Estimating global arthropod species richness: refining probabilistic models using probability bounds analysis

Andrew J. Hamilton[1,*], Vojtech Novotný[2], Edward K. Waters[3], Yves Basset[4], Kurt K. Benke[5], Peter S. Grimbacher[1], Scott E. Miller[6], G. Allan Samuelson[7], George D. Weiblen[8], Jian D. L. Yen[9], Nigel E. Stork[10]

[1]Melbourne School of Land and Environment, The University of Melbourne, Dookie Campus, Victoria 3647, Australia. andrewjh@unimelb.edu.au, petersg@unimelb.edu.au

[2]Biology Centre, Czech Academy of Sciences and Faculty of Science, University of South Bohemia, Branišovská 31, 370 05 Ceske Budejovice, Czech Republic. Novotny@entu.cas.cz

[3]The University of Notre Dame Australia, PO Box 944, Broadway. NSW 2007, Australia. edward.waters@nd.edu.au

[4]Smithsonian Tropical Research Institute, Apartado 0843-03092, Balboa, Ancón, Panamá. bassety@si.edu

[5]Department of Primary Industries, 32 Lincoln Square North, Carlton, Parkville Centre, Victoria 3052, Australia. kurt.benke@dpi.vic.gov.au

[6]National Museum of Natural History, Smithsonian Institution, Washington, DC 20013–7012, USA. millers@si.edu

[7]Bishop Museum, Honolulu, Hawaii, USA. alsam@bishopmuseum.org

[8]Department of Plant Biology, University of Minnesota, 220 Biological Sciences Centre, 1445 Gortner Avenue, St Paul, MN 55108-1095, USA. gweiblen@umn.edu

[9]School of Biological Sciences, Monash University, Clayton, Victoria 3800, Australia. jdyen@iinet.net.au

[10]Griffith School of Environment, Griffith University, 170 Kessels Road, Nathan, Queensland 4111, Australia. nigel.stork@griffith.edu.au

***Corresponding Author:** Andrew J. Hamilton, Current address: Department of Agriculture and Food Systems, Melbourne School of Land and Environment, The University of Melbourne, Dookie Campus, 940 Dookie–Nalinga Road, Dookie College, Victoria 3647, Australia.

E-mail: andrewjh@unimelb.edu.au.

Telephone: +61 3 5833 9252

Fax: +61 3 5833 9201

Author contributions

AJH led the development of the mathematical and conceptual models and wrote the majority of the manuscript, NES, VN, and PSG contributed to the conceptual development of the model and wrote parts of the manuscript, with VN also leading the research team in Papua New Guinea upon whose data the model is in large part based upon, EKW assisted with mathematical detail, especially in the area of copulae, KKB and JDLY were involved with the mathematical formulation and implementation of the model, and YB, SEM, GDW and GAS were all involved in the collection of the entomological data used for host specificity calculations and also provided specific input to considerations of host specificity in the manuscript.

ABSTRACT: A key challenge in the estimation of tropical arthropod species richness is the appropriate management of the large uncertainties associated with any model. Such uncertainties had largely been ignored until recently, when we attempted to account for uncertainty associated with model variables, using Monte Carlo analysis. This model is restricted by various assumptions. Here we use a technique known as probability bounds analysis to assess the influence of assumptions about (i) distributional form and (ii) dependencies between variables, and to construct probability bounds around the original model prediction distribution. The original Monte Carlo model yielded a median estimate of 6.1 million species, with a 90% confidence interval of [3.6, 11.4]. Here we found that the probability bounds (p-bounds) surrounding this cumulative distribution were very broad, owing to uncertainties in distributional form and dependencies between variables. Replacing the implicit assumption of pure statistical independence between variables in the model with no dependency assumptions resulted in lower and upper p-bounds at 0.5 cumulative probability (i.e., at the median estimate) of 2.9–12.7 million. From here, replacing probability distributions with probability boxes, which represent classes of distributions, led to even wider bounds (2.4–20.0 million at 0.5 cumulative probability). Even the 100th percentile of the uppermost bound produced (i.e., the absolutely most conservative scenario) did not encompass the well-known hyper-estimate of 30 million species of tropical arthropods. This supports the lower estimates made by several authors over the last two decades.

KEYWORDS: Host specificity, model, Monte Carlo, uncertainty

Introduction

Extrapolating global estimates of tropical arthropod species richness from samples, as first proposed by Erwin (1982) and revisited by many since (e.g., Thomas 1990; Stork 1988, 1993; Ødegaard 2000; Novotný et al. 2002), is an intriguing exercise because it potentially offers a significant short-cut that would save having to count species one by one, but at the same time it is vulnerable to producing massively misleading estimates, owing ultimately to the need to base extrapolations on host specificity measurements made for a minute proportion of all tropical tree species. The models are typically based upon a sample of beetle species collected from one or several tree species. This is because beetles are the most common taxon, accounting for about 25% and 40% of all described insects and species, respectively (Hammond 1992; Yeates et al. 2003). Then, by making assumptions about host specificity to trees, the number of tropical tree species in the world, the proportions of species in the canopy and ground, and the proportion of all arthropods that are beetles, one can estimate how many tropical arthropod species might exist.

The model described above is a model of mean behaviour, that is, the average state of one parameter (species richness) of a far more complicated system over time. An individual-based model, where individual species, and even individual insects, are represented as discrete units would plainly be an insurmountable undertaking. For example, in this mean-behaviour model, the host specificity is a single parameter, but in an individual-based model it could require consideration of such things as the number of individual trees per tree species and hence the size of the populations per beetle species, the evolutionary life time of individual tree species, the number of months during which each tree carries leaves to be eaten by phytophages, the number of closely related tree species that might serve as a pool of phytophagous species to colonize a focal tree species, and the niche breadth and intraspecific differentiation of the tree species.

