Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 195–205 doi: 10.1111/j.2041-210X.2011.00146.x

Local frequency as a key to interpreting species occurrence data when recording effort is not known

Mark O. Hill*†

Centre for Ecology and Hydrology, Wallingford OX10 8BB, UK

Summary

1. Data on the occurrence of species in grid cells are collected by biological recording schemes, typically with the intention of publishing an atlas. Interpretation of such data is often hampered by the lack of information on the effort that went into collecting them. This is the 'recorder effort problem'.

2. One measure of recorder effort is the proportion of a suite of common species ('benchmark species') found at a given location and time. Benchmark species have in the past been taken as a uniform set across a territory. However, if records are available from a neighbourhood surrounding a given location, then a local set benchmark species can be defined by pooling records from the neighbourhood and selecting the commonest species in the pooled set.

3. Neighbourhoods differ in species richness, so that the list of species that 'ought' to be found in one location may be longer than that for another. If the richness of a neighbourhood can be estimated, then a suite of benchmark species can be standardized to be the commonest of a fixed proportion of the total expected for the neighbourhood. Recording effort is then defined as the proportion of benchmark species that were found.

4. A method of estimating species richness is proposed here, based on the local frequencies f_j of species in neighbouring grid cells. Species discovery is modelled as a Poisson process. It is argued that when a neighbourhood is well sampled, the frequency-weighted mean frequency $\sum f_j^2 / \sum f_j$ of species in the neighbourhood will assume a standard value.

5. The method was applied to a data set of 2 000 000 records detailing the occurrence of bryophytes in 3695 out of the total 3854 hectads (10-km squares) in Great Britain, Ireland, the Isle of Man and the Channel Islands.

6. Three main applications are outlined: estimation of recording effort, scanning data for unexpected presences or absences and measurement of species trends over time. An explicit statistical model was used to estimate trends, modelling the probability of species j being found at location i and time t as the outcome of Poisson process with intensity $Q_{ijt}\, x_{jt}$, where x_{jt} is a time factor for species j, and Q_{ijt} depends on recording effort at location i and time t and on the time-independent probability of species j being found in hectad i.

Key-words: benchmark species, discovery curve, neighbourhood, rank-frequency curve, species richness

Introduction

When the occurrence of species is mapped in grid cells, large-scale patterns of geographical distribution and long-term trends in frequency can be observed. Examples are the widespread decline of arable weeds (Preston *et al.* 2002) and the disappearance of butterflies from parts of their former range (Asher *et al.* 2001). Many changes are less obvious, and to measure these, various statistical methods have been devised. Several of these methods use the proportion of records of a given species relative to the total records of all species as a measure of frequency (Telfer, Preston, & Rothery 2002; Ball & Henshall 2006). Similar methods have been used with museum data (Hedenäs *et al.* 2002; Hofmann *et al.* 2007; Jeppsson *et al.* 2010).

The method of Telfer, Preston, & Rothery (2002) measures relative change only between two time periods, but several of

*Correspondence author. E-mail: moh@ceh.ac.uk
Correspondence site: http://www.respond2articles.com/MEE/
†Present address: 11 Chaucer Road, Cambridge CB2 7EB, UK.

the others can measure trends at intervals. They do this by treating the total number of records as a measure of recording effort. A potential weakness of this approach is that different parts of the territory may be sampled unevenly at different time periods. If sampling at a particular period is concentrated on a region from which a species is absent or very scarce, then its frequency may appear to have declined without this actually being the case.

Unevenness of recording continues to be a problem. Prendergast *et al.* (1993) drew attention to it and called it the 'recorder effort problem'. They observed that virtually all large-scale floral and faunal surveys depend on volunteer recorders, creating data sets that may be biased in favour of recorder, rather than species, distributions. The role of volunteer recorders is increasing (Silvertown 2009). Some volunteers can be persuaded to engage in standardized monitoring or recording programmes (Newson *et al.* 2005; Roy, Rothery, & Brereton 2007), but most cannot. As a consequence, there are large biases in many data sets (Dennis & Thomas 2000; Boakes *et al.* 2010; Hassall & Thompson 2010; Petrik, Pergl, & Wild 2010).

The recorder effort problem can at least partly be solved if there is some measure of regional or local recording effort. One such measure is the number of recorded 'benchmark species', i.e. species that are thought to be ubiquitous and stable in their occurrence. Records of other species can be related to those of the benchmark species (Maes & Swaay 1997). With a relatively small territory, it may be possible to find such ubiquitous species, but with a larger territory such as Great Britain, this may be difficult. Even if there are ubiquitous species, they may differ widely in abundance across the territory, so that their presence in one district does not signify the same recording effort as in another. There is, however, no necessity for the same benchmark species to be used everywhere. Indeed, local recorders conducting systematic surveys often judge what 'ought' to be present in a particular location by comparing the records for that location with others in the neighbourhood. When an area is moderately well recorded, then a simple 'wants list' can be constructed for each recording unit by listing those species that are absent from the recording unit but present in most of its neighbours. Under-recorded units are those with long wants lists.

The method that is outlined below is a generalization of this means of judging which species are likely to be present. It proceeds in four stages at each separate location.

1 Define the neighbourhood of location i as a set of weights for nearby locations, based on both spatial proximity and biological similarity.

2 Based on records of species presence, calculate weighted local species frequencies f_{ij} for all species in the neighbourhood, using weights defined at stage 1.

3 Calculate neighbourhood frequency as the frequency-weighted local mean frequency, i.e. $\sum_j f_{ij}^2 / \sum_j f_{ij}$.

4 Adjust local species frequencies by a 'sampling-effort multiplier' α_i, used to inflate (or occasionally deflate) them to the point where neighbourhood frequency assumes a standard value.

At stage 4, frequencies are assumed to result from a Poisson process, and the multiplier α_i is applied to the intensity of the process, namely $-\log(1\text{-}f_{ij})$. This is converted back to a frequency. The resulting species frequencies f'_{ij} are treated as probabilities and are used to estimate trends or to identify locations with an excess of species ('hotspots').

Data and methods

DATA SETS AND COMPUTER PROGRAMS

Two large data sets held by the Biological Records Centre (BRC) were used as the basis for the analysis. The immediate object of the study was to measure temporal trends in the probability of bryophyte species occurrence, based on the data set of the British Bryological Society. These data were mapped as an atlas in the early 1990s (Hill, Preston, & Smith 1991–1994) and have been steadily added to since then; in July 2010, they comprised 2 038 000 records of 1196 taxa, of which 1053 were species, in 3695 hectads. The other data set was the Vascular Plant Dataset, managed by BRC on behalf of the Botanical Society of the British Isles. In summer 2010, this had 12 675 000 records, from which lists of native and archaeophyte vascular plant species in 3829 out of the total 3854 hectads (10-km squares of the National Grids of Britain and Ireland) were abstracted. Hybrids and taxa not mapped in the most recent atlas (Preston, Pearman, & Dines 2002) were excluded. The abstracted data set comprised 1 666 000 records of 1469 species.

