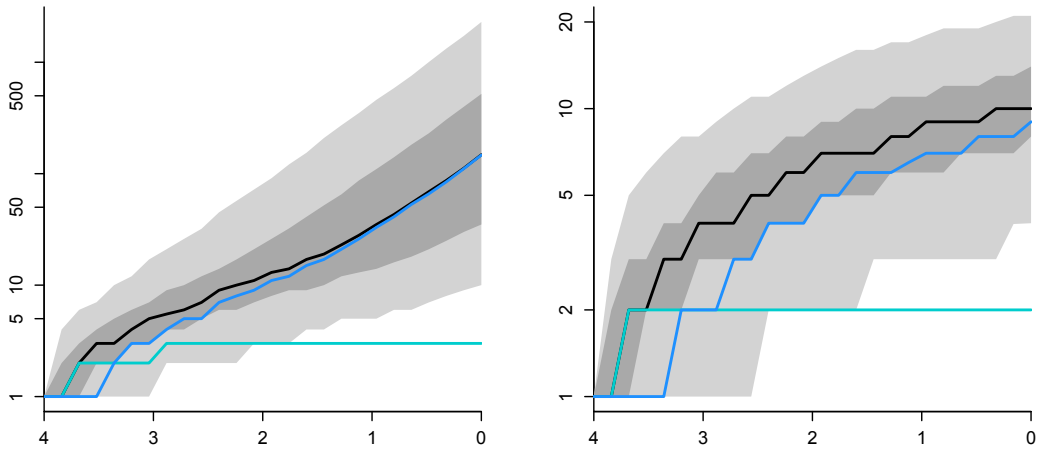
DAISIE – Dynamic Assembly of Island biotas through Speciation, Immigration and Extinction

Authors: Luis M. Valente, Albert B. Phillimore, Rampal S. Etienne

R package



Tutorial
19-May-2015

Citation: Valente LM, Phillimore AB, Etienne RS (2015) Equilibrium and non-equilibrium dynamics simultaneously operate in the Galápagos islands. *Ecology Letters*, In press.

1. Introduction

DAISIE is an island biogeography model that assumes island biota assembly is governed by immigration, extinction and speciation through cladogenesis and anagenesis. This likelihood-based statistical package can simulate islands and estimate parameters of the DAISIE model based on phylogenetic and phylogeographic data. Cladogenesis and immigration rates can be dependent on diversity.

Based on a user-specified table with colonisation and branching times of a given focus group of organisms on an island, DAISIE can estimate island-wide rates of colonisation, extinction, cladogenesis, anagenesis as well as test for the presence of diversity-limits. The package can also simulate species immigration, speciation and extinction on islands and produce species-through-time plots showing the pattern of species accumulation on the island from birth of the island until the present.

A key difference between DAISIE and most other phylogenetic methods to study diversification is that DAISIE estimates rates for a given geographical unit (island, archipelago, lake, etc) rather than for a particular lineage. As an example, DAISIE can reliably estimate overall rates of colonisation and speciation for ferns in Hawaii if the times of colonisation and speciation of Hawaiian ferns are known. However, if the user wants to estimate rates of diversification for a particular lineage of ferns within Hawaii, or wants to compare rates of diversification of ferns within Hawaii with rates in the mainland, then other phylogenetic packages are more appropriate (e.g. DDD, Laser, diversitree, Geiger).

2. DAISIE data requirements

For each focal insular system (island, archipelago, lake, mountain,…) DAISIE requires the times of colonisation of the independent colonisation events of the focal group of organisms (e.g. all snail colonisation events of an island), as well as the branching (speciation) times within the island in case the colonist lineages have radiated within the island. Each colonisation event is treated independently, therefore DAISIE does not require a phylogeny for the entire group linking the different colonisation events.

These are the data required before running DAISIE:

- **Colonisation times** – these are obtained from the stem ages of each island lineage in dated molecular phylogenies, therefore requiring good knowledge of the closest relative of each lineage outside of the island. The time of colonisation is assumed to be the stem age of the island lineage with respect to its sister non-island lineage. Thus, if the island lineage consists of an endemic species or a radiation, DAISIE requires the user to have a divergence-dated phylogeny that samples the island lineage plus the closest related non-island taxa. If the island lineage is a non-endemic species, DAISIE requires the user to have a divergence-dated phylogeny sampling the island population plus the closest related non-island population(s). In case the island population of the non-endemic

species has not been sampled, but the stem age of the species is known, that age can be imputed as an upper bound for the colonisation event.