Until recently, all such extrapolation mean-behaviour models were purely deterministic; that is, despite the considerable uncertainties associated with the various parameters, no attempts were made to account for these. To this end, we recently published a probabilistic model (Hamilton et al. 2010, 2011), which has been seen as a significant step forward because it was the first attempt to explicitly deal with uncertainties in the extrapolation process (May 2010). In line with previous models, the model took the following form:

$$N_{Ai} = \left(xc/p_{cg}\, p_{ba}\right) n_t \,, \tag{1}$$

where N_{Ai} is the estimator of the number of tropical arthropod species under the assumption of independence between variables, x is the average effective specialisation (May 1990) of herbivorous beetle species across all tree species, c is a correction factor for non-herbivorous beetle species, p_{ba} is the proportion of canopy arthropod species that are beetles, p_{cg} is the proportion of all arthropod species found in the canopy, and n_t is the number of tropical tree species. Note the change in notation for p_{ba} and p_{cg} from the original model; this was done because in retrospect the original notation was potentially ambiguous and confusing (see Hamilton et al. 2010, 2011). Probability distributions were assigned to all parameters.

Implementation of our original model was achieved using Latin Hypercube Sampling (LHS), a specialised form of Monte Carlo simulation wherein probability distributions are sampled in a stratified random manner (McKay et al. 1979). As with any modelling technique, Monte Carlo simulation necessitates assumptions. Thus, while this was the first attempt to account for uncertainty, the model (i) made certain assumptions about distributional form used to represent uncertainty and (ii) did not consider potential dependencies between variables.

Before considering the relevance of assumptions about distributional form, we need to appreciate the fundamental nature of uncertainty. While various taxonomies of uncertainty have been proposed (Kahneman and Tversky 1982; Morgan and Henrion 1990; Regan et al. 2002), there are in essence only two basic forms—variability and ignorance (Casti 1990; Benke et al. 2007). Variability represents natural randomness or stochasticity and cannot be reduced, and is often called aleatory uncertainty. Ignorance, on the other hand, is reducible and arises from numerous factors, including, *inter alia*, measurement error, lack of data and small sample-sizes, and personal biases, and is also known as epistemic uncertainty. Theoretically, different methods are required to propagate ignorance and variability (Ferson and Ginzburg 1996). This can be attempted in the Monte Carlo framework using a technique known as Second-Order Monte Carlo, wherein variability is represented by a probability distribution and ignorance is characterised in an outer-loop in one of a number of ways, such as alternative model scenarios or distributional shapes (Vose 2000). However, as pointed out by Regan et al. (2004), Second-Order Monte Carlo still requires a subjective assessment of the realistic range of input distributions. In fact, the process is innately contradictory, because the greater the ignorance, the more data are required to specify bounds on the distribution. Of course, a greater amount of data should lead to narrower bounds for the variable's distribution.

This conundrum often leaves the Monte Carlo analyst with little choice but to construct a one-dimensional model, wherein variability and ignorance are confounded in simple distributions, as we noted in our original paper. For example, very little information existed for variable c in our model, a correction factor for non-herbivorous beetle species. Ødegaard (2000) identified seven different studies relevant to the determination of c. There will of course be true natural variability associated with c, as it would be unreasonable to expect that the ratio of herbivorous to non-herbivorous beetle species would be constant across all tree species throughout the tropics and at all tropical locations. Likewise, ignorance emerges from the facts that the handful of studies used to estimate c use different methods, are all subject to various limitations associated with sampling arthropod faunas, and all have associated biases inherent with site selection (i.e., c would ideally be determined from studies of randomly selected tree species at randomly selected sites across the tropics if their sole purpose was to contribute to the estimation of this variable for this model).

With such limited information available, it was clearly not possible to propagate variability and ignorance separately for c or the other variables in the model (Hamilton et al. 2010). Rather, the approach taken was to consider them together using Uniform distributions. The rationale for using the Uniform distribution to represent highly uncertain environmental variables is that it is the most conservative approach (e.g., Brook et al. 2003; Mara et al. 2007). This makes intuitive sense because we have no reason to favour the selection of any value in the range over another. Upon closer inspection though, the Uniform actually makes some potentially significant assumptions about a variable. Consider c again, which covers the interval [1.79, 2.70]. The Cumulative Distribution Function (CDF) of a Uniform distribution for this interval is a perfectly linear monotonically increasing function where the mean = mode = median = 2.25. In theory, an infinite number of distributions could describe this interval, and these would be bounded within a box defined by two vertical lines extending from 0% to 100% cumulative probability at the minimum and maximum values. But there can be only one true distribution representing variability for this interval, yet its form is unknown to us, and this ignorance needs to be expressed through allowing variation in shape. Using a single distribution (be it Uniform, Triangular or something else) ignores shape uncertainty (a sub-set of ignorance), and therefore leads to an overstatement of confidence.

The second problem associated with the application of Monte Carlo techniques to ecological models is that uncertainties about dependencies between variables cannot be expressed (Ferson 1996, 2002). Knowledge about dependencies is typically very poor in ecological models. Clearly, natural systems are complex and dependencies between variables are likely to exist, and these species richness estimation models are no exception. For example, The Janzen-Connell hypothesis (Janzen 1970; Connell 1971) proposes that predation on plants (in part by arthropods) is one of the mechanisms leading to the high plant richness found in the tropics. But there is also a reciprocal relationship because a greater diversity of plant resources provides opportunities for arthropods to specialize over time and thus diversify (Janz et al. 2006). Therefore, over evolutionary time, plant and arthropod communities interact through positive feedback and this increases the richness of both groups. Of course, the exact details and form of such dependencies are unclear, but the fact that they could exist means that they should not be ignored. It is possible that the dependencies themselves are not stable, and are likely to change over evolutionary time and even as a function of anthropogenic changes. Highly specialised species, for example, are not necessarily destined for an evolutionary dead-end, and can even give rise to generalists (Colles et al. 2009). Despite the complexity of such dependencies, the time is ripe to at least introduce the concept to species richness models so that advances in evolutionary biology can be used to modify these simple sample extrapolation models.

