johnson 2006). Most previous studies of biodiversity surrogates have assessed the performance of particular variables in predicting biodiversity. Here we take a complementary approach with potentially more general applicability. Rather than assessing the predictive power of a particular subset of species, different taxonomic group or reduced

*Correspondence author. E-mail: mvellend@interchange.ubc.ca

taxonomic resolution, we begin with the full set of species in a biodiversity survey and assess the consequences of removing different proportions of that full set. That is, if we are interested in characterizing site-to-site variability in species diversity and composition for a particular group of organisms (e.g. plants, zooplankton or mammals), what proportion of the species pool must we sample to capture patterns in the group as a whole?

This question is of relevance to applied studies given the arguments above, and also of great relevance to basic studies in community ecology. A central theme in community ecology is understanding the drivers of variation in species diversity and composition (MacArthur 1972; Holyoak, Leibold & Holt 2005), in which case the interest is not in knowing the exact number and identity of every species at a given site, but rather how the diversity and composition at that site compares with other sites. In a typical community survey, the bulk of species may be identified quickly, while a relatively small number of species that are quite difficult to identify occupy a disproportionate amount of a researcher's time (Colwell & Coddington 1994). In a temperate plant community, for example, we may quickly identify species of lilies and woody plants, as well as many graminoids that have distinct morphologies, but we might spend hours identifying those dozen or so particularly difficult grasses and sedges. If surveying only a subsample of species allowed us to survey more sites, what would be the influence on our statistical power to detect relationships of biodiversity measures to potential predictors? To our knowledge there has been no general treatment of this critical consideration in the design of ecological studies. Here we analyse 10 empirical data sets and conduct statistical power analyses to evaluate the consequences of surveying only a subset of the regional species pool for capturing patterns in species richness and compositional differences among sites.

Materials and Methods

To test the consequences of surveying only a subset of species in a given study, our approach was to first calculate species richness and measures of species composition for each site using the full data set. Data subsets were created by deleting different numbers of species from the full data set (removing rows from a species × site table), and for each data subset we recalculated species richness and the same measures of species composition in each site. We then calculated the correlation across sites between species richness (or composition) in the full data set vs. species richness (or composition) in each data subset. The strength of these correlations indicates the degree of information loss, and determines the loss of statistical power, when richness and composition are estimated using data subsets (those in which some species are ignored) rather than full data sets (those in which the full species pool is surveyed).

DATA SETS

Our basic unit of analysis was a data set in which the presence or absence of a set of species was recorded in a set of sites. We analysed 10 such data sets, three of which came from our own study systems: plant communities in coastal grasslands in British Columbia, Canada (Roemer 1972); invertebrate communities in moss microecosystems (Starzomski & Srivastava 2007); and forest-herb communities in deciduous forests (Vellend 2004). The present paper was motivated during the design stage of an ongoing study of coastal grasslands on Vancouver Island, Canada. We felt there was a potentially important trade-off between the number of sites we could survey, and the thoroughness of each survey in terms of the statistical power we would have to detect effects of variables such as site isolation or environmental conditions on species diversity and composition. Many species in this ecosystem are difficult to distinguish from close relatives (e.g. some *Bromus* spp. and *Festuca* spp.), or are difficult to detect in dense vegetation (e.g. *Selaginella* spp.). It became obvious that the same or related issues are faced in many, if not most, community studies, perhaps to the greatest degree in surveys of small invertebrates. Our knowledge of the three data sets from our own study systems allowed us to subsample species not only randomly, but also according to the time investment required to include species in a survey.

We selected seven additional data sets from those available in the literature or in public databases with the goal of representing a wide range of taxa, scales of observation and total species pool sizes. These included surveys of zooplankton, ants, butterflies, reptiles, birds and mammals, in addition to the surveys of plants and invertebrates in the data sets from our own study systems. Scales of observation ranged from <1 m^2 in a single local experiment to entire ecoregions in the neotropics, and total species pools ranged from 22 to 1126 species. Each data set represents a thorough survey of the group of interest. Data sets and their characteristics are listed in Table 1 and described in Appendix S1 in Supplementary Material.

CHARACTERIZING SPECIES RICHNESS AND COMPOSITION

For each full data set, we first calculated the number of species in each site: species richness (SR). To characterize species composition, we used two methods. First, we calculated scores for each site on the first axis of a detrended correspondence analysis (DCA1); second, we calculated values of Jaccard's dissimilarity index (J) for all pairs of sites. Our motivation for using these methods is provided below. For the full data sets, our metrics of species richness and composition are denoted SR_{FULL}, $DCA1_{FULL}$, and J_{FULL} and serve as benchmarks with which to compare the same variables calculated when only a subset of the species are included in the analysis (SR_{SUB}, $DCA1_{SUB}$ and J_{SUB}).

Species composition can be characterized in many different ways (Legendre & Legendre 1998). Probably the most common method is to first reduce the dimensionality of a species-by-site data set via ordination, for which there are several classes of approach. Methods based on eigenanalysis are attractive in that they provide singular mathematical solutions for a set of orthogonal axes (regardless of how many axes a researcher decides to look at) on which each site has a score; the first axis describes the maximum possible amount of variation among sites in species composition, with each successive axis describing progressively less variation (McCune & Grace 2002). Correspondence analysis (CA) is attractive to community ecologists because, in contrast to methods based on principal components analysis, it allows for unimodal responses to underlying gradients, thereby providing a clear representation of 'long' gradients (as well as 'short' gradients). Detrended correspondence analysis (DCA) corrects for two artefacts in CA: the 'arch effect' and compression of the two extremes of the first axis (McCune & Grace 2002). Despite some drawbacks of detrending methods (McCune & Grace 2002),