Data base manipulations were performed in Microsoft Access. Other calculations were made in Fortran, using the GNU Fortran G77 v0.5.25 compiler for Windows XP (Free Software Foundation, 1999). Calculations other than those used to define neighbourhoods were performed by a single program, Frescalo (FREquency SCAling LOcal). Source code and executables for Windows XP, together with a worked example, can be downloaded from the BRC website http://www.brc.ac.uk.

SELECTION OF HECTADS FOR DETAILED STUDY

Given the number of hectads and bryophyte species, the total data set of observed neighbourhoods and species was substantial. For this reason, a systematic sample of one in a hundred hectads was selected for a more detailed study. Hectads were numbered alphabetically in two series according to the grid under consideration, starting with the British grid and following it with the Irish grid. The systematic sample (Table 1), comprising hectads 100, 200,…, 3800, includes localities in Scotland (10), England (15), Wales (2) and Ireland (10).

DEFINITION OF NEIGHBOURHOODS

The main use of the vascular plant data was to define neighbourhoods, i.e. sets of hectads that were both physically close to and floristically similar to a given hectad. For any given hectad (the 'target hectad'), the procedure was first to take the 200 spatially closest hectads and then to select the 100 floristically most similar hectads from among the 200. Floristic

Table 1. Systematic sample of one in a hundred hectads in Britain (4-character hectads) and Ireland (3-character hectads)

Hectad	Location	Sea	$\varphi_i[1]$	α_i	Obs	Exp	Absent	Improb
NB13	NW Lewis	39	0·53	3·5	90	292	116	
ND04	Flow country		0·41	5·7	32	295	181	
NG63	Scalpay nr Skye	88	0·63	1·9	–	301	223	
NH66	Cromarty Firth	18	0·47	4·3	100	340	147	1
NJ93	Ellon		0·44	6·7	86	289	124	
NN03	Ben Cruachan	14	0·75	0·9	337	329	10	21
NO03	Bankfoot		0·59	3·0	129	327	117	3
NR59	Jura	88	0·56	2·6	–	331	238	
NS67	nr Glasgow		0·53	3·6	164	349	113	6
NT68	nr Dunbar	86	0·55	3·4	151	334	98	
NX87	W of Dumfries		0·51	4·0	69	320	169	
NY88	Bellingham		0·69	1·3	312	285	10	17
SD33	Blackpool	1	0·51	4·0	155	312	77	1
SE33	E Leeds urban		0·51	4·2	63	293	143	
SH62	Rhinog Mts		0·74	1·0	402	340	10	36
SJ68	Warrington		0·60	3·7	114	264	88	5
SK68	Worksop		0·48	5·2	85	235	90	
SN71	Black Mts		0·69	1·4	384	342	14	17
SO71	nr Gloucester	1	0·62	2·4	172	265	46	2
SP71	Stoke Mandeville		0·60	2·8	72	199	71	1
ST10	Honiton		0·70	1·5	195	269	28	6
SU11	Fordingbridge		0·71	1·3	263	268	8	1
SW54	St Ives	89	0·72	1·1	225	301	51	4
SY67	Portland	77	0·63	2·3	156	282	78	5
TF37	E of Lincoln		0·58	3·3	130	203	33	3
TL36	Boxworth		0·68	1·5	184	187	4	10
TM45	Aldeburgh	39	0·67	1·9	166	205	19	7
TQ93	Weald nr Ashford		0·67	1·8	185	251	38	2
C92	SE of Coleraine		0·17	12·6	–	303	215	
G80	nr Carrick		0·41	6·0	102	300	126	
H80	nr Crossmaglen		0·29	10·2	19	246	159	
L82	Galway Bay	60	0·34	6·7	84	368	179	
M84	W of Athlone		0·27	9·9	70	201	85	
N84	Kilcock, Meath		0·30	9·2	42	204	110	
R21	Mullaghareirk Mts		0·20	12·1	153	265	71	7
S21	Comeragh Mts		0·34	6·3	266	329	28	
T25	Gorey, SE coast	93	0·25	10·3	35	207	112	
W58	nr Cork		0·21	12·6	–	268	196	

Hectad – grid reference with Ordnance Survey lettering for 100-km squares; Sea – sea area km^2; $\varphi_i[1]$ – neighbourhood mean frequency before multiplication; α_i – sampling-effort multiplier (see text); Obs – observed number of bryophyte taxa; Exp – predicted number; Absent – taxa predicted to be present with at least 70% probability that were not recorded; Improb – taxa with $<20\%$ probability that were actually recorded.

similarity of the vascular plant flora in two hectads was measured by Sørensen's similarity coefficient, i.e. twice the number of species in common divided by the sum of the two individual species totals (Legendre & Legendre 1998). The target hectad was included in the list. Weights were applied so that within the 100 selected neighbours, the species lists of those that were physically more distant or floristically less similar received lower weight than those from nearer and more similar hectads. Specifically,

$$w_{ii'} = \left(1 - \frac{(k-1)^2}{200^2}\right)^4 \left(1 - \frac{(l-1)^2}{100^2}\right)^4$$

where $w_{ii'}$ is the weight for hectad i' when considered as a neighbour of the target hectad i, and i' is the kth nearest to i in the order of distance and the lth nearest in the order of floristic similarity. In this scheme, the target hectad gets unit weight, and the others are weighted so that those halfway down the list receive weight about 0·33. For example, if a hectad is 100th in the order of spatial proximity ($k = 100$) and 50th in the order of floristic similarity ($l = 50$), then

$$w_{ii'} = (1 - 0{\cdot}495^2)^4 \times (1 - 0{\cdot}490^2)^4 = 0{\cdot}325 \times 0{\cdot}333 = 0{\cdot}108$$

Five neighbourhoods and their weighting are shown in Fig. 1. See also Data S1 in Supporting information.