With respect to modelling dependencies, the default approach of independence, which is rarely stated explicitly, is intuitively appealing because there is usually not a clear dependency relationship between the various pairs of variables. However, as noted by Tucker and Ferson (2003), independence implies zero correlation but zero correlation does not demand independence. Furthermore, the possibility of higher-order dependencies should not be excluded. That is, not all dependencies will be pair-wise, between two variables: some could be multivariate. Consequently, Ferson (2002) suggests that models should start by making the assumption of dependence between all variables and at all levels, and independence should be assumed only when sound empirical information exists to support it. Vose (2000), whilst acknowledging it is a contentious issue, takes the opposing view, and suggests that one should avoid attempting to model correlation ‘where there is neither a logical reason nor evidence for its existence.’ In line with many Monte Carlo ecological models of systems wherein very little is known about the nature of inter-variable dependencies (Jonzen et al. 2002; Brook et al. 2003), our original species richness model invoked Vose’s philosophy. While dependencies can be specified in the Monte Carlo construct, uncertainty about their nature—magnitude and form—cannot be accommodated. In other words, ignorance about the dependencies cannot be included. Also, as noted by Ferson et al. (2004), the use of correlation coefficients to define dependencies—the typical approach used in Monte Carlo (Vose 2000)—is weak, as a dependency needs to be described by a complete dependency function (a copula), and several copulae can in fact have the same correlation. Finally, Monte Carlo methods do not readily allow for modelling of higher-order dependencies.

Probability bounds (p-bounds) analysis necessitates neither subjective assumptions about distributional form nor the nature of dependencies, and has proved useful in ecological models, where large uncertainties are often associated with these properties (Ferson 2002; Regan et al. 2002). Briefly, p-bounds analysis deals with classes of distributions rather than individual distributions (Frank et al. 1987; Williamson and Downs 1990). It not only offers a method for computing the bounds for a given variable, but also enables the convolution (e.g., multiplication, division, addition, subtraction or exponentiation) of these distributional classes, and thus propagation of ignorance and variability, together, through the model. While confidence intervals or credible intervals set bounds around a statistic for a variable, effectively as a function of its distribution, p-bounds are bounds surrounding the probability distribution itself. P-bounds must be expressed in terms of the CDF, not probability density or mass functions. In essence, p-bounds analysis can be seen as a highly conservative technique for determining the limits of an infinite array of possible CDFs, and it has been described by Burgman (2005) simply as a more honest approach because the analyst is not forced to make unjustified assumptions to satisfy a mathematical framework (*cf* Monte Carlo). Philosophically, p-bounds analysis involves specifying total possible uncertainty and then explicitly removing it, whereas Monte Carlo approaches require uncertainty to be explicitly included. Ferson (2002) describes p-bounds analysis as a useful method for providing 'quality assurance for Monte Carlo results'. Regan et al. (2002), for example, found for a food-web model that the p-bounds analysis was useful for checking the plausibility of a Monte Carlo model. It is also worth noting that they found the p-bounds envelope on the CDF to be markedly broader than one generated by a second-order Monte Carlo analysis.

Here we use probability bounds analysis to explore the implications of assumptions on the independence of variables and distributional forms used to account for uncertainties made by a previous model on the global species richness estimate for tropical arthropods.

Methods

Hamilton et al. (2010, 2011) presented two models, A and B, which were respectively based on the estimated number of tree *species* in the tropics and the number of tropical tree *genera* in New Guinea alone. Here, p-bounds modelling is applied to Model A only (eqn. 1), as this is overwhelmingly the most common approach to the problem (Erwin 1982; Ødegaard 2000; Stork 1988; Thomas 1990). Furthermore, the two models are otherwise analogous. Model A is described in detail in Hamilton et al. (2010), with terminology specifically appropriate to the LHS methodology used.

For the LHS implementation of Model A the following Uniform distributions were used for four variables: c = 1.79–2.70, p_{cg} = 0.25–0.66, p_{ba} = 0.18–0.33, n_t = 43,000–

50,000. The variable x is the product of n_k, the number of herbivorous canopy beetle species on tree species k (k = 1, 2...*l*), and f_k, the proportion of the beetle species effectively specialised on that species (see Hamilton et al. 2010 for calculation of f_k). A distribution for x was then obtained by producing 500,000 non-parametric bootstrap estimates of $n_k f_k$. The reader is also directed to the published corrigendum (Hamilton et al. 2011). It is also important to note that x represents an estimate of the *average* effective specialisation rather than the effective specialisation of a given tree species, $n_k f_k$. A distribution of x is what is required for this model. Drawing realisations from the distribution of $n_k f_k$ would result in a distribution of imprecise estimates of tropical arthropod species richness (i.e., where each estimate is based on a single tree species), rather than a distribution of precise estimates with each being based upon the suite of species. This potential pitfall is common in uncertainty models, as described by Karavarsamis and Hamilton (2010) in a health risk context.

Calculations in p-bounds analysis are made on p-boxes. A p-box is defined as the class of CDFs ($F(y)$) bounded by a pair of CDFs, $\underline{F}(y)$ and $\overline{F}(y)$, such that $\underline{F}(y) \leq F(y) \leq \overline{F}(y)$. In our models, two types of p-boxes were constructed for each variable. First, the entire cumulative probability space within the possible range for the variables was represented using ‘minimum-maximum’ boxes. This is superficially and, perhaps, intuitively analogous to the use of a Uniform distribution in a Monte Carlo analysis, although it is in fact quite different because it permits all possible cumulative distributions between specified 0^{th} and 100^{th} cumulative percentiles. Second, ‘empirical histogram’ p-boxes were used to include all the available estimates of variables, not just the minima and maxima.

