Ecological Monographs, 0(0), 2018, pp. 1–18

Upscaling biodiversity: estimating the species–area relationship from small samples

William E. Kunin,[1,2,20] John Harte,[3] Fangliang He,[4] Cang Hui,[5] R. Todd Jobe,[6,16] Annette Ostling,[7] Chiara Polce,[1,17] Arnošt Šizling,[8] Adam B. Smith,[3,9] Krister Smith,[10] Simon M. Smart,[11] David Storch,[8,12] Even Tjørve,[13,18] Karl-Inne Ugland,[14] Werner Ulrich,[15] and Varun Varma[1,19]

[1]*Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT United Kingdom*
[2]*Stellenbosch Institute for Advanced Studies (STIAS), Wallenberg Research Centre at Stellenbosch University, Stellenbosch 7600 South Africa*
[3]*Energy and Resources Group and Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720 USA*
[4]*Department of Renewable Resources, University of Alberta, Edmonton, Alberta T6G 2H1 Canada*
[5]*Department of Mathematical Sciences, Centre for Invasion Biology, Stellenbosch University, and African Institute for Mathematical Sciences, Stellenbosch 7600 South Africa*
[6]*Department of Geography, University of North Carolina, Chapel Hill, North Carolina 27599-3220 USA*
[7]*Department of Ecology and Evolutionary Biology, University of Michigan, 830 North Avenue, Ann Arbor, MI 48109-1048 USA*
[8]*Center for Theoretical Study, Charles University and the Academy of Sciences of the Czech Republic, Jilská 1, 110 00 Praha 1, Czech Republic*
[9]*Center for Conservation and Sustainable Development, Missouri Botanical Garden, 4344 Shaw Boulevard, St. Louis, Missouri 63110 USA*
[10]*Senkenberg Research Institute and Natural History Museum, Senckenberganlage 25, 60325 Frankfurt am Main, Germany*
[11]*NERC Centre for Ecology and Hydrology, Library Avenue, Bailrigg, Lancaster LA1 4AP United Kingdom*
[12]*Department of Ecology, Faculty of Science, Charles University, Viničná 7, 128 44 Praha 2, Czech Republic*
[13]*Lillehammer University College, P.O. Box 952, NO-2604 Lillehammer, Norway*
[14]*Department of Biology, University of Oslo, PB 1064 Blindern, 0316 Oslo, Norway*
[15]*Faculty of Biology and Environmental Protection, Nicolaus Copernicus University, Lwowska 1, 87-100 Toruń, Poland*

Abstract. The challenge of biodiversity upscaling, estimating the species richness of a large area from scattered local surveys within it, has attracted increasing interest in recent years, producing a wide range of competing approaches. Such methods, if successful, could have important applications to multi-scale biodiversity estimation and monitoring. Here we test 19 techniques using a high quality plant data set: the GB Countryside Survey 1999, detailed surveys of a stratified random sample of British landscapes. In addition to the full data set, a set of geographical and statistical subsets was created, allowing each method to be tested on multiple data sets with different characteristics. The predictions of the models were tested against the "true" species–area relationship for British plants, derived from contemporaneously surveyed national atlas data. This represents a far more ambitious test than is usually employed, requiring 5–10 orders of magnitude in upscaling. The methods differed greatly in their performance; while there are 2,326 focal plant taxa recorded in the focal region, up-scaled species richness estimates ranged from 62 to 11,593. Several models provided reasonably reliable results across the 16 test data sets: the Shen and He and the Ulrich and Ollik models provided the most robust estimates of total species richness, with the former generally providing estimates within 10% of the true value. The methods tested proved less accurate at estimating the shape of the species–area relationship (SAR) as a whole; the best single method was Hui's Occupancy Rank Curve approach, which erred on average by <20%. A hybrid method combining a total species richness estimate (from the Shen and He model) with a downscaling approach (the Šizling model) proved more accurate in predicting the SAR (mean relative error 15.5%) than any of the pure upscaling approaches tested. There remains substantial room for improvement in upscaling methods, but our results suggest that several existing methods have a high potential for practical application to estimating species richness at coarse spatial scales. The methods should greatly facilitate biodiversity estimation in poorly studied taxa and regions, and the monitoring of biodiversity change at multiple spatial scales.

Key words: biodiversity estimation; methods comparison; monitoring; spatial scale; species richness; species–area relationship; upscaling.

Manuscript received 15 September 2017; accepted 16 October 2017.

[16] Present address: Signal Innovations Group, 4721 Emperor Boulevard, Suite 3209 Treewood Lane, Apex, NC 27539 USA
[17] Present address: European Commission, Joint Research Centre, Ispra, 21027 VA Italy
[18] Present address: Inland Norway University of Applied Sciences, Elverum, Norway
[19] Present address: Department of Biosciences, University of Exeter, Exeter, EX4 4QD United Kingdom
[20] E-mail: w.e.kunin@leeds.ac.uk

Introduction

Biological diversity is intrinsically scale-dependent. While the issue of spatial scaling has only recently become prominent in many other areas of scientific research, the appreciation of scale issues in biodiversity research dates back to the foundations of the discipline. The most widely used tool for describing biodiversity scaling remains the species–area relationship (SAR), first devised more than a century ago (Watson 1835, Arrhenius 1921, Gleason 1922). The SAR represents species richness explicitly as a function of sample area, which is to say, as a function of spatial scale. The scale dependence of biodiversity as reflected in the SAR represents the combined effects of statistical sampling and ecological processes. As one examines communities across ever wider expanses, the number of species inevitably rises for a number of reasons: larger samples incorporate more individuals (allowing more species to be sampled), they encompass a wider range of habitats and environmental conditions, and bridge barriers to dispersal (Shmida and Wilson 1985, Drakare et al. 2006), The wide interest in SARs over many decades (e.g., Preston 1960, Connor and McCoy 1979, Rosenzweig 1995, He and Hubbell 2011, Scheiner et al. 2011, Storch 2016) testifies to the long-standing appreciation by ecologists of the centrality of scaling issues.

Classically, SARs have been drawn by conducting intensive biological surveys of different sized areas, which may be nested (e.g., a quadrat within a field, within a county, within a nation) or non-overlapping samples (e.g., a series of islands or political entities of different sizes), and may be ecological isolates (e.g., islands or discrete forest patches) or arbitrarily defined samples from a larger whole (e.g., quadrats or political entities); a great deal of discussion has focused on the properties of SARs composed in these different ways (e.g., Rosenzweig 1995, Scheiner 2003, Tjørve and Turner 2009, Scheiner et al. 2011). The shape of SARs has also been hotly contested, and after decades of debate about the relative merits of power law and logarithmic models (e.g., Connor and McCoy 1979), in recent years a wide range of other functional forms have been explored (reviewed by Tjørve 2003, 2009, see also Scheiner et al. 2011). More than 180 years after its birth, the SAR remains an active topic of ecological research.

The reason for the continued popularity of the SAR is obvious: it provides a clear language for expressing species-richness information across the full range of ecologically relevant scales. As such, it has great potential as a tool for describing and monitoring multi-scale aspects of biodiversity. Policy is often concerned with the preservation of biodiversity at national, continental (e.g., Gothenburg targets, EC 2001) or global (e.g., CBD, UNEP 2002) scales, whereas most biodiversity monitoring is conducted at very fine spatial scales (sometimes <1 m^2). This mismatch between the scales of our policies and of our data creates serious challenges, especially when assessing biodiversity change. It has recently become apparent, for example, that environmental changes may affect biotic diversity differently at different scales (Smart et al. 2006b, Keith et al. 2009, Keil et al. 2011); biotic homogenization, for example, may increase local (α) diversity while decreasing diversity at coarser (β and γ) scales (Socolar et al. 2016); conversely some invasive species may decrease α while increasing γ-scale richness (Rosenzweig 2001, Powell et al. 2013). SARs reflect biodiversity across a wide range of scales (incorporating α, β, γ and coarser scales) and so should provide an efficient tool for examining and communicating such complexities. Global biodiversity monitoring needs have further increased the interest in SARs and biodiversity scaling, due to the need to infer biodiversity patterns from growing global databases of point locations to the regional scale; that is, biodiversity upscaling. Coordinated local sampling schemes, together with reliable/robust upscaling methods, are critical for the integration and generalization of biodiversity information at large scales. Efficient tools for building reliable and accurate SARs may prove increasingly useful for predicting the response of biodiversity to environmental changes across scales, and to assess global conservation policy options (Pereira et al. 2013, Geijzendorffer et al. 2016).

However, one serious problem prevents the widespread application of SARs to multi-scale biodiversity monitoring. The requirement for exhaustive surveys over large areas makes it impractical to survey SARs repeatedly over a short period of time. Indeed, for many poorly studied taxa and regions, it would be difficult to amass sufficient information to provide even a single coarse-scale biodiversity estimate with confidence (e.g., Erwin 1982, May 1990). If the SAR is to fulfil its promise, we need to develop new approaches to parameterizing it with finite investments of surveying effort.

Harte and Kinzig (1997) were the first to explore a method for upscaling biodiversity from local samples. Their approach was based on the idea that the SAR should rise faster with area if dissimilarity in species occurrences in small plots (species turnover or β diversity) increases more rapidly with distance between plots (Harte et al. 1999, Krishnamani et al. 2004). Unfortunately the method involved strong implicit assumptions that limited its applicability. More recently, Harte and colleagues have proposed more sophisticated and general approaches based on the maximum entropy inferential method (Harte et al. 2008, 2009, Harte and Kitzes 2015). The past 15 years have seen a proliferation of other new methods to address this problem, based on approaches ranging from relative abundance distributions (Ulrich and Ollik 2005), species accumulation curves (Shen and He 2008), least distance spanning paths (Smith 2008), multi-site zeta diversity of compositional turnover (Hui and McGeoch 2014), and three-dimensional manifolds (Polce 2009). This sudden flowering of alternative approaches brings with it a new challenge: how do we best choose a method for a particular application? Many of the models have been tested against data, of course, but each against a different data set, and in many cases the tests have been relatively modest: attempting to up-scale by only one or two orders of magnitude or even less. This paper addresses this issue by testing a wide range of biodiversity upscaling approaches on a single high quality data set across a substantial range of scales, within a well studied system. By working in an area with a “known” SAR, we can judge the effectiveness of the various methods in estimating coarse-scale biodiversity.

Methods

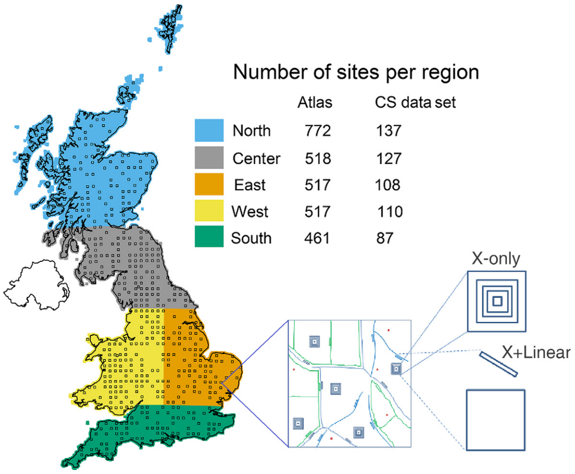
The CS data set

We make use of the GB Countryside Survey (CS), a periodic botanical survey program organized by the NERC Centre for Ecology and Hydrology (CEH). The CS focuses on a stratified random sample of 1-km cells within Britain, chosen to represent the full range of British landscapes (for further details on CS methods, see Firbank et al. [2003]). Specifically, we will rely on the CS survey of 1998–1999 (hereafter “CS1999”), which coincides with the survey period for the *New Atlas of the British and Irish Flora* (Preston et al. 2002), which we can use to generate our “true” SAR (see *Estimating the “True SAR”*). A total of 569 1-km^2 cells were examined in CS1999, scattered over the whole of Britain and its inshore islands (but excluding Northern Ireland and more distant island groups). Within each 1-km cell, a wide range of surveys was conducted, which can be roughly divided into areal surveys (various sized surveys of habitat blocks) and linear surveys (1 × 10 m surveys of linear features such as roadsides, hedgerows, and banks of waterways). For our purposes, the most statistically “representative” surveys were the so-called “X” plots, five of which are sited at random (one in each of five equally sized subsections) within each surveyed 1-km cell. The only departure from truly random placement is that X plots were not allowed to overlap with linear features (but see below). X plots have the added advantage (for this work) in being multi-scaled: each consists of a nested series of quadrats at 4-, 25-, 50-, 100-, and 200-m^2 scales. Species presence/absence is measured at all five scales, and estimates of cover for each species are recorded at the finest (2×2 m = 4 m^2) and coarsest (14.14×14.14 m = 200 m^2) scales. We made data from all five scales available to researchers (in most cases, the authors of upscaling methods), although most used only the coarsest scale (200 m^2) data in fitting their models.