Table 1. Data sets used in this study (described further in Appendix S1)

Taxon	Habitat	Location	Sampling unit	Number of sampling units	Total number of taxa	Reference
Plants	Coastal grasslands	Vancouver Island, Canada	20 × 20-m plot	50	129	Roemer (1972)
Herbaceous forest plants	Deciduous forest	NY State, USA	Discrete habitat patch, 0·5–30 ha	27	72	Vellend (2004)
Invertebrates	Moss patches	Vancouver, Canada	Experimental patch	80	157	Starzomski & Srivastava (2007)
Zooplankton	Lakes	Eastern USA	Lake	350	375	USEPA (1996a)
Ants	Temperate forest	Massachusetts, USA	10 × 10-m plot	16	22	Ellison *et al.* (2005)
Butterflies	Grassland	Boulder, Colorado, USA	'Site'	66	58	Oliver, Prudic & Collinge (2006)
Reptiles	Dry tropical forest	Neotropics	Ecoregion	20	1126	World Wildlife Fund (2006)
Birds	Lakes	Eastern USA	Lake	214	179	USEPA (1996b)
Small mammals	Alpine	Great Basin, USA	Mountain range	21	30	Rickart (2001)
Small mammals	Alpine	Utah, USA	Mountain range or plateau	7	30	Rickart (2001)

the balance of pros and cons favours DCA as our preferred method of eigenanalysis-based ordination. The main criticism of DCA, concerning removal of the arch effect, is relevant only when extracting ≥2 axes (we extract only one), and instability in the solutions that were caused by programming bugs have since been fixed (McCune & Grace 2002). Thus, for each of our data sets we conducted a DCA and extracted the site scores from the first axis so that each site was represented by a single value along this axis (DCA1). Correlating $DCA1_{FULL}$ and $DCA1_{SUB}$ addresses the question of whether each subset reveals the same dominant axis of variation in community composition. A fuller representation of site-to-site variation in community composition can be represented by a raw (dis)similarity matrix, as described below.

In contrast to eigenanalysis-based methods, distance-based methods of ordination (including principal coordinates analysis and non-metric multidimensional scaling, NMDS), begin by calculating a measure of (dis)similarity between each pair of sites, followed by an arrangement of the sites in a predetermined number of dimensions (typically 1–3) so that the distances between pairs of sites in the reduced space most closely reflects the raw dissimilarities. NMDS currently appears to be the method of distance-based ordination preferred by many researchers (McCune & Grace 2002), but it is difficult to implement in a setting with repeated analyses on different subsets of species because each individual analysis needs to be scrutinized for whether an optimal solution was found, and how many axes are appropriate to extract. Unlike DCA axes, the ordering of the NMDS axes carries no interpretable meaning and the results of an NMDS depend on the number of predetermined axes. Rather than employ a distance-based method of ordination, here we look directly at the raw dissimilarities between each pair of sites, which are themselves the focus of analysis in many studies (Tuomisto & Ruokolainen 2006). For each pair of sites in each data set, we calculated Jaccard's dissimilarity (J), which is one of the most commonly used dissimilarity indices for presence–absence data (Legendre & Legendre 1998). In sum, we calculated DCA1 and J as complementary characterizations of species composition, representing the primary axis of variation and the full set of pairwise differences, respectively.

DATA SUBSETTING

For all data sets, we calculated SR, DCA1, and J for subsets of data in which species were chosen randomly for inclusion at each of five to seven species pool sizes (the number of species in the subset) ranging from nearly the full species pool down to 30% or less of the full species pool. At each species pool size, we took 100 random subsets (with replacement) of species with the restriction that no sites have zero species (which makes DCA impossible). To characterize the degree to which patterns of species richness and composition in the subsets reflected patterns in the full data set, for each random draw of species we calculated the Pearson product-moment correlation (Zar 1996) between SR_{SUB} and SR_{FULL}, between $DCA1_{SUB}$ and $DCA1_{FULL}$, and between J_{SUB} and J_{FULL}. These correlations were plotted against the species pool size to assess the consequences of eliminating different numbers of species from each survey. In order to assess worst-case scenarios of information loss due to subsampling the species pool, our interpretation of the results focused largely on the lower 95th percentile of the 100 correlations at each species pool size. As explained in the discussion, we expect this relatively conservative approach to allow inferences to be drawn about the consequences of removing non-random sets of species (e.g. those that are difficult to identify) as well as random subsets.

For each of the data sets from our own study systems, we classified each species or taxon according to the time investment needed for inclusion. For plants in coastal grasslands, this was based on a combination of ease of identification and ease of detection; five categories were used: (1–4) vascular plants of increasing difficulty to identify or detect, and (5) mosses and lichens, which require specialized training to identify. For forest herbs in deciduous forests, three categories were used: (1) easy to identify even at a distance; (2) taxa that require close examination to identify based on vegetative characters, which are often all that is available; (3) spring ephemerals for which an entire second survey is required. For invertebrates in moss patches, four categories were used: (1) taxon identified immediately; (2) key needed for identification, with only a small number of similar taxa; (3) key needed for identification, with a large number of similar species; (4) key needed for identification, always very difficult to identify due to a large number of similar species and very small diagnostic features. For each data set we then removed successive categories of species, starting with the most problematic, and assessed correlations between full data sets and subsets for SR, DCA1 and *J*. All the analyses described above were conducted using R ver. 2·4·0 (R Development Core Team 2004), including VEGAN ver. 1·7-82 by Jari Oksanen (code in Appendix S2).

