Downloaded from rspb.royalsocietypublishing.org on March 28, 2012

PROCEEDINGS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Predicting unknown species numbers using discovery curves

Daniel P Bebber, Francis H.C Marriott, Kevin J Gaston, Stephen A Harris and Robert W Scotland

Proc. R. Soc. B 2007 **274**, 1651-1658
doi: 10.1098/rspb.2007.0464

References This article cites 21 articles, 3 of which can be accessed free
http://rspb.royalsocietypublishing.org/content/274/1618/1651.full.html#ref-list-1

Article cited in:
http://rspb.royalsocietypublishing.org/content/274/1618/1651.full.html#related-urls

Email alerting service Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to *Proc. R. Soc. B* go to: http://rspb.royalsocietypublishing.org/subscriptions

Downloaded from rspb.royalsocietypublishing.org on March 28, 2012

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2007) **274**, 1651–1658
doi:10.1098/rspb.2007.0464
Published online 24 April 2007

Predicting unknown species numbers using discovery curves

Daniel P. Bebber[1], Francis H. C. Marriott[2], Kevin J. Gaston[3], Stephen A. Harris[1] and Robert W. Scotland[1,*]

[1] *Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX3 0EX, UK*
[2] *Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK*
[3] *Department of Animal & Plant Sciences, The University of Sheffield, Sheffield S10 2TN, UK*

A common approach to estimating the total number of extant species in a taxonomic group is to extrapolate from the temporal pattern of known species descriptions. A formal statistical approach to this problem is provided. The approach is applied to a number of global datasets for birds, ants, mosses, lycophytes, monilophytes (ferns and horsetails), gymnosperms and also to New World grasses and UK flowering plants. Overall, our results suggest that unless the inventory of a group is nearly complete, estimating the total number of species is associated with very large margins of error. The strong influence of unpredictable variations in the discovery process on species accumulation curves makes these data unreliable in estimating total species numbers.

Keywords: birds; ants; mosses; monilophytes; lycophytes; gymnosperms

1. INTRODUCTION

Current knowledge of global species richness is based on a small to moderate fraction of all extant species. The actual size of that fraction, and thus the magnitude of global biodiversity, has been much debated (May 1988, 1990, 2000; Gaston 1991, in press; Hammond 1995; Hawksworth 2001; Lambshead & Boucher 2003) and at times the issue has received substantial media attention. Our ability to provide an accurate answer has been argued to provide a valuable test of understanding the structure and composition of global biodiversity, the answer itself to be a basic fact that we should know about the world around us, and one that will facilitate better answers to other important questions, such as the extent to which the carrying capacity of diversity on Earth has been attained, the rate at which species are becoming globally extinct and the scale of the task faced by species conservation (see Gaston (in press) for a review).

The key issue arising from the uncertainty over global species numbers is that of how to extrapolate from the fraction of species that is known to science. By far, the most popular method has involved the use of species discovery curves to estimate the number of species that remain to be discovered in a given taxonomic group, globally or regionally (Steyskal 1965; Frank & Curtis 1979; Soberon & Llorente 1993; Scoble *et al.* 1995; Medellin & Soberon 1999; Ertter 2000; Aravind *et al.* 2004; Gower 2004; Shen Tsung-Jen & Chih-Feng 2003; Solow & Smith 2005; Wilson & Costello 2005; Pimm *et al.* 2006). This approach involves plotting a cumulative frequency curve for the taxon with the expectation that this becomes asymptotic when the inventory is reaching completion and new species are becoming more and more difficult to find. Unfortunately, estimates of species numbers derived in this way tend to lack associated error margins (Steyskal 1965; Frank & Curtis 1979; Soberon & Llorente 1993; Solow & Smith 2005), making it impossible to objectively assess their accuracy. Perhaps in consequence, while some have expressed severe reservations about the application of this approach (Hammond 1995), others have attempted to place it on a more secure statistical foundation (Solow & Smith 2005; Wilson & Costello 2005).

To investigate the confidence with which predictions of total species numbers can be made from species discovery curves, and to determine the effect of incompleteness of discovery curves on confidence limits of predictions, we have compiled and examined datasets for birds of the world and UK flowering plants, which are assumed to be more or less complete. We have also compiled datasets for ants, mosses, lycophytes, monilophytes (ferns and horsetails), gymnosperms and New World grasses as a sample of other major taxa. We developed a generalized linear model for analysing and interpreting the dynamics of these species discovery curves, which provides both point estimates and confidence limits for the number of unknown species using analysis of deviance.

2. MATERIAL AND METHODS

(a) *Datasets*

Datasets of accepted species and their date of publication were assembled for birds of the world, UK flowering plants, ants of the world, mosses of the world, lycophytes of the world, monilophytes (ferns and horsetails) of the world, gymnosperms of the world and New World grasses. The date of the basionym of the accepted name was used where available, namely for UK flowering plants, mosses, lycophytes, monilophytes and gymnosperms. Otherwise, the date used corresponded to the date of publication of the accepted name, namely for birds, ants and New World grasses. The data for birds were supplied by Alan Peterson as a download

[*] Author for correspondence (robert.scotland@plants.ox.ac.uk).