The fact that X plots were not allowed to overlap linear features arguably makes them less diverse in species composition than truly random quadrats would be, as the inclusion of (potentially dissimilar) vegetation from such strips would likely enhance diversity (Smart et al. 2006a). Consequently, we developed a synthetic second set of samples, which we termed “X + Linear” samples (for clarity, the original surveys are hereafter referred to as “X-only” samples). These composite samples were created by choosing the linear feature closest in space to each X plot, and merging its species with those in the coarsest (200 m^2) X plot sample to produce an aggregate sample representing 210 m^2 (see Fig. 1). Where the same linear sample was the nearest neighbor of more than one X plot, it was assigned to the X plot in closest proximity, and others were paired with their second nearest linear surveys. If the X-only analyses arguably underestimate local richness, these X + Linear composite plots are likely to overestimate it, as they tacitly assume that all X plots would have included linear features had they been placed truly at random. We feel confident that a truly representative sample would fall somewhere between these two.

Subsamples

To provide a richer test of the various methods available, we developed a total of 16 test data sets. The largest of these is the “Full” sample, which covers all 569 CS survey cells within the surveyed area, and all five X plots within each. We also developed five regional subsamples, covering the “North,” “Center,” “East,” “West,” and “South” of Britain (Fig. 1). These were non-overlapping regions, chosen to roughly correspond to natural divisions of the area, and as such they were not equal in area. More importantly, they were also not equal in biodiversity, with pronounced regional differences in both α and β diversity between regions (encompassing, e.g., a more than twofold range in mean species richness at the 100-km^2 scale, c.f. Lennon et al. [2001]). We also developed two sets of five statistical subsamples from the full data set. “Wide-shallow” (WS) samples covered



Number of sites per region

		Atlas	CS data set
🟦	North	772	137
⬜	Center	518	127
🟧	East	517	108
🟨	West	517	110
🟩	South	461	87

Fig. 1. The location of GB Countryside Survey (CS) survey sites and Atlas cells, and of the regional subsets used in the analyses. The number of samples in each region are indicated in the legend. A hypothetical 1 × 1 km focal landscape is shown at higher magnification on the right, containing X-plots and Linear samples (not to scale), and the nature of (multi-scaled) “X-only” and (composite) “X + Linear” samples is displayed.

the full set of sampling locations, but included only one X plot (or X + Linear sample) of the five generally available at each site. By contrast, "narrow-deep" (ND) samples included all five X plots at each site, but included only one-fifth of the survey sites, chosen as a stratified random sample following the original CEH landscape stratification. Both WS and ND sample sets were non-overlapping, so that the sum of all five subsamples in either set constituted the Full British CS sample.

Each of the 16 samples (full set + 5 regions + 5 WS + 5 ND) were assessed for both X-only and X + Linear sample strategies, making a total of 32 potential tests for each method employed. However, the stratified nature of the statistical samples tended to make their multiple runs quite similar to each other, and thus treating them as five separate estimates would both overstate their independence and give them undue weight in the overall analysis. Consequently, to simplify reporting, each set of statistical subsamples (WS and ND) were summarized by a single (mean) performance score, thus leaving 16 tests (full set + 5 regions + WS mean + ND mean = 8, for each X-only and X + Linear data sets).

The challenge

The task we set ourselves was to estimate the SAR for scales ranging from 100 km^2 (10 × 10 km, the minimum mapping unit of Preston et al. 2002) to the whole of Britain (or of a specific subregion) using only the CS survey data. Even the finest of these scales was 500,000 times coarser than the 200-m^2 scale of an X-plot survey (or 476,190 times larger than the 210-m^2 scale of an X + Linear sample). For the purpose of this exercise, we will treat the area of Britain as the summed area of all the 100-km^2 cells covering Britain itself and the major outlying islands of the Shetland, Orkney, and Hebridean Islands, a total of 278,500 km^2. This is almost 14 billion times larger than scale of a single X plot, and approximately 500,000 times larger than the full set of survey sites combined (more precisely: 503,799 times the area of the full set of X plots, or 479,808 times the area of the full X + Linear sample). Levels of upscaling in statistical subsamples (with only one-fifth as many samples used) were five times greater still (2,518,995-fold for X-only analyses; 2,399,040-fold for X + Linear). The regional subsamples cover areas between 46,100 and 77,200 km^2, with correspondingly smaller numbers of samples, giving upscaling levels comparable to those for the full national data set. Several of the methods considered here have been tested before, in particular using tropical forest survey data from relatively small (e.g., 50 ha, Shen and He 2008) plots. Such applications involve only relatively modest upscaling; the challenge presented here is substantially more ambitious and more typical of the sort of tasks a practical upscaling approach would be asked to perform in, e.g., regional or national biodiversity estimation. To our knowledge, only a few past papers (Ugland et al. 2003, Krishnamani et al. 2004, Harte et al. 2009) have attempted comparable levels of upscaling, each for only a single model.

Upscaling methods

As noted in *Introduction*, there has been a proliferation of novel methods for upscaling biodiversity in recent years. We have brought together most of the global community of researchers addressing this issue, presenting each with the same CS data sets. To ensure high levels of familiarity with the models employed, most methods were fit by their original proponents, with the exception of the three variants of the Ugland model and the Lomolino model, which were prepared by a working group composed of E. Tjørve, A. Šizling, R. T. Jobe, K. I. Ugland, and W. Ulrich, and the power and logarithmic models, fit by V. Varma and W. E. Kunin. Further details of the models are given in the sections that follow.

Harte MaxEnt method

The maximum entropy theory of ecology (METE) predicts the shape of metrics describing patterns in the spatial distribution, abundance, and energetics of species (Harte et al. 2008, Harte 2011, Harte and Newman 2014). METE is a state variable theory in which the maximum entropy inference procedure (Jaynes 1982), coupled with constraints arising from knowledge of quantities such as the number of species and the number of individuals at plot scale, determine unique and testable macroecological metrics across all scales. METE predicts a non-power law but universal form for the SAR; in particular, if the local log-log slope of the SAR at each spatial scale is plotted against the average abundance per species at each scale, then all SARs are predicted to fall on a universal curve (Harte et al. 2009).

Upscaling species richness can either be carried out from knowledge of the number of species and the number of individuals at any one spatial scale, or alternatively from knowledge of the number of species at two spatial scales (from which information the abundance at each of those scales can be inferred from METE). The CS data set provides abundance information in terms of the percentage of cover, but not the number of individuals (which is hard to assess in many plant species). For that reason, we can upscale using the X-only plot data, which does include measured values of species richness at several plot-sized scales, but we cannot use the X + Linear plot data, as only one scale is available.

The capacity of METE to upscale has been tested successfully for tree species in the Western Ghats, where species richness was upscaled over a scale range of 24 million, from 0.25-ha plots where census data are available to the entire 60,000-km^2 biome (Harte et al. 2009). Other tests of upscaling with this method have been carried out for arthropods and trees in a Panamanian Preserve and trees in the Amazon (Harte and Kitzes 2015). An important limitation of the MaxEnt method, however, is that it is designed only for uspscaling species richness within contiguous blocks of similar habitat. Moreover, accumulating evidence (Harte 2011, Harte and Newman 2014), suggests that due to its reliance on equilibrial statistical outcomes METE's successes are restricted to relatively undisturbed ecosystems, with failures observed in habitats strongly influenced by human activity.

Ugland TS loglinear method

If METE is designed for uniform habitat, the Ugland et al. (2003) TS model was explicitly designed for surveys covering multiple potentially dissimilar communities. Most assemblages have a complex covariance structure between species and subareas. This leads to a largely unrecognized aspect of predicting the number of species by upscaling: with the addition of new subareas or habitats, the observed species accumulation curve (across regions or habitats) will not only extend the previous within-habitat accumulation curve, but also tend to lie above the accumulation curves for smaller subareas. The rate of (vertical) increase of the species-accumulation curves provides the best estimate of total species richness. Ugland et al. (2003) derived an exact analytical expression for the expectance and variance of the species accumulation curve in all random subsets from a given area. In this method, the whole area is divided into subareas, and an increasing sequence of accumulation curves is constructed as follows. The first accumulation curve (the bottom curve) is obtained by taking the average of all single subareas. The second accumulation curve is obtained by taking the average of all accumulation curves based on two randomly chosen subareas. For example, if there are five subareas, the total number of subsets of two subareas is the binomial coefficient $5 \times 4/2 \times 1 = 10$, so the second accumulation curve will be the average of 10 curves. In the same way, the third accumulation curve is the average of accumulation curves based on all possible subsets of three subareas. This procedure is repeated until we end up with the last accumulation curve, which is obtained by randomization of all available samples in the data set.

It is the terminal points of this increasing sequence of species accumulation curves that contain the crucial information of the accumulation rate of new species as sampling effort is increased to new subareas. The total species curve (the TS curve) is then defined as the curve connecting these end points. In a semilogarithmic plot, these curves frequently appear linear, and Ugland's estimator is then simply the linear extrapolation of the TS curve to the whole area in the semilog plot.

Ugland ten-at-a-time method

We also used a variant of the method presented in Ugland et al. (2003), where the mean number of species in a set of samples with the same number of plots is regressed with a semi-log function against the log of summed plot area. In this case, we used 10 groups of 10 plots, 20 plots, 30 plots, and so on, until the last group contained the entire set of plots (of which there is but one group). We examined groups of 50, 100, 150, and so on, plots, but the results were similar to the method using multiples of 10 plots at a time.

Ugland PAM method

A third method of applying the Ugland approach was pioneered by Jobe (2008), using the non-hierarchical clustering method algorithm known as partitioning around medoids (PAM) to determine the subclasses of sites for computing species accumulation curves. The original Ugland estimation method requires an a priori grouping of observations, so the introduction of PAM clustering allows such group assignments to be done on an objective basis in cases where no such classification is available. There are no hard and fast rules for selecting these groups, but the goal is for groups to contain ecologically distinct observations (e.g., communities, assemblages, etc.). PAM makes the grouping process more objective by using compositional similarity among sites as reflected in the clustering algorithm to select both the optimal number of groups and the membership of each group.

Shen and He method

There is a growing literature of methods devoted to estimating species richness in an area from random samples taken from within it (e.g., Palmer 1990, Chao 2005, Magnussen et al. 2006), often using resampling techniques with replacement. While these methods are not designed to estimate the full SAR, they can be used to upscale from a set of point data to estimate the overall species richness of the area from which they were drawn, and thus to estimate at least one point (the top) of the SAR. Many of these methods, however, have been shown to overestimate richness (e.g., Xu et al. 2012). Shen and He (2008) developed a novel approach based on sampling without replacement, using information on presence/absence data on species incidence, based on a modified Beta distribution. The method is not spatially explicit, and provides a single estimate for the species richness of the full sampled area. To derive finer scale estimates, the area to be estimated was shifted downwards (but see Discussion). In the X-only data sets, the Shen and He model was fitted both to data from the full 200-m^2 survey plots, but also to the finest scale (4-m^2) survey data, allowing the model's sensitivity to sample plot size to be assessed.

