Report of programming bugs in the DAISIE R package: consequences and correction

Luis Valente[1,2,*], Nadiah P. Kristensen[3], Albert B. Phillimore[4], Rampal S. Etienne[2]

[1] Naturalis Biodiversity Center, Darwinweg 2, 2333 CR Leiden, The Netherlands

[2] Groningen Institute for Evolutionary Life Sciences, University of Groningen, P.O. Box 11103, Groningen 9700 CC, The Netherlands

[3] Department of Biological Sciences, National University of Singapore, 16 Science Drive 4, Singapore, 117558, Singapore

[4] Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3JT, United Kingdom

*Corresponding author: luis.valente@naturalis.nl

Abstract

DAISIE (Dynamic Assembly of Islands through Speciation, Immigration and Extinction) is a framework for a dynamic stochastic model of island biogeography that can be used to estimate the rates of colonisation, speciation and extinction (CES rates) from phylogenetic trees of insular communities by maximum likelihood, and to simulate such data sets given a set of rates. The framework is available in an R package, and has been used in various studies to estimate CES rates for different insular systems based on molecular phylogenetic data. Simulations with estimated parameters have been used to explore the goodness-of-fit of the model and uncertainty in parameter estimates via bootstrap. We report two minor bugs in the simulation code that was used in the manuscripts published to date that have used DAISIE. One affects the branching times for some species that evolve via cladogenesis, the other affects cases when a single mainland species colonises an island multiple times. While these bugs do not have any effect on maximum likelihood estimation or model selection from empirical (non-simulated) data, they do have the potential to affect analyses based on simulated data. Here, we describe and fix the two bugs. To assess the magnitude of the error that they introduced, we re-run simulations and bootstrap analyses for the published cases that are most likely to be affected. We find that the differences between results generated with the new code and the original code are very minor, equivalent in magnitude to differences that would be obtained when using a different set of random numbers to generate the simulations. We conclude that the qualitative and to a large extent also the quantitative results of previous publications using DAISIE are robust to these bugs, and we provide a new version of the R package (v 3.2.0) with the corrected code.

Introduction

DAISIE is a framework to study insular biodiversity in a phylogenetic context. Focused on community-level phylogenies of species on islands (rather than on phylogenies of specific island lineages) and the times of their colonisation from a mainland pool, it can be used to estimate rates of island colonisation, speciation and extinction (CES rates), as well as island diversity carrying-capacities. Furthermore, it can be used to simulate community-level phylogenies for a given set of CES rates. The model was developed in 2015 (Valente *et al.* 2015) based on a previous mechanistic simulation model (Valente *et al.* 2014). Among other aims, DAISIE has been used to test for equilibrium dynamics in island birds (Valente *et al.* 2015, 2017b), estimate time to recover from anthropogenic extinction in island bats and birds (Valente *et al.* 2017a, 2019) and to test MacArthur & Wilson's theory of island biogeography on a global evolutionary scale in birds (Valente *et al.* 2020). Bootstrap analyses of simulated datasets have shown that parameters can be estimated with high precision and little bias (Valente *et al.* 2015, 2017a, 2017b, 2018, 2020; Hauffe *et al.* 2020).

We here report two bugs in the computer code that has been used to simulate datasets in DAISIE. We know a priori that these bugs do not affect the main findings of any of the manuscripts that have used DAISIE, as they do not affect the maximum likelihood (ML) estimation based on empirical data. They also do not affect the numbers of species that result from the simulations. However, the bugs do affect the distribution of branching times in simulated islands and can affect the reported precision of the parameter estimates from bootstrap analyses as these rely on the simulation code. If the bugs have a strong effect, the estimates of parameters reported in the bootstrap analyses in previous published studies may be incorrect, for example giving the impression that the precision or accuracy is lower or higher than it would be without the bugs. Here, we describe these bugs and how they affect simulations and downstream analyses. We then re-run the most affected analyses using the corrected code and compare them to analyses with the bugs, to assess the magnitude of the error these may have induced.

Methods

Bug 1: Branching times for some cladogenetic species following extinctions in simulations

This bug was discovered by Nadiah Kristensen, who was exploring the open source code of the package on GitHub (github.com/rsetienne/DAISIE). In the DAISIE model, species can speciate via anagenesis (one species becomes a new species without lineage

splitting) or cladogenesis (one species splits into two new cladogenetic species). This bug refers only to cases where cladogenetic species go extinct in the simulations. When a parent species undergoes cladogenetic speciation, both daughters are a new species, receiving new species IDs. Our coding of branching topology follows the AB approach used in CAIC (Purvis & Rambaut 1995). The daughters' branching AB codes and branching times are used in combination to record all divergence times in a clade. If the parent is a mainland species (not yet speciated) or anagenetic species, one daughter has branch code "A" and the other "B". Daughter A's branching time is the mainland ancestor's colonisation time, daughter B's branching time is the time that the branching event took place. If the parent is a cladogenetic species, the A and B codes are appended to the end of the parent's branch code. Daughter B's branching time is again the time when the new branching event took place, but daughter A's branching time is the parent's branching time. Thus, for each cladogenetic species, its branching time is either the branching time resulting in the most recent B in its branching code; or, if there is no B in its branching code, the time that the colonisation took place. An example of this coding system is given in Fig. 1.