Received 23 February 2007
Accepted 3 April 2007

This journal is © 2007 The Royal Society

Downloaded from rspb.royalsocietypublishing.org on March 28, 2012

from http://www.zoonomen.net for the period between 1758 and 2004. UK plant data were compiled from the Oxford University Herbaria database (http://herbaria.plants.ox.ac.uk/bol/?oxford) based on Kent (1992) and Stace (1997) for the period between 1753 and 2005. The ant data were supplied by Donat Agosti as a download from Antbase for the period between 1750 and 2006 (Agosti & Johnson 2005). Moss data were supplied by Marshall Crosby as a download from Crosby *et al.* (2006) for the period between 1753 and 2004. Monilophytes and lycophytes were extracted from World Ferns on CD-ROM for the period between 1753 and 2000 (Hassler & Swale 2001). Data on gymnosperms were compiled from Farjon (2001), World Checklist of Cycads (http://plantnet.rbgsyd.gov.au/PlantNet/cycad/wlist.html) and the TROPICOS database (http://mobot.mobot.org/W3T/Search/vast.html) for the period between 1753 and 2005. New World grasses were supplied by Gerrit Davidse as a download from http://mobot.mobot.org/Pick/Search/nwgc.html for the period between 1753 and 2006.

(b) *The model*

It is assumed that species identification and group membership are generally agreed, and that the total number of species in the group, N_{tot}, is fixed. The problem is to estimate N_{tot}, or equivalently the number hitherto undiscovered. We postulate that the expected number of species discovered in time t, S_t, is some fraction k of the number of undiscovered species at time $t-1$,

$$E(S_t) = k(N_{\text{tot}} - N_{t-1}).$$

The coefficient k depends on several interacting factors, including the effort expended in discovering new species, the visibility of the undiscovered species, the expertise in identifying new species and the proportion of habitat remaining unexplored. We have no independent estimates of how these factors vary through time, but we might suppose that k would decrease if more obvious species are discovered first, and conversely that k would increase with increasing discovery effort and smaller areas of unexplored habitat.

As we have no independent means of separating the effects of discovery effort, the visibility of undiscovered species, etc., we propose fitting the simplest model in which systematic variation in these factors is low compared with the effect of diminishing new species, and k is therefore roughly constant. If a plot of S_t against N_{t-1}, smoothed by local regression or spline interpolation, shows a more or less linear trend, then this is a justification for fitting a generalized linear model of S_t on N_{t-1} to the observations in this linear period. The model then has the form of a linear regression of S_t on N_{t-1}, with intercept $k\hat{N}_{\text{tot}}$ and slope $-k$. The point estimate is then minus the intercept over the slope, i.e. the value of N_{t-1} for which S_t is zero.

S_t could show a nonlinear decline with N_{t-1}. The model could be altered to include, for example, k as a linear function of N_{t-1}, e.g. $k=b+cN_{t-1}$. In this case, $E(S_t)=kN_{\text{tot}}-(b+cN_{t-1})N_{t-1}$. This variation could model both increases and decreases in discovery rates. Once the model has been fitted, this quadratic equation could be solved for $S_t=0$, giving the point estimate $\hat{N}_{\text{tot}}$ as the smallest, positive, real root. Other functions of k are also possible. If S_t increases with N_{t-1}, then increases in discovery effort dominate the decline in the number of species remaining to be discovered, and estimating N_{tot} is impossible.

If species discoveries were independent of one another, the model would have poisson error and identity link. However, in practice, the data are likely to show overdispersion, the residual deviance being greater than the corresponding degrees of freedom. The discovery time is usually defined as the date of the first published description, and such descriptions tend to appear in groups such as monographs and other books (Wilson & Costello 2005; Bebber *et al.* 2007). This type of model is described by McCullough & Nelder (1989) and Venables & Ripley (2002); it is often referred to as a quasi-Poisson model. The model also gives an estimate of the scale factor, or dispersion, of the quasi-Poisson distribution. This is a measure of the overdispersion, being 1 for Poisson errors and larger when data tend to be grouped. The scale factor is best estimated as the residual Pearson χ^2 divided by its degrees of freedom, rather than the mean residual variance (Venables & Ripley 2002).

It is then straightforward to derive confidence limits for N_{tot}. Define a new variable $R=\hat{N}_{\text{tot}}+M-N_{t-1}$, where M can be positive or negative. Fit a generalized linear model for S_t against R with Poisson error and Identity link *without intercept*. This forces the regression through $\hat{N}_{\text{tot}}+M$. The model will give the same residual deviance as the best fit when M is zero, and a larger residual deviance as M diverges from zero. Changes in deviance scaled by the dispersion have an F distribution (McCullough & Nelder 1989), allowing calculation of confidence limits. The fit has one degree of freedom more in the residual term, and thus if the deviance ratio is significant (i.e. if it is greater than 3.84), then—lies outside the 95% CI. A search gives the upper and lower limits for which this happens. Note that if the negative slope is not well defined, the upper limit may be infinity. This model is closely related to Fieller's (1954) theorem, which can also give infinite CIs if the data are uninformative.