Šizling method

Arnošt Šizling and David Storch (Appendix S1) have developed a method using the frequency distribution of species' occupancies to estimate the shape of the SAR between two fixed scales, based on their “finite area model” of the SAR (Šizling and Storch 2004); different species-occupancy distributions produce SARs with different degrees of curvature, with the standard deviation of occupancy playing a key role (see Appendix S1). This approach is a “scaling between” method, rather than an upscaling method per se; that is, it estimates the increase in species richness as one moves from a unit survey plot (here a 200- or 210-m^2 CS sample) up to a predetermined maximum value. Thus it requires an estimate of “known” global species richness for the area in question and information from local samples to estimate species richness at scales in between these two known points on the curve. It would have been unfair to provide this model with more information than its competitors, and so the modeler had to make an arbitrary global richness estimate (1,000) to implement his model; but in practice, the method might best be combined with other methods that make effective global richness estimates in order to estimate the SAR as a whole (see Discussion). The method is based on the fact that if we assume aggregated distributions, the proportional occupancy constrains the size of the maximum gap in a species’ distribution (the “area of saturation”; Šizling and Storch 2004), which in turn determines the number of species sampled within given size window, i.e., in a specific area. As that and occupancy of the unit area together determine the slope of log-SAR (z), one could compose the SAR for any given number of species randomly chosen from the observed frequency distribution of occupancies, and thus estimate species richness of any area between the unit and total areas.

Hui models

Cang Hui developed three additional new approaches for this paper; each will be described briefly here, with full details and computer codes given in Appendix S2.

Hui 1: Occupancy rank curve.—This approach proportionally scales up a sampling occupancy rank curve (ORC) by assuming that the sampling is sufficient and representative of the wider area from which the samples were drawn. Specifically, if one plots the number of sites occupied by species in order of ubiquity, the resulting ORC for samples closely follows a truncated power law (Hui 2012), $O = c_1 e^{c_2 \cdot R} R^{c_3}$, where O and R represent the occupancy and the ranking of a species. This shape consists of two components: a power law function depicting the scale-free relationship between species ranks and their occupancies, and an exponential cut-off depicting a Poisson random process of species occupancy. The power law component is largely applicable to widespread/common species, with their distributions reflecting the spatial partitioning (or sharing) of heterogeneous, often approximately fractal, habitat, while the exponential cut-off reflects the chance events of flickering presence/absence of rare species. This method then scales up the sampled ORC to estimate the true ORC proportionally according to the sampling effort (replacing c_1 from the sampling ORC with $C_1 = c_1/s$, where $0 \leq s \leq 1$ represents sampling effort) and the maximum ranking for the enlarged ORC (i.e., solving $1 = C_1 e^{c_2 \cdot R} R^{c_3}$ for R) then represents the true number of species in the community.

Hui 2: Hypergeometric discovery curve (HDC).—Sampling patterns do not necessarily follow the same shape as the true biodiversity patterns, because the probability of discovering a species in a sample does not correlate linearly with the species’ true occupancy: the probability of encountering very rare species in a moderately sized sample is near zero, with probability rising with occupancy in a sigmoid fashion and approaching an asymptote near 1 for very common species. The sampling theory of species abundances has been extensively studied (Dewdney 1998, Green and Plotkin 2007), and Hui has developed an equivalent sampling theory of species occupancies, together with its continuous approximation for random sampling (Appendix S2). In particular, we need the sampling probability ($\mathrm{prob}(i|j)$) of discovering a species in i samples given a specific true occupancy of j. For random sampling without replacement, this follows a hypergeometric distribution. Importantly, sampling can affect the shape of observed occupancy frequency distribution (OFD), $f(i) = \sum_{j=1}^{m} \mathrm{prob}(i|j)F(j)$, where f is observed OFD, F true albeit unknown OFD, and m the sample extent divided by the grain. This formulation follows the discrete Fredholm equation (also Volterral integral equation) of the first kind (Arfken 1985), with $\mathrm{prob}(i|j)$ the kernel function and F a solvable positive vector. Despite the diverse parametric forms of OFDs (Hui and McGeoch 2007), we reduce the computational demand for parameter optimization by using a lognormal distribution $(F(j) = S \cdot \mathrm{LN}(j|\mu', \sigma'))$ centered at the middle of the possible logarithmic occupancy ($\mu' = \ln(m)/2$) such that its 95% confidence interval encompasses the entire range of occupancy at logarithmic scale ($\sigma' = \ln(m)/3.92$), making species richness the sole variable to be estimated from the parameter optimization.

Hui 3: Zeta diversity.—Zeta diversity represents the overlap in species across multiple samples (Hui and McGeoch 2014). Unlike pairwise beta diversity, which lacks the ability to express the full set of diversity partitions among multiple (three or more) samples, zeta diversity can express and potentially explain the full spectrum of compositional turnover and similarity (Latombe et al. 2017), with power law and negative exponential the most common forms of zeta diversity declines (with increasing number of included samples). We use a truncated power law to ensure a good fit to zeta diversity decline and then estimate the number of new species that are expected to occur when adding extra samples (i.e., the level of completeness) based on fitted zeta diversity decline. The expected number of species in an area can then be estimated according to the generic estimator developed in Hui and McGeoch (2014); note, the Chao II estimator is only a special case for exponentially declining zeta diversity. As the formulation is based on combinatorial probabilities, to reduce the overflow error (a combination of floating-point inaccuracy in any numerical computation platforms and combinatorial explosion [of formulation complexity] with increasing number of samples), we first estimate the number of new species encountered when adding one extra sample and then calculate the expected number of species using integral approximation.

Ulrich and Ollik method

Ulrich and Ollik (2005) made use of a different method based on Relative Abundance Distributions (RADs), which was originally designed to estimate the upper and lower limits of species richness in a focal region. Under the assumption that the occupancy–species-rank-order distribution is either a lognormal or a logseries and that the least abundant species has an occupancy of one cell (200 m^2), they estimated upper species richness boundaries from the logseries by

$$E_{\mathrm{S}} = \frac{\ln(Int) + \ln N_{\mathrm{A1}} - \ln N_{\mathrm{S1}}}{slope} \tag{1}$$

and lower species richness boundaries from the lognormal distribution by

$$E_{\mathrm{S}} = \frac{2\ln(Int) + \ln N_{\mathrm{A1}} - 2\ln N_{\mathrm{S1}}}{slope} \tag{2}$$

where ln(Int) and ln(slope) are natural logarithm of the intercept (Int) and the slope of an exponential regression through the middle 50th percentile of the respective abundance distributions and ln N_{S1} and $\ln N_{\mathrm{A1}}$ are the natural logarithms of the numbers of individuals of the most abundant species of the whole community within the area A_{total} and of the sample of area A_1, respectively. N_{A1} comes from proportional upscaling of the sample area to total area: $N_{\mathrm{A1}} = N_{\mathrm{S1}}A_{\mathrm{total}}/A_1$.

Smith method

A species–distance relationship (SDR) was explored by Smith (2008) as a method for estimating the SAR from point survey data. The SDR slope was found to be highly correlated with the slope of the SAR for the U.S. Breeding Bird Survey data at large geographic scales. The SDR is calculated by estimating the path of shortest length connecting a set of localities, then estimating cumulative distance and cumulative diversity along the path. In the present analysis, data for all X or X + Linear plots were lumped within a given 1-km^2 sampling cell (except for the wide-shallow subsamples, as these only contained one X plot per cell). This is because locality size per se was found not to have a significant influence on the slope of the SDR, whereas sample size (which affects number of individuals surveyed) per locality did.

SDRs were calculated for all subsets of the Countryside Survey data using 1-km^2 cells as localities. No correction was made for sample size. Distance was calculated as Cartesian distance between the midpoints of the cells. Mean slopes of the SDR are based on 200 values (100 paths, each containing 10 cells and measured in forward and reverse directions). To estimate the slope of the SDR, linear regression and standardized major-axis regression were performed. Setting then the slope of the SDR to equal the slope of the SAR, diversity estimates were made for the relevant portions of Britain by assuming two different values for alpha diversity. First, average alpha diversity was calculated for the plots (200 m^2 or 210 m^2 for X and X + Linear plots, respectively). Second, average alpha diversity per cell (1 km^2) was calculated by combining all plots in a sampling cell; this will underestimate diversity for a 1-km^2 area.

Polce and Kunin method

The SAR rises for two reasons (see, e.g., Scheiner et al. 2011): a larger area both encompasses more environmental and spatial diversity than a small area and it includes more total individuals (and thus constitutes a larger sample). These two component processes, increased sample size and increased spatial differentiation, may be expected to behave rather differently with increasing area. In order to factor out these two component processes, we randomly sampled (1) different numbers of quadrat surveys from constant sized “windows” of focal area (to estimate the pure sample size effect), and (2) constant numbers of quadrat samples chosen from different sized windows (to estimate the pure spatial scale effect), and tested the fit of a range of convex and sigmoid curves (from Tjørve 2003) to each component process. Note that in these analyses, total sample size for a set of quadrats is expressed in units of area (total m^2 surveyed), as that is essential for later steps of the analysis. We then constructed a three-dimensional manifold model as a multiplicative combination of the best-fitting sample-size and scale models (see Polce 2009). Pilot work suggested that the MMF model [$Y = (a \times \mathrm{Samplesize}^c)/(b + \mathrm{Samplesize}^c)$] provided the best fit to the pure sample size component (sampled within a fixed window size), whereas a power law ($Y = d \times \mathrm{Scale}^z$) performed best for pure spatial differences (at constant sample size). These two component models could then be combined multiplicatively, to derive a final model

$$Y = (a \times \mathrm{Scale}^z \times \mathrm{Samplesize}^c)/(b + \mathrm{Samplesize}^c) \tag{3}$$

Fitting this three-dimensional model to the data set, the SAR can be estimated as the value of Y over the diagonal line where Samplesize = Scale.

Lomolino model

We also fit a suite of models commonly fit to SARs and to the plot-based species-accumulation curve (SAC) from each data set (see Tjørve [2003] for models). Preliminary results here indicated that in most cases the "Lomolino" model (Lomolino 2001) worked best ($S = a/(1 + b^{\log 10(c/A)})$), where S is number of species, A is area, and a, b, and c are model parameters fit using the Gauss-Newton method for non-linear regression (Myers 1990). In most cases, the AIC weight of the Lomolino model was ~1, and where it was not, it was equally tied with other models that were nested within the Lomolino model. Therefore, we used only the Lomolino model to fit each data set.

Power law and logarithmic models

To complement the range of recently derived methods, we have included a few "old-fashioned" approaches to SAR estimation. Arrhenius (1921) proposed a power law ($S = cA^z$) as the best descriptor of the SAR, and Preston (1962) suggested that the "canonical" SAR would have an exponent (z) of 0.25. Subsequent work (e.g., Connor and McCoy 1979, Rosenzweig 1995) has suggested somewhat less steep z values predominate in many continental systems, with a consensus z of approximately 0.2. Thus, we generated SAR estimates by simply computing mean species richness at the 200-m^2 scale X plot samples (and 210-m^2 for the X + Linear samples) and scaling up to coarser resolutions using power law curves with these two slopes. We also took advantage of the multi-scaled nature of the CS X plot surveys, fitting both power and semi-logarithmic (after Gleason 1922) models to the observed species richness of each plot at the five scales of measurement (4, 25, 50, 100, and 200 m^2), and extrapolating median estimates for each. As the X + Linear data are available only at a single scale, these extrapolations of power law and semi-logarithmic curves can be done only on the X-only data sets.