POWER ANALYSIS

To assess the influence on statistical power of using only a subset of species (the probability of rejecting a false null hypothesis), we conducted two kinds of power analysis. Essentially we are addressing the effect on statistical power of using measurements that act as surrogates for the true values of particular variables, which is akin to adding a degree of measurement error. Data on species richness and composition may be used in a wide variety of analyses, each of which could be subject to its own unique power analysis. When the effect size is very large (e.g. variable *x* is a very strong predictor of SR, DCA1 or *J*), using decent surrogates or different sample sizes will have little influence on statistical power – the effect will be detected regardless. Here we present two generic scenarios with modest effect sizes, one for estimating a correlation (e.g. between one continuous environmental variable and SR) and one for an ANOVA (e.g. testing the effect of one categorical variable).

For testing correlations, we assume that the expected correlation between hypothetical continuous variable ENV and SR (or DCA1 or *J*) is 0·5:

$$E[r(\mathrm{SR_{FULL}} \times \mathrm{ENV})] = 0{\cdot}5 \qquad \text{eqn 1}$$

If instead of measuring $\mathrm{SR_{FULL}}$ we measure $\mathrm{SR_{SUB}}$, then we expect a somewhat weaker relationship, as essentially we are estimating $\mathrm{SR_{FULL}}$ with an added source of error. Specifically, assuming that the errors associated with calculating $\mathrm{SR_{SUB}}$ rather than $\mathrm{SR_{FULL}}$ are normally distributed, the expected correlation between $\mathrm{SR_{SUB}}$ and ENV is (Sokal & Rohlf 1981):

$$\mathrm{E}[r(\mathrm{SR_{SUB}} \times \mathrm{ENV})] = \mathrm{E}[r(\mathrm{SR_{FULL}} \times \mathrm{ENV})] \times \mathrm{E}[r(\mathrm{SR_{FULL}} \times \mathrm{SR_{SUB}})] \qquad \text{eqn 2}$$

For example, if $\mathrm{SR_{SUB}}$ in a particular subset of species shows a correlation of 0·9 with $\mathrm{SR_{FULL}}$, then the expected value of the correlation between $\mathrm{SR_{SUB}}$ and ENV is $0{\cdot}5 \times 0{\cdot}9 = 0{\cdot}45$. To calculate the power of a test of the hypothesis $r \neq 0$, we need to specify the true value of the correlation, the desired alpha level and the sample size. To ask to what degree power is reduced by measuring $\mathrm{SR_{SUB}}$ instead of $\mathrm{SR_{FULL}}$, we can then simply calculate the power of testing for a correlation of $\mathrm{E}[r(\mathrm{SR_{FULL}} \times \mathrm{ENV})]$ vs. a correlation of $\mathrm{E}[r(\mathrm{SR_{SUB}} \times \mathrm{ENV})]$, in this example 0·5 vs. 0·45. For sample sizes of 20–50 (a realistic range for coastal grassland patches in a single field season), we calculated the power of testing for correlations of 0·5 with values of $r(\mathrm{SR_{SUB}} \times \mathrm{SR_{FULL}})$ ranging from 0·8 to 1·0.

For a one-way ANOVA with two groups (e.g. two levels of variable ENV2), the calculation of power requires specification of the ratio of within-group variance to among-group variance, the sample size per group, and the desired alpha level. By measuring $\mathrm{SR_{SUB}}$ instead of $\mathrm{SR_{FULL}}$ (the dependent variable is measured with some error), in effect we increase the ratio of within : among group variance by a factor of $1/r^2(\mathrm{SR_{SUB}} \times \mathrm{SR_{FULL}})$. For this case, we assumed that for $\mathrm{SR_{FULL}}$, the ratio of within : among group variance = 2, and for sample sizes ranging from 20 to 50 (10–25 per group) we calculated power for scenarios corresponding to $r(\mathrm{SR_{SUB}} \times \mathrm{SR_{FULL}})$ ranging from 0·8 to 1·0, as for the correlation example. Power analyses were conducted using SYSTAT ver. 11 (Systat Software 2004) for correlations, and in R for ANOVA.

Results

As the number of species included in data sets with randomly subsampled species is decreased, the strength of correlation in the reduced vs. full data sets declines, typically at an increasing rate as fewer and fewer species are selected (Figs 1–3). Correlations were strongest for species richness (SR; Fig. 1). For subsets of species representing half the full species pool, the median $r(\mathrm{SR_{SUB}} \times \mathrm{SR_{FULL}})$ across random samples was >0·85 in all data sets up to a maximum of 0·98, and the lower 95th percentile was >0·75 in all data sets up to a maximum of 0·98 (Fig. 1). If we take a correlation of 0·9 as indicative of a good surrogate, we can ask what proportion of the species pool can be dropped so that even the lower 95th percentile correlation between the full and reduced data sets is ≥0·9. For SR, a minimum of 20% (Fig. 1f, butterflies in grasslands) and as many as 80% (Fig. 1g, reptiles in the neotropics) of species can be dropped from a data set and show $r(\mathrm{SR_{SUB}} \times \mathrm{SR_{FULL}}) \geq 0{\cdot}9$.

For both DCA1 and *J*, median correlations at half the species pool were lower than for SR, but still ≥0·73 in all data sets, and up to >0·95 (Figs 2 and 3). Particular random subsets of half the species pool can potentially show correlations with the full data set that are much lower, in some cases showing little correlation (e.g. Figure 2f, butterflies in grasslands), although correlations <0·5 were rare. To achieve lower 95th percentile correlations between the full and reduced data sets of ≥0·9 for DCA1 and *J*, all 22 species of ants studied by Ellison *et al*. (2005; Figs 2e and 3e) would be required, but in all other data sets at least 10% of species could be dropped from a data set, up to a maximum of ≈70% for zooplankton in lakes (DCA1; Fig. 2d) and reptiles in the neotropics (*J*; Fig. 3g).

