Conservation Biology

Contributed Paper

Extinction rate of discovered and undiscovered plants in Singapore

Nadiah P. Kristensen,[1] Wei Wei Seah,[2] Kwek Yan Chong,[1] Yi Shuen Yeoh,[3] Tak Fung,[1] Laura M. Berman,[4] Hui Zhen Tan,[1] and Ryan A. Chisholm[1] *

[1]Department of Biological Sciences, National University of Singapore, 16 Science Drive 4, Singapore, 117558, Singapore
[2]Singapore Botanic Gardens, Herbarium Singapore, 1 Cluny Road, Singapore, 259569, Singapore
[3]National Parks Board, Horticulture and Community Gardening Division, 1 Cluny Road, Singapore, 259569, Singapore
[4]Division of Science, Yale-NUS College, 16 College Avenue West, Singapore, 138527, Singapore

Abstract: Extinction is a key issue in the assessment of global biodiversity. However, many extinction rate measures do not account for species that went extinct before they could be discovered. The highly developed island city–state of Singapore has one of the best-documented tropical floras in the world. This allowed us to estimate the total rate of floristic extinctions in Singapore since 1822 after accounting for sampling effort and crypto extinctions by collating herbaria records. Our database comprised 34,224 specimens from 2076 native species, of which 464 species (22%) were considered nationally extinct. We assumed that undiscovered species had the same annual per-species extinction rates as discovered species and that no undiscovered species remained extant. With classical and Bayesian algorithms, we estimated that 304 (95% confidence interval, 213–414) and 412 (95% credible interval, 313–534) additional species went extinct before they could be discovered, respectively; corresponding total extinction rate estimates were 32% and 35% (range 30–38%). We detected violations of our 2 assumptions that could cause our extinction estimates, particularly the absolute numbers, to be biased downward. Thus, our estimates should be treated as lower bounds. Our results illustrate the possible magnitudes of plant extirpations that can be expected in the tropics as development continues.

Keywords: biodiversity loss, generalized fiducial inference, historical extinctions, inferred extinctions, Southeast Asia, species-area relationship, undescribed extinctions, undescribed species

Tasa de Extinción de Plantas Descubiertas y No Descubiertas en Singapur

Resumen: La extinción es un tema importante para la valoración de la biodiversidad global. Sin embargo, muchas medidas de la tasa de extinción no consideran a las especies que se extinguieron antes de que pudieran ser descubiertas. Singapur, la ciudad-estado isleña altamente desarrollada, tiene una de las floras mejor documentadas del mundo. Esto nos permitió estimar la tasa total de las extinciones florísticas en Singapur desde 1822 después de considerar el esfuerzo de muestreo y las criptoextinciones cuando recopilamos los registros de herbarios. Nuestra base de datos incluyó 34,224 especímenes de unas 2,076 especies nativas, de las cuales 464 especies (22%) estaban consideradas como extintas a nivel nacional. Asumimos que las especies no descubiertas tuvieron la misma tasa anual de extinción por especie que las especies descubiertas y que ninguna especie no descubierta permanecía en existencia. Con algoritmos clásicos y bayesianos, respectivamente, estimamos que 304 (95% IC 213–414) y 412 (95% IC 313–534) especies adicionales se extinguieron antes de que fueran descubiertas; las estimaciones correspondientes de la tasa de extinción total fueron 32% y 35% (rango de 30–38%). Detectamos violaciones en nuestras dos suposiciones que podrían causar que nuestras estimaciones de extinción, particularmente los números absolutos, tuvieran un sesgo hacia abajo. Por lo tanto, nuestras estimaciones deberían ser tratadas como límites inferiores. Nuestros resultados ilustran las magnitudes posibles de las extirpaciones de plantas que pueden esperarse en los trópicos conforme el desarrollo continúa.

*Address correspondence to R. A. Chisholm, email ryan.chis@gmail.com
Article impact statement: The number of species that went extinct before they could be discovered can be estimated from museum records.
Paper submitted May 27, 2019; revised manuscript accepted February 28, 2020.

DOI: 10.1111/cobi.13499

Palabras Clave: especies no descritas, extinciones históricas, extinciones inferidas, extinciones no descritas, inferencia generalizada de referencia, pérdida de biodiversidad, relación especie-área, sureste de Asia

摘要: 物种灭绝是全球生物多样性评估的一个关键问题。然而, 许多灭绝率估计中并没有考虑到那些在被发现之前就已经灭绝的物种。高度发达的岛国新加坡拥有世界上记载最完整的热带植物群, 这使得我们能够通过整理草本植物记录来统计采样工作量和隐秘的灭绝事件, 以估计新加坡自1822 年以来植物灭绝的总速率。本研究的数据库包括来自 2076 个本地物种的 34, 224 个标本, 其中有 464 个物种 (22%) 被认为在全国范围内灭绝。我们假设未被发现的物种年均灭绝率与已发现物种相同, 且已没有未被发现的现存物种。利用经典算法和贝叶斯算法, 我们分别估计出 304 个 (95% 置信区间为 213 - 414) 和 412 个 (95% 可信区间为 313 - 534) 新物种在被发现之前就已经灭绝; 相应的总灭绝率估计分别为 32% 和 35% (范围为 30 - 38%) 。我们还发现存在违反上述两个假设的情况, 这可能导致我们对灭绝情况的估计值偏低, 特别是对绝对数值的估计。因此, 我们对灭绝情况的估计值应被视为下限。本研究表明, 随着热带地区的发展, 植物灭绝的规模可能会不断扩大。**【翻译:胡怡思;审校:聂永刚】**

关键词: 生物多样性丧失, 广义置信推断, 历史灭绝, 推断的灭绝, 东南亚, 物种-面积关系, 未描述的灭绝, 未描述的物种, 处死控制, 社会认同的方法, 野生动物管理

Introduction

Globally, and especially in the tropics, many species remain undiscovered, whereas known and unknown species continue to go extinct (Costello et al. 2013). This is also true on national and regional scales. Comprehensive analyses for regions with long survey histories and thorough records are invaluable for shedding light on such processes.

The island city–state of Singapore (103°50′E, 01°20′N; originally 520 km^2; currently 724 km^2 due to land reclamation; population 5.64 million) has been used as a case study of tropical biodiversity loss (Corlett 1992; Turner et al. 1994). It is one of few developed countries in the tropics. Since British colonization in 1819, nearly all its original forest cover has been destroyed, although substantial regrowth has occurred. Primary forest currently occupies 0.28% of its landmass. Forested area (primary, old and young secondary, mangrove, and freshwater swamp) totals 22.47% (Yee et al. 2011). Fortunately, the biota of Singapore, especially plants, has been relatively well documented. Collections began soon after 1819, and tens of thousands of specimens are stored in local and international herbaria.

Documenting historical species extinction presents 2 challenges: inferring that a known species is extinct and accounting for species that went extinct before they could be discovered. Depending on data and resources available, there are many ways to infer extinction of known species, including heuristics (e.g., Davison et al. 2008), combined extinction risk metrics and expert judgement (e.g., Szabo et al. 2012), and statistical analyses of detection records (e.g., Solow 2005). Currently, multiple models are combined in a cost–benefit framework to categorize species (Akçakaya et al. 2017). The extinction of undiscovered species has received comparatively less attention.

Although extinction of undiscovered species may have a large effect on extinction rate estimates (Hawksworth & Cowie 2013), they are difficult to account for. It is simplest to assume that the taxonomic group of interest has the same extinction rate as another better-known group (e.g., extinction rate of birds applied to insects [Dunn 2005]). Another approach is to estimate original species richness based on species composition nearby in relatively undisturbed areas (e.g., Brook et al. 2003; Alcala et al. 2004). Alternatively, a statistical or phenomenological relationship can be used (Turner et al. 1994; Pitman et al. 2002) (e.g., a power-law relationship [Preston 1962] between species richness and habitat area remaining [e.g., Turner et al. 1994]). Where detailed records exist, a mark–recapture-like method on specimen attributes has been used (Pimm et al. 1994; Duncan et al. 2013). If one assumes extinction and discovery rates are constant over time, a parametric statistical model may be used (Tedesco et al. 2014).

Chisholm et al. (2016) introduced a nonparametric method for estimating undiscovered extinctions based on the assumption that extinction probabilities of undiscovered and discovered species are equal within each year. We used 2 algorithms to obtain interval estimates from this model, explored the model's assumptions, and applied it to records of vascular terrestrial plants in Singapore.

Methods

Data and Discovered Species

Electronic records of plant specimens from Singapore were collated and resolved to create a database of native vascular plants (Supporting Information). Species names were resolved with respect to synonymy and redeterminations. Unresolvable names and records without species-specific names, collection year, or collector name were removed.

Table 1. A hypothetical historical record and example calculations of cumulative probability of extinction *P* (see text). The two undiscovered cases are two different potential realizations of outcomes that are consistent with the historical record but differ by the number of extant undiscovered species remaining (marked ∗).

	Year t			
	0	*1*	*2*	*Calculations of* P
Historical record				
discovered extant S_t	100	250	300	
discovered extinct E_t	0	50	100	
discoveries δ_t	200	100	–	
Calculations				
extinction rate μ_t	0.5	0.2	–	$1 - (1 - 0.5)(1 - 0.2) = 0.6$
Assuming $U_2 = 0$				
undiscovered extant U_t	650	125	0∗	
undiscovered extinct X_t	0	325	350	$\frac{100+350}{750} = 0.6$
Assuming $U_2 = 100$				
undiscovered extant U_t	900	250	100∗	
undiscovered extinct X_t	0	450	500	$\frac{100+500}{1000} = 0.6$

Experts in plant identification for field research and conservation classified species as extant or extinct. Unclassified species that had not been collected within the past 30 years were designated extinct, in keeping with standard practice for plants (e.g., Chong et al. 2009). For other species collected within the past 30 years, we assessed current status in Singapore with the Solow (1993) method, which we chose for its simplicity, and a collection effort correction (McCarthy 1998) (Supporting Information). Seventy species had $p < 0.1$ (null hypothesis that they are extant rejected); however, many of these were known to be common. Therefore, experts were consulted a second time to reassess and classify each of the 70 species (Supporting Information).

The method we used to fit redetection effort, $c(t)$, to the detection records also fitted species' intrinsic redetection probabilities:

$$r_i = \frac{\sum_t I_R(i,t)}{\sum_{t \in T_i} c(t)}, \tag{1}$$

where $I_R(i,t) = 1$ if species i was redetected in year t or $I_R(i,t) = 0$ if not, and T_i is the set of all years that the species was known to be extant and therefore available to be redetected. A species' r_i reflects all factors that influenced its relative propensity to be collected, such as conspicuousness, abundance, and research interest.

Simplified extinction rate example

A naive way to estimate extinction rates is to divide the current number of extinct species by the total number of species. Given the hypothetical historical record in Table 1, this gives $100/(300 + 100) = 0.25$. Chisholm et al.'s (2016) method improves on this by accounting for species discoveries and temporal fluctuations in extinction rate. The extinction probability is 1 minus the cumulative probability of persistence. For the example in Table 1, this gives $P = 1 - (1 - 0.5)(1 - 0.2) = 0.6$, which is substantially higher than the naive estimate.

Calculating the cumulative persistence probability effectively accounts for species that went extinct before they were discovered, provided they had the same extinction rates as discovered species. Historical numbers of undiscovered, extinct species can then be estimated by working backward. For the example in Table 1, assuming no undiscovered species remain extant, $U_2 = 0$, then in previous years $U_1 = 100/(1 - 0.2) = 125$ and $U_0 = (125 + 200)/(1 - 0.5) = 650$. Then, the total number of species $N_0 = U_0 + S_0 = 750$, and the total number of undiscovered extinctions $X_2 = N_0 - S_2 - E_2 - U_2 = 350$. This recovers the total extinction rate: $(E_2 + X_2)/N_0 = (100 + 350)/750 = 0.6 = P$. Alternatively, if we assume $U_2 = 100$, then $U_0 = 900$, $X_2 = 500$, and again $(100 + 500)/1000 = 0.6 = P$. With this method, the number of undiscovered extinctions has a linear relationship with the number of undiscovered extant species in the current year T: $X_T = (P(S_T + E_T + U_T) - E_T)/(1 - P)$. This example simplifies the problem by using the observed species extinction rates directly. A more complete description would infer confidence intervals on the estimates.

SEUX model

The SEUX model (Chisholm et al. 2016), which we name for its four variables, tracks the number of species in each year t in 1 of 4 states: discovered and extant, S_t; discovered and extinct, E_t; undiscovered and extant, U_t; and undiscovered and extinct, X_t. The time series $\mathbf{S} = (S_0, \ldots, S_T)$ and $\mathbf{E} = (E_0, \ldots, E_T)$ are known from data (e.g., herbaria records), but $\mathbf{U} = (U_0, \ldots, U_T)$ and $\mathbf{X} = (X_0, \ldots, X_T)$ are unknown. The ultimate goal is to infer the total number of undiscovered extinctions, X_T.

Within each year, species can transition from extant to extinct ($S \rightarrow E$ or $U \rightarrow X$) or undiscovered to

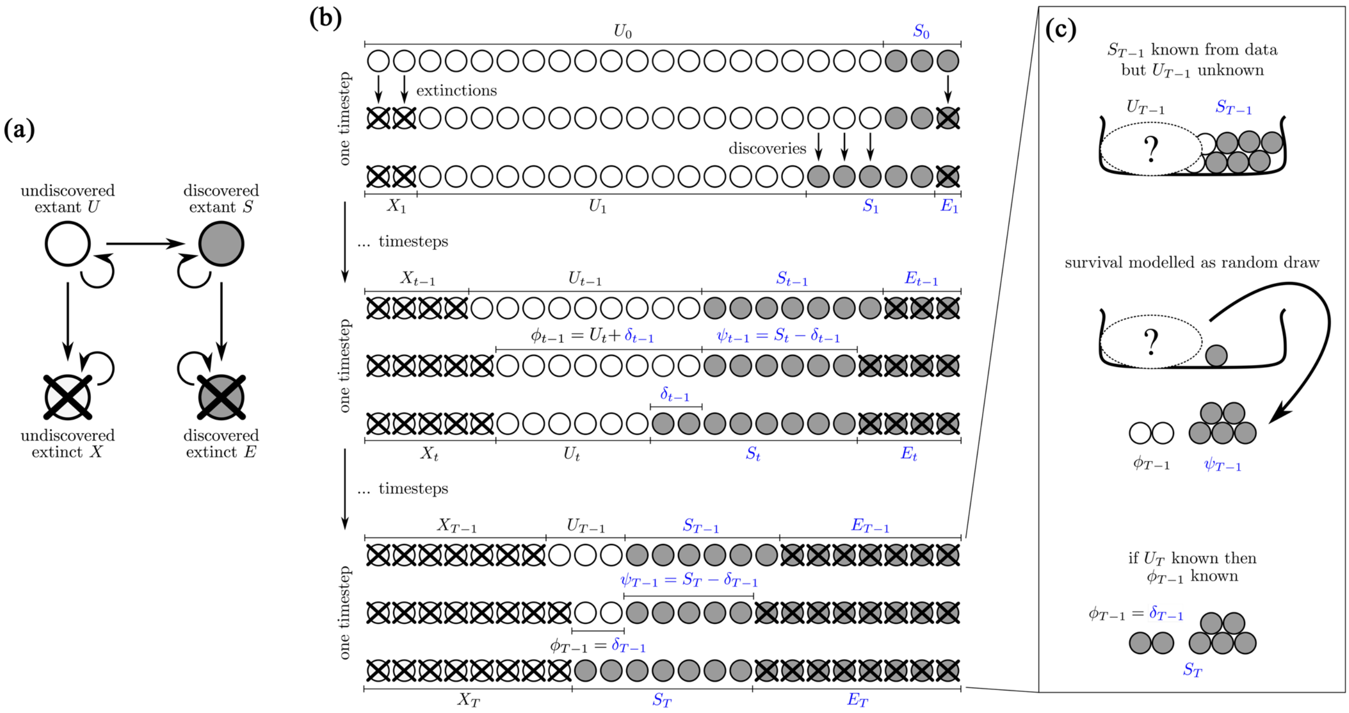


Figure 1. The SEUX model showing (a) possible states and transitions and a hypothetical scenario with (b) a whole time series with variables and (c) the operation of the hypergeometric model within a time step. Within each time step, it is assumed that survival or extinction occurs before discovery, and survival is modeled as a random draw in an urn model. The scenario shown has no undiscovered extant species remaining at the end of the observation period, $U_T = 0$. If U_T is known, then ϕ_{T-1} can be calculated as shown in (b).

discovered ($U \rightarrow S$). We assumed a species discovered in a particular year would not go extinct in that same year (i.e., transition U to E not permitted), which is equivalent to assuming that the process of survival and extinction precedes discovery within each year (Fig. 1).

The initial conditions were S_0 observed, $E_0 = 0$, U_0 unknown, and $X_0 = 0$. Over the observation period $t = 1, \ldots, T$, the variables were updated as follows:

$$S_t = \Psi_{t-1} + \Delta_{t-1}, \qquad (2)$$

$$E_t = E_{t-1} + S_{t-1} - \Psi_{t-1} \qquad (3)$$

$$U_t = \Phi_{t-1} - \Delta_{t-1}, \qquad (4)$$

$$X_t = X_{t-1} + U_{t-1} - \Phi_{t-1}, \qquad (5)$$

where Ψ_{t-1} is the number of surviving discovered species (observed), Φ_{t-1} is the number of surviving undiscovered species (unobserved), and Δ_{t-1} is the number of species discovered (observed) in the previous year.

The key assumption is that extinction probabilities in each year, μ_t, are the same for discovered and undiscovered species. We also assumed a discovery probability, ν_t for each year t. Therefore, the random variables above are binomially distributed:

$$\Psi_{t-1} \sim \text{Bi}\left(S_{t-1}, 1 - \mu_{t-1}\right), \qquad (6)$$

$$\Phi_{t-1} \sim \text{Bi}\left(U_{t-1}, 1 - \mu_{t-1}\right), \qquad (7)$$

$$\Delta_{t-1} \sim \text{Bi}\left(\phi_{t-1}, \nu_{t-1}\right), \qquad (8)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are vectors of unknown parameters.

Let $N_t = \Phi_{t-1} + \Psi_{t-1} = U_t + S_t$ be the total number of species that survive to a given year. Because Ψ_{t-1} and Φ_{t-1} are independent and share parameter $1 - \mu_{t-1}$, N_t is binomially distributed and $P(\mathbf{n} \mid \boldsymbol{\mu}, n_0 = U_0 + S_0) = \prod_{t=1}^{T} P(n_t \mid n_{t-1}, \mu_{t-1})$. The number of discovered (or undiscovered) survivors at each time step is hypergeometric with sample size n_t. So Eqs. 6 and 7 can be replaced with

$$N_t \sim \text{Bi}\left(n_{t-1}, 1 - \mu_{t-1}\right) \qquad (9)$$

$$n_t - \Phi_{t-1} = \Psi_{t-1} \sim \text{Hyp}\left(n_{t-1}, S_{t-1}, n_t\right). \qquad (10)$$

Thus, the joint probability of the data conditional on $\mathbf{n}$ and $\boldsymbol{\nu}$ is

$$P_{\mathbf{n},\boldsymbol{\nu}}(\mathbf{S}, \boldsymbol{\psi}) = \prod_{t=1} T^{P}_{n_{t-1},n_t}(\psi_{t-1} \mid S_{t-1})\, P_{\nu_{t-1},n_t}(S_t \mid \psi_{t-1}), \qquad (11)$$

where

$$P_{n_{t-1},n_t}\left(\psi_{t-1} \mid S_{t-1}\right) = \frac{\binom{S_{t-1}}{\psi_{t-1}}\binom{n_{t-1}-S_{t-1}}{n_t-\psi_{t-1}}}{\binom{n_{t-1}}{n_t}}, \quad (12)$$

$$P_{\nu_{t-1},n_t}\left(S_t \mid \psi_{t-1}\right) = \binom{n_t-\psi_{t-1}}{S_t-\psi_{t-1}} \nu_{t-1}^{S_t-\psi_{t-1}}(1-\nu_{t-1})^{n_t-S_t}. \quad (13)$$

Equation 12 is the probability of obtaining a number of discovered extant species surviving from year $t-1$ to year t (ψ_{t-1}) given the number of discovered extant species in year $t-1$ (S_{t-1}). Equation 13 is the probability of obtaining the number of discovered species in year t after accounting for discoveries during the last year and ϕ_{t-1}.

Algorithms to Infer Undiscovered Extinctions

To estimate the total number of undiscovered extinct species, X_T, Chisholm et al. (2016) assumed that the number of undiscovered species extant in the present day, U_T, was known. Below, we devised algorithms to obtain classical confidence intervals for X_T, which likewise treat U_T as known, and Bayesian credible intervals, which incorporate a U_T prior. Both methods yield interval estimates for U_0 from which intervals for X_T are calculated. We introduced the parameter ω to explore the model's sensitivity to the assumption that extinction probabilities in each year are the same for discovered and undiscovered species.

Classical Confidence Intervals

Classical confidence intervals for $\mathbf{U}$ can be estimated using the quantiles of a large sample from its confidence distribution (Xie & Singh 2013). We sampled a candidate confidence distribution with a step-by-step algorithm, which started with the U_T assumption and worked backward in time, sequentially sampling each U_{t-1} conditional on components in later years (c.f. example in Table 1). The algorithm was derived using the generalized fiducial approach (Hannig et al. 2016). The statistical model implied a data-generating algorithm, which we inverted to obtain a fiducial distribution for $\mathbf{U}$.

The likelihood in Eq. 11 translates directly into a data-generating algorithm. Let us define the cumulative distributions

$$B_{n_t,n_{t+1}}\left(\psi_t, S_t\right) = 1 - P_{n_t,n_{t+1}}\left(\Psi_t \le \psi_t \mid S_t\right) \quad (14)$$

and

$$C_{\nu_t,n_{t+1}}\left(S_{t+1}, \psi_t\right) = 1 - P_{\nu_t,n_{t+1}}\left({S_{t+1}}' \le S_{t+1} \mid \psi_t\right), \quad (15)$$

which correspond to Eqs. 12 and 13, respectively. Then, given parameters $\mathbf{n}$, $\boldsymbol{\nu}$, a sample $\mathbf{S}, \boldsymbol{\psi}$ can be obtained with the following algorithm:

```
Initialise $S_0$
For $t = 0 \dots T-1$
    $\alpha_1 \leftarrow$ rand01()
    $\psi_t \leftarrow B^{-1}_{n_t,n_{t+1}}(\alpha_1, S_t)$
    $\alpha_2 \leftarrow$ rand01()
    $S_{t+1} \leftarrow C^{-1}_{\nu_t,n_{t+1}}(\alpha_2, \psi_t)$        (16)
```

The inverse functions correspond to sampling from discrete distributions (Lemieux 2009).

The data-generating algorithm can be inverted to sample $\mathbf{n}$ given $\mathbf{S}$ and $\boldsymbol{\psi}$ observed:

```
initialize $N_T = U_T + S_T$
For $t = T-1 \dots 0$
    $\alpha \leftarrow$ rand01()
    $n_t \leftarrow B^{-1}_{\psi_t,S_t}\left(\alpha, n_{t+1}\right)$        (17)
```

The sample $\mathbf{U}$ can be calculated from $\mathbf{n}$ ($U_t = n_t - S_t$) or by rearranging the inverse function to sample $\mathbf{U}$ directly. We took the latter approach. We replaced $B_{n_t,n_{t+1}}(\psi_t, S_t)$ with the equivalent formulation

$$H_{U_t,U_{t+1}}\left(\psi_t, S_t\right) = 1 - P_{U_t,U_{t+1}}\left(\Psi_t \le \psi_t \mid S_t\right), \quad (18)$$

where

$$P_{U_{t-1},U_t}\left(\psi_{t-1} \mid S_{t-1}\right) = \frac{\binom{S_{t-1}}{\psi_{t-1}}\binom{U_{t-1}}{U_t+S_t-\psi_{t-1}}}{\binom{S_{t-1}+U_{t-1}}{S_t+U_t}}. \quad (19)$$

Then, we obtained a sample $\mathbf{U}$ with the following algorithm:

```
initialize $U_T$
For $t = T-1 \dots 0$
    $\alpha \leftarrow$ rand01()
    $U_t \leftarrow H^{-1}_{\psi_t,S_t}\left(\alpha, U_{t+1}\right)$        (20)
```

We used a mid-P correction to account for data discreteness (Supporting Information). Via repeated sampling of $\mathbf{U}$ as described above, a probability function $h_{\mathbf{S},\boldsymbol{\psi}}(\mathbf{U})$ is implicitly defined, which is a generalized fiducial distribution for U.