Model summary

Altogether, we have assembled 13 different models for upscaling biodiversity, and several of them (the power law, Shen and He, Ugland's TS and Ulrich and Ollik's methods) have been implemented in multiple forms, for a total of 19 sets of predictions. These methods may be grouped conceptually, based on the approaches they take to the challenge of estimating coarse-scale species richness from fine-scale samples (Fig. 2). Three of the methods (power law, logarithmic, and Lomolino) involve parameterizing and extrapolating a well-studied SAR curve from the observed data. This is an entirely phenomenological approach to upscaling. Two other models (Harte's MaxEnt model and Hui's HDC) also extrapolate functions, but with curves that are built on a strong underlying rationale concerning the patterns expected from random community patterns under constraints. Three models are based on sampling processes from species occupancy (Hui ORC, Šizling) or abundance (Ulrich and Ollik) distributions. Two additional models (Shen and He, Ugland's TS) focus specifically on sampling processes and the resulting accumulation of species. The Polce and Kunin model is similar to Ugland's sampling process approach, but with an explicit emphasis on spatial turnover processes. Such spatial turnover in species is central to Hui's Zeta model, and plays a substantial role in the Smith model as well, which in turn links back to phenomenological curve estimation approaches.

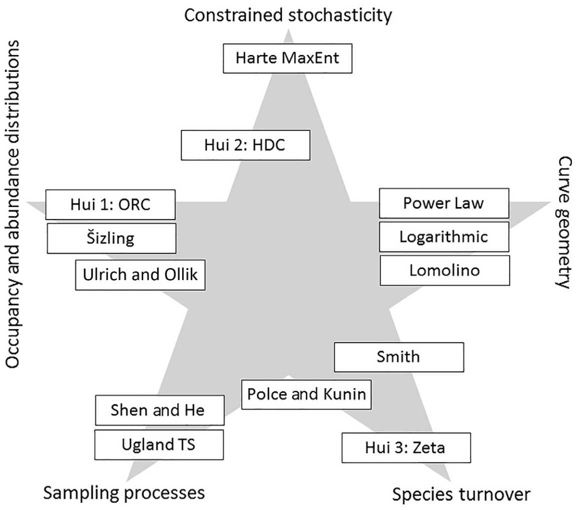


FIG. 2. Conceptual groupings of the methods employed. See *Methods* for further explanation.

Estimating the "True SAR"

The quality of the various SAR predictions can only be tested by comparing them to the "true" SAR for the focal region. This was estimated using data from the *New Atlas of the British and Irish Flora* (Preston et al. 2002; hereafter NABIF), which was compiled based on surveys from the late 1990s, thus approximately at the same time as the CS 1999 sample. In contrast to an earlier attempt at a UK floral atlas (Perring and Walters 1962), the NABIF's compilers made a concerted effort to ensure a relatively even survey effort across the area in a fairly narrow time window, and in particular to avoid false negatives due to the underreporting of common species and the false positives that result from the compilation of records over long periods of time. While no biodiversity survey can be treated as perfect, the NABIF is arguably one of the highest quality biodiversity atlases currently available anywhere. In addition to vascular plants, the CS survey included a predefined set of 160 relatively common and distinctive bryophyte and lichen taxa (species or species groups); consequently distribution maps for these taxa were acquired from the bryophyte and lichen recording schemes, respectively (M. O. Hill, *personal communication*; J. Simkin, *personal* *communication*). The true SAR was composed by superimposing a series of coarser grids (with resolutions from 400 km^2 to 90,000 km^2) over the distributional data set. Only grid cells containing >75% land area were included in our analyses for each scale; at coarse scales, grid cells were shifted somewhat (following Tjørve et al. 2008, Keil et al. 2011) to maximize the area fitting this criterion. Our NABIF SAR calculations are being posted online (Polce and Kunin 2017).

Assessing model performance

To assess the quality of the predictions of each model, we examined two quality criteria, appropriate to somewhat different applications. One goal of diversity upscaling is to estimate the Total Species Richness (TSR) in a focal region, while for other applications, it is valuable to estimate species richness across a range of scales within the region, providing an estimate of the region's species–area relationship (SAR). We assessed model predictions against both of these criteria: SAR and TSR fits.

To assess the quality of SAR fits, we examined the mean absolute value of the difference between predicted and true species richness values at a given scale, expressed relative to the true richness value at that scale, which we term the "mean relative error" (or MRE)

$$\mathrm{MRE} = \left(\frac{1}{n}\right) \sum_{i} \left(\frac{|S_{\mathrm{predicted},i} - S_{\mathrm{true},i}|}{S_{\mathrm{true},i}}\right) \qquad (4)$$

where $S_{\mathrm{predicted},i}$ is the number of species predicted at scale i, $S_{\mathrm{true},i}$ is the number observed at that scale in the true SAR, and the summation is across n observed scales (nine scales in the regional analyses, 10 in the full national and statistical subsample analyses). Note that we normalize errors by dividing them by the true SAR value at each scale, so that, e.g., a 100-species error is deemed to be a larger mistake when the true value is 100 than it is when the true value is 1,000. This has the additional advantage of allowing model fit to be expressed as a dimensionless fraction: the mean proportional error in estimation. We have also calculated model fits using a number of other popular metrics (e.g., RMSE, Pearson χ^2; see Data S1), but there is little qualitative effect on our findings; the same models perform well by any sensible measure, with at most slight rearrangements of the order of the winners.

The quality of Total Species Richness (TSR) predictions was assessed using this same metric, but evaluated only at the coarsest scale considered (278,500 km^2 in national analyses, and the area of each region in regional analyses). In addition, we examined the correlation between true TSR and estimated values across data sets, using the nonparametric Spearman's rank correlation, to test how consistently high richness estimates were provided in highly species-rich regions. A similar correlation test was performed for the full SAR fit, comparing the overall slopes of the estimated SARs (on logarithmic axes) over the range of scales examined (100–278,500 km^2) with the slopes of the true SARs over those scales.

Results

The models tested differed greatly in their predictions for British plant richness; while the true TSR value was 2,326, the model estimates based on the X-only data set ranged from only 62 (median semi-logarithmic curve extrapolation) up to 11,593 (Smith model) species. A somewhat narrower range of predictions for the X + Linear data set (1,136 to 8,647) was largely due to the fact that some of the more extreme value models could not be applied to this data set (e.g., the fitted semi-logarithmic and power law models, which needed multiple scales of diversity surveys). Examples of the true and estimated SARs for the full British data sets are shown in Fig. 3 (full data are provided in Data S1).

Fit scores for Total Species Richness predictions are given in Fig. 4. Three families of models stand out as the most reliable predictors of TSR: the two applications of Shen and He's method (2008; hereafter S&H), the paired upper and lower estimates of Ulrich and Ollik (2005; hereafter U&O), and the Hui ORC models. The best predictive accuracy came from the S&H model, with estimates generally within 10% of the correct TSR value (mean relative error = 0.097 ± 0.085) when parameterized with 200-m^2 (or 210 for X + Linear samples) data; interestingly, the model performed almost as well (mean relative error = 0.110 ± 0.091) when parameterized from much smaller (4-m^2) vegetation samples. The U&O method and Hui's ORC model were the next best approaches: the upper (log-series) U&O model had a mean relative error of 0.155 (±0.083), whereas the lower (log-normal) U&O model had a mean relative error of 0.211 (±0.080). While these two methods are meant to serve as upper and lower estimates, even the upper estimate was usually less than the true TSR. Hui's ORC model performed nearly as well as the best U&O model in accuracy (mean relative error = 0.156 ± 0.089). The Ugland model, applied using the 10-at-a-time algorithm, performed reasonably well (MRE = 0.210 ± 0.162), as did Hui's HDC model (MRE = 0.272 ± 0.173); no other approach came close (the next best was the Polce & Kunin [P&K] model, MRE = 0.375 ± 0.158). Judging by the (Spearman's rank) correlation coefficients between true and predicted species richness across sample sets, a similar picture emerges, with the S&H methods (ρ = 0.825 and 0.805, when parameterized with 200- and 4-m^2 data, respectively) and the Hui HDC, Zeta, and ORC models (ρ = 0.800, 0.752, and 0.697, respectively) showing the highest correlation with true TSR, along with the Ugland (in particular, the 10-at-a-time version with ρ = 0.788), P&K (ρ = 0.728), and U&O (both ρ = 0.655) models.

The full SAR fits of the models are given in Fig 5. Accuracy was not as good as for SDR overall, but one of Hui's models is the clear favorite in predicting the curve as a whole: the Hui ORC model was well within 20% of

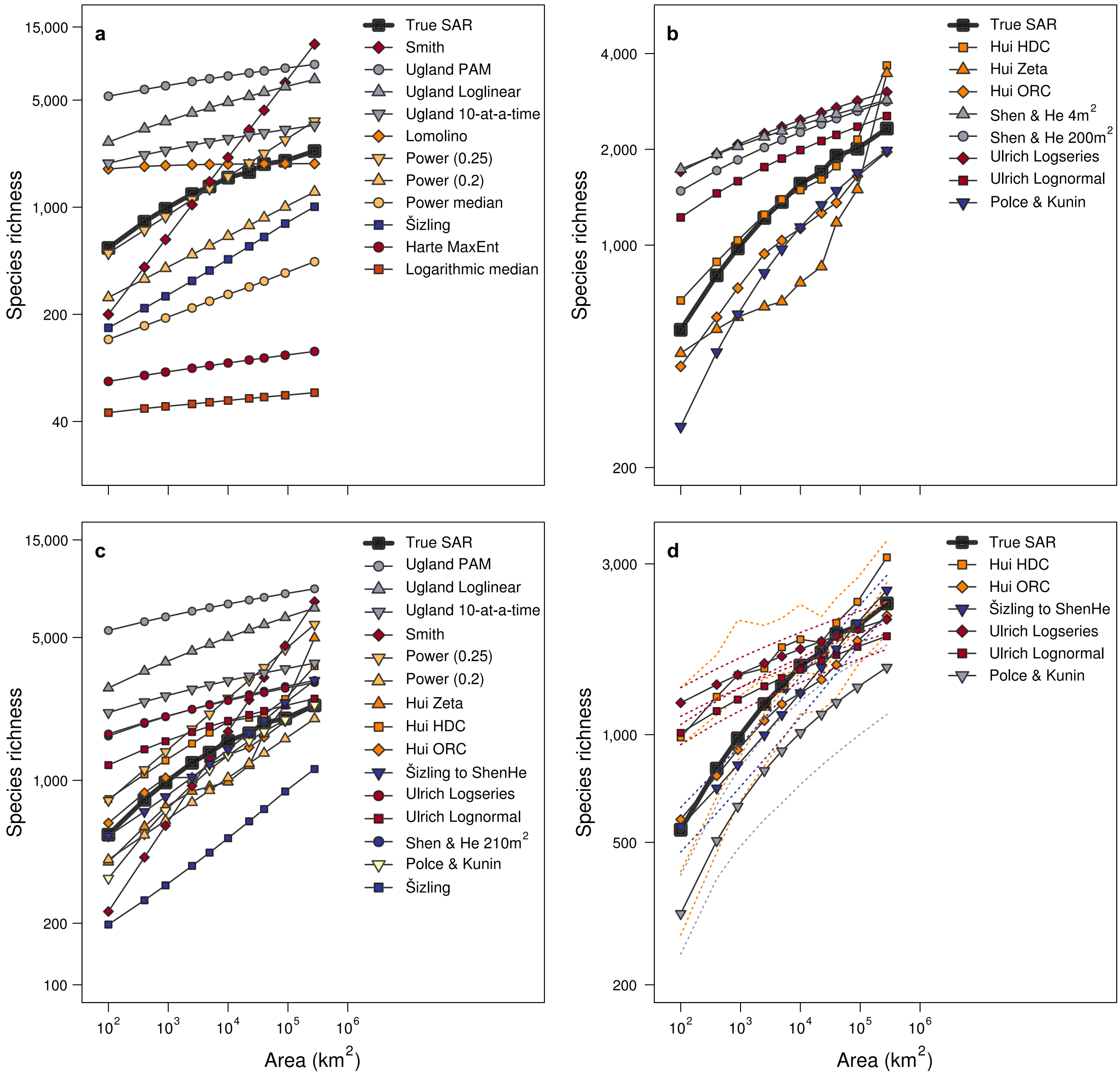


FIG. 3. Model predictions for the full UK data set, based on (a, b) X-only samples, (c) X + Linear samples, and (d) randomized subsets. For clarity, a subset of the best-fitting models are plotted in panel b, with an expanded *y*-axis. Note that several models (MaxEnt and fitted versions of Power and Logarithmic models) could not be estimated on X + Linear samples (see text and Figs. 4, 5). Plots in (d) represent means of X-only and X + Linear data from both wide-shallow (WS) and narrow-deep (ND) samples. Error distributions around each curve (with matching line color) represent trimmed ranges: the central 18 of the 20 data points (roughly corresponding to 90% confidence intervals). The true SAR is indicated by heavy lines in each panel, for clarity.