With up to 20–30% of species removed from a data set, results for DCA1 and *J* were quite similar; but with larger proportions of species removed, correlations for DCA1 showed greater variability among random subsets than

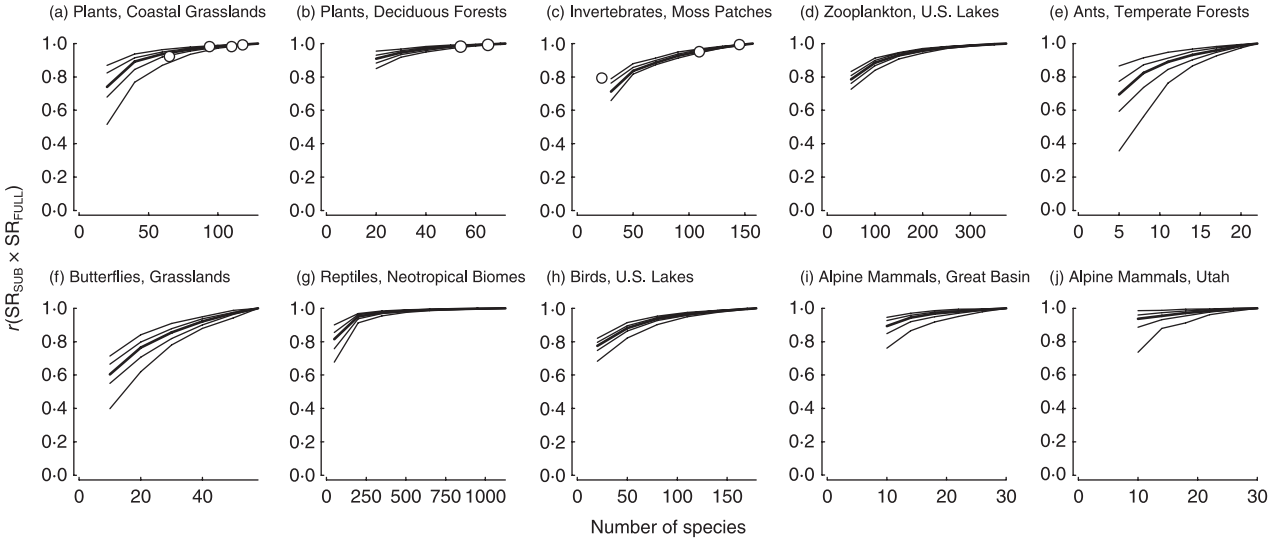

Fig. 1. The magnitude of correlation across sites between species richness (SR) calculated using a subset of species (number of species on *x*-axis) with SR calculated using the full data set. Bold line connects median correlations for 100 randomly chosen subsets at each of five to seven levels of species number; thin lines show upper and lower 75th and 95th percentiles; ❍, non-random subsets of species, with species eliminated based on the amount of resources required to include them in the survey; (a)–(j) correspond to the 10 data sets described in Table 1.

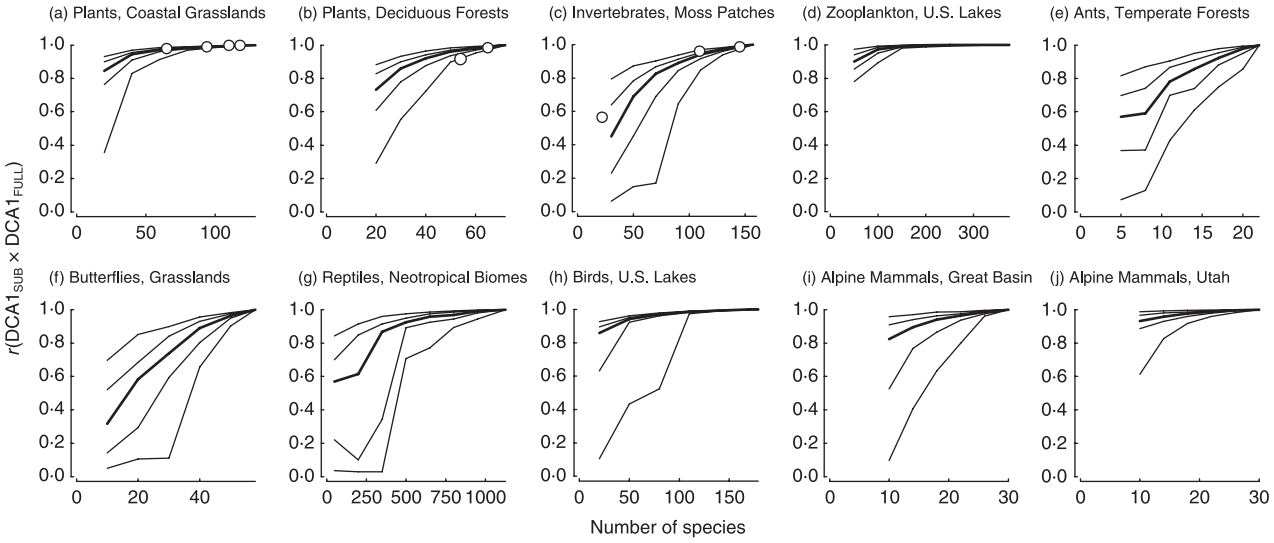

Fig. 2. The magnitude of correlation across sites between ordination scores on the first axis of a detrended correspondence analysis (DCA1) calculated using a subset of species (number of species on *x*-axis) with DCA1 calculated using the full data set. Bold line connects median correlations for 100 randomly chosen subsets at each of five to seven levels of species number; thin lines show upper and lower 75th and 95th percentiles; ❍, non-random subsets of species, with species eliminated based on the amount of resources required to include them in the survey; (a)–(j) correspond to the 10 data sets described in Table 1.