In general, generalized fiducial inference can be used to obtain a parameter distribution, following which the quality of the procedure is evaluated, for example, with simulations (Hannig 2013). Our algorithm (20) performed well in simulations, producing U_0 confidence intervals with coverage that matched or exceeded the nominal value for a wide range of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ scenarios (Supporting Information). To formally prove the algorithm is correct, we would need to prove that $H_{U_t,U_{t+1}}(\psi_t, S_t)$ combined with the discreteness correction satisfies all technical requirements of a confidence distribution. Then, our algorithm would be equivalent to the step-by-step approach of Veronese and Melilli (2018).

To quantify the effect of the assumption that undiscovered and discovered species have equal survival probabilities, we introduced their odds ratio as a parameter

$$\omega = \frac{\left(\frac{1-\mu_{u,t}}{\mu_{u,t}}\right)}{\left(\frac{1-\mu_{s,t}}{\mu_{s,t}}\right)}, \tag{21}$$

where $\mu_{s,t}$ and $\mu_{u,t}$ are the extinction probabilities for discovered and undiscovered species, respectively, and ω is assumed constant in time. Then Eq. 19 is replaced with Fisher's noncentral hypergeometric distribution:

$$P_{U_{t-1},U_t}\left(\psi_{t-1}S_{t-1}\right) = \frac{\begin{pmatrix} S_{t-1} \\ \psi_{t-1} \end{pmatrix} \begin{pmatrix} U_{t-1} \\ U_t + S_t - \psi_{t-1} \end{pmatrix} \omega^{U_t+S_t-\psi_{t-1}}}{\sum_{y=y_{\min}}^{y_{\max}} \begin{pmatrix} U_{t-1} \\ y \end{pmatrix} \begin{pmatrix} S_{t-1} \\ U_t + S_t - y \end{pmatrix} \omega^{y}},$$

$$y_{\min} = \max\left(0, S_t + U_t - S_{t-1}\right), \text{ and}$$

$$y_{\max} = \min\left(S_t + U_t, U_{t-1}\right). \tag{22}$$

Setting $\omega = 1$ in Eq. 22 retrieves Eq. 19, $\omega < 1$ means the extinction probabilities of undiscovered species are higher than discovered species, and $\omega > 1$ means they are lower. We explored the effect of ω on the results of the classical inference.

Bayesian Credible Intervals

We conducted our analysis in a Bayesian framework and used Markov chain Monte Carlo (MCMC) sampling with a Metropolis-within-Gibbs sampler (Gelman et al. 2004). The conditional dependency between the unobserved variables has a linear structure; therefore, a Gibbs sequence is obtained by sampling sequentially from

$$\begin{aligned} \Phi_t^{(j)} &\sim P_{S,\psi}\left(\phi_t \mid \phi_{t+1}^{(j-1)}, \phi_{t-1}^{(j)}\right) \\ &\propto P_{S,\psi}\left(\phi_{t+1}^{(j-1)} \mid \phi_t\right) P_{S,\psi}\left(\phi_t \mid U_t^{(j)}\right), \end{aligned} \tag{23}$$

where j is the sample number, the relationship $\phi_{t-1} = U_t + S_t - \psi_{t-1}$ links unknown parameters, and probabilities on the right side are calculated by rewriting Eq. 12 as a function of ϕ_t and U_t (details in Supporting Information). A sampling distribution must also be defined at the end points of the chain. We imposed a uniform prior on U_0 and, to match the default setting for the classical intervals above, we assumed $U_T = 0$. Alternatively, an informative prior for U_T can be sampled instead.

Two independent chains were monitored for convergence with trace plots and the Gelman–Rubin convergence diagnostic ($\hat{R} < 1.1$ for all U_t [Gelman & Shirley 2011]); a burn-in of 15,000 iterations was consequently discarded. To ensure samples were sufficiently large, we verified that: the combined sample size exceeded the minimum multivariate effective sample size for 95th percentile bounds with a tolerance of $\varepsilon = 0.1$ for every U_t; the multivariate MCMC SE on the mean and 95th percentile bound estimates was small relative to the SD in the posterior (Vats et al. 2017); and the posterior distributions produced by the two chains were visually similar.

Time Step Size and Time Range of Data

To match most herbaria records, a natural time step length for the SEUX model is 1 year; however, this choice is arbitrary and influences model predictions. For example, one could arbitrarily decrease the time step length and thus increase the number of time steps in which no discovered extinctions occur. Because the model allows undiscovered extinctions to occur in time steps with no discovered extinctions, this would arbitrarily increase total extinction rate. To avoid this, we required that every time step have at least 1 discovered extinction. Years with no discovered extinctions were combined with later years within a time step, so time step lengths varied. This procedure may also provide a rough method for accounting for unevenness of collection effort over time.

The starting date chosen for the time series also influences the model results due to the small sample size effect. The decision involves a trade-off between inferring extinctions farther back in time versus reliability of those estimates. The model itself can be used to explore the effects of that decision. We chose a start date of 1822 because it coincides with the first extensive collection of plants in Singapore (Burkill 1927) and by that time 183 species had been collected.

Results

Data and Discovered Species

We collated 34,224 records of plants in Singapore. Periods of intensive collection were 1822, the 1890s, and the 1990s and 2000s. Collection rates were low during WWII and the early 1980s. The cumulative number of species discovered has not yet plateaued (Fig. 2).

The fitted redetection effort reflected collector activity and WWII (Fig. 3a). Effort generally increased through the last century, except for recent years, for which collections are yet to be entered into databases. The distribution of species' intrinsic redetection probabilities was bimodal (Fig. 3b). The relationship of average intrinsic redetection probability to year of discovery was U-shaped, gradually declining over the last 2 centuries but rapidly increasing again after 1980 (Supporting Information).

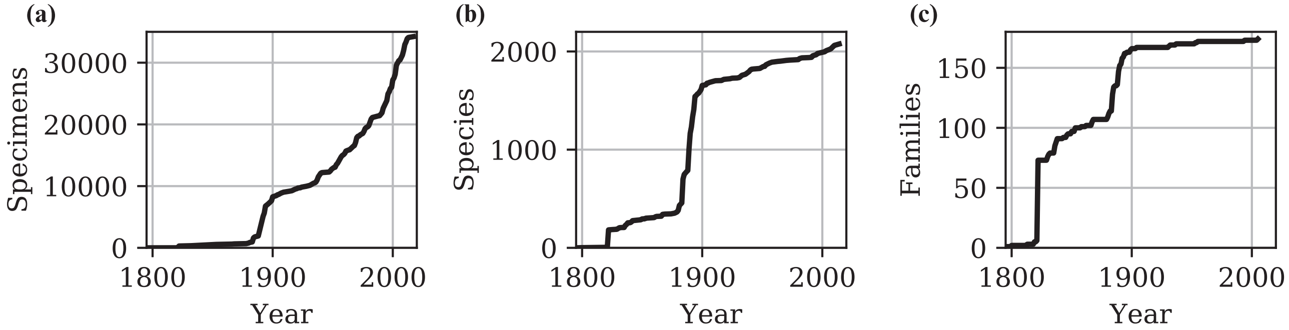


Figure 2. The cumulative number of (a) specimens, (b) species, and (c) families in the combined collections database of plant species of Singapore. The complete database of species detection records is in Supporting Information.

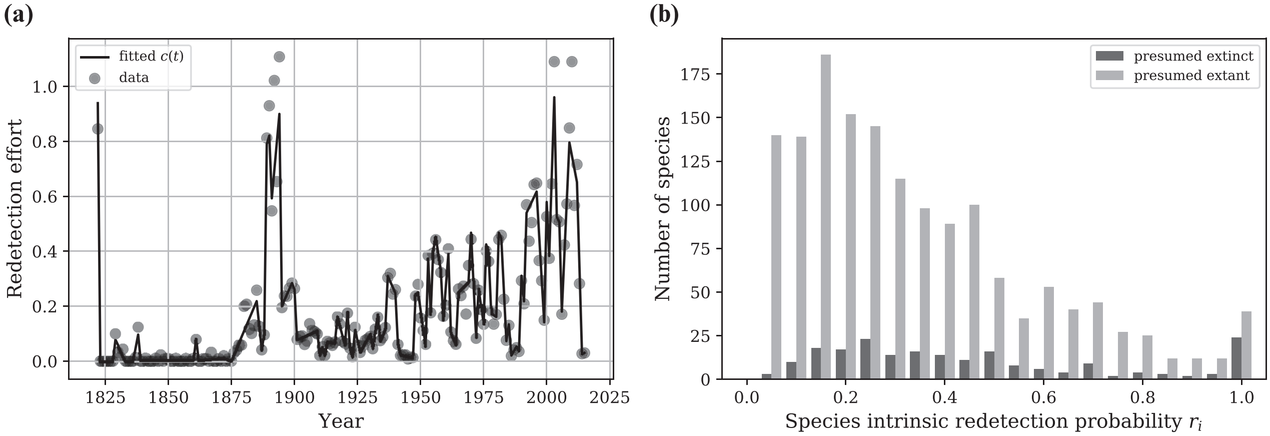


Figure 3. The (a) fitted spline of redetection effort over time for Singapore plants, c(t), and (b) the distribution of inferred species' intrinsic redetection probabilities (r_i) (equation in Supporting Information), each of which gives the number of redetections of a species divided by the total redetection effort over its known lifetime. For this figure, species are presumed extant if they are designated as common in Chong et al. (2009), if they are judged to be extant by experts, or if they were last collected after 1985. Otherwise, they are presumed extinct. The list of species and their corresponding redetection probabilities are in Supporting Information.

Of the 2076 plant species included in the detection records, 464 were inferred extinct (Fig. 4a), for an estimated discovered extinction rate of 22% (Table 1).

Extinction of Undiscovered Species and Total Extinction Rate

Assuming that discovered and undiscovered species have equal extinction probabilities within a time step, and assuming that no undiscovered extant species remain in the present, then the 2 algorithms estimated that 304 (95% confidence interval, 213–414) and 412 (95% credible interval, 313–534) undiscovered extinctions have occurred since 1822 (Fig. 4a), with the Bayesian method giving the higher estimate. The total extinction rate from both methods was higher than the naïve estimate but lower than that of a previous study of Singapore plants (Brook et al. 2003) (Table 2).

As expected, the estimated total extinction rate increased as the survival probability of undiscovered species was experimentally decreased via the parameter ω (Fig. 4b). An odds ratio of $\omega = 0.17$ was required to match Brook et al.'s (2003) high estimate (Supporting Information). The total extinction rate estimate was not sensitive to random species deletions from the record or variation in the current number of undiscovered extant species (Supporting Information). However, the estimated absolute number of undiscovered extinct species increased as U_T increased (Fig. 4c & Supporting Information).

Discussion

Extinction Rates in Singapore and Implications for Southeast Asia

We collated a rich botanical data set from 2 centuries of plant collections and showed how the data can be used to estimate total extinction rates. Accounting

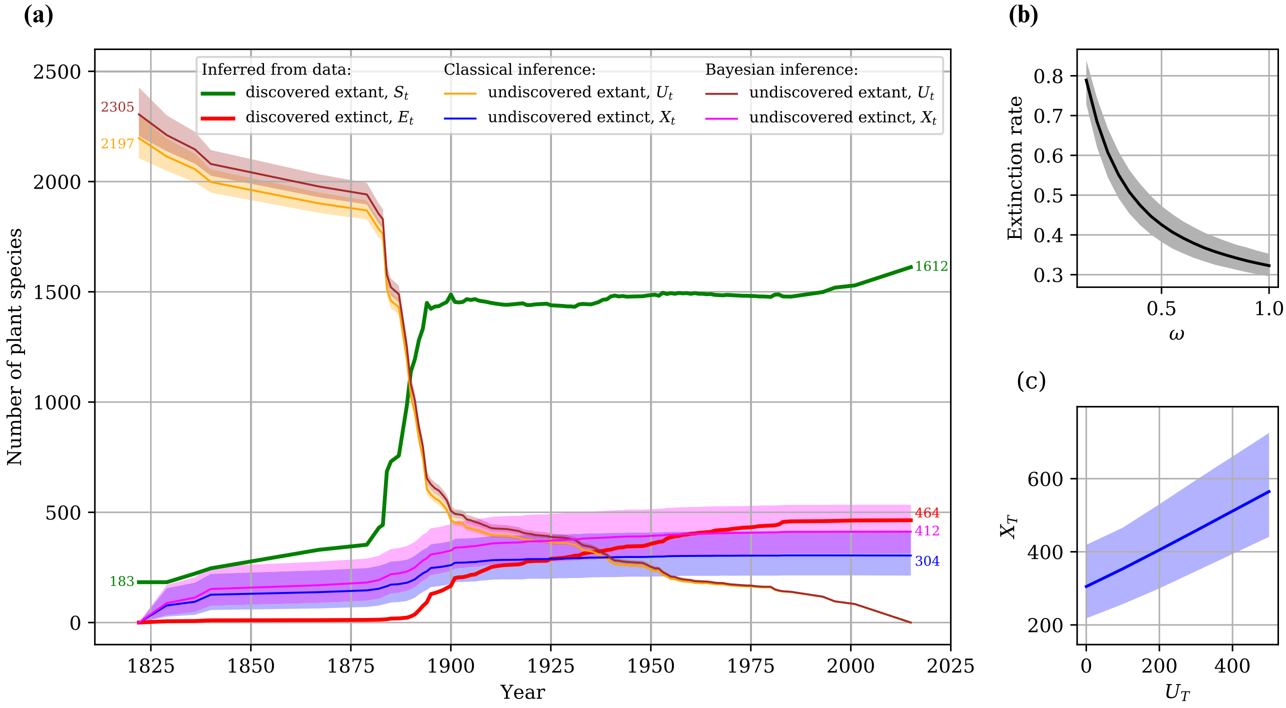


Figure 4. The (a) SEUX model estimates for Singapore plants given the default parameter values $\omega = 1$, $U_T = 0$, and (b and c) responses of the estimates to variation in those parameters (solid lines, mean values; shading, 95% confidence or credible intervals from classical or Bayesian inference). Code is available in Supporting Information.

Table 2. Estimates (95% CI) of extinction rates for plants in Singapore from this study and previous studies.

Description and source	*Total*	*Extinct*	*Extinct (%)*
Discovered species only			
Turner et al. 1994; Brook et al. 2003	2277	594	26
Chong et al. 2009	2145	639	30
This study	2076	464	22
Total: discovered + undiscovered			
Brook et al. 2003	6549	4866	74
This study: classical	2380 (2289,2490)	768 (677, 878)	32 (30, 35)
This study: Bayesian	2488 (2389,2610)	876 (777, 998)	35 (33, 38)

solely for discovered plant extinctions gave an estimated extirpation rate of 22% over 200 years. Accounting for undiscovered plant extinctions as well led to higher total extinction rate estimates of 32–35% (total range 30–38%) (Table 2). The uncertainty range was moderate, which reflected the difficulty of inferring undiscovered extinctions even under our model's simplifying assumptions. Nevertheless, these numbers are similar to the estimates for Singapore birds (Chisholm et al. 2016). Extrapolating our numbers to Southeast Asia under projected deforestation rates (Brook et al. 2003; Chisholm et al. 2016), we estimated that 17–18% of plant species will be extirpated regionally by 2100 (for comparison, 28–33% of plant species are currently classified as threatened [Joppa et al. 2010]).

Although our extinction estimates are high, they are much lower than Brook et al.'s (2003) estimate of 74% Singapore plant extinctions in the present day and 46% regional plant extinctions by 2100. The discrepancy between our estimate and Brook et al.'s (2003) is attributable to their extreme assumption that the original species assemblage in Singapore was similar to that of comparable ecosystems throughout Peninsular Malaysia—an area over 100 times the size (Jain et al. 2018). This assumption violates one of ecology's few laws: the species–area relationship (Lawton 1999; Lomolino & Weiser 2001). If region A has an area that is 1% of region B, then an empirical rule of thumb (a power-law species–area relationship with an exponent z = 0.2 to 0.3) is that region A will have 60–75% fewer species than region B. Brook et al.'s (2003) method attributes this portion of difference in species richness to extinctions rather than a regular manifestation of the species–area relationship. Their method may have further overestimated

Singapore's historic species richness because Singapore (an island) is partially isolated from the mainland species pool and Peninsular Malaysia spans over 5° of latitude, whereas Singapore spans only a fraction of a degree.

Are our estimated plant extinction rates in Singapore consistent with ecological theory? Based on the species–area relationship, if the forested area is reduced to 2.83% of its original extent (current extent of primary, old secondary, mangrove, and freshwater swamp [Yee et al. 2011]), one would expect roughly 51–66% of the species to have gone extinct. Why have these high rates of extinction not been observed? Turner et al. (1994) performed a similar calculation for Singapore plants and suggested that higher theoretical rates were not observed because extinction debt has not yet been paid. Although we agree extinction debt remains, this does not explain the discrepancy because the species–area relationship itself ignores extinction debt. We suggest 2 more likely explanations. First, traditional species–area relationships implicitly assume that remaining habitat is contiguous (Pereira et al. 2012; Chisholm et al. 2018), whereas Singapore's remnant forest is highly fragmented and thus captures some beta diversity. When fragmentation is accounted for, theoretical estimates of tree species extinctions in Singapore accord well with reality (Chisholm et al. 2018). A second explanation is that almost one-fifth of Singapore contains young secondary growth (Yee et al. 2011), which harbors a substantial proportion of the original species.

Further plant extinctions can be expected in Singapore. The extinction curve has not plateaued (Fig. 4a), and there is likely an outstanding extinction debt to be paid in future (Vellend et al. 2006; Hahs et al. 2009), which may be exacerbated by isolation of remnant habitat from immigrants (Drayton & Primack 1996) and impediments to recruitment to secondary forest (Goldsmith et al. 2011). To mitigate future extinctions, a priority is to preserve existing forest remnants, including secondary forests. Increasing connectivity between existing patches will facilitate dispersal and increase effective population sizes. Our plants database may assist species-specific targeted conservation efforts by identifying species that are probably extant but infrequently sighted indicating rarity and vulnerability to extinction.

SEUX Model Assumptions and Recommendations

The SEUX model, like other extinction rate estimation techniques (e.g., E/MSY [Pimm et al. 2014; Tedesco et al. 2014]), assumes that the average extinction rates of discovered and undiscovered species are the same. However, if undiscovered species actually have higher extinction probabilities, then the method underestimates the total extinction rate. We introduced the ω parameter in the SEUX model to quantify this effect and further assess the plausibility of the Brook et al. (2003) estimate. Estimated extinction rates are larger for lower ω. The Brook et al.'s (2003) estimate requires $\omega = 0.17$, indicating that undiscovered species were approximately 5 times as likely to go extinct as discovered species. Although $\omega = 0.17$ is probably unrealistically low, the true odds ratio may differ somewhat from $\omega = 1$.

There are several factors that could lead to $\omega < 1$ (i.e., undiscovered species having higher extinction probabilities than discovered species). However, our understanding of these relationships is largely qualitative. All else being equal, species with low abundance are harder to detect and more extinction prone (e.g., McCarthy et al. 2014). The link between low abundance and extinction risk is a general phenomenon (McKinney 1997) and has been observed specifically for habitat fragmentation (table 2 in Henle et al. 2004) and plants (Matthies et al. 2004; Sutton & Morgan 2009). Similarly, small species (Sutton & Morgan 2009; Marini et al. 2012) and species with restricted geographical ranges (e.g., Scheffers et al. 2012) are harder to detect and more vulnerable to habitat loss. Undiscovered species may also be more extinction prone because they cannot benefit from species-targeted conservation. However, collectors often have a bias toward novel specimens and rare finds (Guralnick & Van Cleve 2005; Pyke & Ehrlich 2010), which we expect to ameliorate $\omega < 1$ to an unknown degree. Obtaining reasonable quantitative estimates for ω is an area for future work.

A particular vulnerability of SEUX is that rare species, which tend to be simultaneously hard to detect and extinction prone, may be preferentially lost soon after habitat destruction begins, leading to underestimation of early undiscovered extinctions. To overcome this, collections need to include a large, early, and representative sample. We judge the size and timing of early plant collections in Singapore to be largely adequate because within 3 years of the arrival of British colonialists, 183 species from 73 families had already been collected. Nevertheless, we found some evidence for low early coverage of extinction-bound species. We observed that early-discovered Singapore species had higher intrinsic redetection probability and lower extinction probability (Supporting Information), which suggests that early collectors sampled easier-to-detect species that were less extinction prone. This implies ω decreases the farther back in time one goes.

We expect low coverage of early extinction-bound species to be the main challenge for future workers. Early collectors generally focus on capturing a wide spectrum of biodiversity, whereas recent collectors tend to focus on species of conservation concern (Boakes et al. 2010). Recently discovered species are generally more likely to be threatened (Giam et al. 2012) and have a narrower range (Treurnicht et al. 2017). However, these general patterns are influenced by the taxonomic group and collectors' biases (Guralnick & Van Cleve 2005; Boakes et al. 2010; Pyke & Ehrlich 2010). For example, early Singapore collections included some species-rich families with high extinction rates (notably orchids [Supporting Information]). Future workers interested in characterizing museum-collection biases may find our new method for inferring redetection effort useful. It addresses some of the shortcomings of previous methods (Duffy et al. 2009) and produced estimates of Singapore plants' intrinsic redetection probabilities as a byproduct (Supporting Information).

To estimate the absolute number of undiscovered extinctions (as opposed to the percentage), SEUX additionally requires information about how many extant species remain undiscovered (U_T); however, U_T is—by definition—unknown. If a prior can be obtained for U_T (e.g., expert opinion), that can be easily incorporated into our Bayesian approach to obtain credible intervals for X_T. Our classical approach, however, requires a U_T value (incorporating a U_T confidence distribution is theoretically possible but it is unclear how it would be obtained). In some circumstances, $U_T = 0$ is a reasonable assumption (e.g., Singapore birds [Chisholm et al. 2016]); otherwise, another method (reviewed in Chao & Chiu [2016]) can be used to estimate U_T. Either way, the X_T estimate must be interpreted as predicated on the U_T value (Fig. 4c). For Singapore plants, we followed Chisholm et al. (2016) and assumed $U_T = 0$. However, given the lack of plateau in Fig. 2b, that new records and rediscoveries continue to be made throughout nature reserves (e.g., Chong et al. 2018; Ho et al. 2018; Khoo et al. 2018), and that existing species have been reassessed recently as new species (e.g., Niissalo et al. 2014), this is almost certainly false. Therefore, the absolute number of undiscovered extinctions we found should be interpreted as a lower bound, additional to the effect of overestimating ω discussed above.

The temporal pattern of extinctions produced by SEUX can be informative, provided it is interpreted with care. We assumed species went extinct the year after last detection, which ensures a conservatively higher estimate of total extinctions, but also means that the extinction pattern is influenced by the pattern of species discovery. Many of the extinctions appearing in the uptick in the late 1800s (Fig. 4a) are more likely to have occurred in the first two decades of the 1900s, when plantations decimated the secondary forests that had replaced the original primary forest after initial deforestation (Corlett 1992).

Future workers may be interested in more sophisticated methods for inferring discovered extinctions. Structured elicitation methods can be used for expert determinations (Keith et al. 2017). We used one of the simpler statistical methods (reviews in Solow [2005], Rivadeneira et al. [2009], and Boakes et al. [2015]); however, if additional data are available (e.g., on species-specific threats and survey quality), then these can be incorporated (Akçakaya et al. 2017). Nevertheless, we found that even our simple method was useful to supplement heuristics and narrow the list of species that required closer scrutiny from experts.