STANDARDIZATION OF NEIGHBOURHOOD MEAN FREQUENCY

When a neighbourhood and its weights have been defined, then f_{ij}, the observed frequency of species j in the neighbourhood of hectad i, is defined as

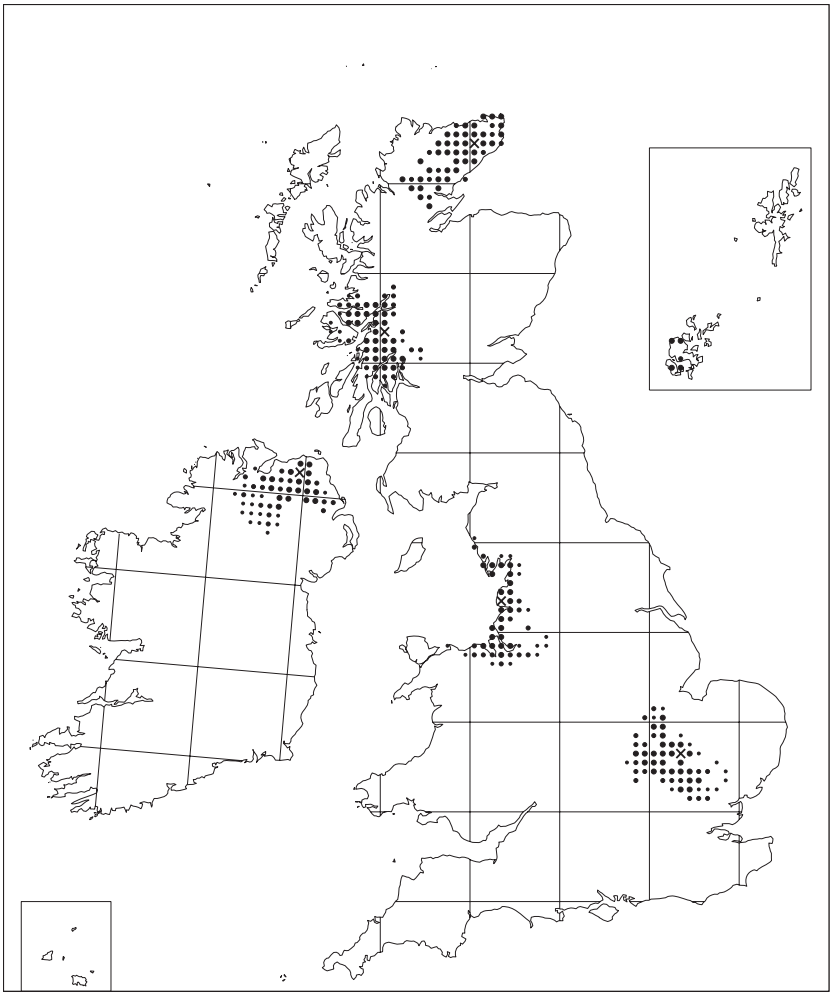

Fig. 1. Neighbourhoods of selected hectads (shown with crosses) using both floristic similarity and spatial proximity. Circles of decreasing size represent neighbouring hectads, with weights 0·7–1·0, 0·2–0·7 and 0·05–0·2. The selected hectads are, from north to south, Flow country, Ben Cruachan, Coleraine, Blackpool and Boxworth.

$$f_{ij} = \sum_{i'} w_{ii'} a_{i'j} / \sum_{i'} w_{ii'}$$

where the summation is taken over the neighbourhood, and $a_{i'j} = 1$ if species j is recorded in hectad i' and $a_{i'j} = 0$ otherwise. In neighbourhoods where there are numerous unsampled hectads, even the most frequent species may have local frequency much less than 1 (Fig. 2).

We assume that discovery is a Poisson process. Let

λ_{ij} = discovery rate for species j in hectad i

s_i = number of searches made in hectad i

Then, the probability of finding species j in hectad i is given by

$$p_{ij} = 1 - \exp(-\lambda_{ij} s_i)$$

Parameter s_i is a measure of sampling effort. In practice, we cannot measure λ_{ij} directly unless separate recording parties are sent in to make independent surveys. We do, however, have information on the frequencies f_{ij} for all species in the neighbourhood. On the assumption that the target hectad is similar to those in its neighbourhood,

$$\lambda_{ij} s_{i(N)} \approx -\log(1 - f_{ij}) \qquad \text{eqn 1}$$

where $s_{i(N)}$ is the weighted mean sampling effort for the neighbourhood, i.e.

$$s_{i(N)} = \sum_{i'} w_{ii'} s_{i'} / \sum_{i'} w_{ii'}$$

The problem of estimating λ_{ij} is therefore solved if we can estimate $s_{i(N)}$. This is done by standardizing the frequency-weighted mean local frequency

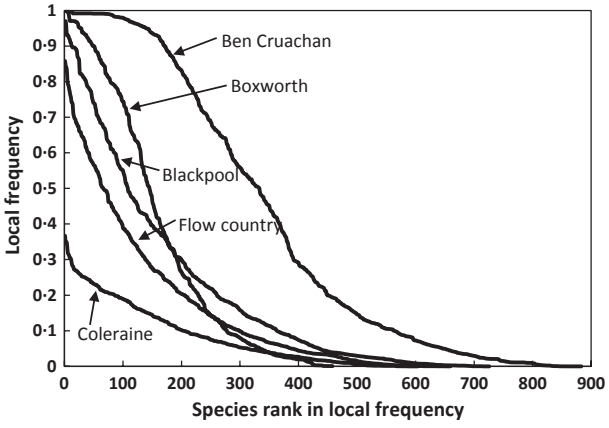

Fig. 2. Rank–frequency curves for bryophytes in the five neighbourhoods shown in Fig. 1.

$$\varphi_i = \sum_j f_{ij}^2 / \sum_j f_{ij}$$

This is a weighted mean frequency, because if weights v_{ij} are set to be equal to the frequencies f_{ij}, then

$$\varphi_i = \sum_j v_{ij} f_{ij} / \sum_j v_{ij}$$

Indeed, φ_i is analagous to the expected value of a frequency. Imagine that all the species are put into a bag, in proportion to the frequencies. Then, select a random species from the bag. The probability that it is of species j is $f_{ij}/\sum_j f_{ij}$. Thus the expectation

$$\mathrm{E(frequency)} = \mathrm{E}(f_{ij}) = \sum_j f_{ij}^2 / \sum_j f_{ij} = \varphi_i$$

An interesting property of φ_i is that it can be expressed as a ratio

$$\varphi_i = \sum_j f_{ij} / \left(\left(\sum_j f_{ij} \right)^2 / \sum_j f_{ij}^2 \right) = \sum_j f_{ij} / N_2$$

where $\sum_j f_{ij}$ is the average species richness of hectads in the neighbourhood and N_2 is the 'effective species number' (reciprocal of Simpson's index) defined by Hill (1973). In other words, φ_i is the ratio of mean species richness per hectad to a dimensionless number, N_2, which depends on the shape of the rank–frequency curve but not on the absolute magnitude of the frequencies.

This property explains the rationale of the method that is described below. If, for example, half of the records from the neighbourhood are deleted, so that the new value of f_{ij} is half the old value, then the numerator goes down by a factor of two, but the denominator is unchanged. Thus, φ_i is in some sense a measure of sampling intensity. The basic assumption of the method is that in a well-sampled neighbourhood, φ_i (hereafter called neighbourhood frequency) is roughly constant, say Φ. This assumption would certainly be correct if the rank–frequency curves were scaled to have the same shape in all neighbourhoods. No such assumption is made. Indeed, the proposed method of rescaling requires fitting only two parameters, one for sampling effort and the other for species richness. It turns out that the resulting curves are remarkably similar.