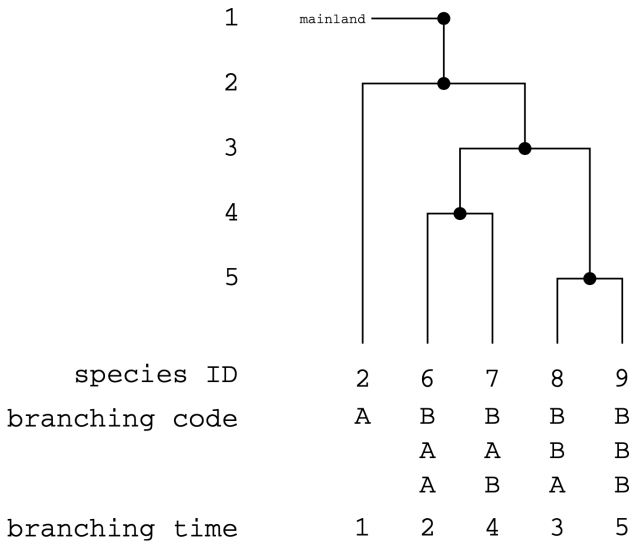

Figure 1 - An illustration of an island lineage with five species resulting from a single mainland ancestor. The species were created from the following sequence of events: (1) colonisation of species 1 from the mainland; (2) cladogenetic speciation of species 1 into species 2 and 3; (3) cladogenetic speciation of species 3 into species 4 and 5; (4) cladogenetic speciation of species 4 into species 6 and 7; (5) cladogenetic speciation of species 5 into species 8 and 9. The lineage is drawn so that A daughters are on the left and B daughters on the right.

The branching times produced by the simulation code are typically compared to empirical divergence times based on extant species data. Reconstructed phylogenetic trees do not record branching events associated with extinct lineages, so these branching events must also be removed from the simulated tree. When a simulated

cladogenetic species goes extinct, then the record of both the branching event in the branching codes and the branching time of that species from its sister must be removed.

If the cladogenetic species that goes extinct has a branching code with its most recent entry as B, then the changes required are relatively simple. The record of its most recent branching event is removed from the branching codes of its sisters (where 'sisters' refers to either the single sister or the set of all of the descendants of the sister). The branching time of that most-recent event is stored with that extinct species, so it is removed when the species is removed.

If the cladogenetic species that becomes extinct has branching code ending in A, then changes also need to be made to the branching time of one of its sisters. The extinct species contains the branching time that corresponds to the branching event (or immigration event) that occurred before its most recent split from a sister, in other words the stem age of the subclade that includes the extinct species and its sisters. One of its sisters holds the branching time of the extinct species from its most recent sister, which is the branching time that we would like to remove following extinction. That sister's branching time must then be replaced with the branching time being held by the extinct species. The relevant sister will be the sister holding the earliest branching time of all the sisters. The bug occurred precisely here: the original code specified that the branching time should be the most recent time (closer to the present), when in fact it should be the earliest time. This has now been corrected, as can be seen in Fig. 2. Where this bug did have an effect, it caused the branching times to be a bit shorter toward the tips rather than the root of the tree.

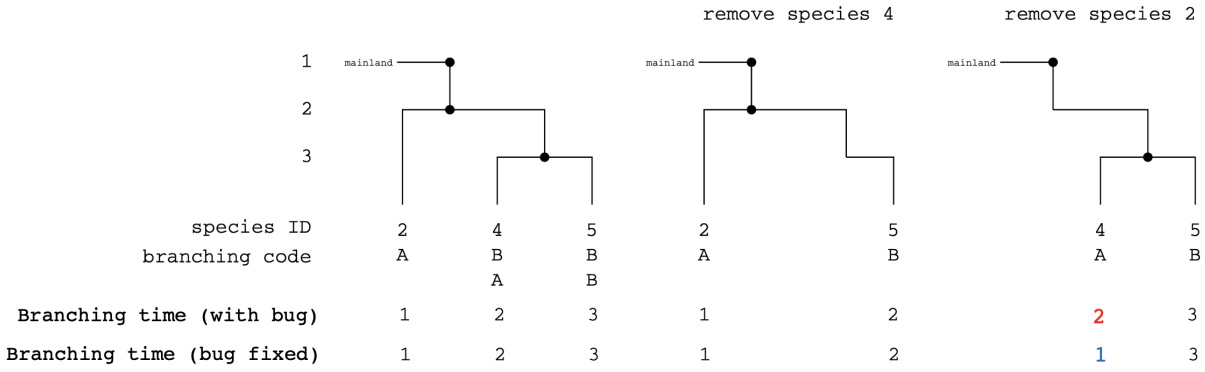

Figure 2 – Example of a scenario where bug 1 occurred, following the extinction of cladogenetic species carrying code "A". The tree on the left shows the original lineage, with three cladogenetic species. The tree in the middle shows the example of extinction of species 4. The resulting branching times were correct in this case. The tree on the right shows the extinction of species 2. In this case, the code with the bug assigned the incorrect branching time to the surviving species 4, giving it a value of 2 (red) when it should have been 1, which was the time held by species 2. This has now been corrected in the package (blue).