In line with the original paper, we used the review of Ødegaard (2000) to obtain estimates of c, p_{cg} and p_{ba}. It is worth noting that the various studies listed by Ødegaard are not all directly comparable, owing to different sampling techniques, and they do not always explicitly represent the variable of interest, but they characterise the best available information and hence have been used by many authors in extrapolating tropical arthropod species richness estimates (May 1990; Thomas 1990; Stork 1988, 1993; Novotný et al. 2002). Therefore, the following p-boxes were constructed, where MM and EH respectively denote variables for which minimum-maximum and empirical histogram boxes have been defined, and 'minmax' and 'histogram' is the respective coding terminology used by Ferson (2002): c_{MM} = minmax (1.79–3.37), c_{EH} = histogram (1.79, 2.70, 1.79, 2.13, 2.27, 2.44, 2.50, 2.70), p_{cgMM} = minmax (0.25, 0.66), p_{cgEH} = histogram (0.25, 0.66, 0.25, 0.33, 0.5, 0.66), p_{baMM} = minmax (0.18, 0.33), p_{baEH} = histogram (0.18, 0.33, 0.18, 0.22, 0.23, 0.23, 0.33), and n_{tMM} = minmax (43,000, 50,000). Note that there were only two estimates available for n_t (see Hamilton et al. 2010), hence only a minimum-maximum box is required. All the empirical histogram p-boxes are shown in fig. 1. Minimum-maximum p-boxes are not shown because they are simply vertical lines extending from zero to 1 cumulative probability at the minima and maxima. x is the only variable for which p-bounds were not constructed (Figure 1). Bootstrapping is a sampling procedure, and therefore the resultant distribution will converge to the Normal with increasing sample size, owing to the Central Limit Theorem. P-bounds are appropriate when uncertainty about distributional form exists, but that is not the case here. As noted in Hamilton et al. (2010), the data-set we used for determining the number of herbivorous beetle species effectively specialised on a given tree species is substantially larger than any other available data-set for this parameter. In any case, it is necessary to use a method that is congruent with the original model, so that the implications of the assumptions stated above can be assessed. Given that 500,000 bootstrap replicates were taken, the CDF of x is Normal. This CDF jointly expresses variability and ignorance. Empirical p-bounds constructed from multiple bootstrap replicates simply converge to the CDF as the number of replicates increases, and are meaningless in the context of representing natural variation and ignorance, reflecting nothing other than the effect of computational sample size.

Figure 1 to appear near here

The original LHS estimate was recalculated (N_{Ai}, eqn. 1). Additionally the following estimators of tropical arthropod species richness were solved for:

$$N_{Ad} = \left(xc / p_{cg}\, p_{ba} \right) n_t \,; \tag{2}$$

where the subscript d denotes that no assumptions are made about the nature of dependencies between variables and all variables are represented by the distributions described for equation 1,

$$N_{AiMM} = \left(x |\cdot| c_{MM} |/| p_{cgMM} |\cdot| p_{baMM} \right) |\cdot| n_{tMM} \,, \tag{3}$$

where all variables are assumed to be statistically independent (i), as marked by the pipes (paired vertical lines) either side of each operator, and most variables are represented by minimum-maximum p-boxes,

$$N_{AdMM} = \left(xc_{MM} / p_{cgMM} p_{baMM} \right) n_{tMM} \,; \tag{4}$$

$$N_{AiEH} = \left(x |\cdot| c_{EH} |/| p_{cgEH} |\cdot| p_{baEH} \right) |\cdot| n_{tEH} \text{; and} \tag{5}$$

$$N_{AdEH} = \left(xc_{EH} / p_{cgEH} p_{baEH} \right) n_{tEH} \,. \tag{6}$$

P-boxes were specified and convolved using RAMAS Risc Calc 4.0 (Ferson 2002). Distributions with infinite tails, such as the Normal, which was used for x, cannot be convolved in p-bounds analysis, and therefore truncation was enforced at 0.005 and 0.995 cumulative probability. Truncation was not necessary for the distributions with finite bounds (i.e., minimum-maximum and empirical histogram). Empirical histogram bounds were constructed using Kolmogorov-Smirnov confidence limits of 95%. The mathematics behind convolving p-boxes is described elsewhere (Frank et al. 1987, Williamson and Downs 1990).

Copulae were used to convolve distributions in solving equations (2)–(6). A copula is a function that describes the dependence relationship between multiple variables by transforming the marginal distributions of each variable to uniform distributions. This works because any variable in the model, y, can be represented by a generalised inverse, $v = F^{-1}(u)$, where F^{-1} is an inverse CDF and u is a uniformly distributed random variable. Thus a copula function, C, is defined as a function, f, of the generalised inverses $U = u_1,...,u_d$ of d variables, $Y = y_1,...,y_d$, in the model so that

$$C(U) = f\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right). \tag{7}$$

Independence is a special copula function. Where independence is assumed $C(U) = \prod_{i=1}^{d} u_I$, which is identical to the form of equation (1). Where no assumptions were made about dependencies, Ferson's (2004) approach of convolving p-boxes within Fréchet bounds was used. Let

$$C^d(U) = \left(\sum_{i=1}^{d} u_i\right) - C(U), \tag{8}$$

where C(U) is a copula fitting within the lower Fréchet copula bounds

$$C^F(U) = \max\left(\left(\sum_{i=1}^{d} u_i\right) - 1, 0\right). \quad (9)$$

Thus both the lower and upper bounds of the dependency function governing Y enclose a copula describing the dependencies in terms of the general inverse U. Both the upper and the lower bounds enclosing this copula can be expressed in terms of the Fréchet lower bounds, which can then be used to elucidate the nature of the dependency function of Y (Ferson 2004). All possible copulae describing Y are enclosed within the Fréchet bounds such that no assumptions about dependency need be made. Kendall's grade correlation was then used to describe the nature of the dependencies identified through this process (Ferson 2004).

Results

The Monte Carlo-LHS model yielded a median estimate of 6.1 million species, with a 90% confidence interval of [3.6, 11.4] million (Fig. 2A). Simply replacing the assumption of pure statistical independence between variables in the model with no dependency assumptions resulted in reasonably broad p-bounds (Fig. 2B), with lower and upper bounds at 0.5 cumulative probability (i.e., at the median estimate) of 2.9–12.7 million. Bounding the input variables had an even larger effect on the bounds for the prediction. In the case of minimum-maximum bounds, the probability envelope was so wide and steep that there was negligible difference with respect to the dependency assumptions, with pure independence and no dependency assumptions yielding bounds at 0.5 cumulative probability of 2.35–19.7 million and 2.4–20.0 million, respectively (Fig. 2C, D). Furthermore, the shapes of the bounds were almost identical in both these cases. The use of empirical bounds on the input variables had negligible impact relative to the min-max bounds, and, in fact, for the case of no dependency assumptions, the lower and upper bounds at 0.5 cumulative probability were identical (2.4–20.0 million) to those for the parallel min-max case, but the bounds did vary slightly in shape from those produced from the min-max model (Fig 2F). Likewise, under the assumptions of pure statistical independence the empirical bounding approach produced bounds of slightly different shape to the min-max model, but the values at 0.5 cumulative probability were very similar (2.7–18.4 million) (Fig 2E).