- **Branching (speciation) times** – For island radiations, the branching times within the island should be inputted. If taxon sampling within the radiation is incomplete and some branching times are missing, this can be accommodated by specifying the number of missing species.
- **Species status** – Whether species are endemic or non-endemic to the focal geographical unit.
- **Maximum age of the island** - An age for the geographical unit needs to be specified. In case of an oceanic island, this should be the age of emergence of the island. In case of an archipelago, this would typically be the age of the oldest currently emerged island or the maximum age of the archipelago considering islands that are now submerged.
- **Mainland pool size estimate** – An estimate of the number of species of the focus group in the source mainland pool is required. This estimate will affect mostly the rate of colonisation, but we recommend the user to try different values.

3. Tutorial: Galápagos land birds

In this tutorial we will estimate rates of colonisation, cladogenesis, anagenesis and extinction for land birds of the Galápagos archipelago using maximum likelihood (ML) in the DAISIE package. We will then use the package to simulate islands with the estimated parameters in order to visualize the pattern of species accumulation in the archipelago.

We will use the dataset of Valente, Phillimore and Etienne (2015, *Ecology Letters*, In press). The dataset includes 25 species corresponding to eight independent colonisations of the archipelago. Three species are non-endemic and 22 are endemic. Six colonisation events have resulted in a single extant species, whereas two others have resulted in radiations (Darwins' finches and Galápagos mockingbirds).

The dataset was compiled by extracting colonisation and branching times from divergence-dated phylogenies trees that include the Galápagos taxa and their closest non-Galápagos relatives.

3.1 Installing and loading the package

DAISIE can be installed by running:

```
install.packages('DAISIE', dependencies=TRUE)
```

To load the package:

```
library(DAISIE)
```

3.2 Data table

The raw dataset is inputted as a table. The Galápagos dataset table can be visualized with:

```
data(Galapagos_datatable)
Galapagos_datatable
```

Each row in the table represents an independent colonisation event. The table has four columns:

- Clade_name - name of the independent colonization event.
- Status - One of the following categories:
  - Endemic: applicable for both anagenetic species and radiations.
  - Non_endemic: If the taxon is not endemic to the island, and the age of colonisation is based on a phylogeny where both island and non-island populations of the species have been sampled.
  - Non_endemic_MaxAge: If the taxon is not endemic to the island, and only an upper bound to the time of colonisation of the island is known. This applies if individuals from the island population of the species have not been sampled, but an age of the species is known.
  - Endemic&Non_Endemic: When an endemic clade is present and the mainland ancestor has re-colonized. For remote islands this is expected to be very rare.
- Missing_species - Number of island species that were not sampled for a particular clade (only applicable for radiations).
- Branching_times – This should be the stem age of the population/species in the case of Non-endemic, Non-endemic_MaxAge and Endemic anagenetic species. For cladogenetic species these should be branching times of the radiation including the stem age of the radiation. **Note** – if there are species within the radiation that are not found on the island (e.g. back-colonisation) the branching times of these species should be excluded, as the mainland species pool is treated as static.

3.3 Formatting table to run in DAISIE

Before running analyses, the datatable needs to be converted to a DAISIE datalist format using the function **DAISIE_dataprep.**

We will prepare two different datalists based on the Galápagos datatable. In the 1st datalist we will treat all taxa as equivalent. We will specify an island age of four million years (*island_age=4*) and a mainland pool size of 1000 (*M=1000*).

```
data(Galapagos_datatable)

Galapagos_datalist <- DAISIE_dataprep(
datatable = Galapagos_datatable,
island_age = 4,
```

M = 1000)

In the 2nd datalist we will allow for the Darwin's finches to form a separate group for which rates can be decoupled from those governing the macroevolutionary process in all other clades (*number_clade_types=2* and *list_type2_clades = "Finches").* We will set the proportion of Darwin's finch type species in the mainland pool to be 0.163. (*prop_type2_pool=0.163*). If *prop_type2_pool* is not specified then by default it is given the value of the proportion of the Galapagos lineages that Darwin's finches represent (1/8=0.125 in this case).

```
data(Galapagos_datatable)

Galapagos_datalist_2types <- DAISIE_dataprep(
datatable = Galapagos_datatable,
island_age = 4,
M = 1000,
number_clade_types = 2,
list_type2_clades = "Finches",
prop_type2_pool = 0.163)
```

The objects Galapagos_datalist and Galapagos_datalist_2types can now be run directly in DAISIE functions.