Bug 2: Re-colonisation by the same mainland species in simulations

In the DAISIE model, species from a mainland pool can colonise the island at a certain rate. Once it colonises the island, a species initially remains the same as the mainland species and is therefore native but non-endemic to the island. This species can eventually go extinct or speciate forming one (via anagenesis) or two (via cladogenesis) new endemic species. If re-colonisation takes place after the mainland species has speciated on the island, the two lineages descending from separate colonisation events of the same mainland ancestor can coexist on the island. These lineages can themselves undergo speciation (e.g. leading to the formation of multiple independent clades on the island whose ancestor is the same mainland species) and extinction. Bug 2 applies exclusively to these cases where there is re-colonisation of the same mainland ancestor *after* speciation has already taken place on the island. In DAISIE, such cases are marked as "re-colonisations" (stac = 3 in the R code), so that they can be interpreted by the DAISIE likelihood functions as such. In the original simulation code (Valente *et al.* 2015, 2017a), for a given mainland species, the information on the first colonisation event - number of species, colonisation time and branching times – was passed on to the likelihood functions, but the information on subsequent re-colonisations of the same mainland species was not. The likelihood optimisation did receive the information that there was a re-colonisation (through the tag "stac = 3"), which could for example suggest higher colonisation rates, but the characteristics of these re-colonisations did not contribute to the likelihood. Therefore, although there was an acknowledgment that there was a re-colonisation, some useful information (number of species, colonisation time and branching times of the re-colonisations) was not being used for likelihood optimization and was ignored. In a later version of DAISIE (Valente *et al.* 2017b, 2018), the branching times of cladogenetic species resulting from re-colonisations were passed on to the likelihood (in addition to the colonisation and any branching times of the initial colonist), but no information on re-colonisations whose descendant lineages had no branching times (those that have not speciated and those that were endemic singletons) was added. In the most recent version of DAISIE (Valente *et al.* 2019, 2020; Hauffe *et al.* 2020), instead of passing the branching times of re-colonisations to the likelihood functions, the code instead considered any endemic species resulting from re-colonisations as missing unsampled species descending from that mainland species. These missing species were used in likelihood optimization as information on the diversity descending from that mainland species.

The different ways to code re-colonisations in the past versions of DAISIE should not necessarily be considered bugs – the choices were deliberately made and are a reflection of our evolving idea of how to best deal with re-colonisations in the DAISIE likelihood framework. However, we have now arrived at a new solution which we believe is currently the most optimal. In the new "corrected" version of the code

presented here, the following is done if a mainland species re-colonises and two or more re-colonisations descending from that species are extant on the island at the end of the simulation. First, as before, that species is tagged as a "re-colonisation" (stac = 3). The time of colonisation and any branching times resulting from the initial colonisation are always passed on to the likelihood function (as they were before). Then, for a single or multiple re-colonisations (of any kind – non-endemic or endemic with a single or multiple species) any colonisation or branching times are passed to the likelihood function, *except* the colonisation time of the last re-colonisation. The latter is ignored so as not to inflate the number of species – the "re-colonisation" tag already indicates to the likelihood function that there was at least one re-colonisation. This also ensures that any diversity-dependence is acting at the right branching times. The information on all branching and colonisations events for the re-colonisations are stored in the simulation file (even the final re-colonisation time that is not used in the likelihood computation).

Re-analysis of previous publications using DAISIE

We corrected the two bugs in the code as described above and updated the simulation scripts of the DAISIE package accordingly. The first bug was corrected in version 2.0, and the second in version 3.2.0, the latter being the recommended version to use (https://github.com/rsetienne/DAISIE/releases/tag/v3.2.0). The two errors in the DAISIE code do not have any effect in the previous analyses on empirical data (e.g. the data extracted from phylogenies of Galápagos birds). They both affect only the simulations, which in previous publications have been used to a) calculate model goodness-of-fit metrics, b) species-through-time plots and c) conduct bootstrap analyses to estimate the precision of DAISIE parameters estimates. Points a) and b) are *not* affected by the bugs because these analyses use the number of species (endemic and non-endemic) from the simulations, and the number of species is not affected by the bugs, as they only influence branching times (bug 1) and the way the likelihood function interprets the simulated data (bug 2). Therefore, here we assess only the error that could have occurred in c), the bootstrap analyses, which were carried out in Valente *et al.* 2015, 2017b, 2017a, 2018, 2020; Hauffe *et al.* 2020.