Let f'_{ij} be the frequency of species j in neighbourhood i when the neighbourhood is sampled to a standard extent, corresponding to a thorough search. Then, the assumption of constant neighbourhood frequency is that for all neighbourhoods i

$$\sum_j f_{ij}'^2 / \sum_j f'_{ij} = \Phi$$

Without loss of generality, set the neighbourhood mean sampling effort corresponding to this extent of sampling to 1. Setting $s_{i(N)} = 1$ in eqn 1

$$\lambda_{ij} \approx -\log(1 - f'_{ij})$$

Now consider the consequences of increasing the actual sampling effort $s_{i(N)}$ by a sampling-effort multiplier α_i. On the assumption of a Poisson process, the frequency of species j for this multiplier is

$$f_{ij}[\alpha_i] = 1 - \exp(\alpha_i \log(1 - f_{ij}))$$

The neighbourhood frequency is defined as

$$\varphi_i[\alpha_i] = \sum_j f_{ij}^2[\alpha_i] / \sum_j f_{ij}[\alpha_i]$$

The problem of estimating $s_{i(N)}$ is then that of finding a sampling-effort multiplier α_i such that

$$\varphi_i[\alpha_i] = \Phi$$

Then, by definition

$$\alpha_i s_{i(N)} = 1$$

Note that $\varphi_i[\alpha_i]$ is a property of the neighbourhood not the target hectad. Indeed, the target hectad may be completely lacking in records, as is the case for the Coleraine hectad indicated in Figs 1 and 2.

Calculation of sampling-effort multipliers requires first that a standard neighbourhood frequency Φ is selected and then that parameters α_i are fitted so that $\varphi_i[\alpha_i] = \Phi$. In the example discussed here, Φ was set to 0·74, which was the 98·5th percentile of observed values of $\varphi_i[1]$. This high percentile was chosen, because bryophytes are not very completely recorded. More details for south-east England is given in Data S1 in Supporting information. In the applications, the precise value of Φ is not critical, but it should correspond to a thorough search of the neighbourhood. Values of α_i were fitted by successive approximation, multiplying trial values α by $\Phi/\varphi_i[\alpha]$ until convergence was achieved.

The curves in Fig. 2 differ not only because the sampling effort varies between neighbourhoods but also because of variation in species richness. It is possible to correct for species richness to compare the shape of the rank–frequency curves after standardization. This is achieved by dividing ranks for frequency by the expected number of species (Fig. 3). In symbols, let R_{ij} be the rank of species j in neighbourhood i. Then R'_{ij},the scaled rank is defined as

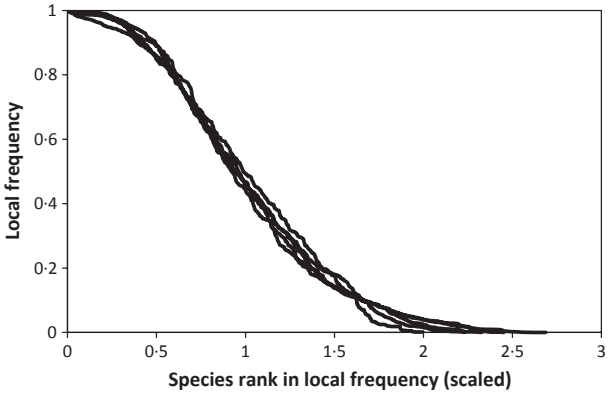

Fig. 3. Rank–frequency curves for bryophytes in the five neighbourhoods shown in Figs 1 and 2 after adjustment of neighbourhood frequency by sampling-effort multipliers and of species rank by division by the expected species number $\sum_j f'_{ij}$.

$$R'_{ij} = R_{ij} / \sum_j f'_{ij}.$$

Note that $\sum_j f'_{ij}$ is the expected number of species after rescaling to standard sampling effort.

ESTIMATING CHANGE IN SPECIES FREQUENCY

For reasons that are explained in the discussion, a simple Poisson model of species discovery on individual visits is not realistic. A general multiplicative model to estimate a time factor x_{jt} for species j is

$$P_{ijt} = 1 - \exp(-Q_{ijt} x_{jt}) \qquad \text{eqn 2}$$

where P_{ijt} is the probability that species j is recorded in hectad i at time t, and Q_{ijt} depends on the recording effort in hectad i at time t and on a time-independent probability of species j being found in hectad i. For the analysis presented here, a simple model of Q_{ijt} is used, based on the proportion of common ('benchmark') species found in hectad i and the estimated probability f'_{ij} of finding species j there. Benchmark species are defined to be those for which

$$R'_{ij} < R^*$$

where R^* is a standard value. In the following example, R^* was chosen as 0·27, which has the interpretation that the top 27% of the expected number of species are treated as benchmark species. If the rescaled curves have the average of the shapes shown in Fig. 3, then for a neighbourhood with standard sampling effort, one would expect 99·2% of benchmark species to be present, and the least frequent benchmark species would have frequency f'_{ij} about 97·5%.

Let s_{it} be the proportion of benchmark species found in hectad i at time t; it is an approximate measure of sampling intensity. Then, for the simple model,

$$\begin{aligned} Q_{ijt} &= -\log(1 - s_{it} f'_{ij}) \quad \text{if } s_{it} f'_{ij} < 0{\cdot}98 \\ &= -\log(1 - 0{\cdot}98) = 3{\cdot}91 \quad \text{otherwise.} \end{aligned} \qquad \text{eqn 3}$$

Q_{ijt} is truncated because for very common species in very well sampled hectads, $s_{it} f'_{ij}$ can assume the value 1. The time factor x_{jt} is estimated as the value of x_{jt} for which

$$\sum_i P_{ijt} = \sum_i Q_{ijt}$$

It can be calculated iteratively, starting with a trial value of 1 and multiplying at each iteration by $\sum_i P_{ijt} / \sum_i Q_{ijt}$.

Results

ESTIMATION OF RECORDING EFFORT

According to the logic of rescaling frequencies, neighbourhood recording effort $s_{i(N)}$ is measured by the reciprocal of the sampling-effort multiplier $1/\alpha_i$ required to standardize $\varphi_i[\alpha_i]$ to Φ. For the hectads in Table 1, this varied from 0·08 for the Coleraine neighbourhood to 1·12 for the Ben Cruachan neighbourhood. When local frequencies were calculated using the recorded species composition of the target hectads, the picture was more complex (Fig. 4). Four of the target hectads had no records and therefore no mean local frequency; they are omitted. The Crossmaglen hectad stands out as unusual, having only 19 recorded species but a relatively low mean frequency. The 19 species are from six separate localities, visited in 1929, 1949, 1950, 1952, 1980 and 1991. Most of them date from before 1960 when the British Bryological Society started systematic recording.