correct SAR values on average (MRE = 0.177 ± 0.059). The lower (log-normal) U&O model performed reasonably well (MRE = 0.272 ± 0.094), as did the Hui HDC model (MRE = 0.304 ± 0.202). The upper (log-series) U&O approach and the P&K method competed for fifth place (P&K, MRE = 0.358 ± 0.118; U&O2, MRE = 0.369 ± 0.217). The only other models that averaged within 50% of the correct SAR were the Hui Zeta model (MRE = 0.408 ± 0.134), the S&H model (MRE = 0.418 ± 0.212), the Lomolino model (MRE = 0.442 ± 0.110), and the power law model with z = 0.2 (MRE = 0.451 ± 0.179) or z = 0.25 (MRE = 0.496 ± 0.444). As noted above, several other models were tested only on X-only data, but none of them performed well enough to challenge the leading methods. The slopes of the estimated SARs were generally uncorrelated with the true SAR slopes over the scales considered here; only the median logarithmic model showed a significant positive correlation (ρ = 0.756, n = 8, P = 0.015).

Sometimes consensus models can be constructed that perform more reliably than any one approach by itself, especially when different models have contrasting weaknesses (e.g., Gritti et al. 2013). The P&K and U&O methods tended to make contrasting errors, with the P&K

		Model:	Harte MaxEnt	Hui HDC	Hui ORC	Hui Zeta	Logarithmic Median	Lomolino	Polce & Kunin	Power 0.2	Power 0.25	Power median	Shen He 4 m^2	Shen He 200/210 m^2	Šizling	Smith	Ugland: Loglinear	Ugland: 10-at-a-time	Ugland: PAM	Ulrich Ollik lognormal	Ulrich Ollik logseries	SH+UO1 Mix	SH+UO2 Mix	OU1+OU2
X-only		Full British	0.951	0.578	0.156	0.490	0.974	0.172	0.148	0.459	0.566	0.810	0.228	0.216	0.567	3.984	1.934	0.471	2.673	0.094	0.301	0.155	0.258	0.198
	Subset	Wide-shallow	0.952	0.282	0.145	0.402	0.973	0.400	0.335	0.454	0.566	0.811	0.049	0.069	0.714	13.533	1.987	0.195	1.629	0.185	0.090	0.058	0.010	0.138
		Narrow-deep	0.951	0.242	0.143	1.144	0.973	0.508	0.425	0.453	0.566	0.808	0.023	0.003	0.735	3.813	1.578	0.111	1.424	0.179	0.084	0.091	0.044	0.132
	Regional	South	0.955	0.282	0.219	0.035	0.976	0.558	0.580	0.617	0.003	0.862	0.051	0.051	0.738	0.848	1.365	0.067	1.391	0.258	0.159	0.155	0.105	0.209
		East	†	0.014	0.135	0.357	0.980	0.571	0.493	0.758	0.363	0.856	0.245	0.148	0.772	0.191	1.330	0.055	1.309	0.327	0.215	0.238	0.182	0.271
		West	0.951	0.033	0.215	0.488	0.974	0.519	0.478	0.560	0.158	0.857	0.166	0.132	0.740	0.673	1.161	0.004	1.282	0.302	0.228	0.217	0.180	0.265
		Centre	0.946	0.026	0.322	0.527	0.971	0.282	0.307	0.511	0.289	0.830	0.091	0.029	0.699	0.313	1.614	0.130	1.634	0.189	0.091	0.109	0.060	0.140
		North	0.916	0.214	0.239	0.389	0.951	0.511	0.551	0.200	1.149	0.704	0.026	0.061	0.686	2.080	0.947	0.109	1.595	0.147	0.052	0.104	0.057	0.099
X + Linear		Full British		0.556	0.015	1.138		0.060	0.004	0.140	1.490			0.325	0.512	2.219	1.992	0.608	2.717	0.079	0.294	0.202	0.309	0.186
	Subset	Wide-shallow		0.436	0.006	0.832		0.317	0.240	0.139	1.490			0.178	0.666	13.767	2.154	0.368	1.823	0.182	0.088	0.002	0.045	0.135
		Narrow-deep		0.417	0.009	0.761		0.437	0.350	0.139	1.490			0.111	0.682	3.673	1.704	0.276	1.732	0.214	0.127	0.051	0.008	0.171
	Regional	South		0.381	0.161	0.424		0.532	0.558	0.369	0.668			0.018	0.688	0.698	1.243	0.182	1.776	0.255	0.157	0.118	0.070	0.206
		East		0.166	0.252	0.500		0.452	0.340	0.540	0.222			0.051	0.720	0.168	1.628	0.127	0.781	0.333	0.219	0.192	0.135	0.276
		West		0.144	0.107	0.386		0.439	0.436	0.331	0.778			0.037	0.692	0.862	1.271	0.128	0.651	0.308	0.231	0.173	0.134	0.269
		Centre		0.287	0.194	0.046		0.194	0.274	0.235	1.035			0.091	0.633	0.540	1.806	0.305	0.900	0.204	0.098	0.056	0.003	0.151
		North		0.300	0.174	0.134		0.439	0.485	0.192	2.234			0.041	0.637	3.574	0.967	0.225	1.898	0.125	0.042	0.042	0.0004	0.084
Overall: Mean (SD)			0.972 (0.145)	0.272 (0.173)	0.156 (0.089)	0.503 (0.328)	0.972 (0.009)	0.400 (0.152)	0.375 (0.158)	0.381 (0.192)	0.817 (0.615)	0.817 (0.050)	0.110 (0.091)	0.097 (0.085)	0.680 (0.067)	3.183 (4.315)	1.543 (0.375)	0.210 (0.162)	1.576 (0.572)	0.211 (0.080)	0.155 (0.083)	0.122 (0.070)	0.100 (0.093)	0.183 (0.063)
Rank correl.			0.074	0.800	0.697	0.752	0.146	0.576	0.728	0.121	0.261	0.122	0.805	0.825	0.600	0.661	0.764	0.788	0.679	0.655	0.655	0.782	0.764	0.655

FIG. 4. Compilation of total species richness fits of the various upscaling models tested. Values represent proportional absolute errors [$|S_{\text{predicted}} - S_{\text{true}}|/S_{\text{true}}$], with underscored numbers indicating the best (solid line) and second-best (dotted line) fitting model for a particular data set. Combined models are underscored relative to the set of individual models. Shading represents fit, with cutoff values 0.05 (no shading), 0.1, 0.25, 0.5, and 1 (darkest). Rank correlation coefficients (Spearman's ρ) for the relationship between true and estimated richness are listed in the final row. The † stands for indicates a case where the model would not converge on a solution.

		Model:	Harte MaxEnt	Hui HDC	Hui ORC	Hui Zeta	Logarithmic Median	Lomolino	Polce & Kunin	Power 0.2	Power 0.25	Power median	Shen He 4 m^2	Shen He 200/210 m^2	Šizling	Smith	Ugland: Loglinear	Ugland: 10-at-a-time	Ugland: PAM	Ulrich Ollik lognormal	Ulrich Ollik logseries	PK+UO1 Mix	U1+U2 mean	Šizling to SH
X-only		Full British	1.296	0.125	0.238	0.400	0.956	0.601	0.294	0.553	0.168	0.809	0.808	0.669	0.683	1.042	2.446	1.094	4.470	0.446	0.848	0.088	0.647	0.168
	Statistical	Wide-shallow	0.927	0.100	0.137	0.473	0.957	0.405	0.335	0.548	0.168	0.807	0.548	0.473	0.765	3.357	2.503	0.725	2.995	0.229	0.336	0.159	0.275	0.137
		Narrow-deep	0.926	0.335	0.218	0.580	0.957	0.381	0.466	0.548	0.168	0.808	0.449	0.377	0.779	0.998	2.063	0.608	2.686	0.232	0.341	0.198	0.281	0.169
	Regional	South	0.944	0.182	0.144	0.378	0.968	0.399	0.520	0.670	0.230	0.864	0.199	0.148	0.779	0.519	1.525	0.302	2.108	0.176	0.170	0.315	0.165	0.200
		East	†	0.249	0.180	0.380	0.972	0.400	0.555	0.786	0.500	0.861	0.216	0.171	0.810	0.580	1.550	0.237	2.072	0.231	0.200	0.367	0.205	0.289
		West	0.932	0.094	0.253	0.573	0.963	0.358	0.441	0.600	0.137	0.813	0.236	0.198	0.769	0.447	1.484	0.330	1.929	0.211	0.214	0.269	0.211	0.206
		Centre	0.922	0.150	0.284	0.642	0.954	0.429	0.244	0.529	0.108	0.837	0.351	0.337	0.721	0.379	2.166	0.600	2.784	0.237	0.328	0.161	0.276	0.102
		North	0.860	0.700	0.252	0.270	0.913	0.440	0.358	0.193	0.873	0.646	0.619	0.494	0.678	0.709	1.741	0.792	2.170	0.411	0.525	0.268	0.463	0.098
X + Linear		Full British		0.293	0.106	0.345		0.770	0.172	0.289	0.726			0.837	0.635	0.646	2.581	1.304	4.556	0.416	0.834	0.123	0.625	0.120
	Statistical	Wide-shallow		0.476	0.132	0.355		0.463	0.229	0.289	0.726			0.638	0.716	3.603	2.751	0.982	3.301	0.231	0.338	0.138	0.278	0.083
		Narrow-deep		0.447	0.128	0.393		0.392	0.383	0.289	0.726			0.547	0.726	0.990	2.267	0.854	3.160	0.225	0.308	0.183	0.257	0.095
	Regional	South		0.209	0.145	0.546		0.377	0.487	0.457	0.285			0.208	0.730	0.451	1.446	0.448	1.384	0.189	0.171	0.303	0.172	0.134
		East		0.130	0.219	0.406		0.337	0.400	0.594	0.144			0.206	0.760	0.526	1.897	0.443	1.396	0.236	0.200	0.294	0.207	0.189
		West		0.210	0.167	0.422		0.338	0.374	0.392	0.432			0.254	0.721	0.446	1.638	0.498	2.176	0.214	0.212	0.249	0.211	0.108
		Centre		0.450	0.103	0.224		0.502	0.175	0.263	0.728			0.506	0.654	0.352	2.434	0.854	3.266	0.230	0.320	0.164	0.267	0.059
		North		0.721	0.125	0.137		0.478	0.289	0.207	1.819			0.634	0.615	1.347	1.829	0.988	4.608	0.440	0.565	0.281	0.497	0.115
Overall: Mean (SD)			0.972 (0.145)	0.304 (0.202)	0.177 (0.059)	0.408 (0.134)	0.955 (0.018)	0.442 (0.110)	0.358 (0.118)	0.451 (0.179)	0.496 (0.444)	0.807 (0.069)	0.428 (0.219)	0.418 (0.212)	0.721 (0.056)	1.024 (1.000)	2.020 (0.438)	0.691 (0.309)	2.816 (1.043)	0.272 (0.094)	0.369 (0.217)	0.222 (0.081)	0.315 (0.155)	0.156 (0.062)
Slope correl.			-0.037	-0.576	-0.497	-0.164	0.756	0.261	-0.146	0	0	0.244	-0.195	-0.176	-0.115	-0.361	-0.036	-0.042	-0.194	-0.194	-0.006	-0.097	-0.152	-0.115

FIG. 5. Quality of SAR fit, as indicated by mean relative absolute error. Underscores indicate the best and second best models for each data set, as in Fig. 4. Shading is as in Fig. 4, to aid comparison. The final row lists Spearman's rank correlation coefficients between true and estimated SAR slopes across the different data sets tested.

model predicting a lower and steeper SAR than was found in many cases, while the U&O method predicted a higher and flatter SAR than that observed over the relevant range of scales, so that there was an inverse correlation between the performance of the two models (Pearson $r = -0.470$). Consequently, the mean of these two estimates often provided a better (and more reliable) SAR estimate than either model by itself (MRE = 0.222 ± 0.081). An even more successful combined SAR model could be constructed by using the S&H estimate of TSR and then downscaling to finer scales using the Šizling method (MRE = 0.156 ± 0.062), combining the strengths of both models. This combination provides our best SAR predictions.