correlations for J (Figs 2 and 3). This suggests a fairly predictable loss of information with increasingly small subsets of species when analysing the full pattern of compositional variability (J), but much less predictability in terms of whether a particular subset of species will reveal the same dominant gradient in species composition (DCA1). Failure to capture a dominant axis of compositional variability with a species subset suggests that subsequent ordination axes might be comparable in importance with the first, and therefore would emerge as the first axis when certain subsets of species are removed. As a rough test of this idea, we calculated the lower 95th percentile of $r(\mathrm{DCA1_{SUB}} \times \mathrm{DCA1_{FULL}})$ at the species pool size in each data set that was closest to representing 30% of the full set of species (27–36% across different data sets), and correlated this with the ratio of eigenvalues between the second and first axes of each DCA, which represents the relative proportion of community composition accounted for by the second vs. first axis. The correlation was significantly negative ($r = -0{\cdot}66$, $P < 0{\cdot}04$), suggesting that when species composition is characterized by multiple gradients of comparable importance, small subsets of species may fail completely in capturing the dominant gradient in the full data set.

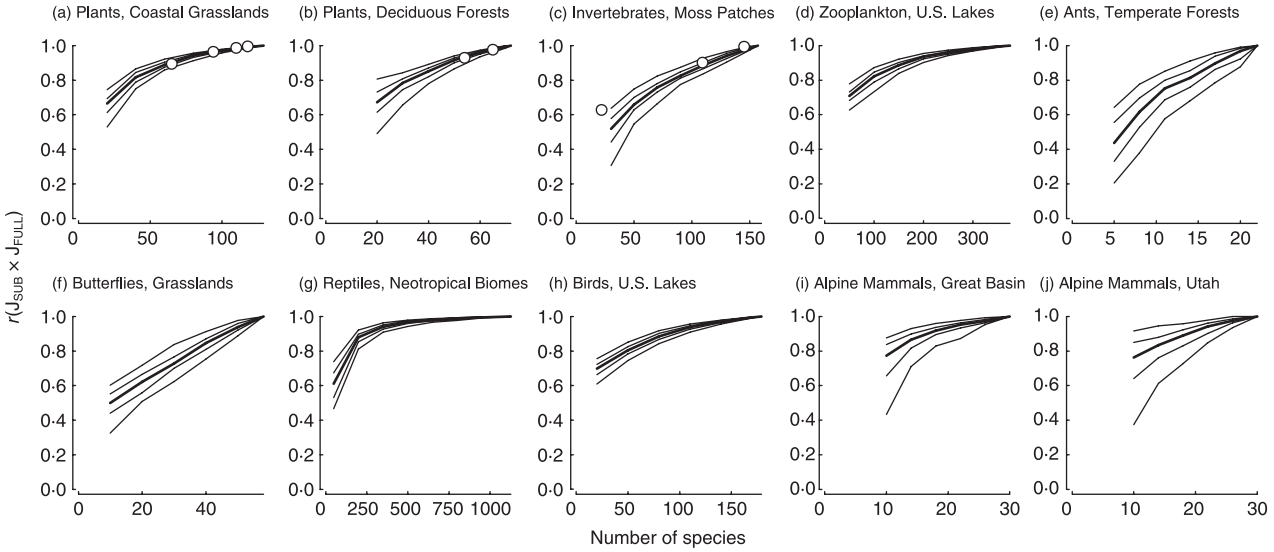

Fig. 3. The magnitude of correlation across site-to-site pairs between Jaccard's dissimilarities in species composition (J) calculated using a subset of species (number of species on x-axis) with J calculated using the full data set. Bold line connects median correlations for 100 randomly chosen subsets at each of five to seven levels of species number; thin lines show upper and lower 75th and 95th percentiles; ❍, non-random subsets of species, with species eliminated based on the amount of resources required to include them in the survey; (a)–(j) correspond to the 10 data sets described in Table 1.

Removing non-random subsets of species based on the time and effort required to survey them revealed correlations within the range expected for random subsets (a–c in Figs 1–3). For plants in coastal grasslands, the three most difficult categories included 35 species, and correlations between the subsets and full data sets were >0·96 for SR, DCA1 and J after removing these species. Removing an additional 29 species in the next most difficult category gave correlations >0·89 for all three variables. For deciduous forest herbs, removing the seven spring ephemerals gave correlations >0·97 for all variables; removing an additional 11 difficult species gave correlations >0·9. For invertebrates in moss patches, removing the 12 and then 36 species in the two most difficult categories resulted in correlations of >0·9, and removing an additional 87 species in the next most difficult category resulted in correlations of 0·79, 0·57 and 0·63 for SR, DCA1 and J, respectively.

The power analyses for testing correlations and factor effects in one-way ANOVAs were quite similar (Fig. 4). If we could hypothetically sample a maximum of 20 sites for a full suite of species, then using a subset of species with r(SUB × FULL) = 0·95 would increase statistical power in these scenarios if the time savings meant we could sample two or more additional sites. If r(SUB × FULL) = 0·9, we would have increased statistical power if we could sample at least four to five additional sites. With either a greater number of sites that could be sampled for the full suite of species, or with a reduced r(SUB × FULL), the number of additional sites needed to make up for the drop in power due to using a subset of species increases. Alternatively, if we could sample a maximum of 50 sites for a subset of species given r(SUB × FULL) = 0·9, then we could increase statistical power by sampling the full subset of species as long as we could sample more than 40 sites.