STANDARDIZED RANK–FREQUENCY CURVES

Neighbourhood frequencies were mostly inflated by the standardization procedure, but in the case of the Ben Cruachan neighbourhood, they were slightly deflated. After rescaling, the originally disparate rank–frequency curves of Fig. 2 appear much more similar (Fig. 3). Curves for the most poorly sampled neighbourhoods had somewhat shorter tails because the rarest species were less likely to be encountered. If the average curve for relatively well-sampled neighbourhoods, i.e. those with $\alpha_i < 2{\cdot}0$, is calculated for the whole data set, average standardized discovery rates can be calculated as

$$\lambda(R') = -\log(1 - f'(R'))$$

where f' is the mean standardized frequency for the average curve. For R' in the range 0·1–1·8, there was a nearly linear relation between R' and log(λ). This allows the average curve to be well approximated by an equation of the form

$$f'(R') = 1 - \exp(-\exp(a + bR'))$$

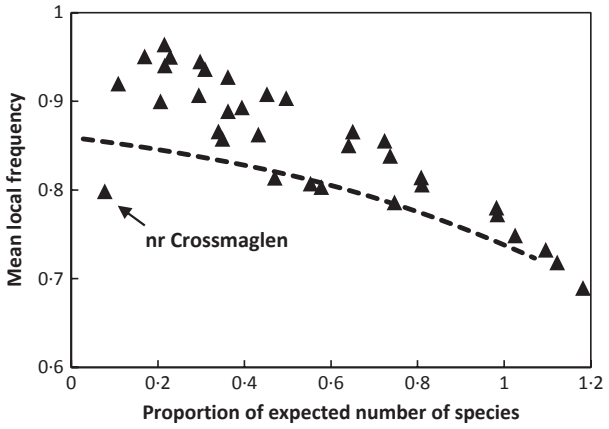

Fig. 4. Mean local frequency of species actually found in target hectads in relation to total number recorded, expressed as a proportion of the expected total. The trend line shows the mean frequencies that would be expected if the probability of discovery followed the average rank–frequency curves after scaling as in Fig. 3.

The fitted constants a and b were in this case 2·005 and −2·545.

RECOGNITION OF UNUSUAL SPECIES OR MISSING SPECIES

The hectad with the largest number of unusual species was SH62, in the Rhinog Mountains (Table 1). This hectad also had the greatest excess of observed species over the number expected. In the lowlands, the Boxworth hectad TL36 also had several locally scarce species, of which the three scarcest were *Racomitrium heterostichum*, *R. ericoides* and *Grimmia trichophylla*, with probabilities of 8%, 8% and 9%, respectively. These Grimmiaceae are normally plants of acid rock in the uplands. All were on artificial imported substrates, the *Racomitrium* species being on clinker of sewage works and *G. trichophylla* on a memorial rock brought from the Cheviot to Boxworth churchyard.

In the other direction, two Scottish and two Irish hectads were completely unrecorded (Fig. 5). Inevitably, they had a large number of omissions. As measured by omissions, the best-recorded hectad was that at Boxworth, with only four unrecorded species falling above the threshold $f'_{ij} = 70\%$.

DOCUMENTING SPECIES CHANGE

The reappearance of epiphytes as a result of cleaner air provides a good example of change in frequency over time (Fig. 6). There is a suggestion of a slight decline over the period 1960–1975, followed by a strong recovery after 1990. Time factors were calculated for the whole of Britain and Ireland; if they were calculated for those districts that had experienced severe air pollution, the increase would appear even more marked.

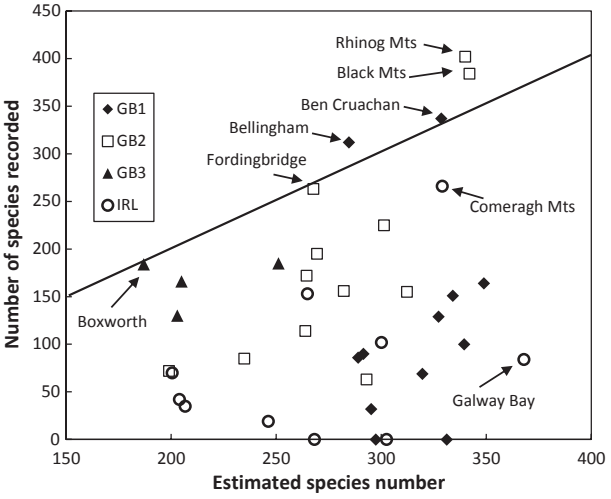

Fig. 5. Observed species numbers in hectads in relation to numbers predicted. The diagonal line shows where observed and predicted are equal. Geographical regions are distinguished by symbols: GB1 – Scotland and Northumberland; GB2 – Wales, southern and western England; GB3 – eastern England; IRL – Ireland.

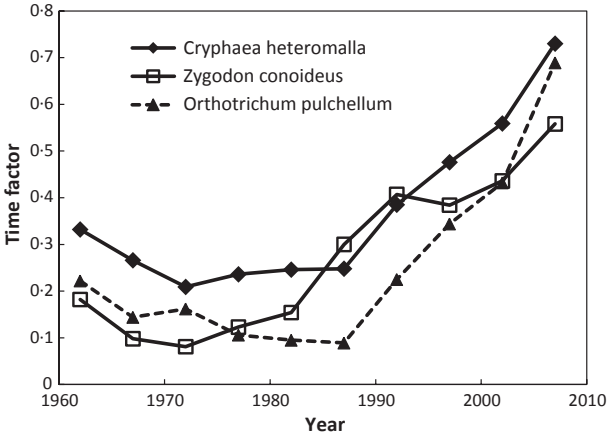

Fig. 6. Change in frequency of three epiphytic mosses over period 1960–2010. Time factors for each species are calculated for quinquennia 1960–64, 1965–69, etc.

Discussion

THE RECORDER EFFORT PROBLEM

The recorder effort problem has vexed many investigators, causing them to make ad hoc adjustments such as rejection of data from locations that were not adequately sampled at both an earlier and a later time. The novel feature of the neighbourhood-frequency method is that it pools data from neighbouring locations, so that the fact that a particular location was not sampled adequately at one date does not mean that the information from the other date is wasted. During the first recording scheme for bryophytes, it took 30 years to achieve adequate cover of Great Britain (Hill, Preston, & Smith 1991–1994). During that period, an individual county might be recorded intensively only for a period of 10 years. Recording was therefore very patchy in space and time. A method that can cope with such patchiness is needed.

Uneven recorder effort is not the only source of bias. Recorders differ widely in competence. Recording may be affected by the weather and by the seasonal apparency of species. The advice of Rich & Smith (1996) to note the time spent in recording is helpful but not sufficient. In the end, what matters is not the effort that went in, but how successful it was. The methods advocated here depend on the outcome of recording, not the input of effort.