The replicate runs of statistically subsampled data sets allow estimates of the variance in index values holding sample effort constant (at one-fifth of the total sample). Fig. 6 shows the coefficients of variation in these replicated analyses. Most models showed acceptable levels of variation in estimates, although the Smith (2008) model, Hui's Zeta model, and approaches based on median fits of classical SAR models (power law and semi-logarithmic) showed much higher variation than the others tested. For many of the models (most strikingly in the two Ulrich and Ollik models), variation between runs was substantially higher in the narrow-deep analyses than in the wide-shallow runs, presumably because the latter allowed higher levels of statistical independence between samples. For some of the models (most notably the Lomolino, Ugland PAM, and Ulrich and Ollik models) these statistical subsamples also tended to produce systematically lower upscaled biodiversity predictions than resulted from the full data set, even though each set of five (non-overlapping) subsamples comprised the full sample set, and all were being used to estimate the same full British SAR.

Discussion

The challenge of upscaling biodiversity from plot to regional or national scale is an important goal of spatial ecology, one with the potential for important practical value. If we could reliably estimate coarse-scale species richness from fine-scale samples, it would allow biodiversity estimation in poorly studied regions and taxa, and facilitate the monitoring of multi-scale biodiversity change and the scaling up of experimental results. A range of methods have been proposed to address this issue, but there has to date been no clear consensus as to their relative strengths and weaknesses. To test these methods, we set a much more ambitious test than has usually been applied, requiring species richness to be estimated at scales some 500,000 times larger than the full data set used and 14 billion times larger than a single sample plot (the scale of resolution from which richness was extrapolated by most of the methods). The models considered varied greatly in their performance in this test, but the best of them did well enough to suggest that they have the potential for useful application in the near term. Nonetheless, further tests of these methods should be attempted on data sets covering other taxa and regions, so that the generality of our conclusions can be ascertained. Many of the models (especially those with relatively inflexible shapes) may be expected to fit much better in some areas than in others; differences in species richness, evenness, habitat diversity and spatial patchiness may all affect the form of SARs (Tjørve et al. 2008), and thus may improve the relative success of some models over others. Similarly, different models may be differentially sensitive to differences in the structure and intensity of sampling (CS is perhaps a best-case scenario), which may again affect relative performance. Only by examining a wide range of data sets with differently diversity patterns can we be certain of the generality of our results.

		Model:	Harte MaxEnt	Hui HDC	Hui ORC	Hui Zeta	Logarithmic Median	Lomolino	Polce & Kunin	Power 0.2	Power 0.25	Power median	Shen He 4 m^2	Shen He 200/210	Sizling	Smith	Ugland: Loglinear	Ugland: 10-at-a-time	Ugland: PAM	Ulrich & Ollik lognormal	Ulrich & Ollik logseries
X-only	CV:	Wide-shallow	0.0133	0.1186	0.1716	0.4156	0.0661	0.0243	0.0522	0.0448	0.0448	0.1271	0.0134	0.0182	0.0477	0.1608	0.0221	0.0160	0.0611	0.0170	0.0073
		Narrow-deep	0.0266	0.0989	0.2063	0.3275	0.0419	0.1389	0.1260	0.0336	0.0336	0.1181	0.0774	0.0780	0.0655	0.1688	0.1081	0.0779	0.0730	0.0742	0.0767
	Ratio ND:WS		1.9938	0.8334	1.2024	0.7881	0.6345	5.7163	2.4123	0.7495	0.7495	0.9289	5.7947	4.2958	1.3751	1.0501	4.8800	4.8769	1.1941	4.3605	10.477
	Rel. to whole	Wide-shallow	0.9901	1.0000	1.1743	0.8662	1.0044	0.7346	0.9629	1.0098	1.0000	1.0073	0.8556	0.8821	0.7470	2.1438	1.0168	0.8216	0.7276	0.7753	0.7093
		Narrow-deep	1.0058	1.1977	1.0261	0.8918	1.0040	0.6041	0.7691	1.0099	1.0001	1.0000	0.7982	0.8237	0.7041	0.9692	0.8878	0.7655	0.6712	0.7817	0.7138
X+Linear	CV:	Wide-shallow		0.1350	0.0829	0.2541		0.0423	0.0443	0.0169	0.0169			0.0200	0.0150	0.4212	0.0199	0.0109	0.0412	0.0086	0.0037
		Narrow-deep		0.1352	0.1449	0.2969		0.0922	0.0973	0.0341	0.0341			0.0595	0.0558	0.1624	0.0783	0.0593	0.1185	0.1327	0.1652
	Ratio ND:WS			1.0021	1.7482	1.1686		2.1787	2.1972	2.0198	2.0198			2.9760	3.7108	0.3857	3.9284	5.4218	2.8795	15.467	44.585
	Rel. to whole	Wide-shallow		1.1440	1.1367	0.9677		0.7394	0.9500	1.0000	1.0000			0.8911	0.7842	2.9627	1.0483	0.8584	0.7714	0.7947	0.7162
		Narrow-deep		1.1190	0.9285	0.8895		0.6105	0.7576	1.0000	1.0000			0.8415	0.7589	1.2934	0.9115	0.8025	0.7461	0.7664	0.6863

Fig. 6. Variation in statistical subsample runs. For each model, the coefficient of variation (standard error/mean) is given for both wide-shallow and narrow-deep subsample sets. Shading reflects CV values, with cutoff values of (no shading) 0.01, 0.03, 0.1 and 0.3 (darkest). "Ratio WS : ND" indicates the CV of narrow-deep divided by that of wide-shallow samples. The mean value of subsample projections relative to those of the full sample set are indicated as "relative."

Specific model performance

Harte and colleagues (Harte et al. 1999, Harte et al. 2005, Harte 2007) pioneered the study of biodiversity upscaling, and their MaxEnt approach (Harte et al. 2008, 2009) is an important conceptual advance. As expected in the fragmented and human-influenced habitats of the United Kingdom, the METE model performed poorly in our trials, greatly underestimating coarse-scale species richness despite its record of success in upscaling within relatively undisturbed and contiguous habitat (Harte et al. 2009, Harte and Kitzes 2015). Harte's MaxEnt approach can be estimated using surprisingly little information (see *Methods*), which makes it a strikingly efficient tool, but also a very inflexible one. That property is a virtue when applying the model to the sort of homogeneous natural community for which it was designed, but it may create difficulties in applying the model to more anthropogenic landscapes. METE relies on natural communities displaying statistical patterns that maximize entropy within ecological constraints, patterns that may be slow to stabilize (Harte 2011). It would be useful to conduct future tests of the METE upscaling method within contiguous extents of UK biomes that are relatively undisturbed by human activity, such as within large areas of heathland.

After Harte et al.'s (1999) paper, the TS method proposed by Ugland et al. (2003) is arguably one of the longest established and best supported methods in the literature. For example, Jobe (2008) found it to have a reasonable predictive accuracy when applied to tree diversity in the southeastern United States. Extrapolation of the semilogarithmic curve fitted to the terminal points of the species accumulation curves is a robust approach that is designed for heterogeneous environments and it is insensitive to shifts in species abundance, as only presence/absence information is taken into account. This is a great advantage in most applications as there is often substantial variability in the assessment of numbers of individuals, and in many data sets (as here) data on population sizes are not available at all. The TS curve estimates the accumulation rate of new species as more subareas are covered; thus only species' spatial distributions affect the curve.

We tested three different implementations of Ugland's approach, but none of them predicted the SAR very well. The approaches showed more than two-fold differences between the highest (PAM) and lowest (10-at-a-time) estimates, but all three curves were substantially higher and flatter than the true SAR over the scales considered here. The discrepancy is probably the result of the large number of species that occur in just a few plots (e.g., 24.6% of all species were found in just one plot in the X-only data set), which causes the TS curve to rise very steeply initially, and then overshoot. This steepness occurs at relatively fine scales (between the 200 m^2 scale of the survey plots and the scale of the finest Atlas grid, 100 km^2), but when extrapolated to the scales investigated here the curves flatten out and have lower slopes than the actual SAR. The differences in performance between the three implementations of Ugland's TS approach were instructive. While the PAM approach formed groups of similar plots, the 10-at-a-time approach assembled sets at random, and predicted fewer species at every scale. This occurred because PAM groups were more divergent in composition between groups, resulting in faster species accumulation curve as groups are combined.

The TS model's prediction of high, shallow SARs over the scales considered here was shared by several other models without explicit spatial structure (e.g., the Ulrich and Ollik [2005] and Shen and He [2008] approaches). Indeed, in the case of S&H, the SAR approached an asymptote at a value close to the true *S* value. By ignoring spatial structure in species occupancy, these approaches tend to bring in more new species with each added sample initially, but rapidly exhaust the species pool, so that few species remain to be added at coarser scales (Scheiner et al. 2011). The spatial structure of natural biotic communities means that expanding the sample continues to bring in new environments and thus new species even at coarse spatial scales.

Another time-honored approach to upscaling is curve extrapolation. We explored a range of options here, including traditional canonical power laws, but also several methods (median power law, logarithmic, and Lomolino curves) that made use of the multi-scale nature of the field survey data to estimate the slope of species accumulation. None performed particularly well in our comparisons, yet some fared almost as well as some of the more complex approaches. The Lomolino model was the best of a suite of 14 models (Tjørve 2003) commonly fit to species–area relationships, but its accuracy was sensitive to the spatial dispersion and density of plots. When extrapolated from the entire data set, the Lomolino model sometimes gave accurate estimates of the total number of species, but underestimated species number by several hundred when data subsets were used. The model displayed asymptotic behavior, rising very little above about 100 km^2. Our results suggest that a cautious approach should be used when fitting asymptotic models to SARs, even when the model fits well at the fine scale of survey plots.

The classical power law relationship provided a surprisingly good fit to some of the data sets, although different values of the exponent z fit different cases. However, the more variable slopes fit using the median value of z fitted from the multiscale X-plot surveys (from 4- to 200-m^2 scales) produced generally lower slopes, with very poor predictive power. These low fitted slopes are probably affected by the uniformity of land management at these fine scales, especially in the X-only plots, which were constrained not to cross linear features; these resulted in particularly low SAR curves for the fitted logarithmic model, which predicted a total of only 62 species for all of Britain, despite the presence of more than 1,000 species in the overall sample set! On the other hand, despite its abysmal performance in estimating total *S*, the fitted logarithmic model was the only one of all those tested that showed a significant positive correlation with the slope of the true SAR across data sets. Linear extrapolation methods may predict unrealistically high total species richness when the true underlying species accumulation curves reach an asymptote within the extrapolation domain. For example, in an investigation of arthropods in the Azorean Laurisilva forests, Hortal et al. (2006) found very low beta diversity and a rapidly saturating total richness, so that linear extrapolation became heavily biased. In the UK, however, underlying heterogeneity is sufficient that 55% of sampled species were found in seven or fewer sampling quadrats. This large fraction of species with a narrow geographical distribution prevents the species accumulation curve from flattening out, and thus favors straight line extrapolation.