Figure 2 to appear near here

Discussion

Probability bounds analysis was used to assess the plausibility of a Monte Carlo model of tropical arthropod species richness. While broad, the bounds rule out the possibility of estimates of 30 million species or greater, with the 100th percentile of the right-hand bound—i.e., the absolutely most conservative scenario—being < 30 million in each case. P-bounds define the cumulative probability space in which the true distribution will lie, but it is important to note that each of the infinite number of CDFs within this space is not equally likely. In fact, it could be argued that this approach is markedly too conservative, as it even allows for highly unlikely distribution forms, such as multimodal, that are probably not appropriate for the parameters in this model (or indeed the prediction). Interestingly, the CDF of the Monte Carlo model was always situated toward the left-hand side of cumulative probability space defined by the p-bounds, regardless of the dependency or distributional form assumptions made in assigning these bounds. The reason for this is unknown.

Removing the assumption of independence between variables in the original model and replacing it with no dependency assumptions resulted in reasonably broad p-bounds (Fig. 1B). While the state of knowledge about the nature of dependencies between the model variables is very poor, it is reasonable to expect that some dependencies will indeed exist. For example, the richness and specialization of insects is not independent of tropical tree species richness, as illustrated through the Janzen-Connell hypothesis (Janzen 1970; Connell 1971), but it is difficult to know the strength of the relationship because the relative contributions of phylogenetic conservativeness and geographic contingency and local mass effects in the assemblage of communities remain unclear (Goßner et al. 2009). It would also be reasonable to hypothesise that the proportion of arthropods—including beetles—that are herbivores is likely to be dependent upon plant species richness. On the whole, plant-feeding arthropods are more specialized and constrained in the diversity of their resource use than non-plant-feeding species, such as carnivores and fungivores (but not parasitoids) (Ross et al. 1982). It could be that this is because dealing with plant physical and chemical herbivore deterrents is more digestively demanding and thus requires a more specialized digestive system. Therefore, it could be hypothesised that over time increasing tree diversity would likely lead to more herbivores and alter the ratio of herbivores to non-herbivores. The relationship between tree species richness and the ratio of herbivores to non-herbivores logically leads to the possibility of secondary dependencies with the canopy to ground ratio, and the proportion of non-beetle arthropods.

The canopy to ground ratio may be dependent upon plant species diversity if there are more plant-feeding arthropods in the canopy (e.g., Grimbacher and Stork 2007). Arthropods associated with the ground are likely to show lower levels of specialisation (e.g., Crutsinger et al. 2008; Donoso et al. 2010). Thus, if there are more plant-feeding arthropods in the canopy, then over evolutionary time an increase in tree species richness might be expected to lead to increasing arthropod richness in the canopy at a greater rate, relative to the ground.

For similar reasons, it may be that the proportion of non-beetle arthropods found in the canopy relative to the ground is related to tree species richness. Unlike the Coleoptera, which are highly diverse in their feeding ecology, species from most insect orders are likely to have one main mode of feeding (Ross et al. 1988). The Lepidoptera, for example, are predominantly herbivorous, while the non-ant Hymenoptera are largely predatory or parasitoid. Because not all arthropod orders contribute equally to global species richness (Nielsen and Mound 2000), and herbivores are likely to have a much tighter association to tree species richness than non-herbivores, the ratio of canopy to ground diversity is likely to alter the relative contribution of non-beetle arthropods.

Of course, these are just some examples of potential dependencies between variables typically used in an extrapolation model of species richness. The exact form of these dependencies is unknown; there are likely to be other dependencies, including those of higher-order. The theoretical arguments given above cover only some processes that could influence dependencies—there may indeed be other processes negating, antagonising, or complementing these. It is for these reasons that Fréchet's (1935) copula was used to convolve the distributions, as it makes no assumptions about the nature of the dependencies. Other copulae can be used to specify other dependencies, such as perfect, opposite, positive, negative, straight-positive, and straight-negative, and these can be implemented in RiscCalc (Ferson et al. 2004). With further ecological and evolutionary insight into the nature of the dependencies, these less conservative copulae could be used in such models. But before any gains are to be made in narrowing the bounds through this means, the more problematic issue of uncertainty associated with the model variables needs to be addressed, as discussed in the original manuscript (Hamilton et al. 2010). This had a larger effect on the breadth of the bounds than the independence assumption (fig. 2C and fig. 2E *cf* fig. 2B).

Since Erwin presented the extrapolation approach to estimating tropical arthropod species richness, overwhelmingly the debate has centred around what values best represent the variables (Erwin 1988; Thomas 1990; Stork 1988, 1993; Ødegaard 2000), but this thinking needs to be broadened to consider the relevance of potential dependencies between variables and the relative merits of different technical approaches to uncertainty modelling in this context. Furthermore, other methods of estimating tropical arthropod species richness would benefit from more thorough use of uncertainty modelling, including, *inter alia*, extrapolations from known faunas and regions, methods using ecological models, eliciting taxonomists' views, and species description rates (Stork 1993). Mora et al. (2011) recently made a step in this direction through accommodating uncertainty in their taxonomic-level based global species richness model, which produced a median estimate of 8.7 million eukaryotic organisms on Earth (± 1.3 million SE), with 6.5 million of these being terrestrial, which accords well with our original median estimate of 6.1 million tropical arthropods (Hamilton et al. 2010, 2011). Another recently published model (Costello et al. in press), based on species description rates, also accounted for uncertainty, and produced a median estimate of 490,960 [95% CI = 449,010, 477,990] terrestrial species remaining to be described, which equates to only 1.6–1.7 million terrestrial species existing globally, a much lower prediction than that of Mora et al. (2011) and Hamilton et al (2010, 2011). The variation in such models highlights the need to consider uncertainty surrounding this important question even more broadly, that is, not just within models but between models. While tropical arthropods species are of primary interest on a global scale, given their high richness and the potentially huge numbers of undescribed species, improved uncertainty modelling can contribute also to other large-scale species richness estimates, be it European marine species (Wilson and Costello 2005) or flowering plants globally (Joppa et al. 2011). Every statistical approach has something to offer but equally has its limitations; the next step in tackling this important question will be to combine models and their associated uncertainties, perhaps using techniques such as Bayesian modelling averaging.