We re-ran analyses which could have been affected by the bugs, for selected publications. We first compiled the datasets simulated with DAISIE used in the main analyses of all previous published studies. The frequency of bug 1 cannot be directly identified from the published data, but it is assumed to be higher when rates of cladogenesis and extinction are both high. We identified instances where bug 2 was present: cases where there are re-colonisations of the same mainland species present on the island at the end of a simulation. The datasets with the highest proportion of islands with at least one re-colonisation were the Lake Biwa fishes (Hauffe *et al.* 2020),

Canary Island birds (part of the Macaronesian bird dataset from Valente *et al.* 2017b) and Greater Antillean bats (Valente *et al.* 2017a). This was to be expected, as these insular systems have high colonisation rates, and thus are likely to have more re-colonisations. Note that the Lake Biwa fish dataset does not have bug 1, because this bug had already been corrected when the analyses for that study were performed. The datasets for Canary Islands birds and Greater Antillean bats are good examples for bug 1, because they both have cladogenesis and were simulated with high rates of extinction (higher than the rates of cladogenesis). We therefore chose these datasets to re-run the analyses as conservative examples of the most affected datasets.

We also re-ran analyses of the global bird dataset used in Valente *et al.* 2020 (note this only has potential to affect the results presented in Extended Data Fig 2. and Extended Data Table 5 of that study). While this dataset has a very low proportion of islands with bug 2, it has a mixture of islands with low and high extinction and cladogenesis rates, and therefore we decided to also re-assess it with the corrected code.

Table 1 – Previous simulated datasets using DAISIE and frequency of datasets with at least one re-colonisation, which has relevance to bug 2.

Publication	**Dataset**	**Number of simulated datasets**	**Number with at least one re-colonisation**	**%**	**Mean proportion of re-colonisations per simulated dataset**
Valente *et al.* 2015	Galápagos birds	5000	83	1.66	0.002
Valente *et al.* 2017b	Azores birds	5000	543	10.86	0.008
	Canary Island birds	5000	3676	73.52	0.032
	Cape Verde birds	5000	545	10.9	0.0081
	Madeira birds	5000	514	10.28	0.0076
Valente *et al.* 2017a	Greater Antilles bats	5000	3032	60.64	0.064
Valente *et al.* 2019	New Zealand birds	5000	852	17.04	0.005
Valente *et al.* 2018	Galápagos birds (4 Myr)	3000	63	2.1	0.003
	Galápagos birds (10 Myr)	3000	146	5.1	0.005
Valente *et al.* 2020	Global bird dataset	41000	1271	3.1	0.002
Hauffe *et al.* 2020	Lake Biwa fishes	1000	871	87.1	0.034

For birds of Macaronesia (Valente *et al.* 2017b), bats of Greater Antilles (Valente *et al.* 2017a), global island birds (Valente *et al.* 2020) and fishes of Lake Biwa (Hauffe *et al.* 2020) we ran analyses on the following datasets:
A – the same dataset used in the publication, which was simulated with the bugs.
B – same as A (the same script with the bug), but with a different random set of simulations.
C – dataset simulated using the same script but with the bugs fixed.
The motivation for running analyses on the datasets A-C was to examine how the effect of the bug (A vs C, or B vs C) compares with the effect of using different sets of simulated data (A vs B).
We fitted DAISIE to each of the simulated datasets to estimate via maximum likelihood the same parameters that were estimated in the original published analyses. Details of each parameter are given in the respective figures 3-6. We then compared the distribution of ML parameter estimates for datasets A, B and C. We have deposited all datasets and simulation scripts used in this study in Mendeley data (http://dx.doi.org/10.17632/x984h4c8c3.1).

We have not revisited the analyses presented in the simulation study of Valente *et al.* 2014, but below we examine the effect of bug 1 (they are unaffected by bug 2) on the results. The main focus of this work was on the ontogeny of species richness (Fig. 2 of that study) and these results were not impacted. This study also examined the age distributions of species on islands of different ages, revealing that species age generally increased with island age, but would then decline again on the oldest islands under high extinction rate (Fig. 3 of that study). The bug may have led the distribution of species to be shifted towards ages that were slightly younger than they actually were, though immigrant species would not have been affected. A second aim of this study was to examine the ability to detect a signature of diversity-dependent diversification. Figure 4 of the study shows the results of fitting diversification models to phylogenies simulated under diversity-dependent and diversity-independent rates of colonisation and cladogenesis, with the main finding being that diversity-dependence is more often detected in islands simulated under diversity-dependence. As bug 1 occurs when there is extinction (which is present in both simulation types), it will have equally affected both diversity-dependent and diversity-independent simulations and it would therefore not have biased these results. In sum, we anticipate that qualitatively the results of the study are robust to both bugs.