Species discovery curves show a variety of trajectories, from those that appear to be increasing exponentially, such as the New World grasses (figure 1*a*), to those that appear to have reached an asymptote, for example, birds of the world (figure 1*h*). By plotting S_t against N_{t-1} and fitting smoothing splines to these data, changes in the discovery rate k can be followed (figure 2). Successful prediction using the model requires that the slope of S_t against N_{t-1} be negative and constant, such that k is positive and constant. If the slope is zero or positive, this shows that more species are being discovered than expected by the model. Zero or positive slopes indicate increases in discovery effort or rates of description, or some other process unrelated to N_{tot}. Smoothing splines were therefore used to identify regions of the data with negative slopes, where model fitting would give sensible predictions.

3. RESULTS

In all cases except for the British flora, S_t increased with N_{t-1} for the early discoveries (figure 2*a–h*). This was reflected in an exponential increase in N_t over time (figure 1*a–h*), and meant that for all groups except the British flora and birds of the world, bounded confidence limits on N could not be estimated when the entire dataset was included (table 1). For all groups except New World grasses (figure 2*a*), the discovery rate declined after this initial ‘start-up’ period. The New World grasses were omitted from further analyses, as there was no indication of decline in S_t, and subsequently no point estimate of N_{tot} could be made. Subsets of the data for the other groups were then fitted, which omitted the early increasing S_t phase.