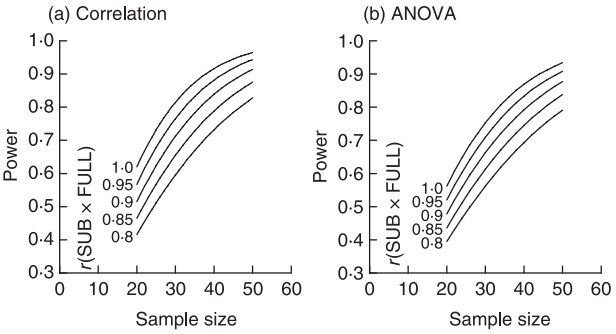

Fig. 4. Statistical power analysis showing the power (probability of rejecting a false null hypothesis) of (a) testing the null hypothesis that $r \neq 0$ for a true correlation of 0·5 between a hypothetical environmental variable and species richness or composition; (b) testing the null hypothesis that two estimated means (for species richness or composition) are the same, with the true ratio of within : among group variance = 2. In both cases we assume that species richness or composition is measured using a subset of the data with a correlation of r(SUB × FULL) with the full data set.

Discussion

The utility of our results depends on knowledge of the trade-off between the resources invested in including additional species in a survey (typically a researcher's time, or money to pay assistants) and other ways in which the same resources could be used to increase statistical power, which we employ here as the relevant currency in a research enterprise. Increasing the sample size at the expense of including all species is an obvious manifestation of this trade-off, although resources could be put to other uses as well, such as making more accurate or meaningful measurements of potential predictors of species richness and composition (e.g. environmental

variables). Although it is generally difficult to accurately quantify such trade-offs, which will probably be highly system-specific, our intuition and experience suggest that, in many cases, statistical power will not be maximized by attempting to include every last species in a survey.

A real example serves to illustrate the potential utility of our results. A huge proportion of ecological research is carried out in the form of graduate student projects, which are quite restricted in terms of time and available resources (Karban & Huntzinger 2006). As a Master's student with a full-time field assistant, one of us (P.L.L.) was able to conduct two vegetation surveys in each of about 40 sites in coastal grassland habitat in one field season, in the same region of Vancouver Island as studied by Roemer (1972). Two surveys, one in spring and one in summer, are needed to capture species with different phenologies. Given the time period over which the surveys were conducted, the phenology of the plant species, and difficulties with identification of certain taxa, it will probably be necessary to eliminate ≈10–15% of the species from the data set due to lack of comparability among sites and uncertain identifications. We estimated that spending the time to examine every difficult species in the field closely, and to collect a voucher specimen of every difficult identification, would have meant reducing the sample size to as few as 30 sites. The analyses of Roemer's (1972) data in the same ecosystem (a in Figs 1–3), and most of the other data sets as well, suggest that eliminating 10–15% of species from a survey provides estimates of species richness and composition that correlate very strongly with species richness and composition in the full data set ($r > 0{\cdot}96$ for Roemer's data). The resulting reduction in statistical power for detecting community–environment relationships in our generic scenarios (Fig. 4) is much smaller than the reduction we would have imposed if only 30 sites were sampled for the full suite of species, instead of 40.

More general application of our results requires additional considerations. In particular, while most of our analyses focused on random subsets of species, trade-offs in terms of time or money come from eliminating difficult (non-random) species from a survey. However, in the three cases for which we also explored species removals based on degree of time investment, the results fell within the range of variation observed among random subsets. Combined with the fact that we generated hundreds of different species subsets for each data set, we feel that the lower 95th percentile of correlations among random subsets represents a conservative estimate of the potential information lost as a consequence of surveying less than the full set of species, even for non-random species removals. In theory, it is possible that a group of species that is particularly informative in differentiating species composition among sites is also particularly difficult to identify, but this seems unlikely to be generally the case, and we expect that most researchers will have sufficient experience with their study system to avoid excluding such a group of species. Before deciding to eliminate a large proportion (e.g. >20%) of species from a survey, pilot data would be needed to determine the trade-off with sample size in terms of statistical power. For smaller numbers of species, our results suggest a general and fairly conservative rule of thumb: if ignoring up to 10% of species provides sufficient time savings to allow a substantial increase in sample size, statistical power is likely to be maximized by ignoring those species. A possible exception would be when the species pool itself is quite small, as for the 22 ant species (the smallest species pool among the data sets used here) surveyed by Ellison *et al*. (2005), in which case the investment in species identification is unlikely to be a major limitation to begin with.

We suspect that the trade-off between the comprehensiveness of a survey and the sample size will apply in many studies, particularly when identifying a small subset of species occupies a disproportionate amount of resources, as is often the case for plants and diverse groups of invertebrates in particular (Colwell & Coddington 1994). In many cases, of course, the trade-off will not be so clear. For most vertebrates, all but a negligible proportion of resources is spent on setting up and monitoring traps or nets (Thompson, White & Gowan 1998), walking transects and stopping to make point counts, etc. (Krebs 1999). In these cases there may be virtually nothing to gain by eliminating species from the survey (we included vertebrate data sets here to ensure broad taxonomic coverage). For data sets put together based on range maps (e.g. the WildFinder database, World Wildlife Fund 2006), resources can be saved by processing data for fewer species, but no particular subset of species takes more time than any other, and the process may be automated anyway. Finally, when a large number of sites can be surveyed for the full suite of species, the trade-off may be such that statistical power is maximized by using the full set rather than using a subset in an even larger number of sites, given the decelerating and parallel power curves at different levels of subsetting (Fig. 4). Nonetheless, our results suggest that for all those data sets with entries for *Carex* spp., *Daphnia* spp., or a few cryptic species that could not be identified, very little information on species diversity and composition is lost as long as the number of such taxa is relatively small.