Acknowledgments

We gratefully acknowledge the various herbaria for making their data freely available online, the participants in GBIF (Global Biodiversity Information Facility), and the Kew Herbarium Catalogue. We specifically thank the Singapore Botanic Gardens for access to the SING database and the physical collections, as well as H. T. W. Tan and K. S. Chua with the Lee Kong Chian Natural History Museum for use of the SINU database and physical collection. We are also grateful to C. C. Ho, P. Leong, J. Leong-Škorničková, R. Lim, J. Loh, L. Neo, M. Niissalo, J. Strijk, and J. Tan for expert opinions and checking and correcting entries in the database and to J. G. Lefevre, C. Hamalainen, M. Waterhouse, A. R. Solow, P. Veronese, E. Melilli, and J. Hannig for comments on the statistical model. N.P.K. was funded by Singapore Ministry of Education Academic Research Fund (Tier 1 WBS R-154-000-A12-114).

Data and Code Accessibility

The Singapore plant collections provide an unusually rich database. We hope future researchers will find this data set and code (Supporting Information) useful as a testing ground for their work.

Supporting Information

Data and code accessibility (Appendix S1), inferring redetection effort (Appendix S2), inferring discovered extinctions (Appendix S3), details of classical inference (Appendix S4), details of Bayesian inference (Appendix S5), additional results for Fisher's extended SEUX (Appendix S6), and additional results for sensitivity of the SEUX model (Appendix S7) are available online. The authors are solely responsible for the content and functionality of these materials. Queries (other than absence of the material) should be directed to the corresponding author.

Literature Cited

Akçakaya H, Keith DA, Burgman M, Butchart SH, Hoffmann M, Regan HM, Harrison I, Boakes E. 2017. Inferring extinctions III: a cost-benefit framework for listing extinct species. Biological Conservation **214:**336–342.

Alcala E, Alcala A, Dolino C. 2004. Amphibians and reptiles in tropical rainforest fragments on Negros Island, the Philippines. Environmental Conservation **31:**254–261.

Boakes EH, McGowan PJ, Fuller RA, Chang-qing D, Clark NE, O'Connor K, Mace GM. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biology **8:**e1000385.

Boakes EH, Rout TM, Collen B. 2015. Inferring species extinction: the use of sighting records. Methods in Ecology and Evolution **6:**678–687.

Brook BW, Sodhi NS, Ng PK. 2003. Catastrophic extinctions follow deforestation in Singapore. Nature **424:**420–426.

Burkill IH. 1927. Botanical collectors, collections and collecting places in the Malay Peninsula. Gardens' Bulletin Straits Settlements **4:**113–202.

Chao A, Chiu C-H. 2016. Species richness: estimation and comparison. Pages 1–26 In Balakrishnan N, Colton T, Everitt B, Piegorsch W, Ruggeri F, Teugels J, editors. Wiley StatsRef: statistics reference online. John Wiley & Sons, Hoboken, New Jersey. https://doi.org/10.1002/9781118445112.stat03432.pub2.

Chisholm RA, Giam X, Sadanandan KR, Fung T, Rheindt FE. 2016. A robust nonparametric method for quantifying undetected extinctions. Conservation Biology **30:**610–617.

Chisholm RA, Lim F, Yeoh YS, Seah WW, Condit R, Rosindell J. 2018. Species–area relationships and biodiversity loss in fragmented landscapes. Ecology Letters **21:**804–813.

Chong K, Lim R, Loh J, Neo L, Seah W, Tan S, Tan H. 2018. Rediscoveries, new records, and the floristic value of the Nee Soon freshwater swamp forest, Singapore. Gardens' Bulletin Singapore **70:**49–69.

Chong KY, Tan HT, Corlett RT. 2009. A checklist of the total vascular plant flora of Singapore: native, naturalised and cultivated species. Raffles Museum of Biodiversity Research, National University of Singapore, Singapore.

Corlett RT. 1992. The ecological transformation of Singapore, 1819–1990. Journal of Biogeography **19:**411–420.

Costello MJ, May RM, Stork NE. 2013. Can we name Earth's species before they go extinct? Science **339:**413–416.

Davison GW, Ng PK, Ho H. 2008. The Singapore red data book: threatened plants and animals of Singapore. Nature Society (Singapore), Singapore.

Drayton B, Primack RB. 1996. Plant species lost in an isolated conservation area in metropolitan Boston from 1894 to 1993. Conservation Biology **10:**30–39.

Duffy KJ, Kingston NE, Sayers BA, Roberts DL, Stout JC. 2009. Inferring national and regional declines of rare orchid species with probabilistic models. Conservation Biology **23:**184–195.

Duncan RP, Boyer AG, Blackburn TM. 2013. Magnitude and variation of prehistoric bird extinctions in the Pacific. Proceedings of the National Academy of Sciences **110:**6436–6441.

Dunn RR. 2005. Modern insect extinctions, the neglected majority. Conservation Biology **19:**1030–1036.

Gelman A, Carlin JB, Stern HS, Rubin D. 2004. Bayesian data analysis. 2nd edition. Chapman & Hall/CRC, Boca Raton, Florida.

Gelman A, Shirley K. 2011. Inference from simulations and monitoring convergence. Pages 163–174 In Brooks S, Gelman A, Jones G, Meng X-L, editors. Handbook of Markov chain Monte Carlo. Chapman & Hall/CRC, Boca Raton, Florida.

Giam X, Scheffers BR, Sodhi NS, Wilcove DS, Ceballos G, Ehrlich PR. 2012. Reservoirs of richness: least disturbed tropical forests are centres of undescribed species diversity. Proceedings of the Royal Society B **279:**67–76.

Goldsmith GR, Comita LS, Chua SC. 2011. Evidence for arrested succession within a tropical forest fragment in Singapore. Journal of Tropical Ecology **27:**323–326.

Guralnick R, Van Cleve J. 2005. Strengths and weaknesses of museum and national survey data sets for predicting regional species richness: comparative and combined approaches. Diversity and Distributions **11:**349–359.

Hahs AK, et al. 2009. A global synthesis of plant extinction rates in urban areas. Ecology Letters **12:**1165–1173.

Hannig J. 2013. Generalized fiducial inference via discretization. Statistica Sinica **23:**489–514.

Hannig J, Iyer H, Lai RC, Lee TC. 2016. Generalized fiducial inference: a review and new results. Journal of the American Statistical Association **111:**1346–1361.

Hawksworth DL, Cowie RH. 2013. The discovery of historically extinct, but hitherto undescribed, species: an under-appreciated element in extinction-rate assessments. Biodiversity and Conservation **22:**2429–2432.

Henle K, Davies KF, Kleyer M, Margules C, Settele J. 2004. Predictors of species sensitivity to fragmentation. Biodiversity and Conservation **13:**207–251.

Ho B, Lua H, Leong P, Lindsay S, Seah W, Ibrahim B, Loo A, Koh S, Ibrahim A, Athen P. 2018. New records and rediscoveries of vascular plants in Bukit Timah Nature Reserve, Singapore. Gardens' Bulletin Singapore **70:**33–55.

Jain A, Khoon KS, Gan CW, Webb EL. 2018. Butterfly extirpations, discoveries and rediscoveries in Singapore over 28 years. Raffles Bulletin of Zoology **66:**217–257.

Joppa LN, Roberts DL, Pimm SL. 2010. How many species of flowering plants are there? Proceedings of the Royal Society B: Biological Sciences **278:**554–559.

Keith DA, Butchart SH, Regan HM, Harrison I, Akçakaya HR, Solow AR, Burgman MA. 2017. Inferring extinctions I: a structured method using information on threats. Biological Conservation **214:** 320–327.

Khoo M, Chua S, Lum S. 2018. An annotated list of new records for Singapore: results from large-scale tree surveys at the Bukit Timah Nature Reserve. Gardens' Bulletin Singapore **70:**57–65.

Lawton JH. 1999. Are there general laws in ecology? Oikos **84:**177–192.

Lemieux C. 2009. Monte Carlo and quasi-Monte Carlo sampling. Springer, New York.

Lomolino M, Weiser M. 2001. Towards a more general species-area relationship: diversity on all islands, great and small. Journal of Biogeography **28:**431–445.

Marini L, Bruun HH, Heikkinen RK, Helm A, Honnay O, Krauss J, Kühn I, Lindborg R, Pärtel M, Bommarco R. 2012. Traits related to species persistence and dispersal explain changes in plant communities subjected to habitat loss. Diversity and Distributions **18:**898–908.

Matthies D, Bräuer I, Maibom W, Tscharntke T. 2004. Population size and the risk of local extinction: empirical evidence from rare plants. Oikos **105:**481–488.

McCarthy MA. 1998. Identifying declining and threatened species with museum data. Biological Conservation **83:**9–17.

McCarthy MA, Moore AL, Krauss J, Morgan JW, Clements CF. 2014. Linking indices for biodiversity monitoring to extinction risk theory. Conservation Biology **28:**1575–1583.

McKinney ML. 1997. Extinction vulnerability and selectivity: combining ecological and paleontological views. Annual Review of Ecology and Systematics **28:**495–516.

Niissalo MA, Wijedasa LS, Boyce PC, Leong-Skornickova J. 2014. *Hanguana neglecta* (Hanguanaceae): a new plant species from a heavily collected and visited reserve in Singapore. Phytotaxa **188:**14–20.

Pereira HM, Borda-de-Água L, Martins IS. 2012. Geometry and scale in species–area relationships. Nature **482:**E3–E4.

Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts CM, Sexton JO. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. Science **344:**1246752.

Pimm SL, Moulton MP, Justice LJ, Collar N, Bowman D, Bond W. 1994. Bird extinctions in the central Pacific. Philosophical Transactions of the Royal Society of London B: Biological Sciences **344:** 27–33.

Pitman NC, Jørgensen PM, Williams RS, León-Yánez S, Valencia R. 2002. Extinction-rate estimates for a modern Neotropical flora. Conservation Biology **16:**1427–1431.

Preston FW. 1962. The canonical distribution of commonness and rarity: part I. Ecology **43:**185–215.

Pyke GH, Ehrlich PR. 2010. Biological collections and ecological/environmental research: a review, some observations and a look to the future. Biological Reviews **85:**247–266.

Rivadeneira MM, Hunt G, Roy K. 2009. The use of sighting records to infer species extinctions: an evaluation of different methods. Ecology **90:**1291–1300.

Scheffers BR, Joppa LN, Pimm SL, Laurance WF. 2012. What we know and don't know about Earth's missing biodiversity. Trends in Ecology & Evolution **27:**501–510.

Solow AR. 1993. Inferring extinction from sighting data. Ecology **74:**962–964.

Solow AR. 2005. Inferring extinction from a sighting record. Mathematical Biosciences **195:**47–55.

Sutton FM, Morgan JW. 2009. Functional traits and prior abundance explain native plant extirpation in a fragmented woodland landscape. Journal of Ecology **97:**718–727.

Szabo JK, Khwaja N, Garnett ST, Butchart SH. 2012. Global patterns and drivers of avian extinctions at the species and subspecies level. PLOS ONE **7**(e47080) https://doi.org/10.1371/journal.pone.0047080.

Tedesco P, Bigorne R, Bogan A, Giam X, Jézéquel C, Hugueny B. 2014. Estimating how many undescribed species have gone extinct. Conservation Biology **28:**1360–1370.

Treurnicht M, Colville JF, Joppa LN, Huyser O, Manning J. 2017. Counting complete? Finalising the plant inventory of a global biodiversity hotspot. PeerJ **5:**e2984.

Turner I, Tan H, Wee Y, Ibrahim AB, Chew P, Corlett R. 1994. A study of plant species extinction in Singapore: lessons for the conservation of tropical biodiversity. Conservation Biology **8:** 705–712.

Vats D, Flegal JM, Jones GL. 2017. Multivariate output analysis for Markov chain Monte Carlo. arXiv:1512.07713v4.

Vellend M, Verheyen K, Jacquemyn H, Kolb A, Van Calster H, Peterken G, Hermy M. 2006. Extinction debt of forest plants persists for more than a century following habitat fragmentation. Ecology **87:**542–548.

Veronese P, Melilli E. 2018. Fiducial, confidence and objective Bayesian posterior distributions for a multidimensional parameter. Journal of Statistical Planning and Inference **195:**153–173.

Xie M-g, Singh K. 2013. Confidence distribution, the frequentist distribution estimator of a parameter: a review. International Statistical Review **81:**3–39.

Yee A, Corlett RT, Liew S, Tan HT. 2011. The vegetation of Singapore – an updated map. Gardens' Bulletin Singapore **63:**205–212.

Supplementary Information

Extinction rate of discovered and undiscovered plants in Singapore

Kristensen, et al. (2020)

Table of Contents

S1 Data and code accessibility

S1.1 Overview

Fig. S1 provides an overview of the repository, which includes the plants database and the code used to obtain the results. Both can be accessed at: `https://github.com/nadiahpk/inferring-undiscovered-species-extinctions`, archived with Zenodo DOI: `10.5281/zenodo.3733920`

Readers who wish to implement the SEUX model on their own data may prefer an R package, which is under development at: `https://github.com/nadiahpk/seux`.

The plants database is stored as a comma-separated values file `merged.csv`. With such a large dataset there are bound to be errors and omissions; during the period of compilation the process of species redetermination and reclassification was ongoing, and we expect errors to be concentrated around issues of synonymisation and redetermination in particular. Therefore we have endeavoured to be as transparent as possible, e.g. the complete species list includes unique identifying barcodes, so that future updates and corrections to the record can be easily made.

The input to the SEUX model was derived from the plants database. After processing to take into account expert knowledge, discovered extinctions within the past 30 years, etc., the first and last detections for each species were stored in `first_last_detns_final.csv`. This processed data is then used as the input to the SEUX model, to obtain the key result of the paper (Fig. 4), which combines results obtained in the `mcmc` and `classical` directories.

Readers may also be interested in intermediate results produced. An estimate of each species' intrinsic redetection probability was produced as a by-product of inferring the collection effort (details in S2), which is stored in `intrinsic_redetection_probabilities.csv`. The *P*-values resulting from the Solow (1993) method of inferring species extinction, after accounting for redetection effort, are found in `solows_pvalue.csv`.

```
repository
├── data
│   ├── cleaned_plants_database
│   │   ├── experts_extant.csv  (which species are common and extant according to experts)
│   │   └── merged.csv  (the plants database merged from different herbaria and museums)
│   ├── processed
│   │   └── first_last_detns_final.csv  (input to SEUX model)
│   ├── chong_2009_checklist
│   └── inferred_extinct_second_opinion  (recheck status of species inferred extinct by Solow method)
├── undetected_extinctions  (shared SEUX model functions)
├── scripts  (scripts to generate results)
│   ├── describe_data  (plot summaries of plants database)
│   ├── redetection_effort  (infer yearly redetection effort to modify Solow method)
│   ├── infer_detected_extinctions  (use redetection effort and Solow method to infer detected extinctions)
│   ├── process_data  (prepare data, incl. detected extinctions, for input into models)
│   ├── mcmc  (Bayesian inference)
│   │   ├── mcmc_basic_results.py  (obtain credible intervals)
│   │   ├── convergence_check.py  (check convergence)
│   │   └── mcmc_se_results.py  (check MCMC error)
│   ├── classical  (classical inference)
│   │   ├── classical.py  (obtain confidence intervals for basic model)
│   │   ├── fisher_variant.py  (obtain confidence inteverals for a particular Fisher's variant model)
│   │   ├── fisher_explore.py  (explore relationship between ω and total extinctions)
│   │   ├── fisher_relationship.py  (plot relationship between ω and total extinctions)
│   │   ├── sensitivity  (sensitivity analysis)
│   │   └── verify  (use simulations to verify that coverage is correct)
│   └── main_figure  (plot main figure combining Bayesian and classical result)
└── results  (intermediate and final results)
    ├── describe_data  (summaries of plants database)
    ├── redetection_effort  (yearly redetection effort to modify Solow method)
    │   └── intrinsic_redetection_probabilities.csv  (inferred species intrinsic redetection probabilities)
    ├── infer_detected_extinctions
    │   └── solows_pvalue.csv  (Solow's P-value for each species)
    ├── mcmc  (Bayesian inference)
    ├── classical  (classical inference)
    │   ├── classical_basic_result.csv  (confidence intervals for basic model)
    │   ├── classical_fisher_*.pdf  (confidence intervals for particular Fisher's variant models)
    │   ├── fisher_explore  (explore relationship between ω and total extinctions)
    │   ├── fisher_relationship  (relationship between ω and total extinctions)
    │   ├── sensitivity  (sensitivity analysis)
    │   └── verify*  (verification that coverage is correct)
    └── main_figure  (main figure combining Bayesian and classical result)
```

Figure S1: Directory structure of the repository, highlighting key data files, results, and code used to obtain the results.

S1.2 Plants database

S1.2.1 Details of data collation

Databases and herbaria records were acquired (dates in Table S1) and collated to create a record of detections of plant species in Singapore. The SING database was updated twice in order to resolve contradictions arising from an earlier tallying of the record (details below).

Table S1: Databases and date of acquisition of collections database.

herbarium	acquisition date
SING	October 2014
SINU	October 2014
KEW	February 2015
GBIF	October 2017
SING	September 2017 (redownload)

For all databases, the following records were removed:

- Species of algae, fungi, bryophytes, and seagrasses (families Cymodoceaceae (=Potamogetonaceae) and Hydrocharitaceae).
- Species that are cultivated only or cryptogenic weeds (following Chong et al., 2009).
- Records without species-specific names, or with names that could not be synonymised with a name in standardised databases such as Taxonstand (version 1.7, R package) and Tropicos and Global Names Index.
- Records with both collector name and date collected missing.
- Records without a barcode or other uniquely-identifying record information (e.g. a combination of barcode and catalogue number that could be used to uniquely identify the record).

  We chose to omit these records to ensure that the database could be updated when specimens are reidentified or redesignated in the future.
- Records with the same barcode but different dates or species names that could not be reconciled (i.e. by consulting the sheet date or by synonymisations)
- Records from non-wild localities and those likely to be cultivated.

  For the Singapore-based collections, this process relied upon expert knowledge, e.g. particular records from the Botanic Gardens Jungle were kept because they were judged to be remnant or natural colonisations.

For each record that was retained, a year of collection and species designation had to be determined. Varieties of species were not lumped together. Records with a collector name but no date were given a date range, based on other collections by the same collector and knowledge of historical collectors. For the purposes of the model, detection year was taken as the mid-point of this range.

The SING database, which contains the largest number of records, was obtained twice to resolve changes to the database in the time intervening. The initial acquisition led to an effort to clean the database with respect to redeterminations and synonymisation of specimens, and to add a by-hand count of physical specimens that were not found in the electronic database. However this process was complicated and contradictions were found between the database and the final species counts obtained. Therefore the SING database was reobtained in September 2017 to resolve the contradictions. This included removing earlier records whose barcode did not reappear in the most recent version, which would have been removed due to removal of duplicates, corrections to name/date, etc.

The final database therefore consists of the information from the various electronic databases, plus two columns that were appended to each record, which were used to create the final input that was used in the model. The two new columns are:

- 'standardised year': either the year in which specimen was collected, or if a year was not given, a range of years based on the collector name.
- 'standardised name': a species name chosen so that all synonyms share the same unique name, and redeterminations are accounted for.

S1.2.2 Summaries of the plants database

Table S2 and Fig. S2–S4 provide an overview of the content of the plants database.

Table S2: Summary of herbaria collections used.

	number of		
collection	specimens	species	families
Singapore Botanic Gardens Herbarium	21692	2020	174
Herbarium of the Lee Kong Chian Natural History Museum	7547	733	101
National Herbarium of the Netherlands	2352	810	137
Herbarium at the Royal Botanic Gardens Kew	1134	548	120
Other	1499	704	124
All combined	34224	2076	174

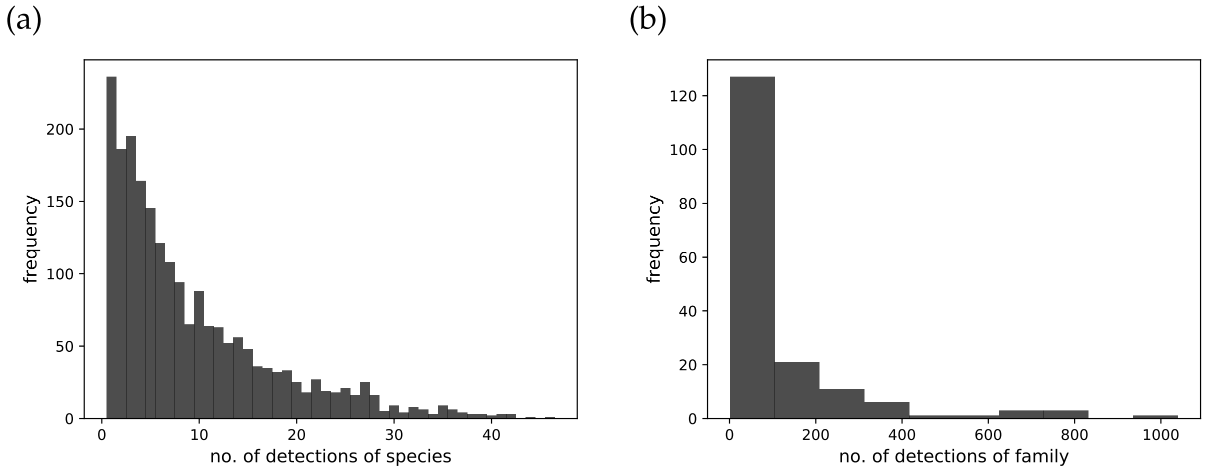


Figure S2: Frequency distribution of the number of detections in the database, by species (a) and by family (b).

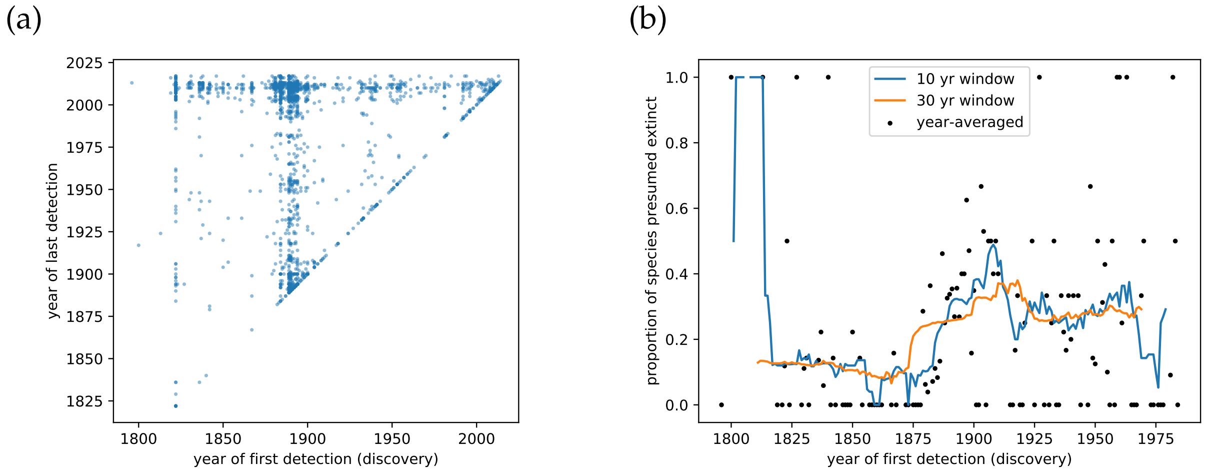


Figure S3: A scatter plot of each species' year of first versus last detection (a), and the relationship between a species' year of discovery and the probability of it being presumed extinct (b), where species were presumed extant based on expert advice, if common (Chong et al., 2012), or if detected since 1985.