Downloaded from rspb.royalsocietypublishing.org on March 28, 2012

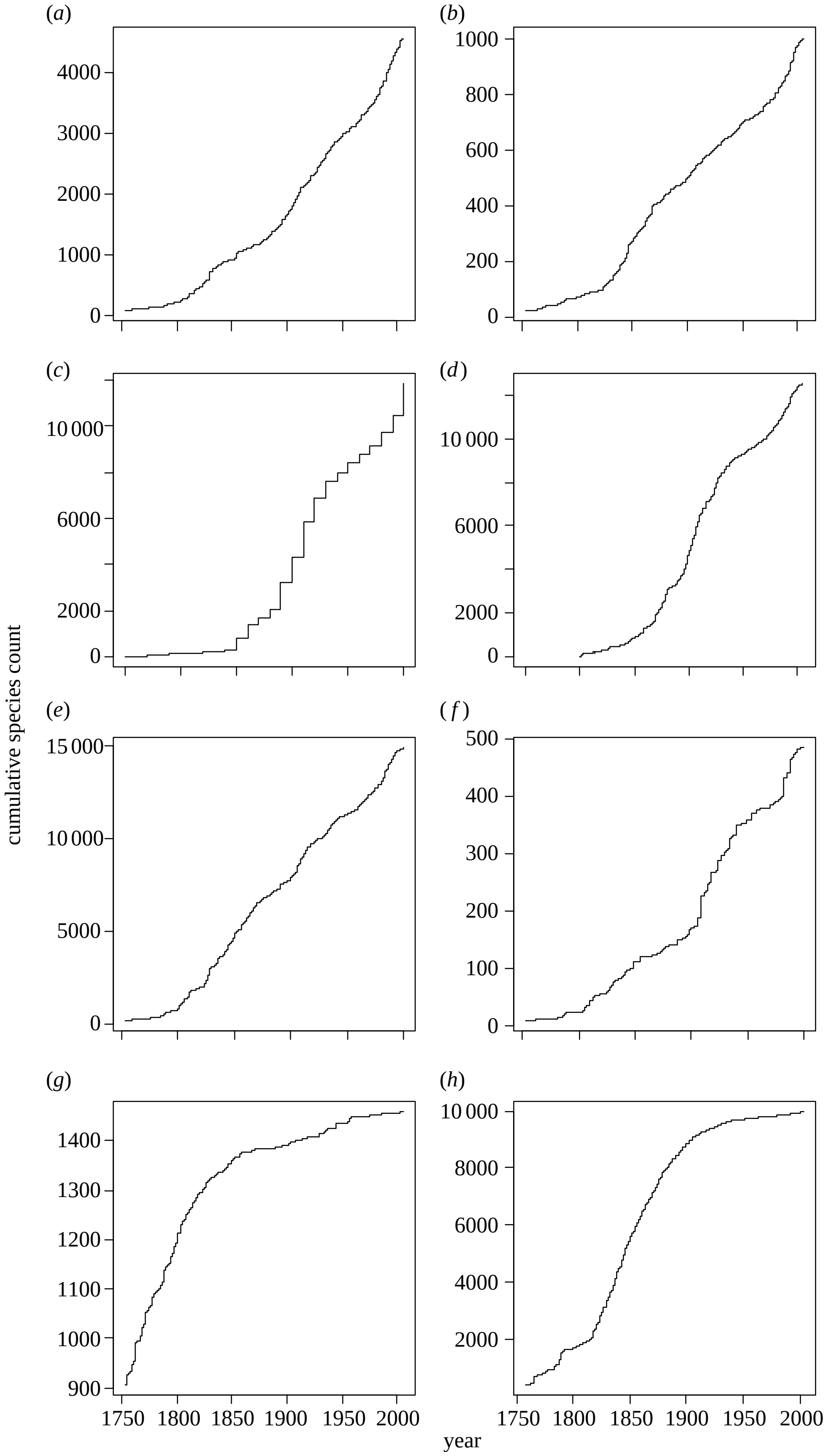

Figure 1. Cumulative species count against year for (*a*) New World grasses, (*b*) gymnosperms, (*c*) ants (by decade), (*d*) mosses, (*e*) ferns, (*f*) lycopods, (*g*) British flora and (*h*) birds of the world.

Gymnosperms, ants and mosses all showed a similar discovery rate dynamic (figure 2*b*–*d*). Although S_t declined after an initial increase, this was followed by another increase for the most recently discovered species. Model fits that included only the central declining S_t phase gave bounded confidence limits for N_{tot} (table 1). However, inclusion of the most recently discovered species either gave very large, or unbounded, confidence limits (table 1).

Ferns and lycopods show a slightly different pattern (figure 2*e*,*f*). In these groups, S_t remained roughly constant, precluding estimation of N_{tot} even when the increasing S_t phase was omitted. Bounded confidence limits could be obtained for the latest 10–20% of discoveries, however, owing to recent declines in discovery rates (table 1). In other words, predictions at some point in the past would have been impossible. For ferns, the upper