3.4 Optimizing parameters using maximum likelihood

The function that conducts maximum likelihood optimization of DAISIE model parameters is called **DAISIE_ML**.

Different models can be specified using *ddmodel* option in DAISIE_ML:
ddmodel = 0 : no diversity-dependence
ddmodel = 1 : linear diversity-dependence in speciation rate
ddmodel = 11: linear diversity-dependence in speciation and immigration rate
ddmodel = 2 : exponential diversity-dependence in speciation rate
ddmodel = 21: exponential diversity-dependence in speciation and immigration rate

Different types of parameters can be optimized or fixed. The parameters are given in the following order: (1) cladogenesis rate, (2) extinction rate, (3) K' or carrying capacity (maximum number of species that a clade can attain within the island), (4) colonisation rate, and (5) anagenesis rate.

The identities of the parameters to be optimized or fixed are specified with *idparsopt* and *idparsfix* within the DAISIE_ML function. For example, to optimize all parameters we set *idparsopt=1:5* and *idparsfix=NULL.* To optimize all parameters but fix the rate of extinction, we set *idparsopt=c(1,3,4,5)* and *idparsfix=2*. To optimize all parameters except cladogenesis and anagenesis we set *idparsopt=c(2,3,4)* and *idparsfix=c(1,5)*.

The values of the parameters to be used as initial values for the optimization are specified

with *initparsopt*, and the values to be fixed are specified with *parsfix*. For example, if we want to optimize all parameters with a starting value of 2 we set *initparsopt=c(2,2,2,2,2)* and *parsfix=NULL*. If we want all starting rates to be 0.1, but K' to be fixed at 20, we use *initparsopt=c(0.1,0.1,0.1,0.1)* and *parsfix=20*.

When running your own data, we strongly recommend that you test multiple initial starting parameters for each model, particularly when optimizing models with multiple free parameters, as there is a high risk of being trapped in local likelihood sub-optima. We also suggest running two rounds of optimization using the optimized parameter set of the 1st round as the initial starting values for the 2nd round. Also note that the initial starting values in the examples of this tutorial may not be appropriate for your data.

Example 1 - Optimizing all parameters, with diversity-dependence in speciation and colonisation

We will now optimize all five parameters for a datalist where all clades share the same parameters. We will set the model with linear diversity-dependence in speciation rate and in immigration rate using *ddmodel=11*. We will set an initial rate of cladogenesis of 2.5, an initial rate of extinction of 2.7, an initial K' value of 20, an initial rate of colonisation of 0.009 and an initial rate of anagenesis of 1.01 (*initparsopt = c(2.5,2.7,20,0.009,1.01)).* We will optimize all 5 parameters (*idparsopt = 1:5*) and we will fix no parameters (*parsfix = NULL*, *idparsfix = NULL*).

```
data(Galapagos_datalist)

DAISIE_ML(
datalist = Galapagos_datalist,
initparsopt = c(2.5,2.7,20,0.009,1.01),
ddmodel = 11,
idparsopt = 1:5,
parsfix = NULL,
idparsfix = NULL)
```

This will take several minutes to run. The parameters optimized and fixed as well as the loglikelihood of the initial starting parameters we have set are shown at the top of the screen output of DAISIE_ML. Once the optimization is completed, the program will output the maximum likelihood parameter estimates and the maximum loglikelihood value. For a given dataset, the likelihood of different DAISIE models can be compared with information criteria such as BIC and AIC.