Acknowledgements

Cindy Hauser provided useful technical comments on a draft of this manuscript. The host specificity studies in New Guinea, upon which this model draws substantially upon, were supported by the National Science Foundation (USA), Christensen Fund (USA), Grant Agency of the Czech Republic, Czech Academy of Sciences, the Swedish Natural Science Research Council, Czech Ministry of Education, Otto Kinne Foundation, Darwin Initiative (UK), International Centre of Insect Physiology and Ecology (ICIPE) and Bishop Museum. Parataxonomists in New Guinea are thanked for their assistance and are listed in Novotný et al. (2002). This paper is dedicated to the late Ken Hamilton, the consummate logician and giver.

References

Benke KK, Hamilton AJ, Lowell K (2007) Uncertainty analysis and risk assessment in the management of environmental resources. Australian Journal of Environmental Management 14:243–249.

Brook BW, Sodhi NS, Ng PKL (2003) Catastrophic extinctions follow deforestation in Singapore. Nature 424:420–423 doi:10.1038/nature01795

Buckland ST (1984) Monte Carlo confidence intervals. Biometrics 40: 811–817.

Burgman M (2005) Risks and Decisions for Conservation and Environmental Management, 1st edn. Cambridge University Press, Cambridge

Casti JL (1990) Searching for Certainty, 1st edn. William Morrow and Company, London

Chapman AD (2009) Numbers of Living Species in Australia and the World (2nd ed.). Department of the Environment, Water, Heritage and the Arts. http://www.environment.gov.au/biodiversity/abrs/publications/other/species-numbers/2009/pubs/nlsaw-2nd-complete.pdf

Colles A, Liow LH, Prinzing A (2009) Are specialists at risk under environmental change? Neoecological, paleoecological and phylogenetic approaches. Ecology Letters. 12: 849–863 doi: 10.1111/j.1461-0248.2009.01336.x

Connell JH (1971) On the role of natural enemies in preventing competitive exclusion in some marine animals and in rain forest trees. – In: Den Boer, P. J. and Gradwell G. (ed.), Dynamics of populations. Centre for Agricultural Publishing and Documentation, pp. 298–312.

Costello MJ, Wilson S, Houlding B (in press). Predicting total global species richness using rates of species description and estimates of taxonomic effort. Systematic Biology.

Crutsinger GM, Reynolds WN, Classen AT, Sanders NJ (2008) Disparate effects of plant genotypic diversity on foliage and litter arthropod communities. Oecologia 158:65–75 doi: 10.1007/s00442-008-1130-y

Donoso DA, Johnston MK, Kaspari M (2010) Trees as templates for tropical litter arthropod diversity. Oecologia 164:201–211 doi:10.1007/s00442-010-1607-3

Erwin TL (1982) Tropical forests: their richness in Coleoptera and other arthropod species. Coleopts Bull 36:74–75

Erwin TL (1988) The tropical forest canopy: the heart of biotic diversity. In: Wilson EO (ed.) Biodiversity. National Academy Press, Washington D.C., pp 123–129

Ferson S (1996) What Monte Carlo methods cannot do. Hum Ecol Risk Assess 2:990–1007 doi:10.1080/10807039609383659

Ferson S (2002) RAMAS Risk Calc 4.0 software: risk assessment with uncertain numbers. Lewis Publishers

Ferson S, Ginzburg LR (1996) Different methods are needed to propagate ignorance and variability. Reliab Eng Syst Safe 54:133–144 doi:10.1016/S0951-8320(96)00071-3

Ferson S, Nelsen RB, Hajagos J, Berleant DJ, Zhang J, Tucker WT, Ginzburg LR, Oberkampf WL (2004) Dependence in probabilistic modelling, Dempster–Shafer theory, and probability bounds analysis. Sandia National Laboratories, New Mexico SAND2004–3072. http://www.ramas.com/depend.pdf

Frank MJ, Nelsen RB, Schweizer, B (1987) Best–possible bounds for the distribution of a sum—a problem of Kolmogorov. Probab Theory Rel 74:199–211 doi: 10.1007/BF00569989

Fréchet M (1935) Généralisations du theorem des probabilities totales. Fund Math 25:379–387

Goßner MM, Chao A, Bailey RI, Prinzing A (2009) Native fauna on exotic trees: phylogenetic conservatism and geographic contingency in two lineages of phytophages on two lineages of trees. Am Nat 173:599–614.

Grimbacher PS, Stork NE (2007) Vertical stratification of feeding guilds and body size in beetle assemblages from an Australian tropical rainforest. Austral Ecol 32:77–85 doi: 10.1111/j.1442-9993.2007.01735.x

Hamilton AJ, Basset Y, Benke KK, Grimbacher PS, Miller SE, Novotny´ V, Samuelson GA, Stork NE, Weiblen GD, Yen JDL (2010) Quantifying uncertainty in tropical arthropod species richness estimation. Am Nat 176:90–95 doi: 10.1086/652998

Hamilton AJ, Basset Y, Benke KK, Grimbacher PS, Miller SE, Novotny´ V, Samuelson GA, Stork NE, Weiblen GD, Yen JDL (2011) Correction: Quantifying uncertainty in tropical arthropod species richness estimation. Am Nat 177:544–545

Hammond PM (1992) Species inventory. In ‘Global Biodiversity: Status of the Earth’s Living Resources, B. Groombridge, ed., pp. 17–39. Chapman and Hall, London.