Downloaded from rspb.royalsocietypublishing.org on March 28, 2012

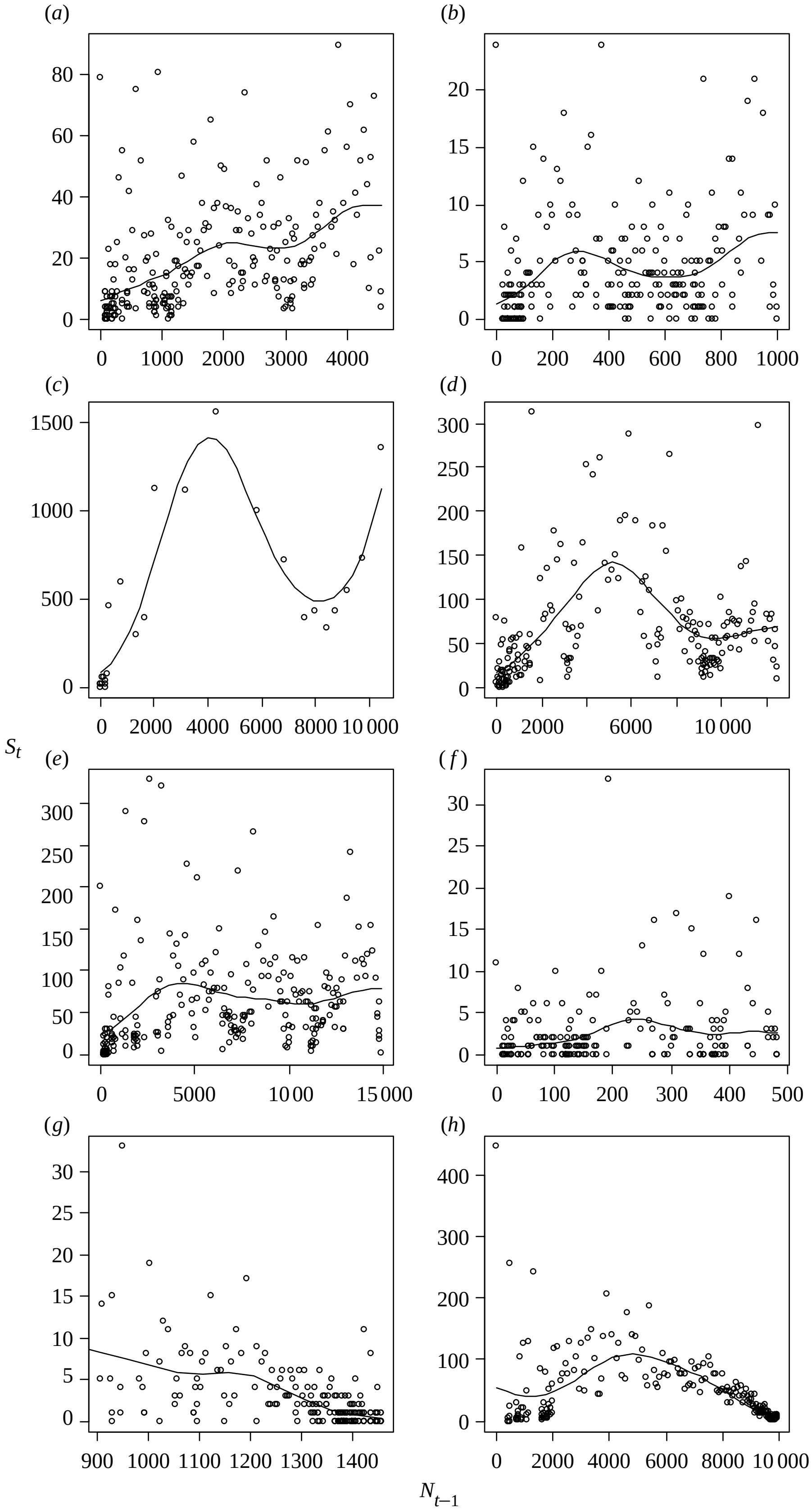

Figure 2. Species discovered per year (S_t) against cumulative species count to previous year (N_{t-1}) for (*a*) New World grasses, (*b*) gymnosperms, (*c*) ants (by decade), (*d*) mosses, (*e*) ferns, (*f*) lycopods, (*g*) British flora (omitting 1753 data) and (*h*) birds of the world.

95% confidence limit of the best estimate that includes the most recent data was only 271 species more than the current total of 14 891 species (table 1). For lycopods, the upper limit was 22 species more than the current total of 484 species (table 1).

The best-behaved groups, in terms of long-term declines in S_t, were the British flora and birds of the world (figure 2*g*,*h*). For the British flora, the first data point (Linnaeus) was omitted from the model fit as it contained many more species than any other point. Fitting

Downloaded from rspb.royalsocietypublishing.org on March 28, 2012

Table 1. Estimates and 95% confidence limits for total species number based on subsets of data for eight groups. (Models were fitted with constant k. Data range gives the lower and upper proportions of known species included in the model. Date range is the range of years in which the subset was discovered. N is the number of species contained in the subset. Dispersion is the estimated dispersion parameter for the quasi-Poisson error distribution. Best estimates (for which the 95% confidence limits were the narrowest) of N_{tot} for subsets that include the most recent data, are italicized. No best estimate is given where a bounded 95% confidence limit could not be found.)

data range	date range	N	dispersion	estimate	95% CI
New World grasses					
0.0–1.0	1753–2006	4559	n.a.	n.a.	n.a.
gymnosperms					
0.0–1.0	1753–2006	1004	n.a.	n.a.	n.a.
0.3–0.6	1856–1924	294	3.6	1022	704–n.a.
0.25–0.75	1848–1969	480	3.7	1001	958–n.a.
0.2–0.8	1843–1980	595	3.7	1296	1001–2883
0.2–0.9	1843–1994	695	4.0	3837	1554–n.a.
0.1–1.0	*1825–2006*	*894*	*4.3*	*202 317*	*2446–202 317*
ants[a]					
0.0–1.0	1750–2000	11 827	n.a.	n.a.	n.a.
0.4–0.8	1910–1950	3677	11.5	9175	8585–10 460
0.3–0.9	1900–1980	5597	64.4	11 326	9966–16 112
0.4–0.9	1910–1980	4879	64.4	10 921	9760–15 720
mosses					
0.0–1.0	1800–2004	12 503	n.a.	n.a.	n.a.
0.4–0.8	1902–1971	5081	23.5	10 937	10 468–11 790
0.4–1.0	1902–2004	7501	38.4	17 977	14 896–29 841
0.31–1.0	*1897–2004*	*8483*	*39.6*	*16 330*	*14 327–21 333*
ferns					
0.0–1.0	1753–2000	14 891	n.a.	n.a.	n.a.
0.3–1.0	1849–2000	10 418	34.7	134 795	31 629–n.a.
0.9–1.0	1984–2000	1373	15.7	15 191	14 952–15 966
0.95–1.0	*1990–2000*	*707*	*12.5*	*14 933*	*14 890–15 162*
lycopods					
0.0–1.0	1753–2000	484	n.a.	n.a.	n.a.
0.5–1.0	1915–2000	241	7.0	859	526–n.a.
0.8–1.0	1974–2000	95	5.9	520	484–n.a.
0.82–1.0	*1981–2000*	*83*	*2.9*	*486*	*484–506*
British flora[b]					
0.6–1.0	1754–2006	552	3.3	1469	1459–1489
0.6–0.9	1754–1827	408	4.1	1651	1419–n.a.
0.6–0.8	1754–1796	264	5.6	1490	1221–n.a.
0.68–1.0	*1765–2006*	*465*	*3.0*	*1469*	*1459–1488*
birds					
0.0–1.0	1758–2004	9961	41.6	11 997	10 979–13 847
0.2–1.0	1816–2004	7953	8.3	10 205	10 102–10 336
0.4–1.0	1838–2004	5847	4.4	10 077	10 028–10 139
0.6–1.0	1856–2004	3979	3.1	10 049	10 013–10 095
0.8–1.0	1883–2004	1974	2.6	10 030	9998–10 072
0.7–0.9	1869–1904	1990	2.8	10 028	9998–10 067
0.4–0.8	1838–1882	3873	8.1	11 985	10 169—17 175
0.75–1.0	*1876–2004*	*2464*	*2.7*	*10 023*	*9994–10 061*

[a] Ant data are grouped by decade rather than year.