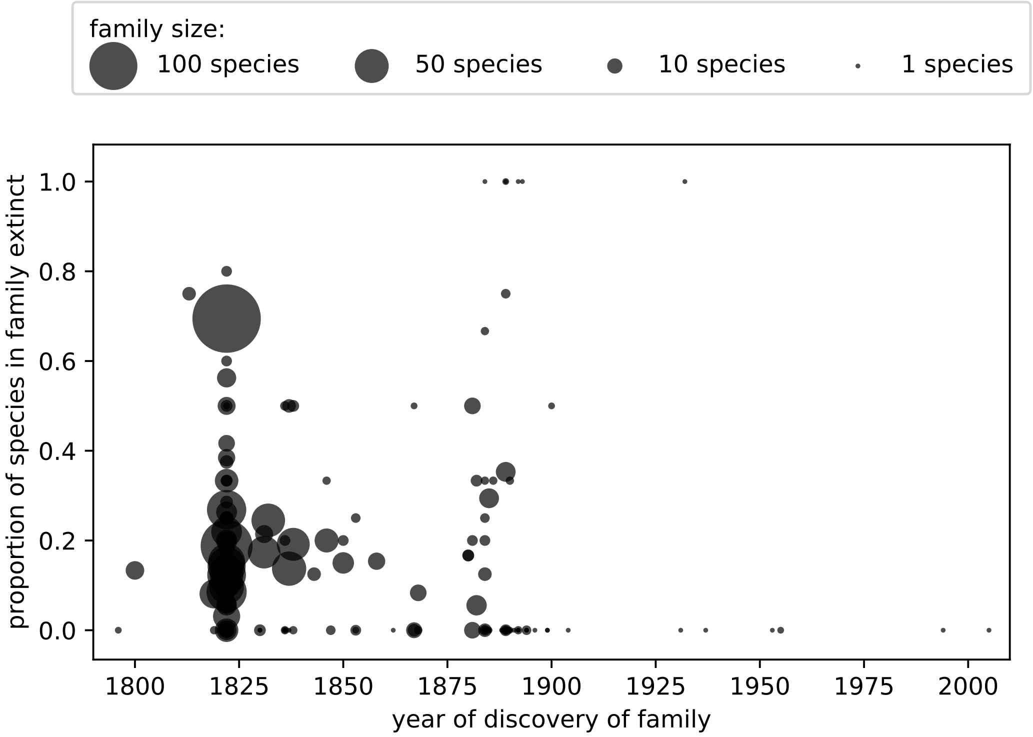


Figure S4: The proportion of extinctions in each family versus its year of discovery (first collection of a member species). The marker size is proportional to the number of species in the family. Status based on expert opinion or last detection since 1985.

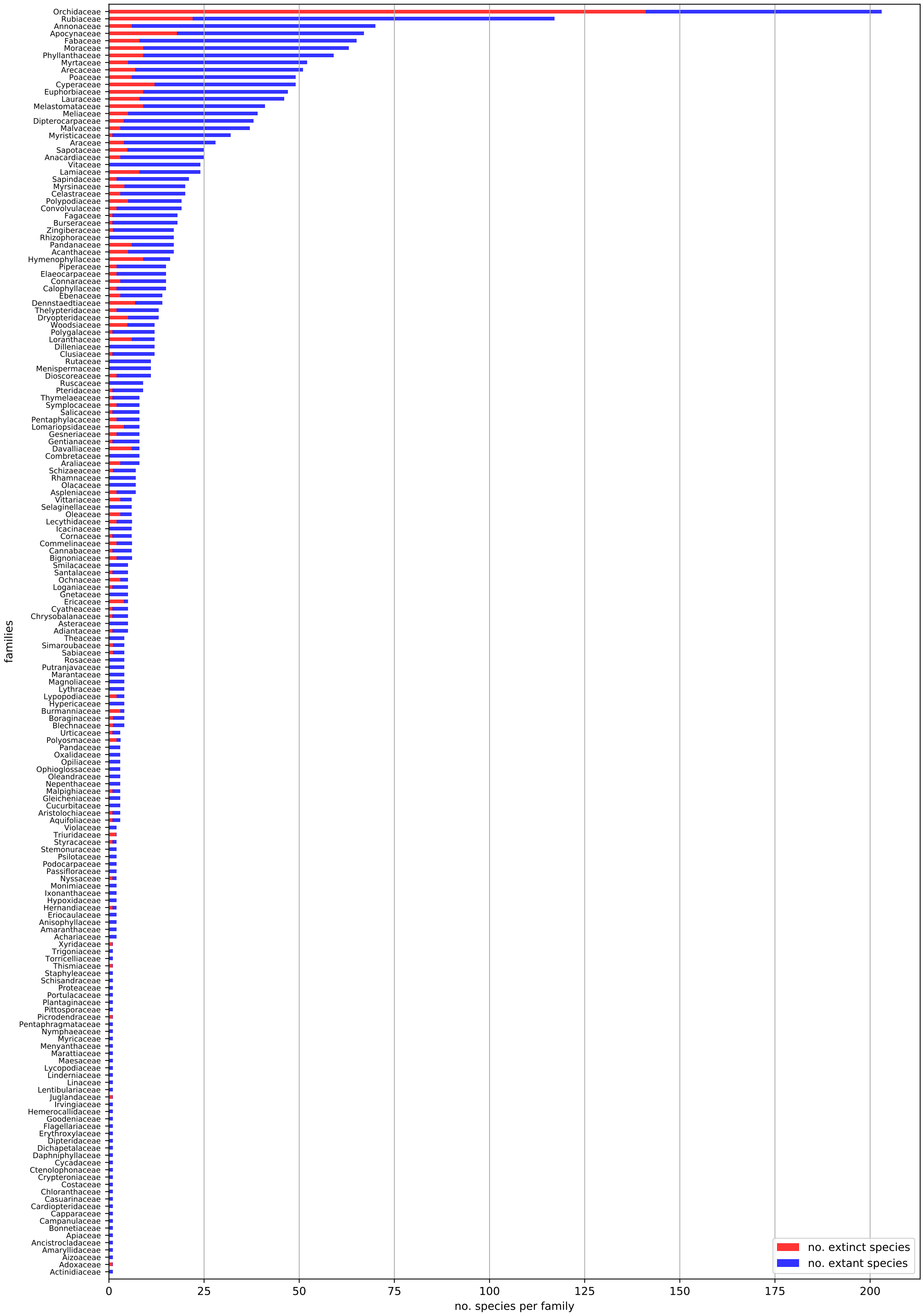


Figure S5: A list of all families in the database, and the total number of species represented, separated into extinct and extant species. Status based on expert opinion or last detection since 1985.

S2 Inferring redetection effort

S2.1 Background

Many statistical methods for inferring extinction, including the Solow (1993) method that we use, neglect the effect of varying collection effort over time. A common way to estimate collection effort is to use some collection effort index that is proportional to the probability of a detection occurring in a given timestep (McCarthy, 1998). For example, collection effort can be calculated using the number of specimens collected (Chong et al., 2012). Where multiple collections of the same species in the same year are pooled (e.g. because they are unlikely to be independent, as in our case), the number of species detected in each timestep can be used as a proxy (Robbirt et al., 2006), that is

$$c(t) = \frac{\text{number of species detected at time } t}{\text{total number of species}} \tag{S1}$$

However, the problems with the above approach are well known: (1) it assumes that the probability of first-detection is equal to redetection, when it is more plausible that species that have never been detected before may be less likely to be detected than those that are known (Duffy et al., 2009); (2) it counts all species as present in all time intervals, including those that may be extinct towards the end of the detection record (Duffy et al., 2009); (3) it assumes that all species will have the same intrinsic detection probabilities; and (4) it assumes that species' detection probabilities are constant in time.

In this appendix, we describe the method we used, which was designed to address the problems above. The resulting collection effort function that we obtain is used to modify the Solow (1993) method for inferring discovered species' extinction, which is detailed in S3.

S2.2 An improved method for inferring redetection effort

S2.2.1 Model and assumptions

Three main changes are made to the McCarthy (1998) approach and Eq. S1. First, the 'collection effort' should measure the effort that is expended in *re*-detecting species that have already been discovered. This is because it is redetection that is used for the estimation of P-values in the Solow (1993) method, and because undiscovered species are likely to have lower detection probability than those who have been discovered (point 1 above). Therefore the first-detection of the species is not included in the calculation.

Second, species who have gone extinct should not be included in the calculation (point 2). Therefore, the denominator in Eq. S1 should only count those species that were known to be extant at each time t.

Third, species will almost certainly differ in their intrinsic redetection probabilities; therefore, the total number of species in the denominator of Eq. S1 should be replaced with the sum of their individual redetection probabilities (point 3). Species intrinsic detection probabilities may also change with time (point 4); however the method was not able to address this assumption (below), and so those probabilities were assumed constant in time.

Define a set $\mathcal{S}_t$ of all species who were available to be redetected at time t, that is, who were first detected at some time earlier than t, and who are certain to be extant because their last detection was after t. For each species i we have a redetection record $\mathcal{R}_i$, which is the set of all times t that it was redetected i.e. all detections apart from the first and last in the detection record.

The record of each species was arbitrarily cut short to exclude the last detection. Ideally, one would include the time up to the date of extinction, which occurs some time after the last detection, in a species' record, but this date is unknown. Including the last detection would have biased the sample because the last detection is always a detection by definition, though our choice to exclude the last detection may also bias the sample to lower redetection effort in the time period since the penultimate detection. However, provided that enough species records are included, then the choice of where to end a particular record should not much influence the overall redetection effort inferred.

Define an indicator function

$$I_R(i,t) = \begin{cases} 1 & \text{if } t \in \mathcal{R}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{S2}$$

Then Eq. S1 can be rewritten

$$c(t) \approx \frac{\sum_{i \in \mathcal{S}_t} I_R(i,t)}{\sum_{i \in \mathcal{S}_t} r_i(t)}, \tag{S3}$$

where $r_i(t)$ is a function (e.g. a declining exponential) describing species i's intrinsic redetection probability at time t.

Eq. S3 makes use of the detection records by summing the number of species redetected at each timestep, however one can also make use of the records by taking sums for each species. Define the set $\mathcal{T}_i$ as the set of all times that the species i was available to be redetected. The sets are related $i \in \mathcal{S}_t \leftrightarrow t \in \mathcal{T}_i$. Therefore the set $\mathcal{T}_i$ starts at the year after first detection and includes all years up to but excluding the year of the last detection. Then for each species, the probability of being redetected at time t is

$$p_i(t) = \begin{cases} r_i(t)\, c(t) & \text{if } t \in \mathcal{T}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{S4}$$

Then the total number of times that a species was redetected should be approximately equal to the sum of the redetection probabilities at each timestep

$$\sum_t I_R(i,t) \approx \sum_t p_i(t) = \sum_{t \in \mathcal{T}_i} r_i(t) c(t). \tag{S5}$$

We are interested in obtaining a function for $c(t)$; however Eq. S5 creates the simultaneous problem of obtaining functions $r_i(t)$ for all of the species. Therefore it is necessary to make a simplifying assumption: $r_i(t) \approx r_i$ is constant in time for all species. Then Eq. S5 can be rewritten

$$r_i \approx \frac{\sum_t I_R(i,t)}{\sum_{t \in \mathcal{T}_i} c(t)}. \tag{S6}$$

Substituting Eq. S6 into Eq. S3

$$c(t) \approx \frac{\sum_{i \in \mathcal{S}_t} I_R(i,t)}{\sum_{i \in \mathcal{S}_t} \frac{\sum_t I_R(i,t)}{\sum_{t \in \mathcal{T}_i} c(t)}}, \tag{S7}$$

so that $c(t)$ is the only unknown.

To fit $c(t)$, we use a likelihood approach. The probability of the detection record at time t given $c(t)$ is a product over all species available to be detected at time t

$$P(R(t) \mid c(t)) = L_t = \prod_{i \in \mathcal{S}_t} (r_i \, c(t))^{I_R(i,t)} \, (1 - r_i \, c(t))^{1 - I_R(i,t)}, \tag{S8}$$

Then the likelihood of the detection record across the entire time series from year t_0 to T is

$$L = \prod_{t=t_0}^{T} L_t. \tag{S9}$$

S2.2.2 Choosing species to include

A subset of the species records were used for the fitting of $c(t)$. Species were excluded if they had:

1. Zero redetections.

   Species with no redetections do not provide information about redetection effort.

2. A last detection between 1985 and 2015.

   The $c(t)$ function fitted will be used to estimate p-values for this set of species.

3. A lifespan shorter than 30 years.

   This choice was made primarily to reduce the number of species included in order to speed up the fitting process.

We checked that the subset of species chosen above still provided good coverage over the years of the time series (Fig. S6). Coverage is lowest at the start of the time series.

S2.2.3 Fitting the model using AIC

The function $c(t)$ is defined using a linear spline, with parameters being the pairs $(t, c(t))$ that define locations of points in the spline. Therefore the AIC for a given spline can be calculated

$$\text{AIC} = 2k - 2\ln(\hat{L}) \tag{S10}$$

where k is the number of parameters (each spline point counts as two parameters), and $\hat{L}$ is the maximum likelihood, which is calculated by optimising Eq. S9.

The minimisation of the negative log-likelihood is computationally expensive because of the number of species and number of years in the time series. Our goal is only to find

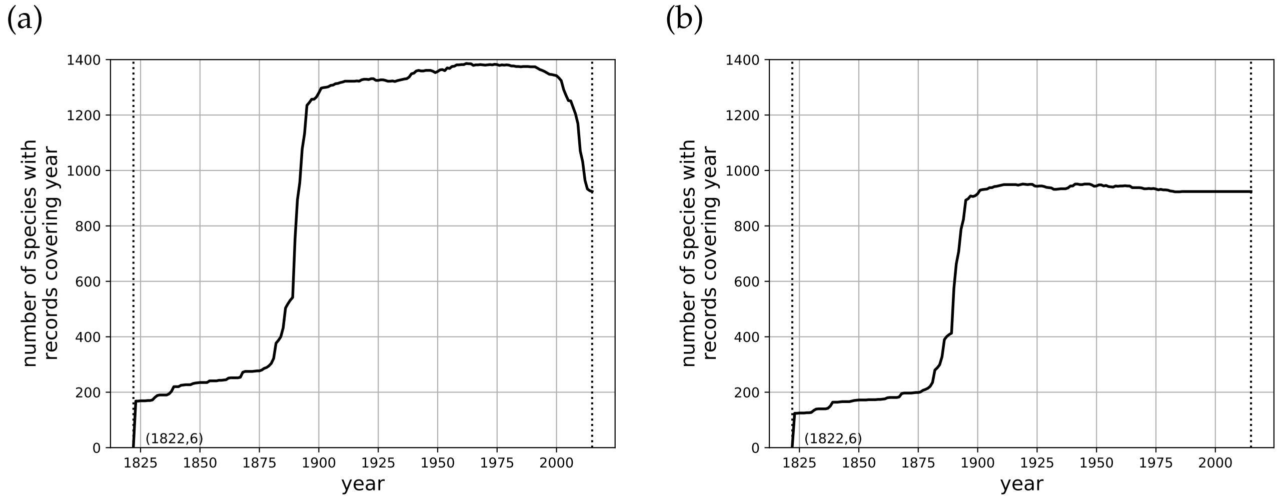


Figure S6: Number of species with redetection records covering each year when all species with record lengths of at least 30 years (1508 species) are included (a), and when only species meeting the conditions in the text (1056 species) are included (b). The 1056 species in (b) were used to fit the redetections effort function.

a reasonably parsimonious function for $c(t)$. Therefore, for computational efficiency, a backwards stepwise algorithm was used. Starting with the best-fitting function, points defining the spline were removed sequentially, choosing the removal that led to the greatest decrease in AIC. This was repeated until a valley of AIC values with respect to number of parameters was found. Then, within the valley, the c_t were fitted using maximum likelihood and the AIC recalculated. The $c(t)$ function with the lowest AIC resulting from the recalculation was chosen.

S2.3 Additional results

S2.3.1 Fitted redetection effort function

The initial best-fitting spline, with a point at every year in the time series, was obtained by initialising every $r_i = 1$ and iteratively solving Eq. S7 and S6 (Fig. S7). Sequential removal of points defining the spline resulted in a valley of AIC near 180 points, and fitting found a minimum $AIC = 62412$ at 182 points (Fig. S8). The resulting $c(t)$ function, and a histogram of species' intrinsic redetection probabilities r_i, are shown in Fig. 3 in the body of the text. The cumulative $c(t)$ function is used to define a mapping from years to effort-years (Fig. S9), which is used to correct the Solow (1993) method for inferring extinction from the detection record.

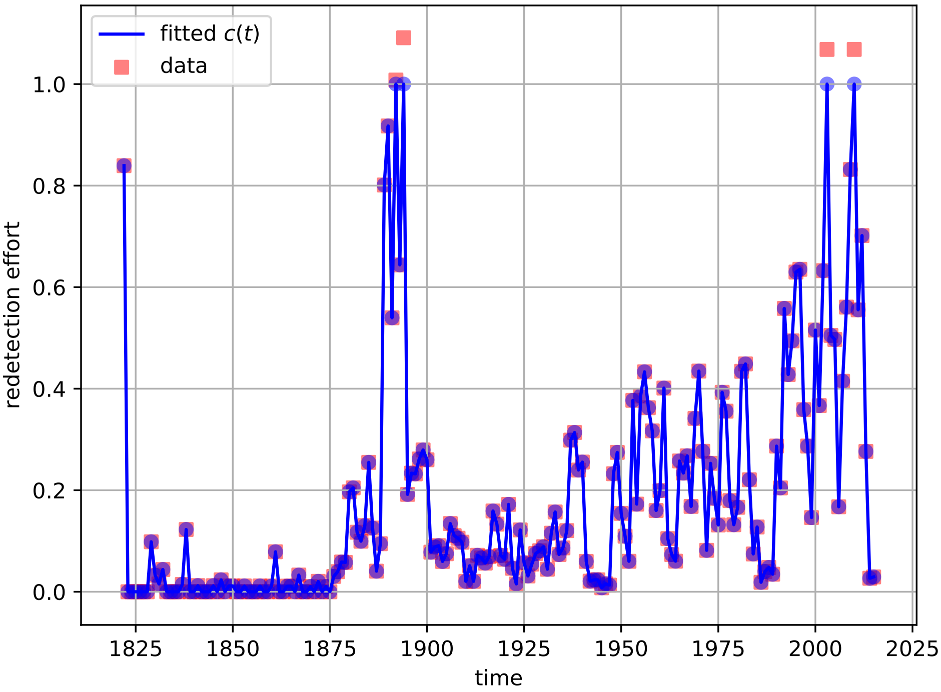


Figure S7: The results of the initial fitting of the redetection effort function $c(t)$, which was a linear spline with a point determined for every year in the time series.

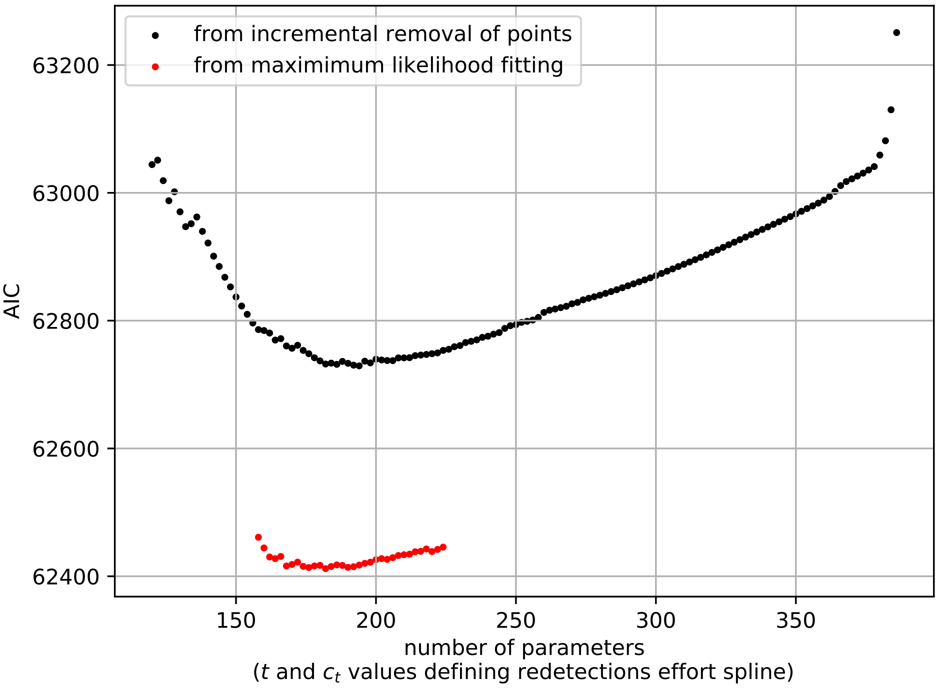


Figure S8: The AIC calculated in the first stage from sequential removal of points from the initial best-fitting spline (black), and in the second stage by fitting c_t values defining the spline using maximum likelihood (red).

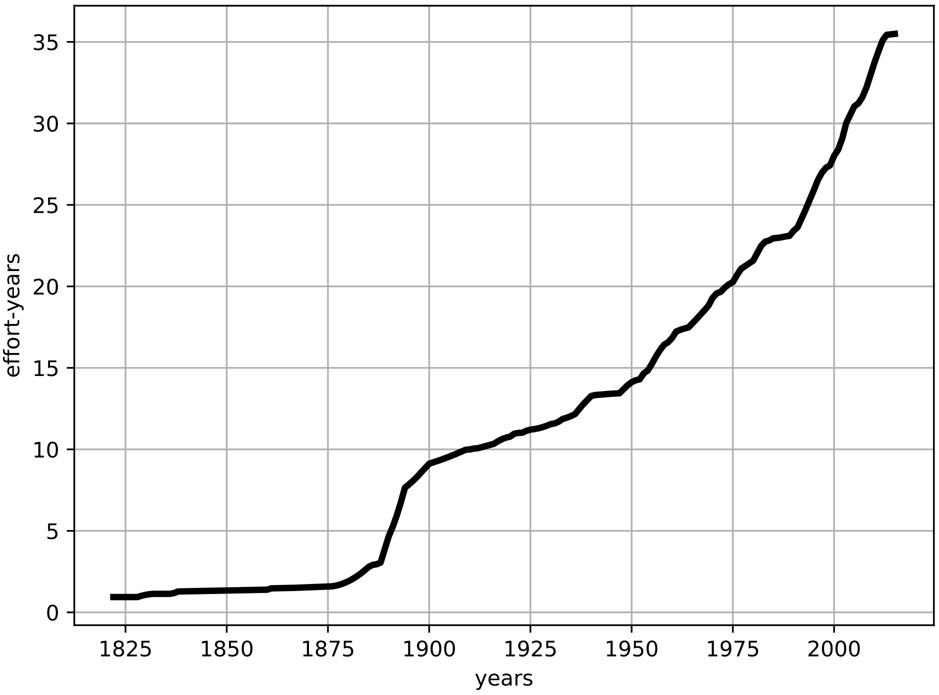


Figure S9: The mapping from years to effort-years using the redetection effort function obtained in Fig. 3a.

S2.3.2 Intrinsic probability of redetection

Intrinsic redetection probabilities r_i can be inferred for species with sufficient number of collection dates (i.e. 3 or more). A redetection probability is different to the detection probabilities that appear in the SEUX model; however we expect them to be correlated. Through the 1800s, there is a trend for later-discovered species to have lower intrinsic redetection probabilities (Fig. S10). Recently-discovered species appear to have high redetection probabilities; however, redetection effort towards the end of the time series may be underestimated in the data due to the time-delay between a specimen's collection and its determination and entry into the database.