Janz N, Nylin S, Wahlberg N (2006) Diversity begets diversity: host expansions and the diversification of plant–feeding insects. BMC Evol Biol 6:1–10 doi: 10.1186/1471-2148-6-4

Janzen DH (1970) Herbivores and number of tree species in tropical forests. Am Nat 104:501–28

Joppa LN, Roberts DL, Pimm SL (2011) How many species of flowering plants are there? Proc R Soc B. 278:554–559

Jonzen N, Cardinale M, Gårdmark A, Arrhenius F, Lundberg P (2002) Risk of collapse in the eastern Baltic cod fishery. Mar Ecol Prog Ser 240:225–233

Kahneman D, Tversky A (1982) Variants of uncertainty. In: Kahneman D, Slovic P, Tversky A (eds) Judgements under uncertainty: heuristics and biases. Cambridge University Press, Cambridge, pp 509–520

Karavarsamis N, Hamilton AJ (2010) Estimators of annual infection risk. J Water Health 80:365–373 doi:10.2166/wh.2010.045

Mara DD, Sleigh PA, Blumenthal UJ, Carr RM (2007) Health risks in wastewater irrigation: comparing estimates from quantitative microbial risk analyses and epidemiological studies. J Water Health 5:39–50 doi: 10.2166/wh.2006.055

May RM (1990) How many species? Philos Trans R Soc Lond Ser B 330:293–304 doi:10.1098/rstb.1990.0200

May RM (2010) Tropical Arthropod Species, More or Less. Science 329:41–42 doi: 10.1126/science.1191058

McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21:239–245

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on earth and in the ocean. PLoS Biol 9:e1001127

Morgan MG, Henrion M (1990) Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis, 1st edn. Cambridge University Press, Cambridge

Nielsen ES, Mound LA (2000) Global diversity of insects: the problems of estimating numbers. – In: Raven PH, Williams T (eds) Nature and Human Society: the Quest for a Sustainable World. National Academy Press, Washington D.C. pp 212 – 222

Novotný V, Basset Y, Miller SE, Weiblen GD, Bremer B, Cizek L, Drozd P (2002) Low host specificity of herbivorous insects in a tropical forest. Nature 416:841–844 doi:10.1038/416841a

Ødegaard F (2000) How many species of arthropods? Erwin's estimate revised. Biol J Linn Soc 71:583–597 doi:10.1006/bijl.2000.0468

Regan HM, Hope BK, Ferson S (2002) Analysis and portrayal of uncertainty in a food web exposure model. HERA 8:1757–1777

Regan HM, Ferson S, Berleant D (2004) Equivalence of five methods for bounding uncertainty. Int J Approx Reason 36:1–30

Ross HH, Ross CA, Ross JPR (1988) A Textbook of Entomology. Fourth Edition. John Wiley and Sons, New York.

Stork NE (1988) Insect diversity: facts, fiction and speculation. Biol J Linn Soc 35:321–337 doi: 10.1111/j.1095-8312.1988.tb00474.x

Stork NE (1993) How many species are there? Biodiversity Conserv 2:215–232 doi: 10.1007/BF00056669

Thomas CD (1990) Fewer species. Nature 347**:** 237 doi:10.1038/347237a0

Tucker WT, Ferson S (2003) Probability bounds analysis in environmental risk assessment 1st edn. Applied Biomathematics, Setauket

Williamson RC, Downs T (1990) Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds. Int J Approx Reason 4:89–158

Wilson SP, Costello MJ (2005) Predicting future discoveries of European marine species by using a nonhomogeneous renewal process. App Stat 54:897–918

Vose D (2000) Risk analysis: a quantitative guide, 2nd ed. John Wiley and Sons Ltd, Chichester

Figure 1. (A) Cumulative Distribution Function of the average effective specialization across all tree species (x). (B, C, D) Empirical probability bounds (solid lines) for the Cumulative Distribution Functions of the correction factor for non-herbivorous arthropods (c_{EH}), the proportion of all arthropods found in the canopy (p_{cgEH}), and the proportion of beetles that are arthropods (p_{baEH}), respectively. The dashed lines in (B), (C), and (D) represent the Cumulative Distribution Function for Uniform distributions defined as follows: c = Uniform (1.79, 2.70), p_{cg} = Uniform (0.25, 0.66), p_{ba} = Uniform (0.18, 0.33).

Figure 2. (A) Cumulative Distribution Function for the original Monte Carlo estimator N_{Ai} (dotted curved line) and associated 5^{th} and 95^{th} confidence limits (dotted vertical lines). The filled circle marks the median. B–F Probability bounds (solid lines) for the estimation of tropical arthropod species richness using the following estimators: (B) N_{Ad}: the original estimator but with no dependency assumptions; (C) N_{AiMM}: all variables assumed to be statistically independent from each other and represented with min-max p-boxes; (D) N_{AdMM}: no dependency assumptions, min-max p-boxes; (E) N_{AiEH}: statistical independence; empirical histogram p-boxes; (F) N_{AdEH}: no dependency assumptions, empirical histogram p-boxes. The dotted line in each plot represents the Cumulative Distribution Function for N_{Ai}.