Most current methods for estimating species richness start from the assumption that there is a definite number of species at each location (Colwell & Coddington 1994; Magnussen *et al.* 2010). However, if records are kept over a long period, this assumption is not realistic. Species come and go. They may be temporarily resident or may opportunistically immigrate from surrounding favourable habitat (Owen 2010). Thus, a probabilistic model of species occurrence and discovery may be just as realistic as a deterministic one.

A strength of the neighbourhood-frequency method is that it requires no extra information beyond species lists from neighbouring localities. It compensates for data

incompleteness by using a species-richness estimator (Hortal 2008). There are two key assumptions, namely that the probability of finding a species at a locality can be estimated by its frequency in the neighbourhood and that in a well-sampled neighbourhood, the weighted mean frequency is a constant, independent of species richness. A potential weakness of the method is that some neighbourhoods may be so poorly recorded that they provide no information about local species frequency. If such neighbourhoods exist, then they should be omitted or neighbourhood size should be increased.

NEIGHBOURHOODS AND THE PROBABILITY OF OCCURRENCE

Neighbouring hectads can give an indication of how likely a species is to be found in the target hectad only if the target is reasonably similar to its neighbours. If the target is not similar to its neighbours, then the key approximation of eqn (1) will not hold. The reason for using multiplicative weights here is simply that purely spatial smoothing may include very dissimilar entities such as a sandy coast and an inland mountain, whereas purely floristic weights may result in a 'neighbourhood' scattered over several hundred kilometres.

The use of neighbourhoods to define a smoothing window is well established, both in physical space to show local frequency (Gibbons, Reid, & Chapman 1993) and in environmental space to define climate response surfaces (Huntley *et al.* 1995). Neighbourhoods do not have to have low weight near their boundaries. Prendergast *et al.* (1993) used unweighted blocks of 25 hectads to define their neighbourhoods. The smoothing kernel $(1\text{-}x^2)^4$ used here is hardly distinguishable in practice from a Gaussian function. It uses ranks rather than actual distances to ensure that remote islands also have a good supply of neighbours. For floristic similarity, Sørensen's coefficient was chosen because it measures overall similarity, so that if the flora of one hectad is a small subset of that of another, then the two hectads have low floristic similarity.

Hectads can be exceptional in various ways. We have already seen that a hectad in the Rhinog Mountains was markedly richer in species than its neighbours (Fig. 5). In the other direction, hectads with little land are likely to have fewer species than their neighbours would indicate. These factors can be mitigated by using floristic similarity as well as spatial proximity to define the neighbourhood. Thus, the hectads that are similar to Boxworth lie mainly to the west of it (Fig. 1); those to the north and east of it are in the East Anglian fen country, which is very flat and lacks ancient woodland. Likewise, the neighbourhood of Blackpool lies mostly along the coast and not in the Lancashire uplands.

There is undoubtedly scope for improving the definition of neighbourhoods, especially in the mountains, where maximum altitude and the number of wet days per annum (Ratcliffe 1968) can give useful information. There is, of course, no necessity to use floristic similarity in defining neighbourhoods. Similarity of physical attributes might be preferable, but a suitable data set for both Britain and Ireland was not readily available.

STANDARDIZATION OF LOCAL FREQUENCIES

The method of standardizing local frequencies depends on three assumptions: that the target hectad resembles its neighbours; that species discovery is basically a Poisson process; and that in a well-worked neighbourhood, there is a characteristic mean species frequency. The first of these assumptions has already been discussed. The second is not true in detail but is plausible enough at the relatively large scale of recording visits to the neighbourhood. Recorders visiting an area can search it in various ways, but they are always constrained by their inability to make a random search of the whole hectad. Thus, on any one visit, a recorder would not ordinarily expect to cover the whole range of habitats in the hectad, let alone search its whole area. This will apply to visits to neighbouring hectads just as much as to the target hectad. A more thorough study of recorder behaviour would no doubt reveal that visits come in many kinds, ranging from the quick examination of a wall to a full day's systematic recording of a range of habitats. Thus, the chance of a given species being discovered is the outcome of a two-stage stochastic process. The first stage concerns the type and duration of the visit, while the second concerns the frequency with which a given species is encountered during a visit of a certain type. When these processes are combined, each species will have a standard probability of being recorded on a visit. Under most assumptions about the nature of this two-stage process, the discovery of less common species will be a rare event, so that their records can indeed be taken as the outcome of a Poisson process.

The third assumption is at present based on limited data, namely that the untransformed neighbourhood frequency in well-recorded neighbourhoods was in the range $\varphi_i[1] = 0{\cdot}67$–$0{\cdot}75$. On closer inspection, many of these neighbourhoods were found to include some hectads that were not particularly well recorded. Thus, the standard value $\Phi = 0{\cdot}74$ was selected. In most applications, the exact choice of Φ is not critical. If frequencies follow the average curve found here, then lowering Φ to 0·7 reduces the expected number of species by 2·2%, while raising it to 0·9 increases the expected number by 7·6%. However, the question of whether the proportion of locally rare to locally common species is the same in different parts of the country deserves further investigation. The fact that observed species totals for well-sampled hectads Ben Cruachan in western Scotland, Fordingbridge in southern England and Boxworth in eastern England are in good agreement with predictions (Fig. 5) is encouraging but amounts to rather limited evidence.

Multipliers α_i for sampling effort are closely related to neighbourhood frequency $\varphi_i[1]$ (Fig. 7). The scale for $\varphi_i[1]$ is shown as a reciprocal to emphasize the discrepancy between $1/\varphi_i[1]$, which is the naïve estimate of α_i, and the relatively large values of α_i that apply in poorly recorded neighbourhoods. By definition, a sampling-effort multiplier of 1 corresponds to a neighbourhood frequency of Φ. The trend line in Fig. 7 is calculated from data for all 3854 neighbourhoods. It crosses the x-axis ($\alpha_i = 0$) at $1/\varphi_i[1] = 1{\cdot}164$. This implies a value $\varphi_i[1] = 0{\cdot}86$ if the neighbourhood were exhaustively sampled. This is not

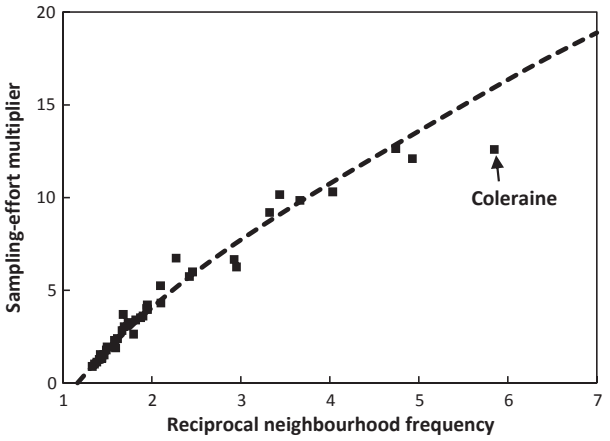

Fig. 7. Sampling-effort multipliers α_i in relation to reciprocal mean neighbourhood frequencies $1/\varphi_i[1]$. The trend line is a quartic curve calculated for the full hectad data set, not just for the 38 selected as a sample.