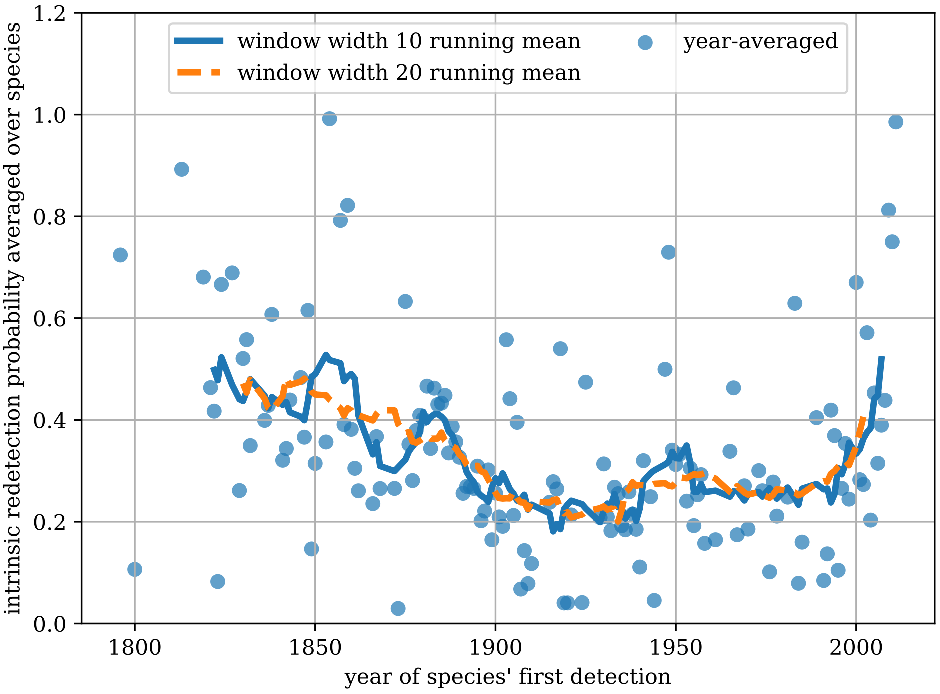


Figure S10: The relationship between the average intrinsic redetection probability of species and their year of discovery.

S2.4 Testing the new method on simulated data

S2.4.1 Method

We used simulated redetection data to test our method. The simulated data was created as follows:

1. A 'true' $c(t)$ function, $c^*(t)$, was created that was qualitatively similar to the $c(t)$ that fitted to the real data (blue lines in Fig. S11).
2. The 'true' $c^*(t)$ was used to infer the intrinsic redetection probabilities of each species.
3. Thirty redetection records were simulated, where the probability that a species was detected in each year was equal to its intrinsic redetection probability, calculated above, multiplied by the redetection effort in that year (i.e. Eq. S4).
4. Each species' redetection record was trimmed so that the first and last redetection in the simulation was within that species' first and last detection years in the real data.

The simulated data was used to test our method in two ways:

- For each simulation, the best-fitting $c(t)$ was obtained using the same iterative method that was used for the initial fitting (i.e. that was used to obtain Fig. S7). Each fitting gives a value of c_t for each year t, which defines the spline $c(t)$. The range of these spline points was compared to the 'true' $c^*(t)$ values that were used to generate the simulation.
- One simulated redetection record was chosen at random, and the backwards stepwise algorithm (Section S2.2.3) was used to find a function $c(t)$ that minimised AIC. The estimated function was visually compared to the 'true' function $c^*(t)$.

S2.4.2 Result

The 30 $c(t)$ functions that best fitted the 30 simulated data sets generally conformed closely to the 'true' $c^*(t)$ used to generate the data (Fig. S11a). The exception is the point $c(t = 1822)$ at the start of the time series, where the number of species available to be redetected is lowest (i.e. 6 species). When the backwards stepwise algorithm (Section S2.2.3) was used to fit a $c(t)$ to one of the simulated datasets, the best-fitting $c(t)$ showed good agreement with the 'true' $c^*(t)$ that was used to create the simulated data (Fig. S11b). Both results suggest that our method was adequate given the assumptions underlying it.

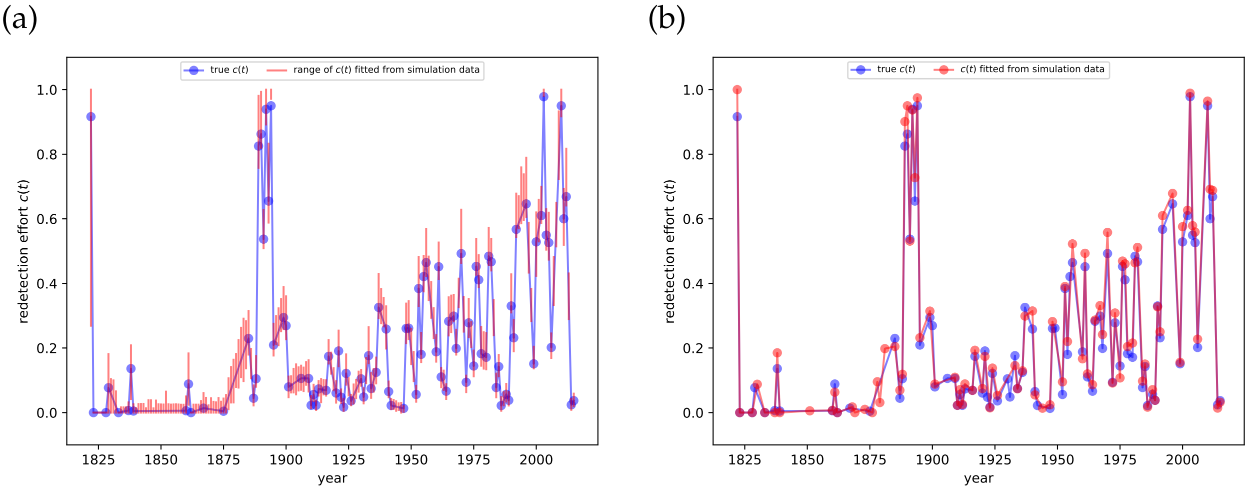


Figure S11: Comparisons between a 'true' $c(t)$ function that was used to create simulated redetection records (blue), and the $c(t)$ function that was subsequently inferred from that simulated data (red).

S3 Inferring discovered extinctions

S3.1 Background

Solow (1993) presented a simple way to infer if a species is extinct, based on how frequently it has been detected in the past compared to how long ago it was last detected. It calculates the P-value, under the null hypothesis that the species is extant, of the most recent sighting occurring no later than its observed value

$$P = \left(\frac{\tau_f - \tau_0}{T - \tau_0} \right)^n, \tag{S11}$$

where τ_0 is the particular species' year of discovery, τ_f is the year of final detection, T is the current year, and n is the number of years in which the species was redetected (i.e. the number of detections minus the first detection).

One shortcoming of the Solow (1993) method is that it assumes a constant pre-extinction sighting rate (but see e.g. Solow, 1993b), which is not a plausible assumption for our dataset. Therefore McCarthy (1998) proposed that Eq. S11 be corrected by using collection-effort-years C instead of calendar years as the basis for calculating the P-value

$$P = \left(\frac{C(\tau_f) - C(\tau_0)}{C(T) - C(\tau_0)} \right)^n, \tag{S12}$$

where the collection-effort-year is calculated as the sum of collection efforts from the start of the timeseries t_0 up to and including the particular year of interest τ

$$C(\tau) = \sum_{t=t_0}^{\tau} c(t). \tag{S13}$$

In S2, we developed a new method to obtain $c(t)$, the results from which were used for our modified Solow (1993) method. In this supplement, we provide more details on that analysis.

S3.2 Additional results and discussion

If species status is determined using only the detection records and expert knowledge, then the temporal pattern in the number of discovered extant and extinct species shows a large discontinuity towards the end of the time series (dashed line, Fig. S12a), with a sudden increase in the number of extinct species. This suggests that detection records and expert knowledge alone will classify many extant species as extinct. The 30-year heuristic removes the discontinuity, however it forces a plateau onto the end of the discovered extinct species curve (solid red line, Fig. S12a).

When the modified Solow (1993) method was applied to all species seen within the past 30 years, it smoothened the temporal pattern in discovered extinct species, and in a way that is consistent with the time-period before 1985 (Fig. S12b). The pattern of discovered extant species showed an increase since 1985, implying that the rate of discovery of new

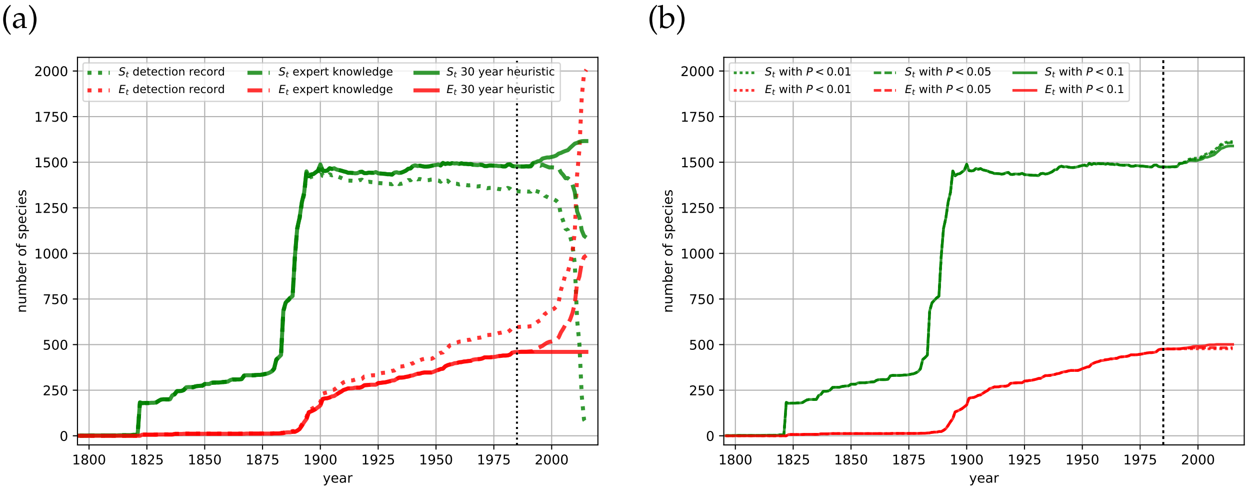


Figure S12: Time series of discovered extant and extinct species, where species statuses were determined from the detection record, expert judgement, and the 30-year heuristic (a), or where the Solow (1993) P-value was used to designate some species last-detected in the past 30 years as extinct (b).

species has recently increased. This result is consistent with the results that were obtained when fitting a redetection effort function to the data. First, the redetection effort is highest towards the end of the time series (Fig. S7). Second, while there has been an overall decline in time in the average intrinsic redetection probability of species that are discovered, since approximately the 1980s the average intrinsic redetection probability of species increased in time (Fig. S10). While redetection and discovery probabilities are separate concepts, we might expect them to be correlated (e.g. small and obscure species are both harder to discover and harder to find again once discovered).

The choice of P-value criterion in the Solow (1993) method is somewhat arbitrary; however we found that the choice does not make a substantial difference to the relative outcome (Fig. S12b). In order to be conservative, i.e. overestimate rather than underestimate extinction rates, we chose $P < 0.1$. This also provided a longer list of species for experts to examine in the second pass.

Of the species that were last detected after 1985 and with long-enough detection records for an inference to be made, 70 species had $P < 0.1$ indicating that they were extinct (Table S3). We then asked experts to thoroughly check the list 70 species and indicate which they could verify as extant and which in their opinion were likely to be extinct. There was some expert uncertainty about the status of some species (see below), however we estimate that, of those 70 species that the Solow (1993) method designated extinct, approximately 65 were in fact extant. This result implies a much higher misclassification rate – of at least 0.93 – than the P-value criterion that was used.

Experts were uncertain about the following species:

- *Pogonanthera pulverulenta*

  Unable to obtain detailed information, however a verified record from 1997 contributed to the opinion that the species was likely extant.

- *Gymnacranthera bancana*

  Needs further verification. The taxonomic expert could not confirm the identity of a recently photographed tree. The last verified collection was in 1953. Therefore the species was classified as extinct.

- *Hedyotis pinifolia*

  Unsure. The 2003 collection was not traced and the taxonomic expert did not determine any recent specimens of the species from Singapore.

- *Baccaurea brevipes*

  Unable to get information. Only two specimens exist and no specialists have confirmed their identity, therefore it was designated likely extinct.

Table S3: Species with detections since 1985 that were designated as extinct using the Solow (1993) method with criterion $P < 0.1$. Of these 70 species, expert opinion classified 65 of them as likely still extant. These expert opinions were used for the final input into the SEUX model. An asterisk marks species with particular uncertainty about status (see text).

		detection		Solow		Chong et al. (2009)		expert
family	species	first	last	P	lifeform	status	cultivated	opinion
Pentaphragmataceae	*Pentaphragma ellipticum*	1889	1995	0.0004	herb	vulnerable	cultivated	extant
Myrsinaceae	*Embelia lampanii*	1884	1992	0.0077	climber	common		extant
Myristicaceae	*Horsfieldia irya*	1897	1992	0.0078	tree	crit. endangered	cultivated	extant
Euphorbiaceae	*Excoecaria agallocha*	1890	2003	0.0111	tree	common	cultivated	extant
Malvaceae	*Commersonia bartramia*	1857	2004	0.0145	tree	common		extant
Malvaceae	*Neesia synandra*	1890	1992	0.0163	tree	vulnerable		extant
Araceae	*Homalomena griffithii*	1887	1993	0.018	herb	vulnerable		extant
Apocynaceae	*Alstonia angustifolia*	1822	2003	0.0189	tree	common	cultivated	extant
Fabaceae	*Albizia splendens*	1893	1994	0.0195	tree	endangered	cultivated	extant
Arecaceae	*Myrialepis paradoxa*	1888	1996	0.0206	climber	crit. endangered		extant
Rhizophoraceae	*Bruguiera gymnorhiza*	1885	2006	0.023	tree	common	cultivated	extant
Rubiaceae	*Saprosma glomerulatum*	1981	1998	0.023	shrub	crit. endangered		extant
Euphorbiaceae	*Macaranga heynei*	1881	2003	0.0234	tree	common		extant
Arecaceae	*Daemonorops longipes*	1889	1996	0.0256	shrub	vulnerable		extant
Melastomataceae	*Pogonanthera pulverulenta*	1884	1997	0.0276	epiphyte	crit. endangered		extant*
Ochnaceae	*Campylospermum serratum*	1822	2005	0.0279	tree	common		extant
Elaeocarpaceae	*Elaeocarpus pedunculatus*	1822	2004	0.0285	tree	common	cultivated	extant
Fabaceae	*Archidendron jiringa*	1822	2003	0.0317	tree	vulnerable	cultivated	extant
Arecaceae	*Daemonorops didymophylla*	1890	1996	0.0321	climber	vulnerable		extant
Lythraceae	*Sonneratia alba*	1949	2007	0.0343	tree	common	cultivated	extant
Myristicaceae	*Gymnacranthera bancana*	1822	1993	0.0348	tree	crit. endangered		extinct*
Euphorbiaceae	*Macaranga trichocarpa*	1884	2005	0.0356	shrub	endangered		extant
Dipterocarpaceae	*Anisoptera megistocarpa*	1894	2005	0.0367	tree	crit. endangered		extant
Chrysobalanaceae	*Licania splendens*	1884	2005	0.0384	tree	common	cultivated	extant
Acanthaceae	*Justicia vasculosa*	1889	1995	0.0389	herb	crit. endangered		extant
Annonaceae	*Meiogyne virgata*	1885	1995	0.0441	tree	crit. endangered		extant

Continued on next page

Table S3 – continued from previous page

		detection		Solow		Chong et al. (2009)		expert
family	species	first	last	P	lifeform	status	cultivated	opinion
Arecaceae	*Calamus lobbianus*	1889	1992	0.0451	tree	crit. endangered		extant
Arecaceae	*Daemonorops angustifolia*	1890	1996	0.0453	climber	vulnerable		extant
Dioscoreaceae	*Dioscorea pyrifolia*	1822	2004	0.0454	climber	common		extant
Arecaceae	*Korthalsia rigida*	1894	1999	0.046	climber	crit. endangered		extant
Rhizophoraceae	*Ceriops tagal*	1889	2006	0.0479	tree	vulnerable		extant
Arecaceae	*Calamus oxleyanus*	1890	1999	0.0481	climber	crit. endangered		extant
Rubiaceae	*Hedyotis capitellata*	1992	1993	0.0494	herb	crit. endangered		extant
Moraceae	*Ficus subgelderi*	1890	1990	0.0506	strangler	crit. endangered		extant
Fabaceae	*Peltophorum pterocarpum*	1890	2004	0.051	tree	crit. endangered	cultivated	extant
Combretaceae	*Combretum tetralophum*	1996	1997	0.0511	climber	crit. endangered		extant
Sapindaceae	*Lepisanthes senegalensis*	1996	1997	0.0511	tree	crit. endangered	cultivated	extant
Rubiaceae	*Hedyotis pinifolia*	1950	2003	0.0518	herb	vulnerable		extant*
Plantaginaceae	*Adenosma javanicum*	1894	1995	0.0521	herb	common		extant
Meliaceae	*Aphanamixis polystachya*	1884	2005	0.0531	tree	endangered	cultivated	extant
Moraceae	*Ficus heteropleura*	1884	2004	0.0555	climber	common		extant
Rhizophoraceae	*Rhizophora mucronata*	1822	2005	0.0555	tree	common	cultivated	extant
Fagaceae	*Castanopsis wallichii*	1884	2003	0.0579	tree	crit. endangered	cultivated	extant
Sapotaceae	*Planchonella linggensis*	1996	2007	0.0589	tree	crit. endangered		extant
Rubiaceae	*Lasianthus cyanocarpus*	1966	2007	0.0627	shrub	crit. endangered		extant
Menyanthaceae	*Nymphoides indica*	1890	1992	0.0642	herb	endangered	cultivated	extant
Phyllanthaceae	*Baccaurea brevipes*	1995	1996	0.0644	tree	crit. endangered		extinct*
Nepenthaceae	*Nepenthes rafflesiana*	1819	2007	0.0647	climber	vulnerable	cultivated	extant
Orchidaceae	*Hylophila mollis*	1822	1996	0.0668	herb	crit. endangered		extinct
Melastomataceae	*Memecylon pseudomegacarpum*	1892	2003	0.0696	tree	endangered		extant
Zingiberaceae	*Elettariopsis latiflora*	1890	1995	0.0737	herb	crit. endangered		extant
Erythroxylaceae	*Erythroxylum cuneatum*	1889	2006	0.0739	tree	common	cultivated	extant
Asteraceae	*Gynura procumbens*	1890	1993	0.076	climber	crit. endangered		extinct
Euphorbiaceae	*Macaranga hypoleuca*	1881	2002	0.077	tree	common		extant

Continued on next page

Table S3 – continued from previous page

		detection		Solow		Chong et al. (2009)		expert
family	species	first	last	P	lifeform	status	cultivated	opinion
Rubiaceae	*Oldenlandia affinis*	1930	2004	0.0782	herb	vulnerable		extant
Menispermaceae	*Fibraurea tinctoria*	1889	2009	0.0787	climber	common	cultivated	extant
Pandanaceae	*Pandanus odorifer*	1890	2003	0.0787	shrub	common	cultivated	extant
Asteraceae	*Pluchea indica*	1853	2007	0.0794	shrub	common		extant
Magnoliaceae	*Magnolia villosa*	1939	2003	0.0809	tree	crit. endangered		extant
Annonaceae	*Xylopia malayana*	1867	2009	0.081	tree	common	cultivated	extant
Selaginellaceae	*Selaginella roxburghii*	1830	2002	0.0834	herb	vulnerable		extant
Acanthaceae	*Acanthus volubilis*	1889	2003	0.0847	shrub	vulnerable		extant
Clusiaceae	*Calophyllum sundaicum*	1822	1994	0.0867	tree	crit. endangered		extant
Arecaceae	*Daemonorops periacantha*	1890	1996	0.0901	climber	crit. endangered		extant
Fagaceae	*Lithocarpus hystrix*	1884	1997	0.0927	tree	crit. endangered		extant
Annonaceae	*Polyalthia jenkinsii*	1938	1989	0.0942	tree	crit. endangered		extant
Apocynaceae	*Dyera costulata*	1882	2003	0.0968	tree	common	cultivated	extant
Malvaceae	*Scaphium linearicarpum*	1894	1996	0.097	tree	crit. endangered		extant
Annonaceae	*Anaxagorea javanica*	1916	2001	0.0978	tree	crit. endangered	cultivated	extinct
Rutaceae	*Merope angulata*	1893	2003	0.0978	shrub	crit. endangered		extant

S4 Classical inference

S4.1 Algorithm for obtaining confidence intervals

In this section, we provide details for the algorithm summarised in (17) of the main text. Pseudocode for the algorithm is presented in Algorithm 1.

Denote the i^{th} sample from the fiducial distribution of U_0 by a superscript, $U_0^{[i]}$. The algorithm obtains $U_0^{[i]}$ by working backwards in time. It starts at $t = T$ and the assumed value of U_T. Then it obtains sample $U_{T-1}^{[i]}$ conditional on U_T. It repeats this, stepping backwards in time, sequentially sampling each $U_{t-1}^{[i]}$ conditional on $U_t^{[i]}$, until $U_0^{[i]}$ is obtain. The procedure is repeated many times to obtain a large sample of $U_0^{[i]}$. Then classical confidence intervals for U_0 were estimated from the quantiles of the large sample.

Each $U_{t-1}^{[i]}$ found at each step of the algorithm can be interpreted as the $1 - \alpha_{t-1}^{[i]}$ upper confidence bound for U_{t-1} given $U_t = U_t^{[i]}$. We draw $\alpha_{t-1}^{[i]} \sim \mathcal{U}(0,1)$. Then we wish to find $U_{t-1}^{[i]}$ that solves Eq. 18 (i.e. $\alpha_{t-1}^{[i]} = H_{U_{t-1}^{[i]},U_t^{[i]}}(\boldsymbol{\psi}, \mathbf{S})$). Drawing $\alpha_{t-1}^{[i]}$ from the uniform distribution imposes Requirement 2 of the confidence distribution's definition (Sect. 2.2 Xie and Singh, 2013), which ensures that confidence intervals of each U_t can be approximated from the quantiles of the corresponding sample of bounds.

However, because the data are discrete, it is not generally possible to find $U_{t-1}^{[i]}$ that gives coverage exactly matching the $1 - \alpha_{t-1}^{[i]}$ nominal level. We overcame this by conceiving of the bounds as fuzzy (Geyer and Meeden, 2005). It can be shown (S4.2) that random sampling of fuzzy bounds is equivalent to random sampling using the mid-P method (Agresti and Gottard, 2007). Therefore, the sample bounds were obtained by numerically solving

$$\begin{aligned}&U_{t-1}^{[i]} \text{ is the greatest value of } U_{t-1}^{*} \text{ such that}\\ &P_M(U_{t-1}^{*} \mid \phi_{t-1}^{[i]}) = P(\Phi_{t-1} < \phi_{t-1}^{[i]} \mid U_{t-1}^{*}) + 0.5\, P(\Phi_{t-1} = \phi_{t-1}^{[i]} \mid U_{t-1}^{*}) \geq \alpha_{t-1}^{[i]},\end{aligned} \qquad \text{(S14)}$$

where $\phi_{t-1}^{[i]}$ is calculated from $U_t^{[i]} + \delta_{t-1}$.

As discussed in Agresti and Gottard (2007), the mid-P method can result in physically impossible bounds. In the SEUX model, the minimum value of U_{t-1} that is physically possible is obtained when every undiscovered species survives the timestep, i.e. $U_{t-1} \geq U_t + \delta_{t-1}$. When using the mid-$P$ method at a particular timestep, it can occasionally sample the physically impossible bound $U_{t-1}^{[i]} = U_t^{[i]} + \delta_{t-1} - 1$. It is necessary to include these impossible values in the sample, and to sample bounds for earlier timesteps predicated upon them, in order for the confidence bounds to match the nominal confidence level. However, if the physically impossible bound for U_{t-1} is sampled when the value of U_t is already physically impossible (due to previous sampling steps), then the bounds sampled may drift further into the region of values that are physically impossible. We found that permitting this to occur will give the wrong coverage. Therefore, the algorithm must also keep track of whether or not the current sample is already predicated upon an impossible value, and use that to determine the constraint on $U_{t-1}^{[i]}$. The variable `impossibleFlag` is used for this purpose in the code. This prevents the sample from moving more than one value into the region of physically impossible values.