[b] Data for British flora omit Linnaeus' 1753 descriptions which contain 907 species.

the remaining data gave very small confidence limits for N_{tot}, but omission of just the most recent 10% of discoveries lead to unbounded confidence limits (table 1). The best estimate for British flora gave 95% confidence limits of 1459–1488 species, with a current known total of 1458 species. Implementing k as a linear function of N_{t-1} (the 'varying k' model) gave a best estimate of 1493 species with 95% confidence limits of 1459–1559 species, using the 131 most recent discoveries (table 2).This interval is wider than the best estimate from the constant k model. The varying k model was able to give bounded, though wide, confidence limits when the most recent 10% of discoveries were omitted (table 2). Omission of more than 10% of the most recent discoveries made prediction impossible.

For the birds of the world, omission of earlier data gave progressively smaller estimates of the dispersion index, and tighter confidence limits on N_{tot} (table 1). However, omission of just a few late discoveries widened the confidence limits dramatically. Once again, prediction in the absence of just a few species was impossible. The 95% confidence limits on the best estimate for birds were 9994–10 061, with a current total of 9968 species. This estimate included only the most recent 25% of species in

Downloaded from rspb.royalsocietypublishing.org on March 28, 2012

Table 2. Estimates and 95% confidence limits for total species number with k as a linear function of N_{t-1}. (Only British flora and birds of the world are shown, as the other groups showed no decline in S_t.)

data range	date range	N	dispersion	estimate	95% CI
British flora[a]					
0.6–1.0	1754–2004	552	n.a.	n.a.	n.a.
0.9–1.0	1828–2006	144	3.1	1502	1459–2222
0.6–0.9	1754–1827	408	4.2	1526	1343–2178
0.6–0.8	1754–1796	264	n.a.	n.a.	n.a.
0.91–1.0	1835–2006	131	3.2	1493	1459–1558
birds					
0.0–1.0	1758–2004	9961	47.9	10 024	9957–n.a.
0.6–1.0	1856–2004	3979	n.a.	n.a.	n.a.
0.8–1.0	1883–2004	1974	2.4	10 118	10 032–19 988
0.6–0.9	1856–1904	3003	5.6	10 316	9605–15 734
0.83–1.0	*1889–2004*	*1665*	*2.3*	*10 298*	*10 067–13 115*

[a] Data for British flora omit Linnaeus' 1753 descriptions which contain 907 species.

the model. Use of the varying k model did not lead to improvements in prediction over the constant k model (table 2).

4. DISCUSSION

(a) *Methodological issues*

The modelling of species discovery curves and the prediction of total species numbers are complicated by three features of the data. The model we have proposed, and the analyses we have conducted, explicitly address these features. Firstly, the discovery rate is governed not only by the number of species remaining to be found, but also by the effort employed in finding and reporting them. The variability of discovery effort is best illustrated by the early exponential increase in discoveries, which is independent of the number of species remaining to be found. All but one of the datasets presented here contain this feature. The UK plants dataset is anomalous because Linnaeus described more than 800 species in 1753. This dataset therefore does not suffer from the problem of erratic early data collection. Wilson & Costello (2005) attempted to model early rate increases using logistic curves. However, because these early discoveries are largely uninformative of N_{tot}, there seems to be little reason to include them. In the worst case, their inclusion could bias estimates of N_{tot} or overstate the informativeness of the dataset. We found no support for the use of models of varying k, and would instead recommend limiting the data to subsets in which S_t declines linearly with N_{t-1}.

The second problematic feature of the data is also due to variability in discovery effort, namely the occurrence of false plateaux that leads to underestimates of N_{tot}. The ant and moss datasets demonstrate this issue. Both curves apparently begin to flatten at approximately 80% of the current total. However, for both ants and mosses, the subsequent rate of discovery increases, and analyses that include these most recent data cannot provide upper confidence limits for N_{tot}.

The third feature of the data regards the error distribution. Solow & Smith (2005) regard discovery dates as having independent Poisson errors, while Wilson & Costello (2005) recognize that the assumption of independence cannot be maintained for these data. Estimates of the dispersion parameter for our data are much greater than unity, and assumption of Poisson errors would lead to underestimates of the range of the confidence limits on N_{tot}. The model also avoids problems of time-series autocorrelations, as the fits are not functions of time.