Results

The results of the bootstrap analyses are presented in Figures 3-6. They show that the distribution of parameter estimates is very similar with and without the reported bugs (dataset A vs C). In addition, the differences between A and C are similar to the differences that are introduced when one uses a different random set of simulations (that is, simulated with the same code – e.g. A versus B).

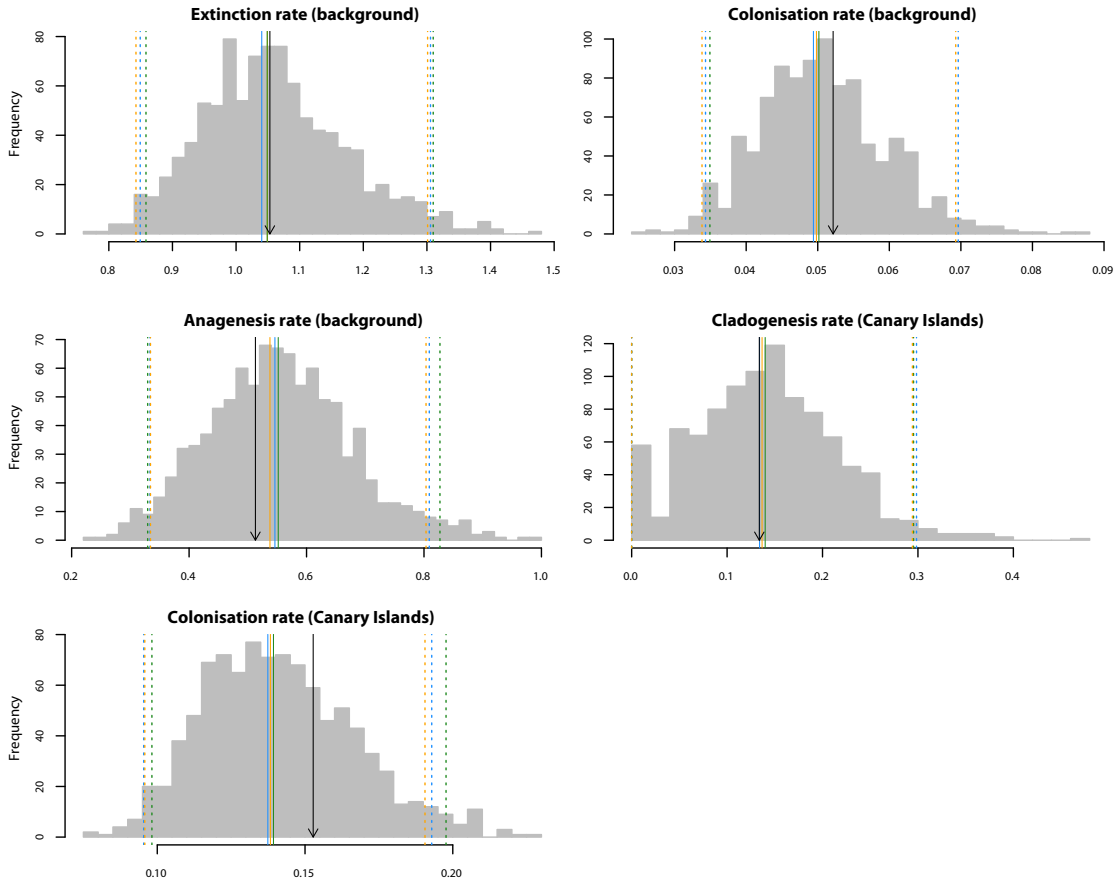

Figure 3 – Macaronesian birds. Bootstrap precision estimates of the parameters of the preferred model in Valente *et al.* 2017b, the M17 model. The plots are frequency histograms of parameters estimated in DAISIE based on dataset C (the same as in the published article, but with the bugs fixed). Solid lines show the median estimated value of the parameter across 1,000 simulated datasets. Dashed lines show 2.5 – 97.5 percentiles. Line colours: blue – dataset A (same as in the published article with the bugs); orange - dataset B (simulated with the same code as A but a different random set of simulations); green – dataset C (bugs fixed). Black arrow: simulated value. Units are the number of events per lineage per million years.