Algorithm 1 Step-by-step random sampling algorithm to obtain a large sample of U_0 bounds

```
input: $\mathbf{S}, \mathbf{E}, U_T$    ▷ Data, assumption
$\mathbf{U_0} = []$    ▷ Initialise an empty list to receive the $U_0$ bounds sampled

for $[i] = 1 \ldots$ manySamples do

    impossibleFlag = 0    ▷ Handles physically impossible bounds of $U_t$
    $\mathbf{U}^{[i]} = [0, \ldots, 0, U_T]$    ▷ A list to store bounds sampled, initialised with $U_T$

    for $t = T \ldots 1$ do    ▷ Working backwards in time

        $\delta_{t-1} = S_t - S_{t-1} + E_t - E_{t-1}$
        $\phi_{t-1}^{[i]} = \mathbf{U}^{[i]}[t] + \delta_{t-1}$    ▷ Obtained from previously sampled bound
        $U_{t-1}^{\min} = \phi_{t-1}^{[i]}$    ▷ Minimum physically possible bound

        Draw $\alpha \sim \text{Uniform}(0,1)$

        if $P_M(U_{t-1}^{\min} \mid \phi_{t-1}^{[i]}) < \alpha$ then

            if impossibleFlag then    ▷ Only sample bounds one step into the impossible region
                $U_{t-1}^{[i]} = U_{t-1}^{\min}$    ▷ Do not step further into impossible region
            else
                $U_{t-1}^{[i]} = U_{t-1}^{\min} - 1$    ▷ Sample one step into impossible region
                impossibleFlag = 1
            end if

        else

            Solve $U_{t-1}^{[i]}$ is greatest value of $U_{t-1}^{*}$ s.t. $U_{t-1}^{*} \geq U_{t-1}^{\min}$ and $P_M(U_{t-1}^{*} \mid \phi_{t-1}^{[i]}) \geq \alpha$

        end if

        if $U_{t-1}^{[i]} > U_{t-1}^{\min}$ then    ▷ Algorithm has moved out of impossible region
            impossibleFlag = 0
        end if

        $\mathbf{U}^{[i]}[t-1] = U_{t-1}^{[i]}$    ▷ Append bound to list storing sampled bounds at each timestep

    end for

    Append $\mathbf{U}^{[i]}[0]$ to $\mathbf{U_0}$

end for

return $\mathbf{U_0}$
```

S4.2 The technical challenge of obtaining confidence intervals for discrete data

The objective is to obtain confidence intervals for the hypergeometric SEUX model on the unknown U_t values. Obtaining confidence intervals for discrete data presents a technical challenge: it is generally not possible to obtain intervals that give coverage probability exactly equal to any arbitrary nominal confidence level. This is a particular problem for the SEUX model, because the overarching approach is to sample confidence bounds at each timestep to make inferences about the timestep before; therefore errors will propagate and biases will accumulate in the solution process.

In this section, we discuss fuzzy confidence intervals (Geyer and Meeden, 2005), which can obtain coverage that exactly matches the nominal level. We verify that random sampling of the fuzzy confidence bounds is equivalent to random sampling using the mid-P method (Berry and Armitage, 1995; Agresti and Gottard, 2007). We then describe how a random sampling scheme can be used to estimate confidence intervals on U_t, and particularly U_0 at the start of the timeseries, which is then used to estimate the number of undiscovered extinctions.

S4.2.1 Preliminaries

Following the explanation given by Feldman and Cousins (1998)[1], the confidence set for confidence level $\gamma = 1 - \alpha$ may be obtained using the method of confidence belts (Neyman, 1937). For each value of the unknown parameter μ, one selects an acceptance interval $[x_1, x_2]$ such that

$$P(x \in [x_1, x_2] \mid \mu) = \gamma. \tag{S15}$$

Then the confidence set is the union of all values of μ for which the corresponding acceptance interval contains the sample value; typically an interval $[\mu_1, \mu_2]$.

There are potentially many different ways to choose an acceptance interval that would satisfy Eq. S15; therefore in order to give a unique solution, one must also specify some auxiliary criteria. A common choice for obtaining a one-sided upper confidence limit on μ is

$$P(x \geq x_1 \mid \mu) = \gamma. \tag{S16}$$

When the data is discrete, as in the SEUX model, then Eq. S15 becomes

$$\sum_{x \in [x_1, x_2]} P(x \mid \mu) = \gamma. \tag{S17}$$

However, because the data is discrete, it is not possible in all cases to find an acceptance interval that satisfies Eq. S17 exactly. As a consequence, it is not possible to obtain a confidence interval that has coverage probability (calculated using the left-hand side of Eq. S17) exactly matching any arbitrary nominal level (i.e. the right-hand side of Eq. S17).

[1] Please note that this paper swaps the notation of α and γ from its most common usage.

In order to obtain a match between coverage and nominal level, an alternative approach is needed. Geyer and Meeden (2005) introduced the fuzzy confidence set, where the degrees of membership can be chosen so that an exact match between actual and nominal coverage is obtained. We apply their approach to the SEUX model.

Consider just one timestep of the hypergeometric SEUX model. Then omitting the timestep subscripts

$$P(\Phi = \phi \mid U) = \frac{\binom{U}{\phi}\binom{S}{\psi}}{\binom{U+S}{\phi+\psi}}. \tag{S18}$$

where S, ϕ, and ψ are known and U is the unknown parameter of interest.

Let $d(U,\phi,\gamma)$ be the degree of membership of U and ϕ of the respective confidence and acceptance intervals. If we assume that membership degrees are commensurate with probabilities (i.e. so that they can be multiplied and summed together), then

$$\sum_{\phi} d(U,\phi,\gamma) P(\Phi = \phi \mid U) = \gamma, \tag{S19}$$

where $0 \leq d(U,\phi,\gamma) \leq 1$.

Eq. S19 is analogous to Eq. S15 in the method of confidence belts, except that instead of choosing the acceptance interval (i.e. some interval of ϕ values), the solution is found by choosing values for the membership degrees $d(U,\phi,\gamma)$. Consequently the confidence interval is obtained by finding the union of all U values for which the membership degree is non-zero, describing a fuzzy set.

To uniquely specify the values of $d(U,\phi,\gamma)$, we need some auxiliary criteria. Inspired by Eq. S16, we create the following auxiliary criterion for obtaining the upper confidence limit

$$d(U,\phi,\gamma) = \begin{cases} 0 & \text{if } \gamma \leq P(\Phi > \phi \mid U), \\ \frac{\gamma - P(\Phi > \phi \mid U)}{P(\Phi = \phi \mid U)} & \text{if } P(\Phi > \phi \mid U) < \gamma < P(\Phi \geq \phi \mid U), \\ 1 & \text{if } P(\Phi \geq \phi \mid U) \leq \gamma. \end{cases} \tag{S20}$$

In Eq. S20, for a given U, because $d(U,\phi,\gamma) \leq d(U,\phi+1,\gamma)$, then the fuzzy acceptance interval is a convex set. In Subsection S4.2.2 below we will also show that, for a given ϕ

$$d(U,\phi,\gamma) \leq d(U-1,\phi,\gamma). \tag{S21}$$

Showing that Eq. S21 is true is useful for showing that random sampling of the fuzzy confidence bounds is equivalent to sampling using the mid-P method; this topic is continued in Subsection S4.2.3.

S4.2.2 Proving Eq. S21

In order to show $d(U,\phi,\gamma) \leq d(U-1,\phi,\gamma)$, we first show that

$$P(\Phi \geq \phi \mid U) \geq P(\Phi \geq \phi \mid U-1), \tag{S22}$$

and

$$P(\Phi > \phi \mid U) \geq P(\Phi > \phi \mid U-1). \tag{S23}$$

It can be shown that

$$\binom{U}{\phi} = \frac{U}{U-\phi}\binom{U-1}{\phi}, \tag{S24}$$

and

$$\binom{U+S}{\phi+\psi} = \frac{U+S}{U+S-(\phi+\psi)}\binom{U+S-1}{\phi+\psi}, \tag{S25}$$

therefore substituting Eq. S24 and S25 into Eq. S18 we can obtain the relationship

$$\begin{aligned} P(\Phi = \phi \mid U) &= \frac{\frac{U}{U-\phi}\binom{U-1}{\phi}\binom{S}{\psi}}{\frac{U+S}{U+S-(\phi+\psi)}\binom{U+S-1}{\phi+\psi}} \\ &= \frac{U(U+S-\phi)-U\psi}{U(U+S-\phi)-S\phi}P(\Phi = \phi \mid U-1). \end{aligned} \tag{S26}$$

Call

$$\phi^* = \frac{U\psi}{S}. \tag{S27}$$

Then, because both the numerator and denominator in Eq. S26 are positive, Eq. S26 implies

$$P(\Phi = \phi \mid U) > P(\Phi = \phi \mid U-1) \text{ if } \phi > \phi^*, \tag{S28}$$
$$P(\Phi = \phi \mid U) = P(\Phi = \phi \mid U-1) \text{ if } \phi = \phi^*, \tag{S29}$$
$$P(\Phi = \phi \mid U) < P(\Phi = \phi \mid U-1) \text{ if } \phi < \phi^*. \tag{S30}$$

We split the analysis into two cases: (1) $\phi \geq \phi^*$; and (2) $\phi < \phi^*$.

Case 1: $\phi \geq \phi^*$

By Eq. S28 and S29, $\phi \geq \phi^*$ implies

$$P(\Phi = \phi_i \mid U) \geq P(\Phi = \phi_i \mid U-1)\ \forall\, \phi_i \geq \phi, \tag{S31}$$

which implies

$$\sum_{\phi_i=\phi}^{U} P(\Phi = \phi_i \mid U) \geq \sum_{\phi_i=\phi}^{U-1} P(\Phi = \phi_i \mid U-1),$$
$$P(\Phi \geq \phi \mid U) \geq P(\Phi \geq \phi \mid U-1), \tag{S32}$$

and also implies

$$\sum_{\phi_i=\phi+1}^{U} P(\Phi = \phi_i \mid U) \geq \sum_{\phi_i=\phi+1}^{U-1} P(\Phi = \phi_i \mid U-1),$$
$$P(\Phi > \phi \mid U) \geq P(\Phi > \phi \mid U-1). \tag{S33}$$

Case 2: $\phi < \phi^*$

By Eq. S30, $\phi < \phi^*$ implies

$$P(\Phi = \phi_i \mid U) < P(\Phi = \phi_i \mid U-1) \, \forall \, \phi_i \leq \phi, \tag{S34}$$

which implies

$$\sum_{\phi_i=0}^{\phi-1} P(\Phi = \phi_i \mid U) < \sum_{\phi_i=0}^{\phi-1} P(\Phi = \phi_i \mid U-1),$$
$$P(\Phi < \phi \mid U) < P(\Phi < \phi \mid U-1),$$
$$1 - P(\Phi < \phi \mid U) > 1 - P(\Phi < \phi \mid U-1),$$
$$P(\Phi \geq \phi \mid U) > P(\Phi \geq \phi \mid U-1), \tag{S35}$$

and also implies

$$\sum_{\phi_i=0}^{\phi} P(\Phi = \phi_i \mid U) < \sum_{\phi_i=0}^{\phi} P(\Phi = \phi_i \mid U-1),$$
$$P(\Phi \leq \phi \mid U) < P(\Phi \leq \phi \mid U-1),$$
$$1 - P(\Phi \leq \phi \mid U) > 1 - P(\Phi \leq \phi \mid U-1),$$
$$P(\Phi > \phi \mid U) > P(\Phi > \phi \mid U-1), \tag{S36}$$

Combining Eq. S32 and S35 we have shown that Eq. S22 is always satisfied, and combining Eq. S33 and S36 we have shown that Eq. S23 is always satisfied. Now we can use Eq. S22 and S23 to show Eq. S21.

To determine the value of $d(U,\phi,\gamma)$ in Eq. S21, we use Eq. S20, which involves three conditions. Therefore we split the analysis into the cases resulting from these conditions.

Case 1: If $d(U,\phi,\gamma) = 0$ then $d(U,\phi,\gamma) \leq d(U-1,\phi,\gamma)$ holds trivially.

Case 2: If $d(U,\phi,\gamma) = 1$, then Eq. S20 implies $\gamma > P(\Phi \geq \phi \mid U)$. By Eq. S22, $P(\Phi \geq \phi \mid U) \geq P(\Phi \geq \phi \mid U-1)$, therefore $\gamma > P(\Phi \geq \phi \mid U-1)$ also. Therefore by Eq. S20

$d(U-1,\phi,\gamma)=1$.

Case 3: If $d(U-1,\phi,\gamma)=0$, then Eq. S20 implies $P(\Phi>\phi \mid U-1)>\gamma$. By Eq. S23, $P(\Phi>\phi \mid U)\geq P(\Phi>\phi \mid U-1)$, therefore $P(\Phi>\phi \mid U)>\gamma$ also. Therefore by Eq. S20 $d(U,\phi,\gamma)=0$.

Case 4: If $d(U-1,\phi,\gamma)=1$ then $d(U,\phi,\gamma)\leq d(U-1,\phi,\gamma)$ holds trivially.

Case 5: If both $0<d(U,\phi,\gamma)<1$ and $0<d(U-1,\phi,\gamma)<1$, then combining the conditions for the case in Eq. S20 with Eqs. S22 and Eq. S23 shown above, the relationships between γ and the probabilities in Eq. S38 are as illustrated in Fig. S13.

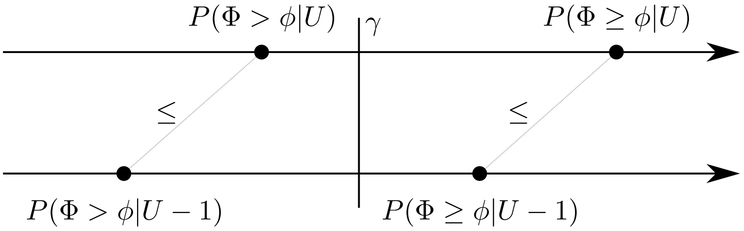


Figure S13: Relationships between γ and the reverse cumulative densities in Case 5.

By Eq. S20, to show $d(U,\phi,\gamma)\leq d(U-1,\phi,\gamma)$ is to show

$$\frac{\gamma-P(\Phi>\phi \mid U)}{P(\Phi=\phi \mid U)}\leq\frac{\gamma-P(\Phi>\phi \mid U-1)}{P(\Phi=\phi \mid U-1)} \tag{S37}$$

which (because $\gamma>P(\Phi>\phi \mid U-1)$) can be rearranged

$$\begin{aligned}\frac{\gamma-P(\Phi>\phi \mid U-1)}{\gamma-P(\Phi>\phi \mid U)}&\geq\frac{P(\Phi=\phi \mid U-1)}{P(\Phi=\phi \mid U)}\\ 1+\frac{P(\Phi>\phi \mid U)-P(\Phi>\phi \mid U-1)}{\gamma-P(\Phi>\phi \mid U)}&\geq\frac{P(\Phi\geq\phi \mid U-1)-P(\Phi>\phi \mid U-1)}{P(\Phi\geq\phi \mid U)-P(\Phi>\phi \mid U)}.\end{aligned} \tag{S38}$$

By Eq. S23 $P(\Phi>\phi \mid U)\geq P(\Phi>\phi \mid U-1)$; therefore, the minimum value of the left-hand side of Eq. S38 occurs when γ is as large as possible. As seen in Fig. S13, the maximum value for γ is $\gamma=P(\Phi\geq\phi \mid U-1)$. Substituting $\gamma=P(\Phi\geq\phi \mid U-1)$ into Eq. S38 and cancelling and rearranging gives the condition

$$P(\Phi\geq\phi \mid U)\geq P(\Phi\geq\phi \mid U-1), \tag{S39}$$

which is true by Eq. S22.

S4.2.3 Random sampling of fuzzy confidence intervals is equivalent to sampling using the mid-P method

What does a random sample of the fuzzy confidence bounds look like? For each randomly sampled bound, we do not obtain a single value, but rather a set of values with partial membership of the corresponding confidence interval. The values define a fuzzy set that is used to determine a fuzzy confidence bound. However, if we assume that membership degrees are commensurate with probabilities, then we can interpret the fuzzy bounds probabilistically, by equating their membership degrees to a probability of being sampled.

One perspective is to think of the fuzzy intervals as the result of random sampling of *crisp* intervals whose members occur within the interval with a probability equal to their membership degree. Then the probability that the bound of a randomly sampled crisp interval takes a certain value can be calculated from those membership degrees. Ideally the crisp intervals should have the interval property: when considering an upper bound, if U_k is a member of the interval, then all $U < U_k$ should be members as well. Conveniently, because we have shown that $d(U,\phi,\gamma_i) \leq d(U-1,\phi,\gamma_i)$ (Subsection S4.2.2 above), then the probability of sampling bound value U_b can be calculated

$$P_S(U_b \mid \phi,\gamma_i) = d(U_b,\phi,\gamma_i) - d(U_b+1,\phi,\gamma_i). \tag{S40}$$

Randomly sampling the bounds in this way results in a probability distribution of bounds given by

$$P_S(U_b \mid \phi) = \int_{\gamma_i=0}^{\gamma_i=1} d(U_b,\phi,\gamma_i) - d(U_b+1,\phi,\gamma_i)\, .d\gamma_i \tag{S41}$$

$$= P(\Phi < \phi \mid U_b) + 0.5P(\Phi = \phi \mid U_b) - \{P(\Phi < \phi \mid U_b+1) + 0.5P(\Phi = \phi \mid U_b+1)\}. \tag{S42}$$

Eq. S42 has a reverse cumulative density function

$$P_S(U_b \geq U^* \mid \phi) = \sum_{U_k=U^*}^{N} P_S(U_k \mid \phi) \tag{S43}$$

$$= P(\Phi < \phi \mid U^*) + 0.5P(\Phi = \phi \mid U^*). \tag{S44}$$

Sampling according to Eq. S44 is equivalent to sampling using the mid-P method (Agresti and Gottard, 2007). That is, for randomly sampled $\alpha^{**} \sim \mathcal{U}(0,1)$, the corresponding sampled bound value U^{**} can be found by:

$$U^{**} \text{ is the greatest value of } U^* \text{ such that } P_M(U_b \geq U^* \mid \phi) \geq \alpha^{**}, \tag{S45}$$

where

$$P_M(U_b \geq U^* \mid \phi) = 0.5[P(\Phi \leq \phi \mid U^*) + 1 - P(\Phi \geq \phi \mid U^*)]. \tag{S46}$$

S4.3 Coverage checks

In this SI, we verify that the confidence intervals obtained by our method have coverage close to the nominal value. The general approach to calculating coverage is given Algorithm 2.

Note that the coverage checks assume that the U_T assumption is true and that U_T is truly known. The true U_T value resulting from the simulation is taken as observed and is used to obtain the sample bounds (Lines 8 and 9).

Algorithm 2 General approach for calculating coverage

```
 1: Choose $\gamma$                                   ▷ Confidence level desired
 2: Choose nsams   ▷ No. of bound samples taken to approximate confidence distribution
 3: Choose nsims              ▷ No. of simulations over which to calculate coverage

 4: Choose $U_0, S_0, T$   ▷ Initialise unobserved and observed parameters for simulations
 5: Choose $\boldsymbol{\mu}, \boldsymbol{\nu}$                     ▷ Functions to return $\mu_t$ and $\nu_t$

 6: cnt_within$=0$
 7: for nsim $=1\ldots$ nsims do

 8:     $\mathbf{S}, \mathbf{E}, U_T =$Simulate$(S_0, U_0, T, \boldsymbol{\mu}, \boldsymbol{\nu})$   ▷ Observe a simulated data **including** $U_T$
 9:     $\mathbf{U_0} =$SampleBounds$(\mathbf{S}, \mathbf{E}, U_T,$nsams$)$          ▷ Uses Algorithm 1 “step-by-step”

10:     $U_{0,\text{lo}} = \text{Percentile}(\mathbf{U_0}, \frac{1-\gamma}{2})$      ▷ $U_0$ bounds approximated using percentiles of sampled bounds
11:     $U_{0,\text{hi}} = \text{Percentile}(\mathbf{U_0}, 1 - \frac{1-\gamma}{2})$

12:     if $U_{0,\text{lo}} \leq U_0 \leq U_{0,\text{hi}}$ then
13:         cnt_within$+=1$
14:     end if

15: end for

16: Coverage $=$ cnt_within/nsims
```

S4.3.1 Simple coverage checks for a variety of scenarios

To check coverage in a simplified case, we considered a timeseries of only $T = 5$ years, where and $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ were governed by a predefined function. Algorithm 3 gives pseudocode for how records were simulated.

Example results are shown in Fig. S14–S17 (more results in Github repository). Comparing Fig. S14 and S15, the coverage improved as the number of species increases. This is behaviour we would expect given that the data is discrete and therefore the confidence distribution is an asymptotic confidence distribution. We explored scenarios in which μ_t and ν_t were sampled from a fixed distribution or varied in time; in both cases, coverage was close to nominal (Fig. S16 and S17.

Algorithm 3 `Simulate` — Simulate a single data record

```
input: $S_0$, $U_0$, $T$, $\boldsymbol{\mu}$, $\boldsymbol{\nu}$

$E_0 = 0$; $X_0 = 0$;                                   ▷ Initialise parameters for simulation

for $t = 1 \dots T$ do

    Draw $N_t \sim \mathrm{Bi}(S_{t-1} + U_{t-1}, 1 - \mu_{t-1})$        ▷ Number of species surviving year

    Draw $\Phi_{t-1} \sim \mathrm{Hyp}(S_{t-1} + U_{t-1}, U_{t-1}, n_t)$   ▷ Survival
    $\psi_{t-1} = n_t - \phi_{t-1}$

    Draw $\Delta_{t-1} \sim \mathrm{Bi}(\phi_{t-1}, \nu_{t-1})$              ▷ Discovery
    $S_t = \psi_{t-1} + \delta_{t-1}$                                ▷ Update records
    $E_t = S_{t-1} - \psi_{t-1} + E_{t-1}$
    $U_t = \phi_{t-1} - \delta_{t-1}$

end for

return $\mathbf{S}, \mathbf{E}, U_T$
```

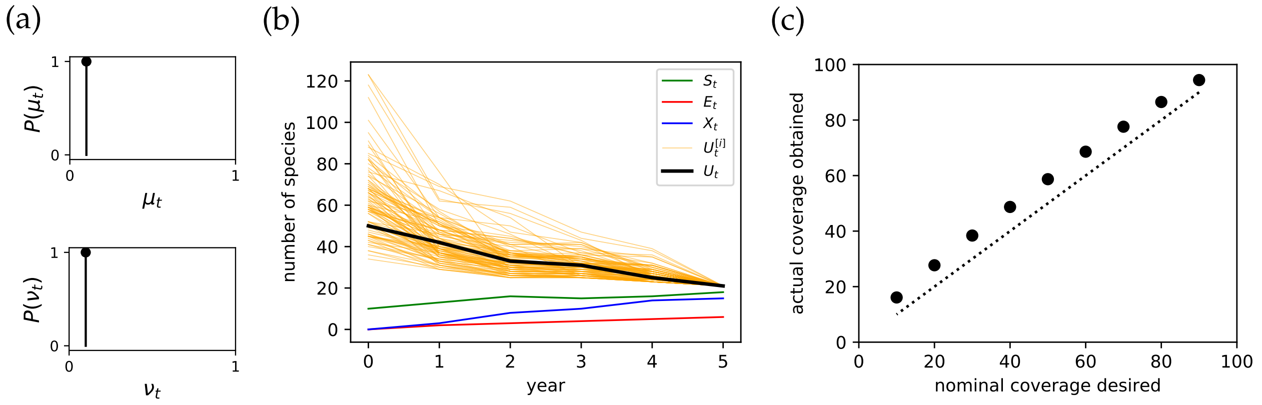


Figure S14: Extinction and detection probability parameters (a), example of one simulation (true U_t in black) and 100 sample bounds (orange) collected by the algorithm (b), and resulting coverage over 1000 simulations with 1000 U_0 bounds collected per simulation (c). Parameters were: $U_0 = 50$, $S_0 = 10$, $T = 5$, with constant $\mu_t = 0.1$ and $\nu_t = 0.1$.