Several other models showed relatively poor performance. The Smith (2008) model not only showed a low predictive accuracy for both TSR and SAR shape, it also displayed extreme variability in richness predictions across the multiple replicate subsamples, suggesting that its estimates are unstable. Unless those problems can be addressed, there is little to recommend it for future applications. On the other hand, the poor performance of the Šizling model (see Appendix S1) is not surprising, as it has been used here for a task rather different from the one for which it was designed. The Šizling model is designed to downscale the SAR from a known value of total species richness, based on the species–occupancy distribution observed within a sample of cells. As such, its application here required the choice of an arbitrary estimate of total richness (1,000), which was not very accurate. The method is included here, however, as it provides a valuable component of a mixed modeling framework, if used together with a companion model for estimating total richness (see *Combining models* below).

The best performance in our tests came from a series of relatively recent models: the Shen and He (2008), Ulrich and Ollik (2005), and Polce (2009) models, and the three Hui models and Šizling model introduced here. Each had distinctive strengths and weaknesses. The Shen and He model performed both well and consistently in estimating total *S*, but proved to be ill-suited to assessing the shape of the SAR, presumably because it ignores the spatial structure of samples. Clearly, the development of a spatially explicit version of this model should be a priority for future research. The Hui ORC and HDC models performed more consistently, providing credible TSR estimates and the best estimates of the SAR as a whole (ORC) of any model considered; they certainly merit further attention. HDC requires reliable numbers of observed rare species in samples, while ORC relies on robust/representative estimates of sampling occupancies for common species. The CS data obviously fulfill the latter of these requirements (sampling common species) very well, but even a survey of this scale (and expense) samples only a tiny fraction of rare species. This may help explain the superior performance of the ORC model in our analyses. The Ulrich and Ollik method proved third-best in total richness estimation, and provided the second best SAR fit of the models tested, suggesting it may be a useful alternative. However, its performance was only moderate in either regard, and the two versions of the model did not consistently bracket the true value, as they were meant to do (in most cases, both estimates were above the true value of species richness). This suggests that the true occupancy–species-rank-order distribution is not a symmetric lognormal but is skewed in the lower part to have more rare than abundant species.

The S&H and U&O methods are both examples of a broader literature devoted to estimating overall species richness in an area based on representative samples (see also, e.g., Palmer 1990, Chao 2005, Magnussen et al. 2006). These methods have been designed to estimate TSR, but they are not explicitly aimed at SAR estimation; thus it is not surprising that they both perform the former task more effectively than the latter. Many of the methods developed for TSR estimation require large proportions of the focal biota to be observed (see Ulrich and Ollik 2005), making them inappropriate for large-scale applications such as the one attempted here. Moreover, systematic biases in most such estimates have been documented in the past (reviewed in Shen and He 2008), further undermining their applicability. The two methods employed here were both explicitly developed with an aim to increasing the accuracy and range of such projections. While these models differ fundamentally in their approaches (with S&H using sampling theory, whereas U&O extrapolate relative abundance distributions), our results here suggest that they have both been quite successful in this respect.

The Polce & Kunin model was explicitly designed for the more difficult task of SAR estimation. While it performed moderately well in our tests, its finer scale estimates (in particular) were often substantially lower than expected. One potential reason for this is the clustered nature of the CS sample set, with five samples taken in each focal 1-km^2 site. The P&K method involved sampling random sets of observations from varying sized sampling windows; when small numbers of samples were drawn from relatively small areas (e.g., 400 km^2 or less), there was consequently a high probability of drawing multiple samples in close proximity to one another, sampling less diversity than expected of a truly random sample of that size. While the logic of the method (separating pure sample size and pure spatial extent effects) is compelling, there clearly remains considerable scope for improvements.

Two of the most accurate individual methods for SAR estimation were developed for this paper: Hui's ORC and HDC methods. Both made use of the distribution of occupancy values across species in the sample. The models differed in what they did with those values: the ORC method extrapolated the curve of species occurrence frequencies using a truncated power law to assess how many species would be expected to occupy one or more 200-m^2 plot, had all of Britain been surveyed; the HDC method examines the number of species represented by different levels of occupancy in the sample, and estimates from observation probabilities how many other such species were likely to have been missed. The SAR downscaling approach developed by Šizling and Storch, which provided even better SAR estimates when married to the Shen and He (2008) TSR estimate, was also based on species occupancy distributions. The success of these three model here spotlights this general approach as one of great promise for future SAR research.

Considering the diverse classes of models tested here (Fig. 2), shows a high level of performance for those based on species occupancy (Hui ORC, Šizling) and related (Ulrich & Ollik, Hui HDC) approaches. Conversely, methods based around extrapolating specific curves (power law, logarithmic, Lomolino, and even MaxEnt) were far less successful. There was mixed success in approaches based on subsampling and spatial species turnover, and there remains significant potential for further developing such approaches.

Combining models

As noted above, consensus models combining more than one of the more promising approaches often outperformed any single "best" model for predicting the total species richness or SAR shape. This generally occurred because different methods showed contrasting errors. Such combinations come at a cost (Levins 1966); there is often a trade-off in modeling between precision (which requires complexity) and insight (which requires simplicity). Developing hybrids of multiple incommensurate approaches runs the risk of producing a method that works well, but which has no compelling logic. Such approaches may prove useful, but they are intellectually ugly. We can only hope that they will be supplanted in time by models that are both accurate and meaningful.

There are additional unexplored opportunities for methodological hybrids amongst the methods presented here, given the wide differences in approach set out above. Note, for instance, that the Šizling model requires the user to have a prior estimate of S_0, the total species richness in the focal region (as does the original Harte et al. [2008] MaxEnt approach), while the Shen and He (2008) model estimates that quantity but cannot estimate diversity at finer scales with any accuracy. Feeding the Shen and He (2008) TSR estimate into the new Šizling or Harte et al. (2008) model would then provide credible estimates of both. Thus for example, if we incorporate the Shen & He estimate of S_0 into the Šizling approach and then downscale, the resulting SAR has a mean relative error score substantially better than any of the individual models tested (Fig. 5).

Reducing survey effort

Our focal data set may represent a tiny fraction of the whole British land surface (roughly one part in 500,000), but it nonetheless requires an impressive investment in time and money to survey. It would obviously be advantageous to have methods that could be nearly as effective with much lower survey effort. We explored this issue at three spatial scales: (1) reducing the total number of 1-km cells surveyed (represented by the narrow-deep subsamples), (2) reducing the number of quadrats sampled in each focal 1-km cell (represented by the wide-shallow subsamples), and in one case (3) surveying a smaller total area for each quadrat (Shen and He's 4-m^2 analysis compared to the 200-m^2 analyses of the same model). Our results clearly suggest that reducing local sampling intensity is far less serious than reducing the number of sites examined. Wide-shallow sub-samples showed much less variation in estimates and (in many cases) notably less bias (relative to the full data set) than did the equally large (but coarse-scale) narrow-deep samples (Fig. 6). Reducing sample size at still finer scales (by changing the size of the local sample plot) may have even less impact: for the one model that was tried at multiple scales (Shen and He 2008), the predictive accuracy of the model was virtually identical when fit using 4-m^2 scale occupancy data than when fit using 200-m^2 data, despite the 50-fold smaller area surveyed.

One issue with reduced sampling intensity in many models was the introduction of a bias: many of the methods made systematically lower species richness predictions when fit to random subsamples of the data set than when fit to the set as a whole, despite the fact that each combined set of five subsamples comprised the full data set. This behavior was displayed by most methods considered, with the exception of the power law and logarithmic extrapolations and the Hui ODC model (where subsample estimates and full set estimates were virtually identical), and the Smith and Hui Zeta models (which behaved inconsistently in this regard). Two possible explanations for the general trend suggest themselves: one statistical, the other biological. On one hand, the smaller data sets may be noisier (relative to their information content), and this will tend to flatten the regression relationships for small samples (a possible solution would be to use Model II regression or equivalent techniques). A more biologically meaningful explanation is that one needs relatively large samples to encounter rare species, and it is the rarer species that cause the SAR to rise, especially at the coarser scales (see, e.g., Tjørve et al. 2008).

Ideal and empirical models

Looking back over the full set of methods explored here, one useful albeit post hoc distinction is between "ideal" and "empirical" SAR models. Ideal models are based on theoretical attempts to understand the appropriate shape that the SAR should be expected to take in natural communities. As such, they have the potential to provide mechanistic insight into potential processes underlying SAR shape, but they tend to be most appropriately applied to natural diversity patterns (rather than anthropogenic ones) where such mechanisms may be thought to determine diversity patterns. Ideal SAR model predictions tend to be relatively inflexible in shape, and as a consequence, they require relatively little data to parameterize; examples range from the canonical power law SAR (Arrhenius 1921, Preston 1962) to the recent development of Maximum Entropy models (Harte et al. 2008, 2009). The inflexibility of such models makes them intrinsically ill-suited to monitoring, e.g., changes of biodiversity in response to management or other human interventions, since they are insensitive (by design) to precisely the sorts of shifts in SAR shape that we would wish to detect. At the other extreme are models designed to assess the empirical SAR whatever its shape happens to be. Such approaches pay for their flexibility by requiring substantially more information. Nonetheless, this flexibility is needed for some applications; for example, if upscaling methods are to be used for multi-scale biodiversity monitoring (see Introduction), they will need to be flexible enough to allow anthropogenic shifts in biodiversity scaling to be reflected in their results.

It is not surprising, given the highly anthropogenic nature of the British landscape, that the best performing models in this analysis (Shen and He 2008, Hui's HDC and ORC models, Ulrich and Ollik 2005) were all empirical approaches. It would be interesting to see how the relative performance of the various approaches explored here would shift were they to be tested on data from more natural landscapes. Several of the methods that performed relatively poorly here have already been shown to behave quite well in such applications (e.g., Ugland et al. 2003, Krishnamani et al. 2004, Jobe 2008). Indeed, the contrasts between ideal and empirical models may be instructive if well tested methods for each can be employed. In well studied areas with good historical species richness records, a reasonable estimate of the natural SAR might be computed using an ideal model (such as that of Harte et al. 2008). This may then be compared to a current SAR computed using one of the empirical models based on current monitoring data. The difference between the two could be interpreted as the "footprint" of anthropogenic activities on biodiversity across spatial scales.

Conclusions

The topic of biodiversity upscaling has been largely of theoretical interest to date, but it is an area that has tremendous potential practical value. Robust and tested upscaling methods would allow the assessment of species richness in poorly studied regions and taxa; they would also make it possible to monitor multi-scale biodiversity change over time, and might allow the coarse-scale implications of environmental or management changes to be inferred from (necessarily fine-scale) experimental results if replicated across multiple sites. To do so we need methods that can be fit using sets of point survey data, and that will be responsive to any anthropogenic changes in local richness and spatial turnover, giving robust and accurate predictions. To test these methods, we need excellent ground-truthed biodiversity survey data from diverse natural and anthropogenic communities across the globe. We have brought together most existing methods for biodiversity upscaling, and have set them an ambitious target: to estimate the total species richness and species–area relationship of a sizeable land mass, using scattered point biodiversity samples from only a tiny fraction of the total area. While methods differed dramatically in their performance, the best of them did reasonably well. Despite an ~500,000-fold increase in scale from the total area surveyed to the area to be assessed, the best of the approaches reliably predicted total species richness within about 10%, and estimated the full species–area relationship within about 18% of the true values. Combining contrasting methods allowed even better accuracy, allowing the SAR to be estimated within 16%. While there is still substantial room for improvement (in particular, in estimating SAR slope) and additional tests on other data sets (ideally involving contrasting regions and taxa) would be welcome, our results suggest that biodiversity upscaling has begun to come of age. It is notable that of the three best methods for SAR estimation, 2.5 (Hui's ORC and HDC and methods, and Šizling's downscaling) are novel methods published here for the first time, suggesting that the field is progressing rapidly. Additional tools are still in development, but our results suggest that existing methods can begin being applied with some confidence.