intrinsically unlikely. When rank–frequency curves and corresponding statistics were calculated for vascular plants, which are much better recorded than bryophytes, the average value of $\varphi_i[1]$ was 0·80 and the largest value, for Stoke Mandeville, was 0·86. This neighbourhood included 293 vascular plant species with local frequency exceeding 0·99, at which point $R'_{ij} = 0{\cdot}45$. Bryophytes with standardized local frequency exceeding 0·99 were relatively few (Fig. 3), and the boundary for such species on the average curve was $R'_{ij} = 0{\cdot}18$.

IMPROBABLE AND MISSING SPECIES

The estimated local frequency of the rarest species is raised by the inclusion of target hectads in their own neighbourhoods. Prendergast *et al.* (1993) excluded target hectads (which they called reference hectads) from their neighbourhoods. For the present study, target hectads were included because the resulting smoothed frequencies are used for other purposes, notably interpreting trends. It would be paradoxical for locally rare species such as *Amblystegium radicale* in the Rhinog Mountains to have a positive probability in nearby hectads but a zero probability in the target.

Calculated probability values are used for routine screening of bryophyte records received by the recording scheme of the British Bryological Society. Records that are new for their vice-county (Hill *et al.* 2008) are automatically queried. Records of locally scarce species are now also scrutinized and may be queried unless the species is already known from the hectad.

ESTIMATION OF TRENDS

The big advantage of the neighbourhood-frequency method in estimating trends is that it corrects for uneven spatial recording over time. Methods that are based on global frequency as a proportion of all records can be misleading if sampling is spatially uneven. Indeed, species can apparently decrease when they are actually increasing (an example is given in Data S1 in Supporting information). The advantage of the method over ones that use the same benchmark species throughout the territory is that local benchmark species are selected to be suitable for each neighbourhood.

Trend estimation depends on the choice of benchmark species, and therefore on the parameter R^*, the benchmark limit. In a sensitivity analysis, R^* was varied from 0·14 to 0·41. Time factors x_{jt} increased when R^* was larger. This is because sampling intensity s_{it} is measured by the proportion of benchmark species found in location i at time t. If R^* is increased, then relatively rarer species are added to the set of benchmark species, and s_{it} accordingly decreases. To make up for the fall in s_{it}, time factors x_{jt} need to be increased, roughly in proportion. However, the pattern of proportional change was almost completely unaffected by variation in R^*, with time factors merely being multiplied by a constant (see Data S1 in Supporting information).

In the introduction, it was asserted that benchmark species should be ones that are fairly stable over time. When the data were re-analysed after exclusion of strongly increasing or decreasing species from the list of possible benchmarks, the new time factors hardly differed from the old ones. This revised analysis is therefore not reported here, but with other data sets, the exclusion of such species might make a difference.

The basic model (2)

$$P_{ijt} = 1 - \exp(-Q_{ijt}x_{jt})$$

can be varied in many ways depending on the definition of Q_{ijt}. One variant is

$$P_{ijt} = 1 - \exp(-\lambda_{ij}s_{it}x_{jt}), \qquad \text{eqn 4}$$

which is based on multiplication of three parameters. λ_{ij} is the encounter rate for species j in hectad i; s_{it} is the recording effort in hectad i at time t; and x_{jt} is a time factor for species j. If species j is ubiquitous in the neighbourhood of hectad i, then $f_{ij} = f'_{ij} = 1$, and the conversion

$$\lambda_{ij} = -\log(1 - f'_{ij})$$

makes λ_{ij} infinite, so that $P_{ijt} = 1$, regardless of the values of the other parameters. In practice, no species is so common that it is sure to be found on a very short visit. Moreover, recorders do not always make complete lists and sometimes omit the commonest species. Thus, for analysing trends, the assumption of ubiquity should not be made, and model (3) may in fact be preferable to model (4).

RECORDING EFFORT AND DISCOVERY CURVES

In the background of much of the above discussion are three important concepts: visits, discovery curves and recording effort. Prendergast *et al.* (1993) corrected for recording effort by means of discovery curves for individual visits. These were obtained by rarefaction. In the present study, recording intensity for an individual locality and time period is measured by the proportion of benchmark species found at that time.

Discovery is assumed to be a random process, derived by sampling a neighbourhood-frequency distribution.

Species lists from individual visits provide information at a higher resolution than the summarized hectad data analysed here. When data from individual visits are available, the total number of records can be treated as a measure of recording effort (Ball & Henshall 2006). This measure is rarely available for historical data sets. Until about 1990, many vascular plant and bryophyte records were summarized on master cards, which did not distinguish visits. Even in more recent years, botanists have sometimes summarized their local flora using tetrads (2-km squares) or quadrants (5-km squares) and have not necessarily distinguished individual visits within these units. Therefore, the summary of occurrences by hectad and quinquennium to underpin the analysis for Fig. 6 may give the best detail that could be recovered without serious loss of data.

Except in comprehensive and systematic surveys, the problem of variable habitat coverage cannot be ignored. For example, there may be a targeted habitat survey such as the Scottish Loch Survey 1984–1994, data from which were incorporated in the BRC data base in the late 1990s (Preston & Croft 1997). This resulted in aquatic plants seeming to increase in Scotland (Preston *et al*. 2003). Likewise, the Survey of Bryophytes of Arable Land (Preston *et al*. 2010) resulted in a large increase in bryophyte records from this habitat during the period 2001–2005. Most such biases are less obvious, with recorders during a particular period tending, for example, to avoid towns and villages or being reluctant to record non-native plants.

A fruitful topic for future study will be to understand and correct for such changes in recorder behaviour. This will require individual visits to be analysed in more detail, distinguishing the depth of recording within visits and the coverage achieved by the overall spread of visits. This should allow the probability of species being found to be modelled better than by model 3 above. There is no necessity to use benchmark species. For example, if a list of 10 rare species is received from a visit, then the recorder has clearly omitted to make a full list or has selected a subset of what was found for a museum or private herbarium. Such partial lists can provide information. Given the habitat and local frequency of the species in such a list, the probability of a particular species being included could in principle be estimated, but that would require a much more complicated model than the one used here.