(b) *Predictions*

The approach of using virtually completed curves for birds and UK flowering plants has demonstrated that our model can yield predictions with a high degree of confidence, but only when the vast majority of species are already described. The bird dataset shows that the earliest 50% of the data are not helpful in prediction, but thereafter, subsets including the most recent discoveries provide consistent estimates of N_{tot}. Even when discovery rates show a long period of decline, predictions from incomplete datasets can be highly uncertain. For birds, omission of the last 10–20% of species greatly increases the confidence limits. Analysis of the whole UK plants dataset provides a realistic estimate with small confidence limits; however, if 90% of the total dataset is used, then no upper confidence limit can be set on N_{tot}.

We have presented datasets for most major lineages of land plants with the exception of hornworts and liverworts. For mosses, lycophytes, monilophytes (ferns and horsetails) and gymnosperms, we have complete world coverage. These data are not available for angiosperms, so we have used New World grasses as a surrogate for angiosperms. Our results for these five datasets demonstrate that although plants are generally considered as relatively well known, there is no evidence that any of the major lineages of land plants are reaching or nearing an asymptote, with the exception of monilophytes and lycophytes. For both these lineages, there is a point estimate and small associated error only if a very recent subset of the data are used (5% for monilophytes and 18% for lycophytes), but this is over such a short period of time that it is impossible to distinguish an asymptotic curve from what might be a false plateau. The problem of false plateaux is well illustrated by the ant data that yield small-bounded estimates between the 40th and 80th percentile, but subsequent discoveries, i.e. the most recent 20%, clearly show this to be misleading. We consider New World grasses to be a fair surrogate for all angiosperms due to its size and the extent of its geographical distribution covering many distinct habitats. Nonetheless, for New World grasses, we interpret the fact that there is no

Downloaded from rspb.royalsocietypublishing.org on March 28, 2012

indication of decline in S_t, and subsequently no point estimate of N_{tot}, as a strong indication that the inventory of angiosperms is far from nearing completion.

5. CONCLUSION

In conclusion, these results are significant in two respects. First, unless an inventory is more or less complete (e.g. 90% complete for birds), extrapolations based on existing data are associated with very large margins of error. This, in addition to issues relating to synonymy, partly explains current levels of uncertainty about species numbers even for relatively well-known taxa such as plants (Scotland & Wortley 2003; Wortley & Scotland 2004). Unfortunately, the completeness of an inventory cannot be known until all species have been found. Second, any extrapolation from existing data is sensitive to the dynamics of the discovery process over time, as well as to the proportion of known species used in the extrapolation. It is clear that species discovery curves, governed both by the number of species remaining to be discovered, and by the vagaries of discovery effort, are largely unable to provide statistically rigorous estimates of total species numbers in a group, unless long periods with near-zero discovery have elapsed. Changes in discovery effort appear to be arbitrary and are unlikely to be predictable, thereby apparent plateaux in discovery curves cannot be relied upon to indicate the final approach to completeness of the inventory. Even when data are well-behaved, confidence limits on N become very large when just a few of the most recently discovered species are omitted.

Recent literature (Solow & Smith 2005; Wilson & Costello 2005) including this paper, attempts to place analyses of species discovery curves on a more secure statistical foundation by proposing improved models, dealing with the error distribution and making the interpretation of results more transparent. We consider that the approach used here focuses on the essential element of flattening of these curves when species become harder to find and deals appropriately with the error distribution. In addition, plotting the data as number of species discovered per year versus number discovered up to that year (figure 2) reveals the noisy and unpredictable nature of the discoveries, which may be obscured by traditional accumulation curves (figure 1).

Our results suggest that prediction for incomplete datasets is problematic because, unless a curve has flattened for some considerable time, it contains little appropriate information. There are many reasons why species continue to be described for many taxa such as the use of new analytical techniques, new species concepts, new areas of the world being explored, publication of a long-term monographic study, etc. Thus, even for apparently completed curves, it only takes effort in any one of these variables to discover new species. This suggests that prediction using discovery curves for incomplete groups is largely futile. We suggest that biologists shift focus from species discovery curves to other methods (Gaston in press) that are immune to the problems caused by temporal variations in the discovery process.

We thank Donat Agosti, Marshall Crosby, Gerry Davidse and Alan Peterson for making data available. We thank Donat Agosti, Bob May, Alan Peterson and two anonymous reviewers for comments on various drafts of the manuscript.

REFERENCES

Agosti, D. & Johnson, N. F. 2005 Antbase. See http://antbase.org/.

Aravind, N. A., Shaanker, R. U. & Ganeshaiah, K. N. 2004 Croak, croak, croak: are there more frogs to be discovered in the Western Ghats? *Curr. Sci.* **86**, 1471–1472.

Bebber, D. P., Harris, S. A., Gaston, K. J. & Scotland, R. W. 2007 Ethnobotany, plant discovery and the first written records of UK flowering plants. *Glob. Ecol. Biogeogr.* **16**, 103–108. (doi:10.1111/j.1466-8238.2006.00266.x)

Crosby, M. R., Magill, R. E., Allen, B. & He, S. 2006 *A checklist for the mosses*. St Louis, MO: Missouri Botanical Garden.