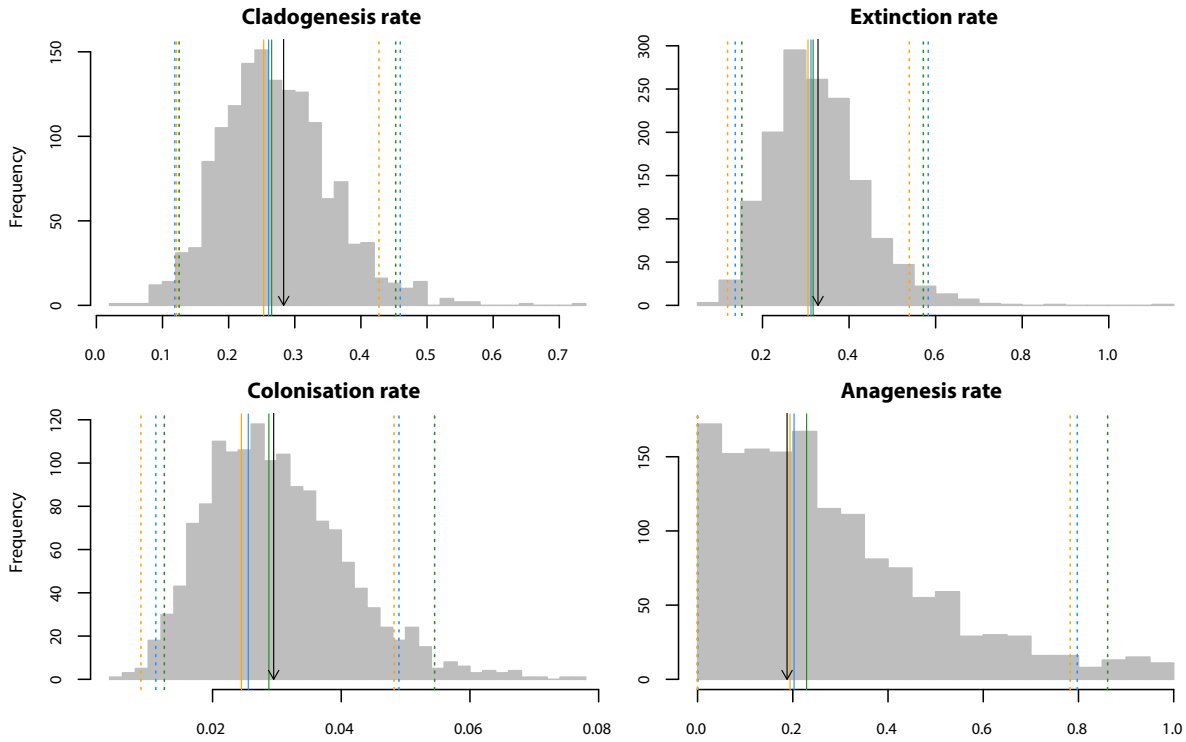

Figure 4 – Noctilionoid bats of the Greater Antilles. Bootstrap precision estimates of the parameters of the preferred model in Valente *et al.* 2017a, the DI model. The plots are frequency histograms of parameters estimated in DAISIE based on dataset C (the same as in the published article, but with the bugs fixed). Solid lines show the median estimated value of the parameter across 1,000 simulated datasets. Dashed lines show 2.5 – 97.5 percentiles. Line colours: blue – dataset A (same as in the published article with the bugs); orange - dataset B (simulated with the same code as A but a different random set of simulations); green – dataset C (bugs fixed). Black arrow: simulated value. Units are the number of events per lineage per million years.