Our results apply to the measurement of relative variation among sampling units in species diversity and composition. This covers a huge number of studies in community ecology (Magurran 1988; Colwell & Coddington 1994), but to the extent that a biological survey study is aimed at providing distributional information for all the species in a given group, or for estimating the actual value of species richness, there is obviously no substitute for attempting to survey all species (Lawton *et al*. 1998; Maurer 2000). Importantly, surrogates of species richness or composition, especially correlations across distinct taxonomic groups, are often used to justify large-scale conservation actions (Moritz *et al*. 2001; Oertli *et al*. 2005; Grenyer *et al*. 2006; Lamoreux *et al*. 2006; Pawar *et al*. 2007). Our results differ in being more specific to the group of organisms directly under study, and they suggest that conservation prioritization decisions based on patterns of species diversity and composition (species richness in particular) probably need not wait for comprehensive surveys to be completed if existing surveys already cover a large

proportion of the regional species pool. In addition, by reducing the need for a high level of taxonomic expertise, sampling a subset of species can reduce the complexity and cost of monitoring programmes, which are critical to their effectiveness (Elzinga *et al*. 2001).

Although we have emphasized so far the potential for the benefits of increased sample size to outweigh the costs of including fewer species in a survey, it is important to note that this is because SR_{SUB}, $DCA1_{SUB}$ and J_{SUB} are often excellent surrogates for SR_{FULL}, $DCA1_{FULL}$ and J_{FULL}, sometimes even with up to 50% of species retained (Figs 1–3). However, in the conservation literature, variables used as surrogates for biodiversity are often considered 'effective' even when correlations are relatively low. For example, Su *et al*. (2004) considered correlations of 0·29–0·77 for site-to-site community dissimilarity values across different taxa to be useful in predicting species composition in one group based on another. The effectiveness of a surrogate depends on the application, but for testing the relationship of a given variable such as species diversity or composition with others, as in the examples used here, using a surrogate that correlates with the variable of true interest at $r < 0{\cdot}7$ may result in a worrying loss of statistical power (Fig. 4). Strong effects of particular predictors on biodiversity may be picked up using a surrogate, but more subtle effects will probably be missed or spuriously found.

In summary, we feel that a wide range of ecological studies could gain from considering the potential benefits of surveying a subset of the total species pool and redirecting resources to increasing the study's sample size, or some other aspect of the study that increases statistical power. It is almost never the default approach to survey all possible sites where a focal group of species might be present, so why should the default approach be to survey every last species in the same focal group? In conservation, the use and study of biodiversity surrogates is widespread (Caro & O'Doherty 1999), and in fundamental community ecology there is almost always some degree to which the taxa recorded represent a subset of the full suite of species (not all individuals are identified to species; e.g. Pik, Oliver & Beattie 1999). Our results suggest that some surrogates considered useful (e.g. if $r = 0{\cdot}6$) may in fact provide misleading statistical results; that using a species subset as a surrogate for the full set may ultimately maximize statistical power; and that ecologists probably need not worry about all those times they needed to drop a handful of species from an analysis.

Acknowledgements

We thank Tara Martin, Robin Naidoo, Luke Harmon, Dolph Schluter, Marc Cadotte and three anonymous reviewers for insightful comments and advice that helped improve the manuscript. This research was supported by the Natural Sciences and Engineering Research Council, Canada.

References

Andelman, S.J. & Fagan, W.F. (2000) Umbrellas and flagships: efficient conservation surrogates or expensive mistakes? *Proceedings of the National Academy of Sciences, USA*, **97**, 5954–5959.

Baldi, A. (2003) Using higher taxa as surrogates of species richness: a study based on 3700 Coleoptera, Diptera, and Acari species in Central Hungarian reserves. *Basic and Applied Ecology*, **4**, 589–593.

Balmford, A., Green, M.J.B. & Murray, M.G. (1996) Using higher-taxon richness as a surrogate for species richness. I. Regional tests. *Proceedings of the Royal Society of London B*, **263**, 1267–1274.

Caro, T.M. & O'Doherty, G. (1999) On the use of surrogate species in conservation biology. *Conservation Biology*, **13**, 805–814.

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

Ellison, A.M., Chen, J., Díaz, D., Kammerer-Burnham, C. & Lau, M. (2005) Changes in ant community structure and composition associated with hemlock decline in New England. *Proceedings of the Third Symposium on Hemlock Woolly Adelgid in the Eastern United States* (eds B. Onken & R. Reardon), pp. 280–289. US Department of Agriculture, US Forest Service, Forest Health Technology Enterprise Team, Morgantown, WV, USA.

Elzinga, C.L., Salzer, D.W., Willoughby, J.W. & Gibbs, J.P. (2001) *Monitoring Plant and Animal Populations*. Blackwell Science, Malden, MA, USA.

Favreau, J.M., Drew, C.A., Hess, G.R., Rubino, M.J., Koch, F.H. & Eschelbach, K.A. (2006) Recommendations for assessing the effectiveness of surrogate species approaches. *Biodiversity and Conservation*, **15**, 3949–3969.

Grenyer, R., Orme, C.D.L., Jackson, S.F. *et al*. (2006) Global distribution and conservation of rare and threatened vertebrates. *Nature*, **444**, 93–96.

Holyoak, M., Leibold, M.A. & Holt, R.D. (2005) *Metacommunities: Spatial Dynamics and Ecological Communities*. University of Chicago Press, Chicago, IL, USA.