Ertter, B. 2000 Floristic surprises in North America north of Mexico. *Ann. Mo. Bot. Gard.* **87**, 81–109. (doi:10.2307/2666211)

Farjon, A. 2001 *World checklist and bibliography of conifers*. Kew, UK: Royal Botanic Gardens.

Fieller, E. C. 1954 Some problems in interval estimation. *J. Roy. Stat. Soc. B* **16**, 175.

Frank, J. H. & Curtis, G. A. 1979 Trend lines and the number of species of Staphylinidae. *Coleopt. Bull.* **33**, 133–149.

Gaston, K. J. 1991 The magnitude of global insect species richness. *Conserv. Biol.* **5**, 283–296. (doi:10.1111/j.1523-1739.1991.tb00140.x)

Gaston, K. J. In press. Global species richness. In *Encyclopedia of biodiversity* (ed. S. A. Levin). San Diego, CA: Academic Press.

Gower, D. J. B., Giri, G., Oommen, V., Ravichandran, O. V. & Wilkinson, M. 2004 Biodiversity in the western Ghats: the discovery of new species of caecilian amphibians. *Curr. Sci.* **87**, 739–740.

Hammond, P. M. 1995 Described and estimated species numbers: an objective assessment of current knowledge. In *Microbial diversity and ecosystem function* (eds D. Allsopp, D. L. Hawksworth & R. R. Colwell), pp. 29–71. Wallingford, UK: CAB International.

Hassler, M. & Swale, B. J. 2001 World ferns on CDROM. [Published by the authors.]

Hawksworth, D. L. 2001 The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycol. Res.* **105**, 1422–1432.

Kent, D. H. 1992 *List of vascular plants of the British Isles*. London, UK: Botanical Society of the British Isles.

Lambshead, P. J. D. & Boucher, G. 2003 Marine nematode deep-sea biodiversity—hyperdiverse or hype? *J. Biogeogr.* **30**, 475–485.

May, R. M. 1988 How many species are there on earth. *Science* **241**, 1441–1449. (doi:10.1126/science.241.4872.1441)

May, R. M. 1990 How many species? *Phil. Trans. R. Soc. B* **330**, 293–304. (doi:10.1098/rstb.1990.0200)

May, R. M. 2000 The dimensions of life on Earth. In *Nature and human society* (eds P. H. Raven & T. Williams), pp. 30–45. Washington, DC: National Academy Press.

McCullough, P. & Nelder, J. A. 1989 *Generalized linear models*. New York, NY: Chapman and Hall.

Medellin, R. A. & Soberon, J. 1999 Prediction of mammal diversity on four land masses. *Conserv. Biol.* **13**, 143–149. (doi:10.1046/j.1523-1739.1999.97315.x)

Pimm, S., Raven, P., Peterson, A., Sekercioglu, C. H. & Ehrlich, P. R. 2006 Human impacts on the rates of recent present and future bird extinctions. *Proc. Natl Acad. Sci. USA* **103**, 10 941–10 946. (doi:10.1073/pnas.0604181103)

Scoble, M. J., Gaston, K. J. & Crook, A. 1995 Using taxonomic data to estimate species richness in Geometridae. *J. Lepid. Soc.* **49**, 136–147.

Downloaded from rspb.royalsocietypublishing.org on March 28, 2012

Scotland, R. W. & Wortley, A. H. 2003 How many species of seed plant are there? *Taxon* **52**, 101–104. (doi:10.2307/3647306)

Shen Tsung-Jen, A. C. & Chih-Feng, L. 2003 Predicting the number of new species in further taxonomic sampling. *Ecology* **84**, 798–804.

Soberon, J. M. & Llorente, J. B. 1993 The use of species accumulation functions for the prediction of species richness. *Conserv. Biol.* **7**, 480–488. (doi:10.1046/j.1523-1739.1993.07030480.x)

Solow, A. R. & Smith, W. K. 2005 On estimating the number of species from the discovery record. *Proc. R. Soc. B* **272**, 285–287. (doi:10.1098/rspb.2004.2955)

Stace, C. A. 1997 *New flora of the British Isles*. Cambridge, UK: Cambridge University Press.

Steyskal, G. C. 1965 Trend curves of the rate of species description in zoology. *Science* **149**, 880–882. (doi:10.1126/science.149.3686.880)

Venables, W. N. & Ripley, B. D. 2002 *Modern applied statistics with S*. New York, NY: Springer.

Wilson, S. P. & Costello, M. J. 2005 Predicting future discoveries of European marine species using non-homogenous renewal process. *Appl. Stat.* **54**, 897–918.

Wortley, A. H. & Scotland, R. W. 2004 Synonymy, sampling and seed plant numbers. *Taxon* **53**, 478–480.