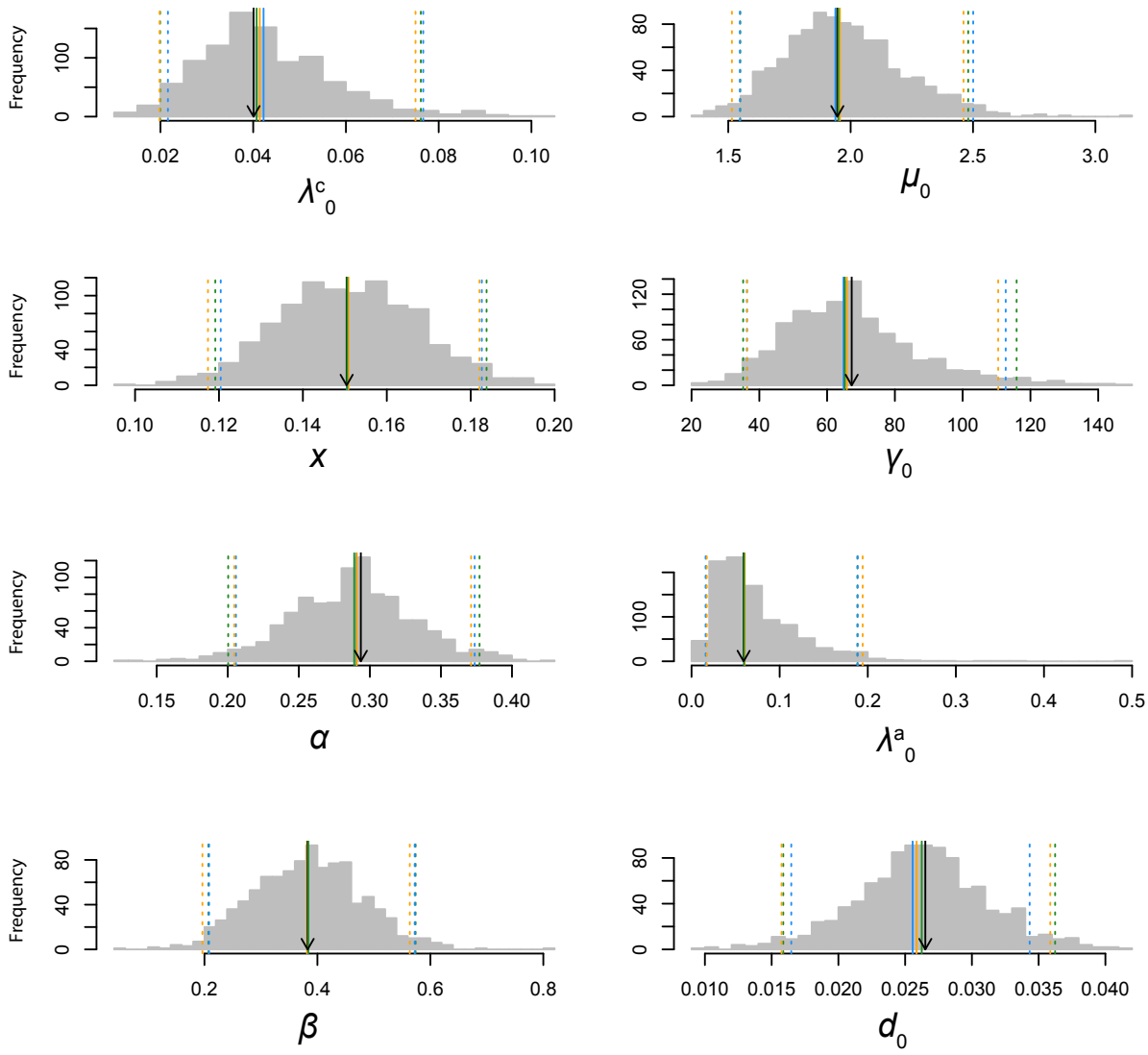

Figure 5 – Global birds. Bootstrap precision estimates of the parameters of the preferred model in Valente *et al.* 2020, the M19 model. The plots are frequency histograms of hyper-parameters estimated in DAISIE based on dataset C (the same as in the published article, but with the bugs fixed). Solid lines show the median estimated value of the parameter across 1,000 simulated datasets. Dashed lines show 2.5 – 97.5 percentiles. Line colours: blue – dataset A (same as in the published article with the bugs); orange - dataset B (simulated with the same code as A but a different random set of simulations); green – dataset C (bugs fixed). Black arrow: simulated value. Parameters are explained in Supplementary Table 1 of Valente *et al.* 2020.