Acknowledgements

This study was funded jointly by the Joint Nature Conservation Committee (JNCC) and the Natural Environment Research Council as a part of the Biological Records Centre (BRC) work programme for 2010. I am grateful to Chris Cheffings (JNCC), David Roy (BRC) and Lawrence Way (JNCC) for their encouragement and to the British Bryological Society and the Botanical Society of the British Isles for assembling such a wealth of data. Colin Harrower of BRC helped with data base queries and prepared the map of neighbourhood weights, which was plotted using the DMAP program written by Dr Alan Morton. My colleagues Nick Isaac, Tom Oliver and Chris Preston read over and commented on draft reports as the work progressed. Finally, I would like to thank five referees and an associate editor for valuable and constructive comments on earlier drafts.

References

Asher, J., Warren, M., Fox, R., Harding, P., Jeffcoate, G. & Jeffcoate, S. (2001) *The Millennium Atlas of Butterflies in Britain and Ireland*. Oxford University Press, Oxford.

Ball, S. & Henshall, M. (2006) Using data to interpret changes in the UK's biodiversity. *Bulletin of the British Ecological Society*, **37**, 51–54.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Ding, C.Q., Clark, N.E., O'Connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *Plos Biology*, **8**(6), e1000385.

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

Dennis, R.L.H. & Thomas, C.D. (2000) Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. *Journal of Insect Conservation*, **4**, 73–77.

Free Software Foundation (1999) *GNU Fortran G77 v0.5.25*. Downloaded 16 December 2010 from http://kkourakis.tripod.com/g77.htm.

Gibbons, D.W., Reid, J.B. & Chapman, R.A. (1993) *The New Atlas of Breeding Birds in Britain and Ireland: 1988–1991*. T. & A.D. Poyser, London.

Hassall, C. & Thompson, D.J. (2010) Accounting for recorder effort in the detection of range shifts from historical data. *Methods in Ecology and Evolution*, **1**, 343–350.

Hedenäs, L., Bisang, I., Tehler, A., Hamnede, M., Jaederfelt, K. & Odelvik, G. (2002) A herbarium-based method for estimates of temporal frequency changes: mosses in Sweden. *Biological Conservation*, **105**, 321–331.

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

Hill, M.O., Preston, C.D. & Smith, A.J.E. (1991–1994) *Atlas of the Bryophytes of Britain and Ireland, Vol. 1 (1991), Vol. 2 (1992), Vol. 3 (1994)*. Harley Books, Colchester.

Hill, M.O., Blackstock, T.H., Long, D.G. & Rothero, G.P. (2008) *A Checklist and Census Catalogue of British and Irish Bryophytes Updated 2008*. British Bryological Society, Middlewich.

Hofmann, H., Urmi, E., Bisang, I., Müller, N., Küchler, M., Schnyder, N. & Schubiger, C. (2007) Retrospective assessment of frequency changes in Swiss bryophytes over the last two centuries. *Lindbergia*, **32**, 18–32.

Hortal, J. (2008) Uncertainty and the measurement of terrestrial biodiversity gradients. *Journal of Biogeography*, **35**, 1335–1336.

Huntley, B., Berry, P.M., Cramer, W. & McDonald, A.P. (1995) Modelling present and potential future ranges of some European higher plants using climate response surfaces. *Journal of Biogeography*, **22**, 967–1001.

Jeppsson, T., Lindhe, A., Gärdenfors, U. & Forslund, P. (2010) The use of historical collections to estimate population trends: a case study using Swedish longhorn beetles (Coleoptera: Cerambycidae). *Biological Conservation*, **143**, 1940–1950.

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

Maes, D. & Swaay, C.A.M.v. (1997) A new methodology for compiling national Red Lists applied to butterflies (Lepidoptera, Rhopalocera) in Flanders (N-Belgium) and the Netherlands. *Journal of Insect Conservation*, **1**, 113–124.

Magnussen, S., Smith, B., Kleinn, C. & Sun, I.F. (2010) An urn model for species richness estimation in quadrat sampling from fixed-area populations. *Forestry*, **83**, 293–306.

Newson, S.E., Woodburn, R.J.W., Noble, D.G., Baillie, S.R. & Gregory, R.D. (2005) Evaluating the breeding bird survey for producing national population size and density estimates. *Bird Study*, **52**, 42–54.

Owen, J. (2010) *Wildlife of a Garden: A Thirty-year Study*. Royal Horticultural Society, London.

Petrik, P., Pergl, J. & Wild, J. (2010) Recording effort biases the species richness cited in plant distribution atlases. *Perspectives in Plant Ecology, Evolution and Systematics*, **12**, 57–65.

Prendergast, J.R., Wood, S.N., Lawton, J.H. & Eversham, B.C. (1993) Correcting for variation in recording effort in analyses of diversity hotspots. *Biodiversity Letters*, **1**, 39–53.

Preston, C.D. & Croft, J.M. (1997) *Aquatic Plants in Britain and Ireland*. Harley Books, Colchester.

Preston, C.D., Pearman, D.A. & Dines, T.D. (2002) *New Atlas of the British and Irish Flora*. Oxford University Press, Oxford.

Preston, C.D., Telfer, M.G., Arnold, H.R., Carey, P.D., Cooper, J.M., Dines, T.D., Hill, M.O., Pearman, D.A., Roy, D.B. & Smart, S.M. (2002) *The Changing Flora of the UK*. DEFRA, London.

Preston, C.D., Telfer, M.G., Roy, D.B., Carey, P.D., Hill, M.O., Meek, W.R., Rothery, P., Smart, S.M., Smith, G.M., Walker, K.J. & Pearman, D.A.

(2003) *The Changing Distribution of the Flora of the United Kingdom: Technical Report*. CEH Monks Wood, Huntingdon.
Preston, C.D., Hill, M.O., Porley, R.D. & Bosanquet, S.D.S. (2010) Survey of the bryophytes of arable land in Britain and Ireland 1: a classification of arable field assemblages. *Journal of Bryology*, **32**, 61–79.
Ratcliffe, D.A. (1968) An ecological account of Atlantic bryophytes in the British Isles. *New Phytologist*, **67**, 365–439.
Rich, T.C.G. & Smith, P.A. (1996) Botanical recording, distribution maps and species frequency. *Watsonia*, **21**, 155–167.
Roy, D.B., Rothery, P. & Brereton, T. (2007) Reduced-effort schemes for monitoring butterfly populations. *Journal of Applied Ecology*, **44**, 993–1000.
Silvertown, J. (2009) A new dawn for citizen science. *Trends in Ecology & Evolution*, **24**, 467–471.
Telfer, M.G., Preston, C.D. & Rothery, P. (2002) A general method for the calculation of relative change in range size from biological atlas data. *Biological Conservation*, **107**, 99–109.

Received 16 September 2010; accepted 24 May 2011
Handling Editor: David Murrell

Supporting Information

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

Data S1. Tests and sensitivity analysis for Frescalo output.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.