Acknowledgments

This work was originally conceived and shaped by W. E. Kunin, with all authors contributing to the analyses and manuscript preparation and hence listed alphabetically. This work was supported by a UKPopNet grant (Dispersed platforms for biodiversity research: developing methods and networks for multi-scale research) to W. E. Kunin, with further support from the EU FP7 SCALES (Securing the Conservation of biodiversity across Administrative Levels and spatial, temporal and Ecological Scales ENV-2008-226852) and EU BON (Building the European Biodiversity Observation Network; ENV-2012-308454) projects, and by a fellowship at the Stellenbosch Institute for Advanced Studies. Atlas data were provided by Chris Preston of CEH's Biological Records Centre, Mark Hill of the Bryophyte Recording Scheme, and Janet Simkin of the Lichen Recording Scheme. C. Hui was supported by the National Research Foundation of South Africa (nos. 81825 and 76912) and the Australian Research Council (DP150103017). C. Polce's research was supported by the Marie Curie BIOCONS (European Centre for Biodiversity and Conservation Research) EST programme (MEST-CT-2004-514350). Work by A. Šizling and D. Storch was further supported by grant from the Czech Science Foundation no. 14-36098G. W. Ulrich was supported by grants from the Polish Science Committee (KBN 3 P04F 03422 and KBN 2 P04F 039 29).

Literature Cited

Arfken, G. 1985. Mathematical methods for physics. Third edition. Academic Press, Orlando, Florida, USA.

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

Chao, A. 2005. Species richness estimation. Pages 7907–7916 *in* N. Balakrishnan, C. B. Read, and B. Vidakovic, editors. Encyclopedia of statistical sciences. Second edition. Volume 12. Wiley, New York, New York, USA.

Connor, E. F., and E. D. McCoy. 1979. The statistics and biology of the species–area relationship. American Naturalist 113:791–833.

Dewdney, A. K. 1998. A general theory of the sampling process with application to the “veil line”. Theoretical Population Biology 54:294–302.

Drakare, S., J. J. Lennon, and H. Hillebrand. 2006. The imprint of the geographical, evolutionary and ecological context on species–area relationships. Ecology Letters 9:215–227.

Erwin, T. L. 1982. Tropical forests: their richness in coleopteran and other arthropod species. Coleopterists Bulletin 36: 74–75.

Firbank, L. G., C. J. Barr, R. G. H. Bunce, M. T. Furse, R. Haires-Young, M. Hornung, D. C. Howard, J. Sheail, A. Sier, and S. M. Smart. 2003. Assessing stock and change in land cover and biodiversity in GB: an introduction to Countryside Survey 2000. Journal of Environmental Management 67:207–218.

Geijzendorffer, I. R., et al. 2016. Bridging the gap between biodiversity policy data and policy reporting needs: an essential biodiversity variables approach. Journal of Applied Ecology 53:1341–1350.

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

Green, J. L., and J. B. Plotkin. 2007. A statistical theory for sampling species abundances. Ecology Letters 10:1037–1045.

Gritti, E. S., A. Deputie, F. Massol, and I. Chuine. 2013. Estimating consensus and associated uncertainty between inherently different species distribution models. Methods in Ecology and Evolution 4:442–452.

Harte, J. 2011. Maximum entropy and ecology: a theory of abundance, distribution, and energetics. Oxford University Press, Oxford, UK.

Harte, J. 2007. Toward a mechanistic basis for a unified theory of spatial structure in ecological communities at multiple spatial scales. Pages 101–126 *in* D. Storch, P. A. Marquet, and J. H. Brown, editors. Scaling biodiversity. Cambridge University Press, Cambridge, UK.

Harte, J., and A. P. Kinzig. 1997. On the implications of species–area relationships for endemism, spatial turnover, and food web patterns. Oikos 80:417–427.

Harte, J., and J. Kitzes. 2015. Inferring regional-scale species diversity from small-plot censuses. PLoS ONE. https://doi.org/10.1371/journal.pone.0117527

Harte, J., and E. Newman. 2014. Maximum entropy as a framework for ecological theory. Trends in Ecology and Evolution 29:384–389.

Harte, J., E. Conlisk, A. Ostling, J. L. Green, and A. B. Gaston. 2005. A theory of spatial structure in ecological communities at multiple spatial scales. Theoretical Ecological Monographs 75:179–197.

Harte, J., S. McCarthy, A. Taylor, A. Kinzig, and M. L. Fischer. 1999. Estimating species–area relationships from plot to landscape scale using spatial-turnover data. Oikos 86:45–54.

Harte, J., T. Zillio, E. Conlisk, and A. B. Smith. 2008. Maximum entropy and the state-variable approach to macroecology. Ecology 89:2700–2711.

Harte, J., A. B. Smith, and D. Storch. 2009. Biodiversity scales from plots to biomes with a universal species–area curve. Ecology Letters 12:789–797.

He, F. L., and S. P. Hubbell. 2011. Species–area relationships always overestimate extinction rates from habitat loss. Nature 473:368–371.

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

Hui, C. 2012. Scale effect and bimodality in the frequency distribution of species occupancy. Community Ecology 13:30–35.

Hui, C., and M. A. McGeoch. 2007. Modelling species distributions by breaking the assumption of self-similarity. Oikos 116:2097–2107.

Hui, C., and M. A. McGeoch. 2014. Zeta diversity as a concept and metric that unifies incidence-based biodiversity patterns. American Naturalist 184:684–694.

Jaynes, E. T. 1982. On the rationale of maximum-entropy methods. Proceedings of the IEEE 70:939–952.

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

Keil, P., J. C. Biesmeijer, A. Barendregt, M. Reemer, and W. E. Kunin. 2011. Biodiversity change is scale-dependent: An example from Dutch and UK hoverflies (Diptera, Syrphidae). Ecography 34:392–401.

Keith, S. A., A. C. Newton, M. D. Morecroft, C. E. Bealey, and J. M. Bullock. 2009. Taxonomic homogenization of woodland plant communities over 70 years. Proceedings of the Royal Society B 276:3539–3544.

Krishnamani, R., A. Kumar, and J. Harte. 2004. Estimating species richness at large spatial scales using data from small discrete plots. Ecography 27:637–642.

Latombe, G., C. Hui, and M. A. McGeoch. 2017. Multi-site generalised dissimilarity modelling: Using zeta diversity to differentiate drivers of turnover in rare and widespread species. Methods in Ecology and Evolution 8:431–442.

Lennon, J. J., P. Koleff, J. J. D. Greenwood, and K. J. Gaston. 2001. The geographical structure of British bird distributions: diversity, spatial turnover and scale. Journal of Animal Ecology 70:966–979.

Levins, R. 1966. The strategy of model building in population biology. American Scientist 54:421–431.

Lomolino, M. V. 2001. The species–area relationship: New challenges for an old pattern. Progress in Physical Geography 25:1–21.

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

May, R. M. 1990. How many species? Philosophical Transactions of the Royal Society B 330:292–304.

Myers, R. H. 1990. Classical and modern regression with applications. PWS-Kent Publishing Co., Boston.

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

Pereira, H. M., et al. 2013. Essential biodiversity variables. Science 339:277–278.

Perring, F. H., and S. M. Walters. 1962. Atlas of the British flora. EP Publishing, Wakefield, UK.

Polce, C. 2009. Dynamics of native and alien plant assemblages: the role of scale. Dissertation. University of Leeds, Leeds, UK.

Polce, C., and W. E. Kunin. 2017. SAR dataset for British plants. University of Leeds, Leeds, UK. https://doi.org/10.5518/264

Powell, K. I., J. M. Chase, and T. M. Knight. 2013. Invasive plants have scale-dependent effects on diversity by altering species–area relationships. Science 339:316–318.

Preston, F. W. 1960. Time and space and the variation of species. Ecology 41:612–627.

Preston, F. W. 1962. The canonical distribution of commonness and rarity. Ecology 43:185–215.

Preston, C. D., D. A. Pearman, and T. D. Dines. 2002. New atlas of the British and Irish flora. Oxford University Press, Oxford, UK.

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

Rosenzweig, M. L. 2001. The four questions: What does the introduction of exotic species do to diversity? Evolutionary Ecology Research 3:361–367.

Scheiner, S. M. 2003. Six types of species–area curves. Global Ecology and Biogeography 12:441–447.

Scheiner, S. M., A. Chiarucci, G. A. Fox, M. R. Helmus, D. J. McGlinn, and M. R. Willig. 2011. The underpinnings of the relationship of species richness with space and time. Ecological Monographs 81:195–213.

Shen, T. J., and F. L. He. 2008. An incidence-based richness estimator for quadrats sampled without replacement. Ecology 89:2052–2060.

Shmida, A., and M. V. Wilson. 1985. Biological determinants of species-diversity. Journal of Biogeography 12:1–20.

Šizling, A. L., and D. Storch. 2004. Power-law species–area relationships and self-similar species distributions within finite areas. Ecology Letters 7:60–68.

Smart, S. M., R. H. Marrs, M. G. Le Duc, K. Thompson, R. G. H. Bunce, L. G. Firbank, and M. J. Rossall. 2006*a*. Spatial relationships between intensive land cover and residual plant species diversity in temperate, farmed landscapes. Journal of Applied Ecology 43:1128–1137.

Smart, S. M., K. Thomspon, R. H. Marrs, M. G. Le Duc, L. C. Maskell, and L. G. Firbank. 2006*b*. Biotic homogenization and changes in species diversity across human-modified ecosystems. Proceedings of the Royal Society B 263:2659–2665.

Smith, K. T. 2008. On the measurement of beta diversity: an analog of the species–area relationship for point sources. Evolutionary Ecology Research 10:987–1006.

Socolar, J. B., J. J. Gilroy, W. E. Kunin, and D. P. Edwards. 2016. How should beta-diversity inform biodiversity conservation? Trends in Ecology and Evolution 31:67–80.

Storch, D. 2016. The theory of the nested species–area relationship: geometric foundations of biodiversity scaling. Journal of Vegetation Science 27:880–891.

Tjørve, E. 2003. Shapes and functions of species–area curves: A review of possible models. Journal of Biogeography 30:827–835.

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

Tjørve, E., and K. M. C. Tjørve. 2008. The species–area relationship, self-similarity, and the true meaning of the *z*-value. Ecology 89:3528–3533.

Tjørve, E., and W. R. Turner. 2009. The importance of samples and isolates for species–area relationships. Ecography 32: 391–400.

Tjørve, E., W. E. Kunin, C. Polce, and K. M. C. Tjørve. 2008. The species–area relationship: separating the effects of species-abundance and spatial distribution. Journal of Ecology 96:1141–1151.

Ugland, K. I., J. Gray, and K. E. Ellingsen. 2003. The species-accumulation curve and estimation of species richness. Journal of Animal Ecology 72:888–897.

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

Watson, H. C. 1835. Remarks on the geographical distribution of British plants. Longman, Rees, Orme, Brown, Green and Longman, London, UK.

Xu, H., S. Liu, Y. Li, R. Zang, and F. L. He. 2012. Assessing non-parametric and area-based methods for estimating regional species richness. Journal of Vegetation Science 23:1006–1012.

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecm.1284/full

Data Availability

Data associated with this study are available from the Research Data Leeds Repository: https://doi.org/10.5518/264.