Example 2 - Optimizing model without diversity-dependence

To optimize the parameters of a model with no diversity-dependence, we use the default model (*ddmodel=0*), and fix the parameter number 3 which corresponds to K' to infinity (Inf).

```
data(Galapagos_datalist)

DAISIE_ML(
datalist = Galapagos_datalist,
initparsopt = c(2.5,2.7,0.009,1.01),
idparsopt = c(1,2,4,5),
parsfix = Inf,
idparsfix = 3)
```

Example 3 - Optimizing model with no diversity-dependence and no anagenesis

To optimize the parameters of a model with no diversity-dependence and no anagenesis, we use the default model (*ddmodel=0*), and fix parameters number 3 and 5, which correspond, respectively to K' and rate of anagenesis.

```
data(Galapagos_datalist)

DAISIE_ML(
datalist=Galapagos_datalist,
initparsopt = c(2.5,2.7,0.009),
idparsopt = c(1,2,4),
parsfix = c(Inf,0),
idparsfix = c(3,5))
```

Example 4 - Optimizing all parameters, but allowing Darwin's finches to have a separate rate of cladogenesis.

For this example we will use the datalist with Darwin's finches specified to be of a separate type - *Galapagos_datalist_2types*.

If two types of species are considered, then the parameters of the second type of species are in the same order as the first set of parameters, but start at number 6 - (6) cladogenesis rate of type 2 species, (7) extinction rate of type 2 species, (8) K' of type 2 species, (9) colonisation rate of type 2 species, and (10) anagenesis rate of type 2 species. There is also an additional parameter when 2 types of species are considered - the proportion of species of type 2 in the mainland pool. This is parameter number 11.

Here we will optimize all parameters, but allow the finches to have a separate rate of cladogenesis. We will fix the proportion of type 2 species in the mainland pool at 0.163 (therefore fixing parameter 11 with *idparsfix=11* and *parsfix=0.163*). Note that because we are only allowing the rate of cladogenesis of Darwin's finches to vary from the background rate, we need to specify that the other rates for Darwin's finches remain the same as the background – using *idparsnoshift = c(7,8,9,10))*.

```
data(Galapagos_datalist_2types)

DAISIE_ML(
ddmodel=11,
datalist=Galapagos_datalist_2types,
initparsopt= c(0.38,0.55,20,0.004,1.1,2.28),
idparsopt = c(1,2,3,4,5,6),
parsfix = 0.163,
idparsfix = c(11),
idparsnoshift = c(7,8,9,10))
```

Example 5 - Optimizing a model with no diversity-dependence, but allowing Darwin's finches to have a separate rate of cladogenesis and extinction.

```
data(Galapagos_datalist_2types)

DAISIE_ML(
ddmodel=0,
datalist=Galapagos_datalist_2types,
initparsopt = c(0.38,0.55,0.004,1.1,2.28,2),
idparsopt = c(1,2,4,5,6,7),
parsfix = c(Inf,0.163),
idparsfix = c(3,11),
idparsnoshift = c(8,9,10))
```

3.5 Simulating islands

The function **DAISIE_sim** allows simulation of DAISIE models and plots the results. The user specifies the parameters to be simulated, the number of replicates, the length of the simulation (typically the island age), and the number of species in the mainland pool.

When the *plot_sims* option is set to the default (TRUE) the function will produce a species-through-time plot showing the accumulation of total, endemic and non-endemic species through time, as well as confidence intervals for the total number of species.

Example 5.1 – Simulating 100 islands with no diversity-dependence, all species sharing the same parameters, and plotting the results

```
pars = c(2.550687345,2.683454548,Inf,0.00933207,1.010073119)

island_replicates = DAISIE_sim(
  time = 4,
  M = 1000,
  pars = pars,
  replicates = 100)
```