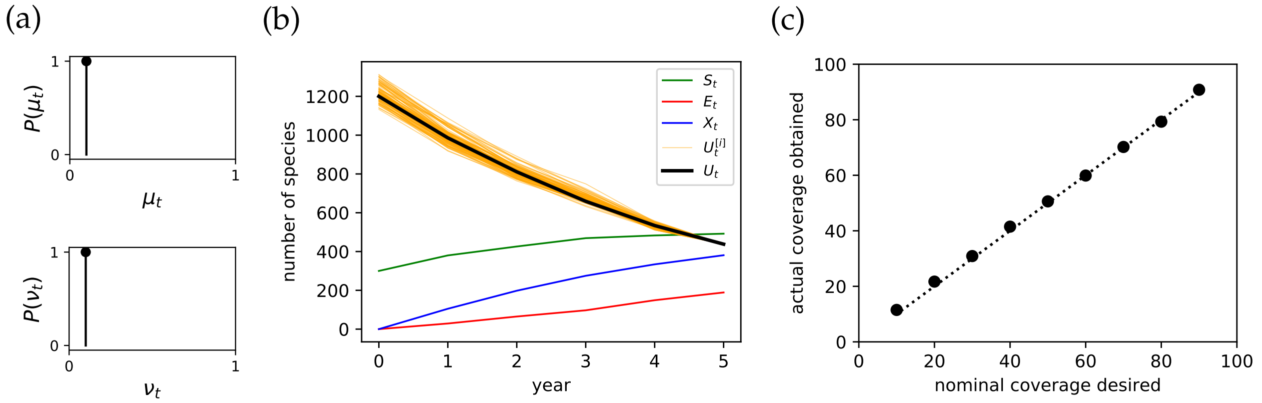


Figure S15: Extinction and detection probability parameters (a), example of one simulation (true U_t in black) and 100 sample bounds (orange) collected by the algorithm (b), and resulting coverage over 1000 simulations with 1000 U_0 bounds collected per simulation (c). Parameters were: $U_0 = 1200$, $S_0 = 300$, $T = 5$, with constant $\mu_t = 0.1$ and $\nu_t = 0.1$.

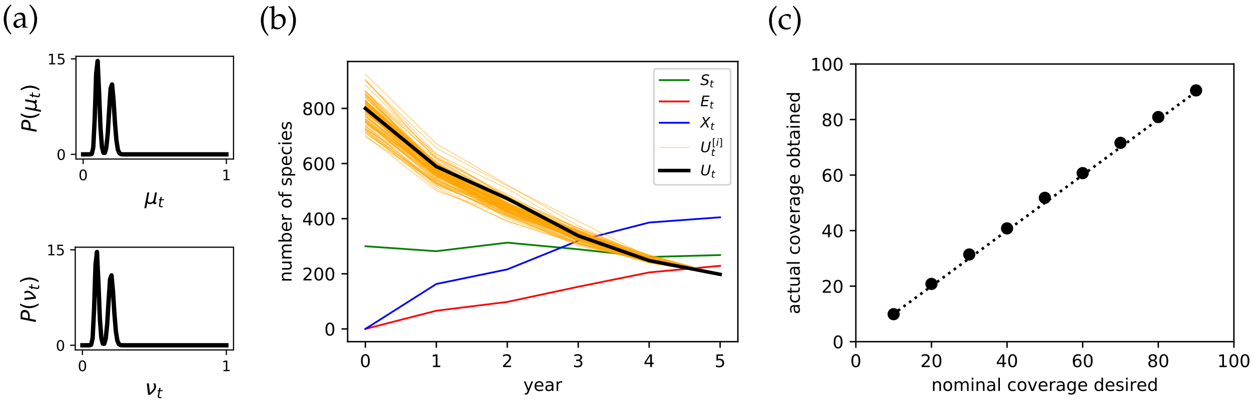


Figure S16: Extinction and detection probability parameters (a), example of one simulation (true U_t in black) and 100 sample bounds (orange) collected by the algorithm (b), and resulting coverage over 1000 simulations with 1000 U_0 bounds collected per simulation (c). Parameters were: $U_0 = 800$, $S_0 = 300$, $T = 5$, with $\mu_t \sim 0.5\,\text{Beta}(50, 450) + 0.5\,\text{Beta}(100, 400)$ and $\nu_t \sim 0.5\,\text{Beta}(50, 450) + 0.5\,\text{Beta}(100, 400)$.

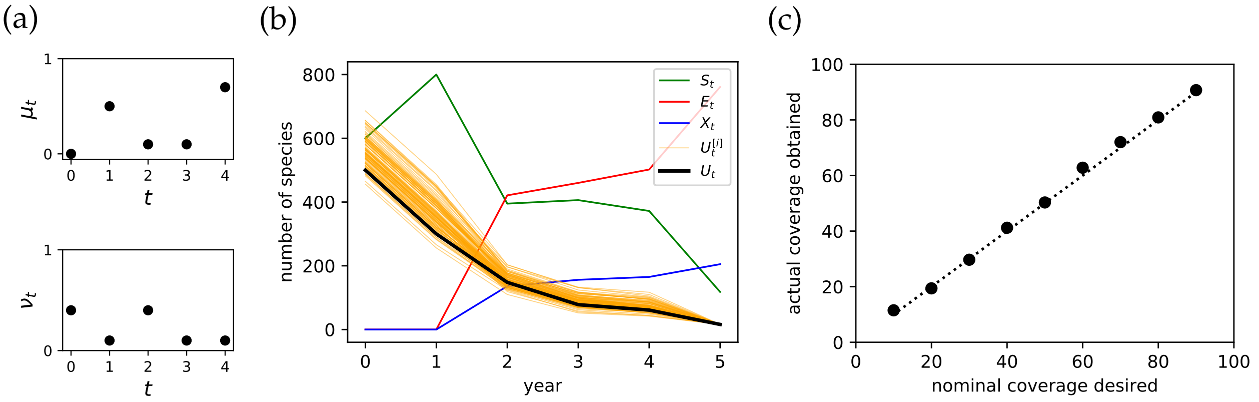


Figure S17: Extinction and detection probability parameters (a), example of one simulation (true U_t in black) and 100 sample bounds (orange) collected by the algorithm (b), and resulting coverage over 1000 simulations with 1000 U_0 bounds collected per simulation (c). Parameters were: $U_0 = 500$, $S_0 = 600$, $T = 5$, with $\boldsymbol{\mu} = [0, 0.5, 0.1, 0.1, 0.7]$ and $\boldsymbol{\nu} = [0.4, 0.1, 0.4, 0.1, 0.1]$.

S4.3.2 Coverage checks for scenarios similar to Singapore plants

In the Singapore plants case study, we defined a timestep as a period of years in which at least one discovered extinction occurred (i.e. the lengths of timesteps vary), and we worked under the assumption that $U_T = 0$. To check coverage under a similar scenario, we modified the simulations so that they were run for as many timesteps as needed until U_T reached zero. Then, before sampling the bounds, we merged all timesteps in which no discovered extinctions occurred.

Our method obtained close agreement between the actual and nominal coverage for the confidence intervals on U_0 for a variety of confidence levels and simulation parameter values (example results in Fig. S18a-b). However, because the value of U_t is bounded below but not above, the distribution of estimates $\bar{U}_0$ has no upper limit and so the method tends to overestimate the true value of U_0 (Fig. S18c-d).

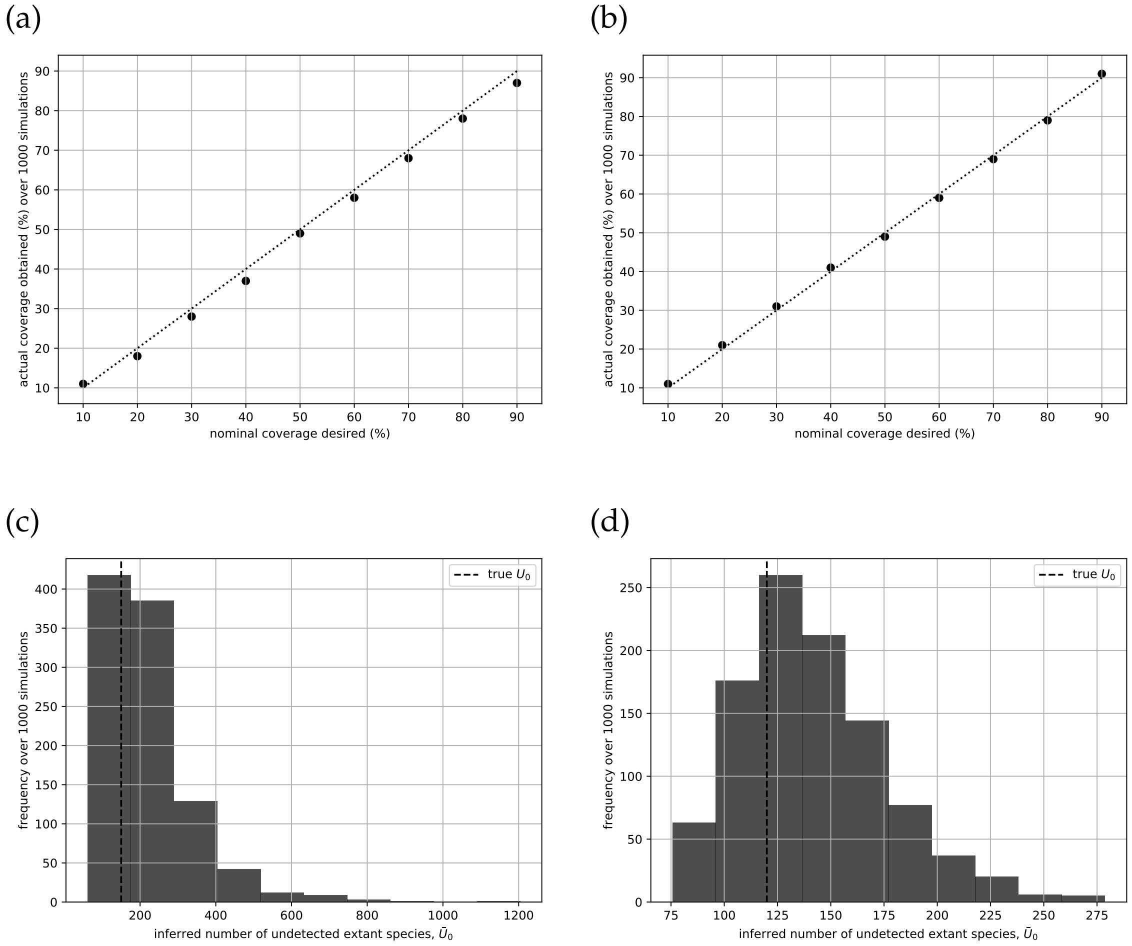


Figure S18: Two examples of the coverage (a-b) and estimates $\bar{U}_0$ (c-d) obtained using our method on simulated data. Results are from 1000 simulations, with confidence intervals estimated from the percentiles (with linear interpolation) obtained from 500 samples of fuzzy confidence bounds. Parameters are: (a) $U_0 = 150$, $S_0 = 20$, $\mu = 0.1$, $\delta_t = \delta = 2$. (b) $U_0 = 120$, $S_0 = 20$, $\mu = 0.05$, $\delta_t = \mathcal{U}(\{1,2,3\})$.

S5 Bayesian inference

S5.1 Sampling scheme used

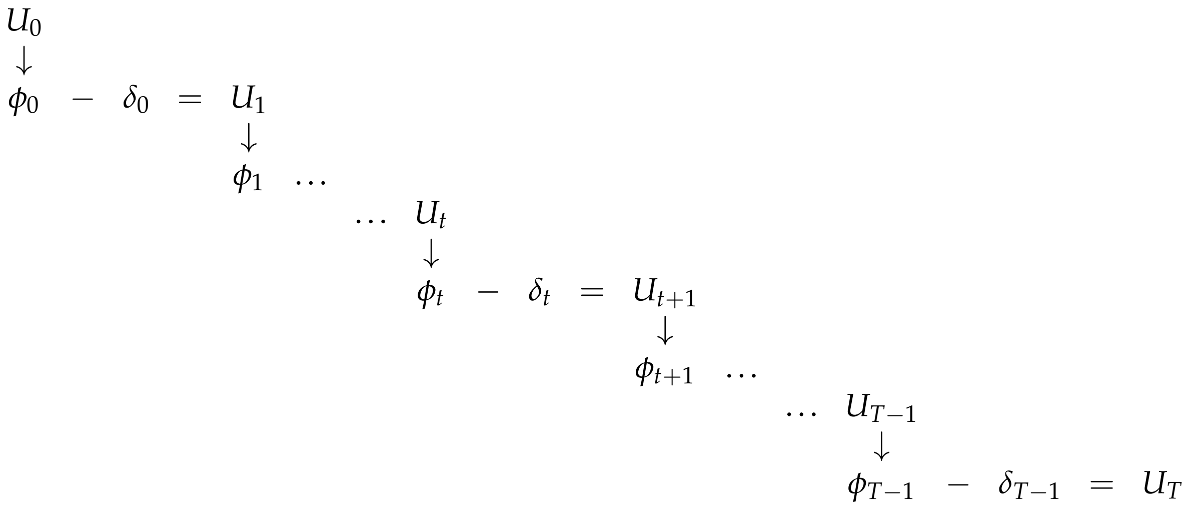
Eq. 12 can be rewritten

$$P_{\mathbf{n}}(\phi_t \mid U_t) = \frac{\binom{U_t}{\phi_t}\binom{n_t - U_t}{n_{t+1}-\phi_t}}{\binom{n_t}{n_{t+1}}}, \tag{S47}$$

where, in the next step, U_{t+1} will be determined by species detections $U_{t+1} = \phi_t - \delta_t$. Thus the model entails a conditional dependency network between ϕ_t values at each timestep (or, equivalently, U_{t+1} values at each timestep) that is a linear chain (Fig. S19). The linear chain structure means that the Markov blanket for ϕ_t (i.e. the node's parents, children, and parents of those children, in the network; see Definition 4.3 of Darwiche (2009)) contains only ϕ_{t-1} and ϕ_{t+1}. Therefore the Gibbs sequence can be obtained by sampling sequentially from

$$\Phi_t^{(j)} \sim P_{\mathbf{S},\boldsymbol{\psi}}(\phi_t \mid \phi_{t+1}^{(j-1)}, \phi_{t-1}^{(j)}) \tag{S48}$$

where the superscript (j) is the sample number (Resnik and Hardisty, 2010).



U_0
$\phi_0 \quad - \quad \delta_0 \quad = \quad U_1$
$\phi_1 \quad \ldots$
$\ldots \quad U_t$
$\phi_t \quad - \quad \delta_t \quad = \quad U_{t+1}$
$\phi_{t+1} \quad \ldots$
$\ldots \quad U_{T-1}$
$\phi_{T-1} \quad - \quad \delta_{T-1} \quad = \quad U_T$

Figure S19: A schematic of the conditional dependency relationship between the number of undiscovered species that survive each timestep ϕ_t, the number of undiscovered species at the start of the timestep U_t, and the number of undiscovered species that survive the next timestep ϕ_{t+1}. Arrows represent the conditional dependencies between variables, and the number of species discovered at each timestep (δ_t) is fixed and obtained from the data. The network structure of the conditional dependencies is a linear chain.

Note first that S48 can also be written

$$\Phi_t^{(j)} \sim P(\phi_t \mid \phi_{t+1}^{(j-1)}, U_t^{(j)}). \tag{S49}$$

The intuitive meaning of Eq. S49 is that it asks for the probability of taking the ϕ_t pathway between ϕ_{t+1} and U_t. Define Q_t as the set of all possible ϕ_t values given the constraint of the U_t before and ϕ_{t+1} after it, i.e. $Q_t = \{\phi_{t+1} + \delta_t, \ldots, U_t\}$. Then Eq. S48 is:

$$\begin{aligned} P(\phi_t \mid U_t, \phi_{t+1}) &= \frac{P(U_t, \phi_t, \phi_{t+1})}{P(U_t, \phi_{t+1})} \\ &= \frac{P(U_t, \phi_t, \phi_{t+1})}{\sum_{\phi'_t \in Q_t} P(U_t, \phi'_t, \phi_{t+1})} \\ &= \frac{P(\phi_{t+1} \mid \phi_t) P(\phi_t \mid U_t) P(U_t)}{\sum_{\phi'_t \in Q_t} P(\phi_{t+1} \mid \phi'_t) P(\phi'_t \mid U_t) P(U_t)} \\ &= \frac{P(\phi_{t+1} \mid \phi_t) P(\phi_t \mid U_t)}{\sum_{\phi'_t \in Q_t} P(\phi_{t+1} \mid \phi'_t) P(\phi'_t \mid U_t)} \end{aligned} \tag{S50}$$

The denominator in Eq. S50 does not have a simple analytic form, and evaluating the hypergeometric for every $\phi'_t \in Q_t$ is computationally expensive. Therefore the denominator (which acts as a normalisation factor) was removed, and the Metropolis algorithm was used for sampling within each Gibbs step instead, with the proposal distribution of a uniform distribution across Q_t.

The terms in the numerator in Eq. S50 are evaluated

$$P(\phi_t \mid U_t) = \frac{\binom{U_t}{\phi_t}\binom{S_t}{\psi_t}}{\binom{U_t + S_t}{\phi_t + \psi_t}} \tag{S51}$$

and

$$P(\phi_{t+1} \mid \phi_t) = \frac{\binom{\phi_t - \delta_t}{\phi_{t+1}}\binom{S_{t+1}}{\psi_{t+1}}}{\binom{\phi_t - \delta_t + S_{t+1}}{\phi_{t+1} + \psi_{t+1}}} \tag{S52}$$

where δ_t are also known from the data ($\delta_t = S_{t+1} - \psi_t$).

At the first step, in order to draw ϕ_0, a value of U_0 is needed. For convenience let $u_0 = U_0 - \phi_0$, so that the constraint $u_0 \geq 0$. One wishes to sample u_0, however only the likelihood of u_0 is known

$$P(\Phi_0 = \phi_0 \mid u_0) = L(u_0) = \frac{\binom{u_0 + \phi_0}{\phi_0}\binom{S_0}{\psi_0}}{\binom{u_0 + \phi_0 + S_0}{\phi_0 + \psi_0}} \tag{S53}$$

In order to sample u_0, an improper uniform prior distribution was assumed. The normalisation factor is

$$L_0 = \sum_{u_0 = 0}^{\infty} L(u_0) = \frac{\binom{S_0}{\psi_0}}{\binom{S_0 + \phi_0}{\phi_0 + \psi_0}} \, {}_2F_1 \left(\begin{matrix} \phi_0 + 1, S_0 - \psi_0 + 1 \\ S_0 + \phi_0 + 1 \end{matrix} ; 1 \right) \tag{S54}$$

By definition

$$\frac{\binom{S_0}{\psi_0}}{\binom{S_0 + \phi_0}{\phi_0 + \psi_0}} = \frac{S_0! (\phi_0 + \psi_0)!}{\psi_0! (S_0 + \phi_0)!} \tag{S55}$$

and by Gauss's Theorem

$$
{}_2F_1\left(\begin{matrix}\phi_0+1, S_0-\psi_0+1\\ S_0+\phi_0+1\end{matrix};1\right)=\frac{(S_0+\phi_0)!(\psi_0-2)!}{(S_0-1)!(\phi_0+\psi_0-1)!} \tag{S56}
$$

therefore substituting Eq. S55 and S56 into Eq. S54, a simple analytic expression for the normalisation factor is found

$$
L_0=\frac{S_0(\phi_0+\psi_0)}{\psi_0(\psi_0-1)} \tag{S57}
$$

and the normalisation factor can be used to define a sampling probability for u_0

$$
P_s(u_0)=\frac{\binom{u_0+\phi_0}{\phi_0}\binom{S_0}{\psi_0}}{\binom{u_0+\phi_0+S_0}{\phi_0+\psi_0}}\frac{\psi_0(\psi_0-1)}{S_0(\phi_0+\psi_0)} \tag{S58}
$$

S5.2 Diagnostics

S5.2.1 Convergence

Following Cowles and Carlin (1996) and Gelman and Shirley (2011), convergence of the Gibbs scheme was monitored using a combination of trace plots and the Gelman-Rubin statistic $\hat{R}$ (obtained using the "PyMC3" module in Python3 on the log-transformed U_t), with heuristic criterion $\hat{R} < 1.1$. Two independent chains with separate starting points were used, and the trace-plots and Gelman-Rubin criterion suggests that a burn-in time of 15,000 iterations is adequate (Fig. S20).

S5.2.2 Adequate effective sample size

The samples obtained from the MCMC are autocorrelated; therefore the precision of the estimates obtained from them is lower than it would be if obtained from the same number of independent samples. To estimate the effective size of the MCMC sample in terms of an equivalent number of independent samples, we measure the effective sample size (ESS), which takes this autocorrelation into account. We need to ensure that the ESS is larger than the minimum ESS that is needed to obtain estimates with a reasonable precision.

The minimum ESS can be calculated *a priori* using the method of Vats et al. (2017). The minimum ESS is a function of the dimension of the problem (i.e. how many unknown U_t values we are trying to estimate), the level of confidence at which we want to obtain confidence region bounds (in our case 95-percentiles), and a tolerance ε. The tolerance ε gives a measure of how large the variance is that is due to the MCMC itself compared to the variance in the posterior distribution. Estimates will be approximately $1/\varepsilon$ more accurate than the standard deviation in the posterior distribution (Gong and Flegal, 2016).

We used the 'mcmcse' package in R (Flegal et al., 2017; Vats, 2017) to determine the minimum multivariate effective sample size that was required in order to estimate the 95-percentile bounds with a tolerance of $\varepsilon = 0.1$. For our $p = 108$ variables, and at confidence level $\alpha = 0.05$, we calculated a minimum ESS of 1997.

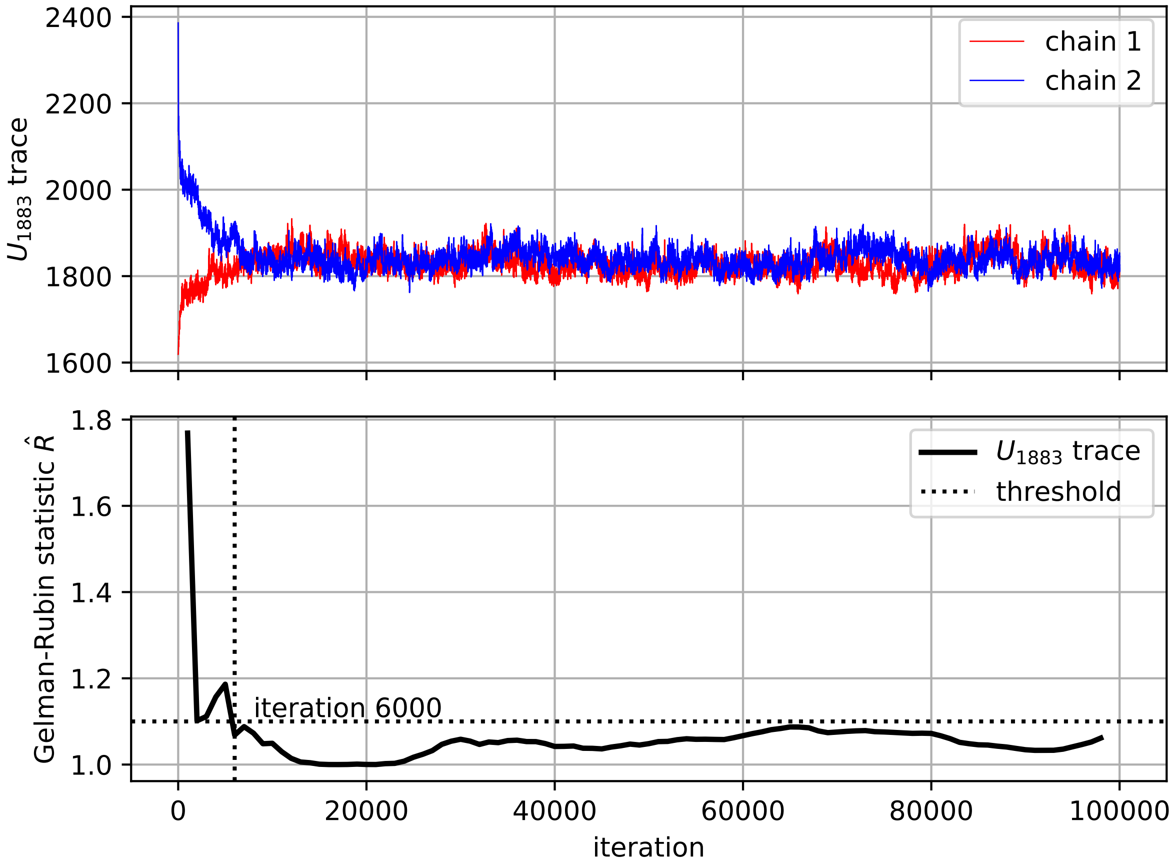


Figure S20: Trace plots and the Gelman-Rubin statistic resulting from the Metropolis-within-Gibbs scheme for the number of undiscovered extant species in 1883. The year 1883 was chosen because U_{1883} was the last variable to cross the threshold $\hat{R} < 1.1$. The threshold was cross by iteration 6000, and the trace-plots were visually similar soon after, therefore a burn-in time of 15,000 was chosen as a conservative overestimate.