Karban, R. & Huntzinger, M. (2006) *How to do Ecology: A Concise Handbook*. Princeton University Press, Princeton, NJ, USA.

Krebs, C.J. (1999) *Ecological Methodology*, 2nd edn. Addison-Welsey, Menlo Park, CA, USA.

Lamoreux, J.F., Morrison, J.C., Ricketts, T.H. *et al*. (2006) Global tests of biodiversity concordance and the importance of endemism. *Nature*, **440**, 212–214.

Lawton, J.H., Bignell, D.E., Bolton, B. *et al*. (1998) Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. *Nature*, **391**, 72–76.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*, 2nd English edn. Elsevier, Amsterdam.

MacArthur, R.H. (1972) *Geographical Ecology: Patterns in the Distribution of Species*. Harper & Row, New York.

Magierowski, R.H. & Johnson, C.R. (2006) Robustness of surrogates of biodiversity in marine benthic communities. *Ecological Applications*, **16**, 2264–2275.

Magurran, A.E. (1988) *Ecological Diversity and its Measurement*. Princeton University Press, Princeton, NJ, USA.

Maurer, D. (2000) The dark side of the taxonomic sufficiency (TS). *Marine Pollution Bulletin*, **40**, 98–101.

McCune, B. & Grace, J.B. (2002) *Analysis of Ecological Communities*. MjM Software Design, Gleneden Beach, OR, USA.

Moritz, C., Richardson, K.S., Ferrier, S. *et al*. (2001) Biogeographical concordance and efficiency of taxon indicators for establishing conservation priority in a tropical rainforest biota. *Proceedings of the Royal Society of London B,* **268**, 1875–1881.

Oertli, S., Muller, A., Steiner, D., Breitenstein, A. & Dorn, S. (2005) Cross-taxon congruence of species diversity and community similarity among three insect taxa in a mosaic landscape. *Biological Conservation*, **126**, 195–205.

Oliver, J.C., Prudic, K.L. & Collinge, S.K. (2006) Boulder County Open Space butterfly diversity and abundance. *Ecology*, **87**, 1066.

Pawar, S., Koo, M.S., Kelley, C., Ahmed, M.F., Chaudhuri, S. & Sarkar, S. (2007) Conservation assessment and prioritization of areas in northeast India: priorities for amphibians and reptiles. *Biological Conservation*, **136**, 346–361.

Pik, A.J., Oliver, I. & Beattie, A.J. (1999) Taxonomic sufficiency in ecological studies of terrestrial invertebrates. *Australian Journal of Ecology*, **24**, 555–562.

R Development Core Team. (2004) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rickart, E.A. (2001) Elevational diversity gradients, biogeography and the structure of montane mammal communities in the intermountain region of North America. *Global Ecology and Biogeography*, **10**, 77–100.

Roemer, H.L. (1972) *Forest vegetation and environments of the Saanich Peninsula, Vancouver Island*. PhD thesis, University of Victoria, Victoria, BC, Canada.

Sergio, F., Newton, I., Marchesi, L. & Pedrini, P. (2006) Ecologically justified charisma: preservation of top predators delivers biodiversity conservation. *Journal of Applied Ecology*, **43**, 1049–1055.

Sokal, R.R. & Rohlf, F.J. (1981) *Biometry*, 2nd edn. W.H. Freeman, San Francisco, CA, USA.
Starzomski, B.M. & Srivastava, D.S. (2007) Landscape geometry determines community response to disturbance. *Oikos*, **116**, 690–699.
Su, J.C., Debinski, D.M., Jakubauskas, M.E. & Kindscher, K. (2004) Beyond species richness: community similarity as a measure of cross-taxon congruence for coarse-filter conservation. *Conservation Biology*, **18**, 167–173.
Systat Software (2004) *SYSTAT for WINDOWS*, ver. 11. Systat Software, San Jose, CA, USA.
Thompson, W.L., White, G.C. & Gowan, C. (1998) *Monitoring Vertebrate Populations*. Academic Press, San Diego, CA, USA.
Tognelli, M.F. (2005) Assessing the utility of indicator groups for the conservation of South American terrestrial mammals. *Biological Conservation*, **121**, 409–417.
Tuomisto, H. & Ruokolainen, K. (2006) Analyzing or explaining beta diversity? Understanding the targets of different methods of analysis. *Ecology*, **87**, 2697–2708.
USEPA (1996a) *EMAP Surface Waters Lake Database, 1991–94, Northeast Lakes, Lake Zooplankton Count Data Summarized by Lake*. US Environmental Protection Agency, National Health Environmental Effects Research Laboratory, Western Ecology Division, Corvallis, OR, USA.
USEPA (1996b) *EMAP Surface Waters Lake Database, 1991–94, Northeast Lakes, Breeding Bird Count Data Summarized by Lake*. US Environmental Protection Agency, National Health Environmental Effects Research Laboratory, Western Ecology Division, Corvallis, OR, USA.
Vellend, M. (2004) Parallel effects of land-use history on species diversity and genetic diversity of forest herbs. *Ecology*, **85**, 3043–3055.
World Wildlife Fund (2006) *WildFinder: Online Database of Species Distributions*, January 2006. http://www.worldwildlife.org/wildfinder
Zar, J.H. (1996) *Biostatistical Analysis*. Prentice Hall, Englewood Cliffs, NJ, USA.

Received 29 January 2007; accepted 20 August 2007
Handling Editor: Marc Cadotte

Supplementary Material

The following supplementary material is available for this article.

Appendix S1. Data set descriptions

Appendix S2. R code

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/full/10.1111/j.1365-2664.2007.01413.x.
(This link will take you to the article abstract.)

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.