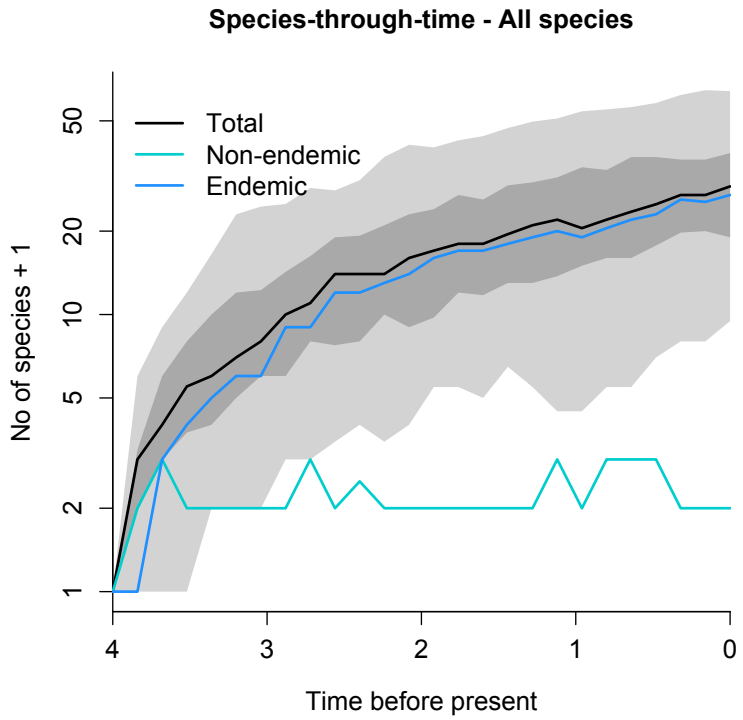


Fig. 1 – Species-through-time plot for model with homogeneity of rates across all lineages, and no diversity-dependence.

The simulation will produce a plot similar to the one on Fig. 1. The object *island_replicates* contains the results of the simulation in DAISIE format. 100 islands are stored in the object, and each island replicate can be viewed separately. For example type *island_replicates[[23]]* to view replicate number 23. The element of the list relating to each island contains a table with the number of species through time, as well as branching time information for each independent colonisation event extant at the end of the simulation.

Example 5.2 – Simulating 100 islands with diversity-dependence (K'=10), all species sharing the same parameters, and plotting the results

```
pars = c(2.550687345,2.683454548,10,0.00933207,1.010073119)

island_replicates_K = DAISIE_sim(
  time = 4,
  M = 1000,
  pars = pars,
  replicates = 100)
```

Example 5.3 – Simulating 100 islands allowing Darwin's finches to have a higher rate of cladogenesis:

```
pars_type1 = c(0.38,0.55,Inf,0.004,1.10)

pars_type2 = c(2.28,0.55,Inf,0.004,1.10)

island_replicates_2types = DAISIE_sim(
  time = 4,
  M = 1000,
  pars = c(pars_type1,pars_type2),
  replicates = 100,
  prop_type2_pool = 0.163)
```

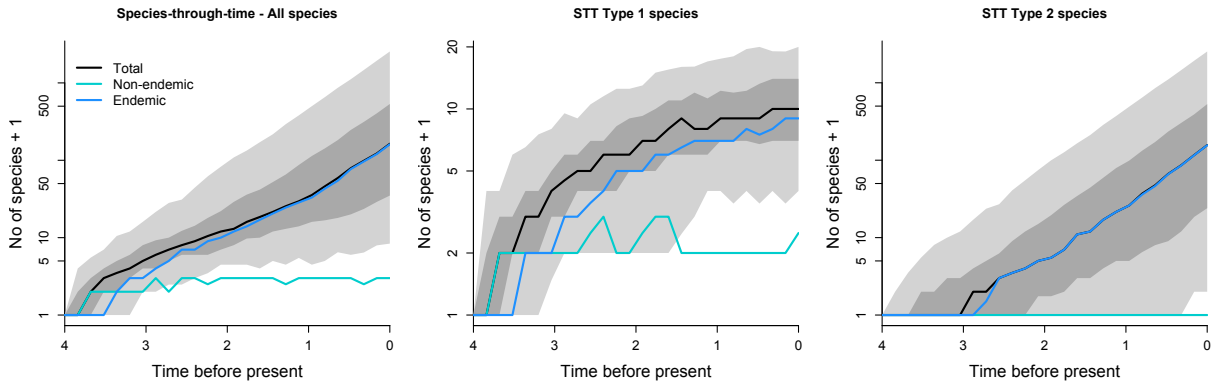


Fig. 2 – Species-through-time plot for model with different rate of cladogenesis for Darwin's finches.

This produces a figure similar to Fig. 2, with three plots: one for the total number of species, one for species of type 1 and one for species of type 2. Accessing each island replicate individually (e.g. *island_replicates_2types[[15]]*) shows information on branching times and species-through-time tables for total, type 1 species and type 2 species.