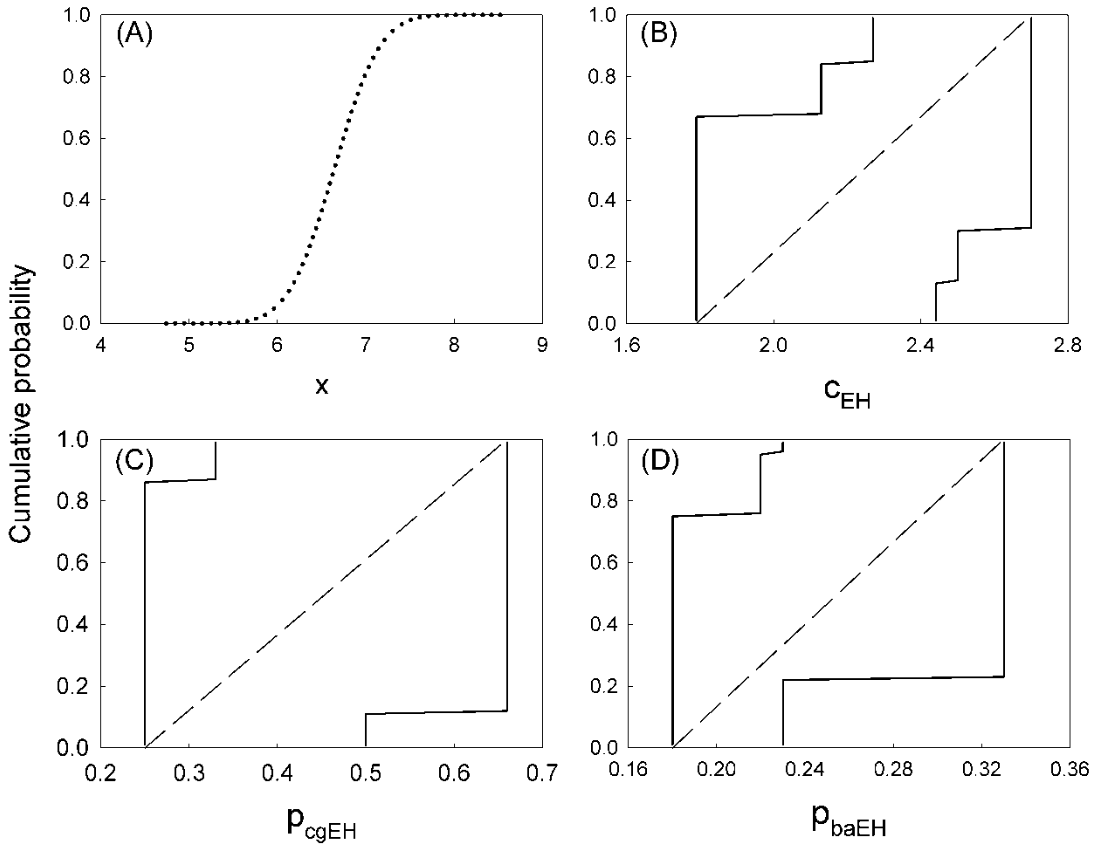
(A)
(B)
(C)
(D)
Cumulative probability
x
c_EH
p_cgEH
p_baEH
1.0
0.8
0.6
0.4
0.2
0.0
4 5 6 7 8 9
1.6 2.0 2.4 2.8
0.2 0.3 0.4 0.5 0.6 0.7
0.16 0.20 0.24 0.28 0.32 0.36

Figure 1.

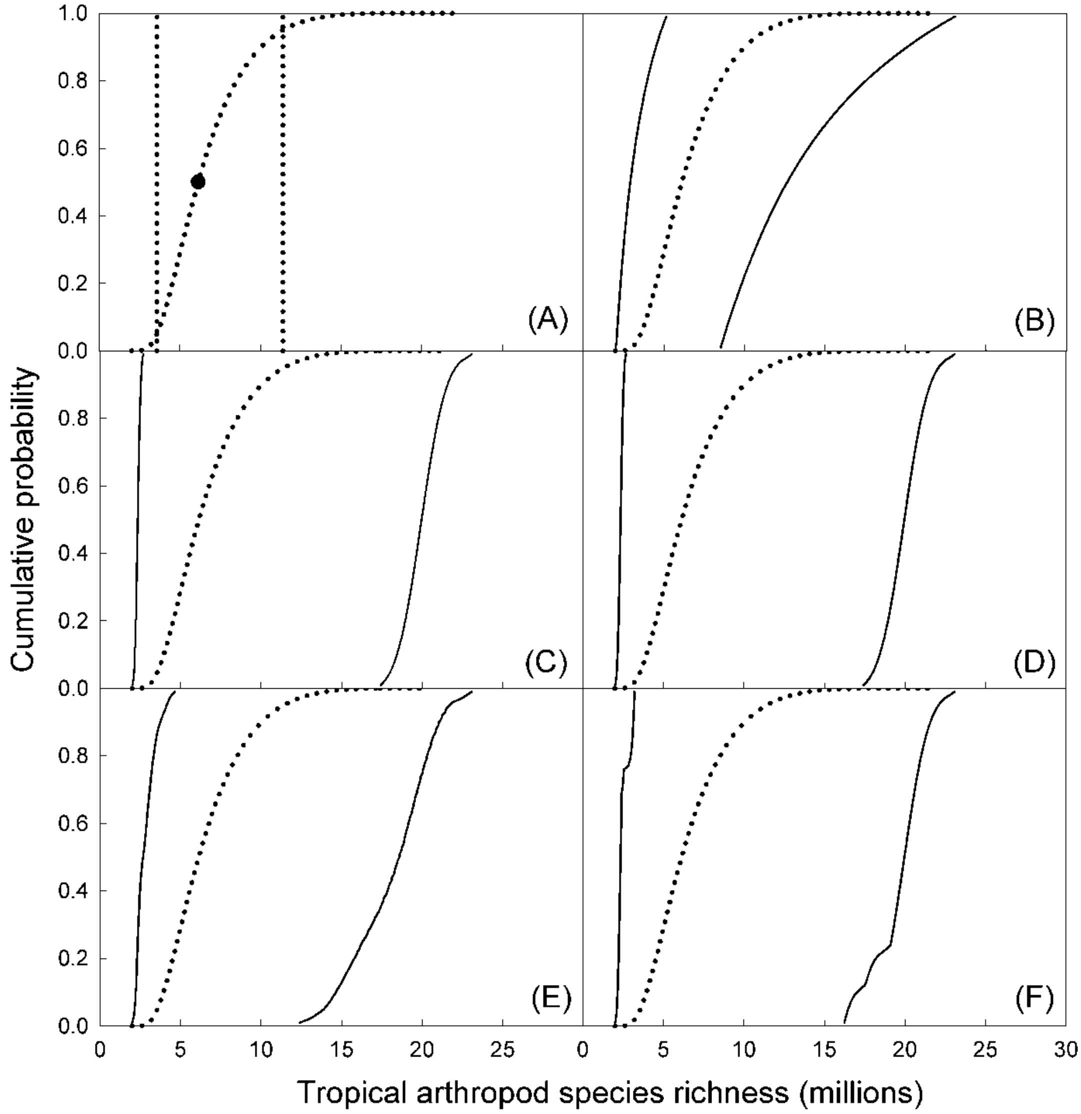
Cumulative probability
1.0
0.8
0.6
0.4
0.2
0.0
(A)
(B)
(C)
(D)
(E)
(F)
0
5
10
15
20
25
30
Tropical arthropod species richness (millions)

Figure 2.