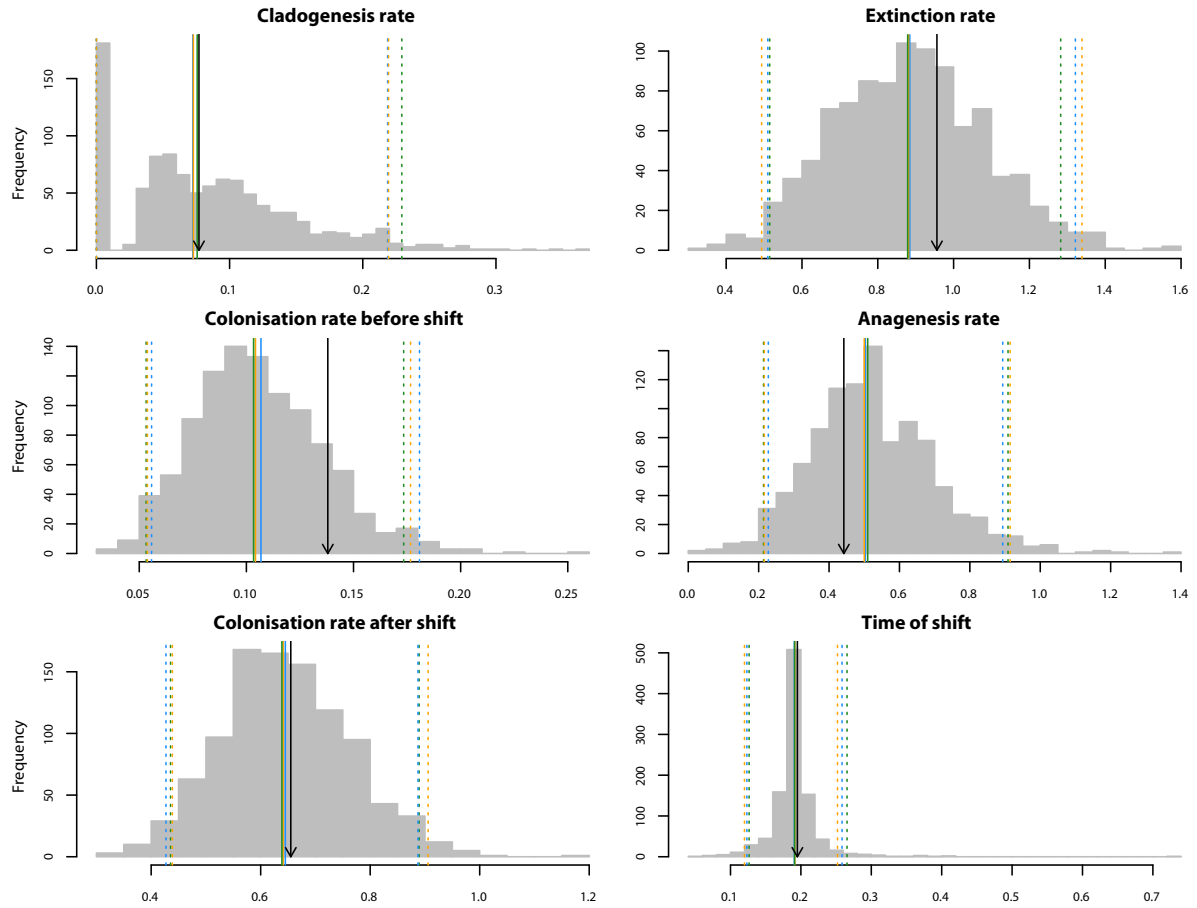

Figure 6 – Lake Biwa fishes. Bootstrap precision estimates of the parameters of the preferred model in Hauffe *et al.* 2020, the M23 model. The plots are frequency histograms of parameters estimated in DAISIE based on dataset C (the same as in the published article, but with the bugs fixed). Solid lines show the median estimated value of the parameter across 1,000 simulated datasets. Dashed lines show 2.5 – 97.5 percentiles. Line colours: blue – dataset A (same as in the published article with the bugs); orange - dataset B (simulated with the same code as A but a different random set of simulations); green – dataset C (bugs fixed). Black arrow: simulated value. Units are the number of events per lineage per million years, except for time of shift, which is time in million years from present.

Discussion

The differences in parameter estimate distribution between the various datasets lead us to conclude that the results of previous studies that have used DAISIE are reliable both qualitatively, but also quantitatively, as the recorded differences are very minor. Indeed, the differences between A and C are similar to the differences that one would find using a set of simulated data obtained with a different random seed. As DAISIE is

a stochastic model, the simulated datasets can vary for the same parameter set and therefore the minor differences between A and B are an expected feature of these types of models. While we did not re-run analyses for all papers, the selected analyses were those most likely to be affected by both errors. We can thus conservatively state that the effect on the bootstrap analyses of the other studies will have been even more minor. Finally, as we mentioned before, the results of the study on island ontogeny (Valente *et al.* 2014) are also robust to both bugs.

Our finding that our results are even quantitatively hardly affected by the first bug (which affects branching times of phylogenies) seems to stand in contrast to the argument we have made that phylogenies provide more information than only species richness in island biogeography (Valente *et al.* 2018). However, that argument was particularly made for island colonisation times extracted from phylogenies, which are not affected by bug 1. Furthermore, we argued that branching times do add to higher accuracy of the parameter estimates, and we believe they still do because bug 1 affects only a small proportion of branching events.

DAISIE release

https://github.com/rsetienne/DAISIE/releases/tag/v3.2.0

Acknowledgements

We thank Pedro Neves Santos, Shu Xie and Joshua Lambert for help with implementation of the bug fixes in the R package.

References

Hauffe, T., Delicado, D., Etienne, R. & Valente, L. (2020). Lake expansion elevates equilibrium diversity via increasing colonisation. *J. Biogeogr.*, 47, 1849–1860.

Purvis, A. & Rambaut, A. (1995). Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Bioinformatics*, 11, 247–251.

Valente, L., Etienne, R.S. & Dávalos, L.M. (2017a). Recent extinctions disturb path to equilibrium diversity in Caribbean bats. *Nat. Ecol. Evol.*, 1, 26.

Valente, L., Etienne, R.S. & Garcia-R, J.C. (2019). Deep Macroevolutionary Impact of Humans on New Zealand's Unique Avifauna. *Curr. Biol.*, 29, 2563–2569.

Valente, L., Illera, J.C., Havenstein, K., Pallien, T., Etienne, R.S. & Tiedemann, R. (2017b). Equilibrium bird species diversity in Atlantic islands. *Curr. Biol.*, 27, 1660–1666.

Valente, L., Phillimore, A. & Etienne, R.S. (2018). Using molecular phylogenies in island biogeography: it's about time. *Ecography*, 41, 1684–1686.

Valente, L., Phillimore, A.B., Melo, M., Warren, B.H., Clegg, S.M., Havenstein, K., *et al.* (2020). A simple dynamic model explains the diversity of island birds

worldwide. *Nature*, 579, 92–96.
Valente, L.M., Etienne, R.S. & Phillimore, A.B. (2014). The effects of island ontogeny on species diversity and phylogeny. *Proc. Biol. Sci.*, 281, 20133227.
Valente, L.M., Phillimore, A.B. & Etienne, R.S. (2015). Equilibrium and non-equilibrium dynamics simultaneously operate in the Galápagos islands. *Ecol. Lett.*, 18, 844–852.