We then monitored the ESS of our MCMC sample, to ensure that our final sample size exceeded the minimum ESS value. Effective sample size can be calculated for each variable independently (Gong and Flegal, 2016), or for all variables combined (Vats et al., 2017). We found that the former gave lower values than the latter, so we used them instead to give a conservative estimate of the effective sample size. We verified that the ESS of our MCMC sample produced by the two chains combined exceeded the minimum ESS for every variable (Fig. S21).

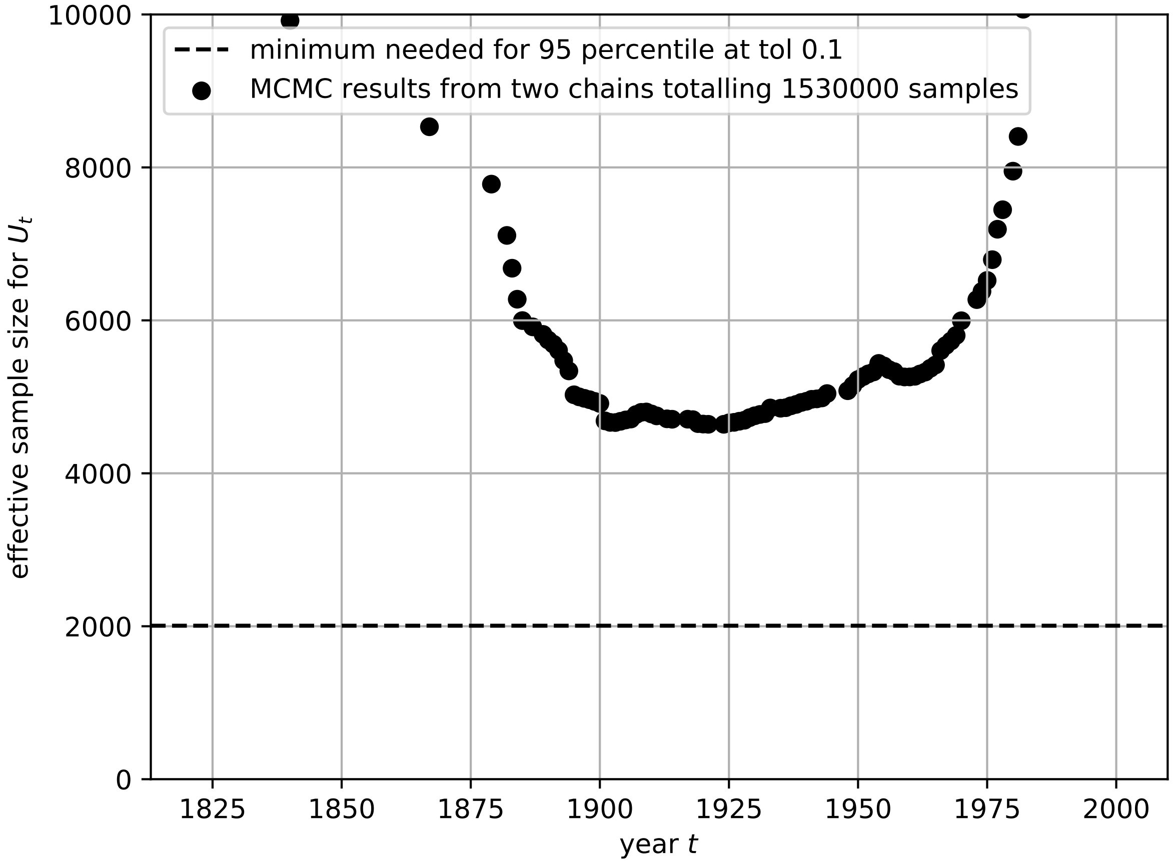


Figure S21: The effective sample sizes for each U_t at timestep t (points) are above the minimum effective sample size needed in order to estimate the 95-percentile bounds with a tolerance of $\varepsilon = 0.1$ (dotted line).

Finally we visually compared the posterior distributions for U_{1822} obtained from the two independent MCMC chains, and verified that they were similar (Fig. S22a). Even for the variables with low ESS, the shape of the posterior and the central estimates were similar between the two chains (e.g. U_{1910} in Fig. S22b).

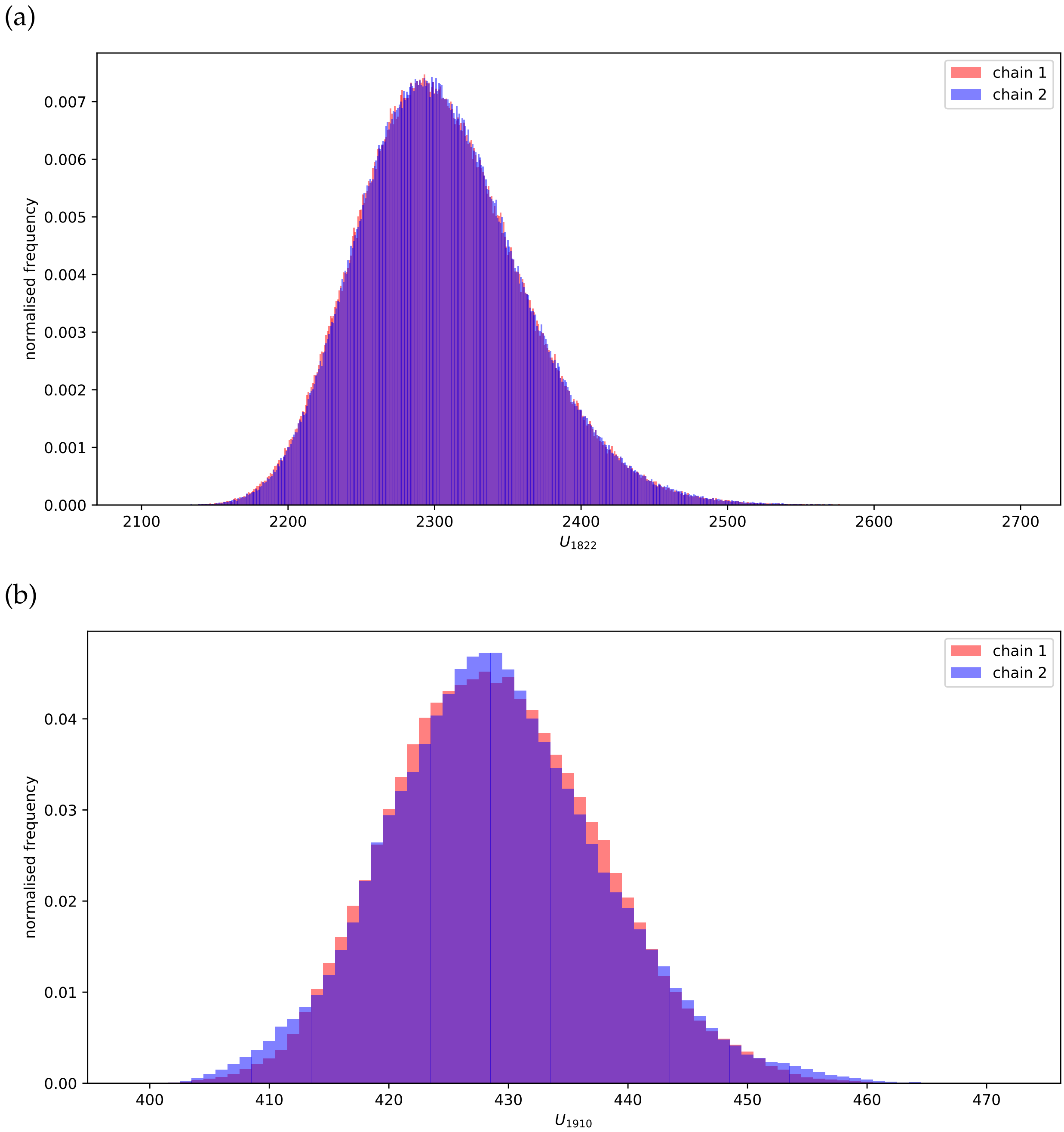


Figure S22: Histograms of the relative frequencies of U_t obtained from the two separate chains (15,000 burn-in iterations excluded), for the the years 1822 and 1910. Comparing the frequency histograms between the two chains gives an intuitive sense of how well the MCMC has converged and whether the sample size is large enough. The year 1822 is an important result because it is the first year used in the SEUX model, and therefore U_{1822} determines the total number of species which in turn determines the total number of undiscovered extinctions. The year 1910 is also included as an example because the effective sample size is lowest around this year (Fig. S21).

S5.3 Monte Carlo standard error

The batch means method (Vats et al., 2017), also implemented in the 'mcmcse' R package, was used to obtain the multivariate MCMC standard error on the mean and 95-percentile bound estimates. The MCMC standard error quantifies how much the posterior estimates will vary between MCMC runs. We verified that the MCMC standard error was small compared to the standard deviation in the posterior (examples in Table S4).

Table S4: Examples of U_t estimates, percentile bounds, and MCMC standard error for estimates and bounds. Results for every year available in code archive.

year	$\bar{U}_t$		95 percentiles			
	estimate	SE	lower bound	SE	upper bound	SE
1822	2305.0	0.8	2206	1.1	2427	1.3
1840	2079.8	0.7	2026	0.8	2143	1.2
1879	1941.5	0.7	1896	0.8	1994	1.0
1900	507.5	0.6	486	0.9	531	0.9
1925	387.9	0.4	374	0.6	404	0.8
1950	253.0	0.2	244	0.3	263	0.4
1975	167.67	0.07	164	0.07	173	0.08
2001	84.122	0.003	84	0.0004	85	0.0009

S6 Fisher's extended SEUX

The purpose of this Appendix is to explain how we estimated the odds ratio ω that was necessary in order for our model (Eq. 22) to reproduce the high total extinction rate that was inferred in the Brook et al. (2003) study. The Fisher's noncentral hypergeometric distribution used in Eq. 22 was implemented using the R package "BiasedUrn" (Fog, 2015) via the R-to-Python interface "rpy2" (Gautier, 2016) in Python3.

Brook et al. (2003) inferred a total extinction rate of 0.74 (extinct/total species). In order to match this result in our model, we first explored how the total extinction rate predicted by our model varied with the parameter ω (Fig. 5a). This estimate was then refined by gradually increasing the number of samples (used to derive the confidence intervals and expected number of undiscovered extinctions) to increase the precision of ω. The final result was obtained using classical inference with a sample size of 10^4 (the same sample size as was used on the plants data for the standard SEUX model). It was found that, when $\omega = 0.17$, the expected value of the total number of species was inferred to be 6293 (c.f. 6549 in Brook et al. (2003)), and the total extinctions inferred to be 4503 (c.f. 4681 in Brook et al. (2003)), giving a total extinction rate of 0.74 (Fig. S23).

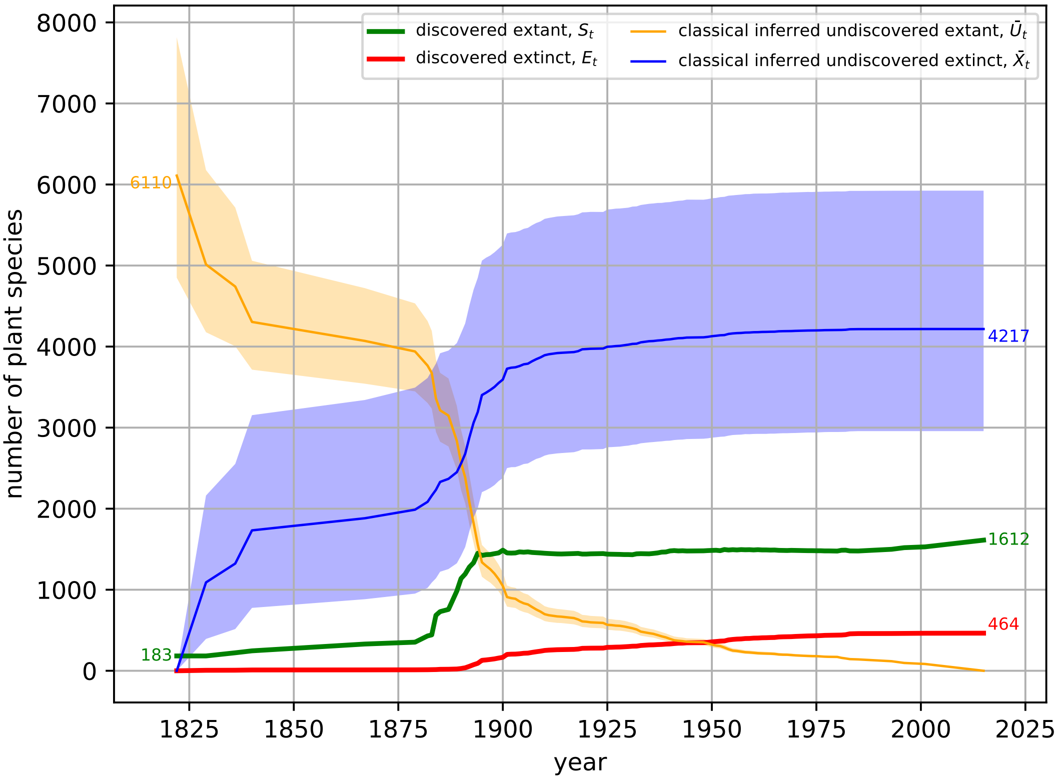


Figure S23: The SEUX plot obtained when $\omega = 0.17$, which results in an expected extinction rate approximately equal to that inferred by Brook et al. (2003) (i.e. rate of 0.74).

S7 Sensitivity of the SEUX model: additional results

The total extinction rate estimated was not sensitive to random deletion of species from the record, nor to the assumed number of undiscovered extant species assumed to remain at the end of the record U_T (Fig. S24). In contrast, the absolute number of undiscovered extant and extinct species can be affected (Fig. S25).

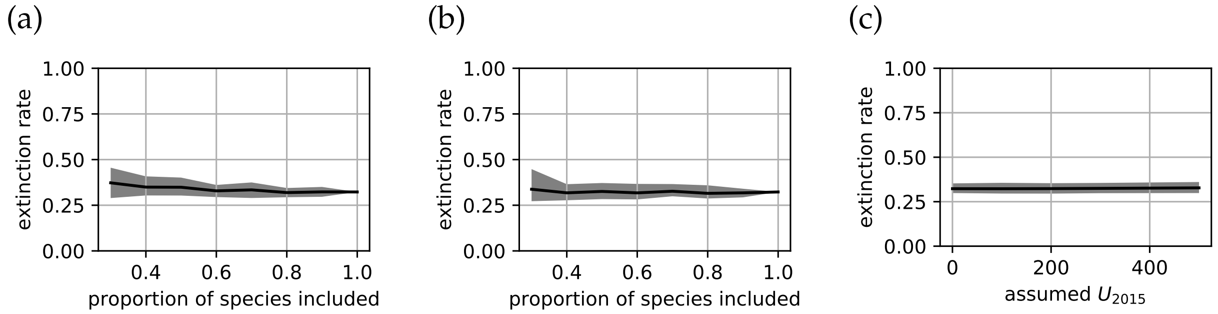


Figure S24: The effect of random deletion of species from the record (a-b), and the number of undiscovered extant species assumed to remain at the end of the record U_T (c), on the extinction rate estimated using classical inference. In panel (a), U_T was calculated from the number of extant species that were deleted from the record; whereas in panel (b), the missing species were treated as unknown and the inference assumed $U_T = 0$.

When species are deleted from the record but the true value of U_T is known, then the total number of species inferred is not much changed, and the main effect is to shift species from the discovered to undiscovered portions (Fig. S25a). However if the incorrect value of U_T is assumed, either because extant species that are missing from the record are not accounted for (Fig. S25b), or because the number of undiscovered extant species is not $U_T = 0$ (Fig. S25c), then the total number of species will be underestimated.

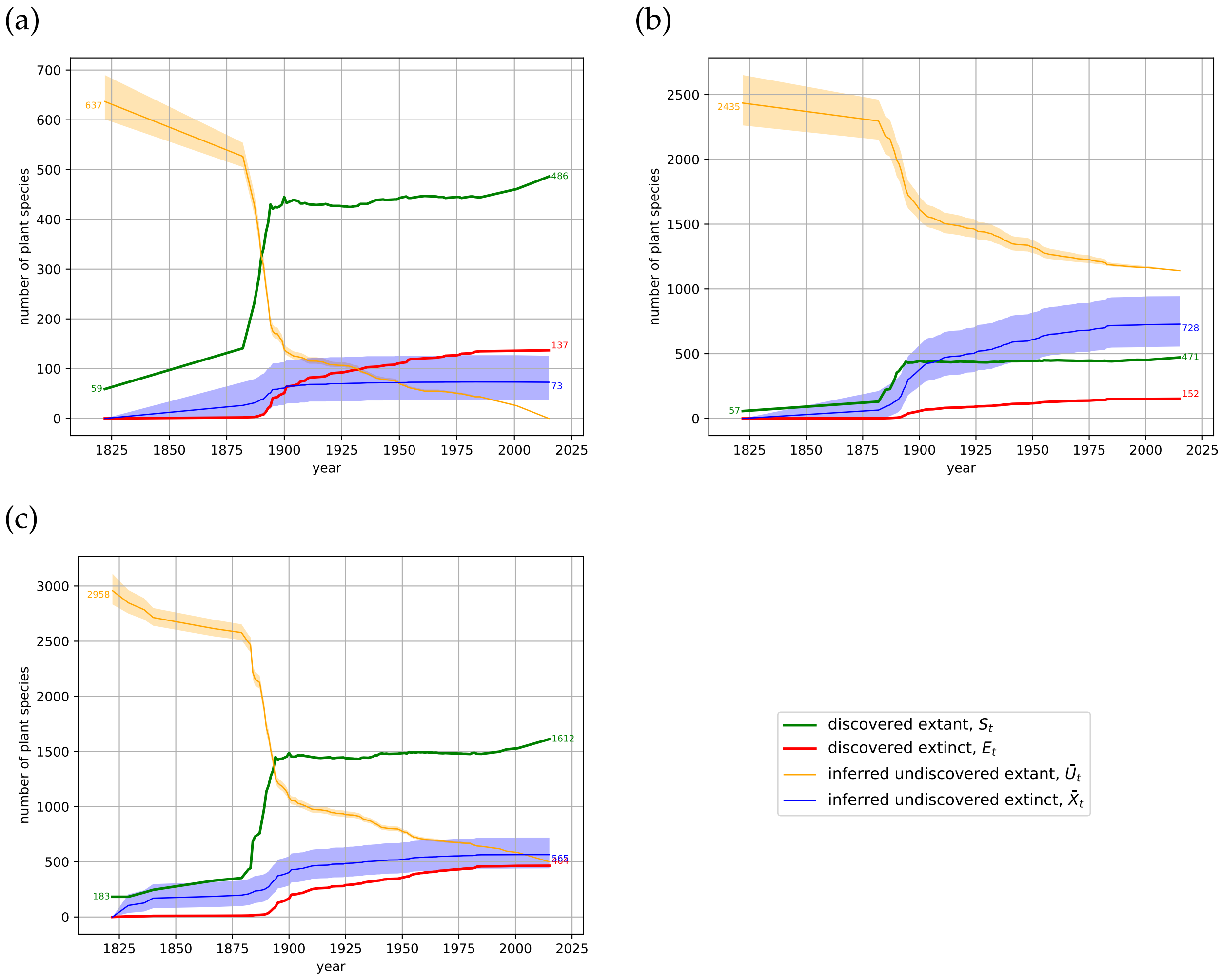


Figure S25: Examples showing the effect upon the SEUX results of: random deletion of 30% of species from the record (a-b), and the number of species assumed to remain undiscovered in the current year, $U_T = 500$ (c). In panel (a), we assume that we can make the standard assumption that $U_T = 0$, whereas in panel (b) we assume that we know the correct U_T given the 30% missing species.

Supplementary references

Agresti, A. and Gottard, A. (2007). Nonconservative exact small-sample inference for discrete data, *Computational Statistics and Data Analysis* **51**(12): 6447–6458.

Berry, G. and Armitage, P. (1995). Mid-p confidence intervals: a brief review, *The Statistician* **44**(4): 417–423.

Brook, B. W., Sodhi, N. S. and Ng, P. K. (2003). Catastrophic extinctions follow deforestation in Singapore, *Nature* **424**(6947): 420–426.

Chong, K. Y., Lee, S. M., Gwee, A. T., Leong, P. K., Ahmad, S., Ang, W. F., Lok, A. F., Yeo, C. K., Corlett, R. T. and Tan, H. T. (2012). Herbarium records do not predict rediscovery of presumed nationally extinct species, *Biodiversity and Conservation* **21**(10): 2589–2599.

Chong, K. Y., Tan, H. T. and Corlett, R. T. (2009). *A checklist of the total vascular plant flora of Singapore: native, naturalised and cultivated species*, Raffles Museum of Biodiversity Research, National University of Singapore, Singapore.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association* **91**(434): 883–904.

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*, Cambridge University Press.

Duffy, K. J., Kingston, N. E., Sayers, B. A., Roberts, D. L. and Stout, J. C. (2009). Inferring national and regional declines of rare orchid species with probabilistic models, *Conservation Biology* **23**(1): 184–195.

Feldman, G. J. and Cousins, R. D. (1998). Unified approach to the classical statistical analysis of small signals, *Physical Review D* **57**(7): 3873.

Flegal, J. M., Hughes, J., Vats, D. and Dai, N. (2017). *mcmcse: Monte Carlo Standard Errors for MCMC*, Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN. R package version 1.3-2.

Fog, A. (2015). Biasedurn: Biased urn model distributions. R package version 1.07.
URL: *https://cran.r-project.org/web/packages/BiasedUrn/*

Gautier, L. (2016). rpy2. r package version 2.8.
URL: *http://rpy2.readthedocs.io/en/version_2.8.x/*

Gelman, A. and Shirley, K. (2011). Inference from simulations and monitoring convergence, *in* S. Brooks, A. Gelman, G. Jones and X.-L. Meng (eds), *Handbook of Markov Chain Monte Carlo*, pp. 163–174.

Geyer, C. J. and Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and p-values, *Statistical Science* **20**(4): 358–366.

Gong, L. and Flegal, J. M. (2016). A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo, *Journal of Computational and Graphical Statistics* **25**(3): 684–700.

McCarthy, M. A. (1998). Identifying declining and threatened species with museum data, *Biological Conservation* **83**(1): 9–17.

Neyman, J. (1937). X – outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transactions of the Royal Society of London A* **236**(767): 333–380.

Resnik, P. and Hardisty, E. (2010). Gibbs sampling for the uninitiated, *Technical report*, Maryland University College Park Institute for Advanced Computer Studies.

Robbirt, K. M., Roberts, D. L. and Hawkins, J. A. (2006). Comparing IUCN and probabilistic assessments of threat: do IUCN red list criteria conflate rarity and threat?, *Biodiversity & Conservation* **15**(6): 1903–1912.

Solow, A. R. (1993a). Inferring extinction from sighting data, *Ecology* **74**(3): 962–964.

Solow, A. R. (1993b). Inferring extinction in a declining population, *Journal of Mathematical Biology* **32**(1): 79–82.

Vats, D. (2017). An introduction to estimating Monte Carlo standard errors with R package mcmcse.
URL: *https://cran.r-project.org/web/packages/mcmcse/vignettes/mcmcse_vignette.pdf*

Vats, D., Flegal, J. M. and Jones, G. L. (2017). Multivariate output analysis for Markov chain Monte Carlo, *arXiv preprint arXiv:1512.07713v4* .