Statistical comparison of DEC and DEC+J is identical to comparison of two ClaSSE submodels, and is therefore valid

Running title: DEC/DEC+J comparison is valid

Nicholas J. Matzke[1]

[1]Senior Lecturer, School of Biological Sciences, University of Auckland.

n.matzke@auckland.ac.nz

Acknowledgements

The initial development of *BioGeoBEARS* occurred while N.J.M. was supported by NSF Award Number 0919124, and by the National Institute for Mathematical and Biological Synthesis (NIMBioS), an Institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Awards #EFJ0832858 and DBI-1300426, with additional support from The University of Tennessee, Knoxville. Work on this paper was supported by Discovery Early Career Researcher Award DE150101773, funded by the Australian Research Council, and by The Australian National University. Recent work was completed under Marsden Grants 16-UOA-277 and 18-UOA-034 and U. Auckland FRDF Project #3722433.

Abstract

Aim. Statistical model comparison has become common in historical biogeography, enabled by the R package *BioGeoBEARS*, which implements several models in a common framework, allowing models to be compared with standard likelihood-based methods of statistical model comparison. Ree and Sanmartín (2018) critiqued the comparison of Dispersal-Extinction-Cladogenesis (DEC) and a modification of it, DEC+J, which adds the process of jump dispersal at speciation. DEC+J provides highly significant improvements in model fit on most (although not all) datasets. They claim that the comparison is statistically invalid for a variety of reasons. I analyze the key claims made by the critique.
Location. Simulated data.
Taxon. Simulated data.
Methods. Likelihood calculations are checked by comparison between programs and by-hand calculations, and by summing likelihoods across all possible

datasets. Model adequacy of DEC vs. DEC+J is checked by a simulation/inference experiment.

Results. Mistakes in the critique's example likelihood calculations are demonstrated. DEC+J fits better on datasets because the DEC model is statistically inadequate in the common situation when most species have geographic ranges of single areas; the DEC model requires long residence times of multi-area ranges, and when these are not observed, a model that does produce such data patterns, such as DEC+J, prevails. More fundamentally, statistical comparison of DEC and DEC+J produces identical log-likelihood differences to statistical comparison of two submodels of ClaSSE where extinction rates are fixed to 0.

Main conclusions. DEC fails a basic model adequacy check for understandable reasons, while DEC+J does not. As Ree and Sanmartín recommend ClaSSE models as valid for comparison, the comparison of DEC and DEC+J is statistically valid according to their own criteria.

Keywords

BioGeoBEARS, dispersal, Dispersal-Extinction-Cladogenesis, extinction, Lagrange, jump speciation, statistical model comparison

Significance Statement

This paper shows that a commonly-used comparison of biogeographical models is statistically valid, using validation calculations and simulations.

Introduction

Probabilistic models of geographic range evolution have become key tools in historical biogeography. Hundreds of published analyses have used the Dispersal-Extinction-Cladogenesis (DEC) model from the program *Lagrange* (Ree, R. H. & Smith, 2008). Matzke (2013) proposed that, rather than relying on the assumptions of a single model, multiple models representing different assumptions should have their fit to the data compared with standard methods of statistical model comparison, such as AIC (Burnham & Anderson, 2002). Matzke (2014) proposed and tested DEC+J, which adds a free parameter *j* to DEC to represent the relative weight of jump dispersal events during cladogenesis. DEC+J is a substantially better fit than DEC on many (but not all) datasets, and simulation results indicated that the better fit model will result in more accurate inference of parameters and ancestral ranges. Statistical comparisons of DEC, DEC+J, and other models have become common in the literature. However, a Critique of this procedure by Ree and Sanmartín (2018; henceforth R&S) alleges that these statistical results should be ignored, reporting unintuitive inferences in two cases, and claiming that DEC+J is a "degenerate" model where dispersal is not a product of a continuous-time process, leading to an "unfair" advantage for DEC+J.

The Critique is now often cited in biogeographical studies, and researchers often opt to run only DEC, or run both models but ignore the likelihood-based comparison of model fit, leaving the choice of model to subjective decision. Unfortunately, the Critique is marred by errors large and small, several of them obvious, and several of them technical but revealing of the fundamental principles involved in discrete models of range evolution, and the use of statistical model comparison in general.

The most obvious error in the Critique has been briefly noted by Matzke & Klaus (2020), namely, its total failure to address (or even mention) the extensive simulation results validating the DEC/DEC+J model comparison, published in Matzke (2014). The Critique instead relies on two examples of unintuitive inference observed in two tiny datasets (2- and 4-taxa), but there is no reason to expect that statistical models with 2 or 3 inferred parameters will exhibit reliable or intuitive inference on such small datasets. Simulation-based testing has long been the industry standard for methods in phylogenetics, and it should be considered surprising that a critique which relies on tiny example datasets, and fails to conduct simulations, or engage with simulations already published, has gained traction.

However, there are a number of less obvious errors in the Critique which deserve detailed exploration, in order to clarify the assumptions behind the likelihood calculations in the DEC and DEC+J models, the principles of likelihood-based statistical model comparison, and the relationship between DEC, DEC+J, and ClaSSE, the model that the authors endorse as valid for statistical model comparison. I summarize these errors below, and then explore each in detail:

(Part 1) The likelihoods in the Critique's two worked examples (with 2-species and 4-species phylogenies) are not replicated by *Lagrange* or *BioGeoBEARS*, which agree with each other but disagree with R&S's claimed likelihoods, suggesting the Critique relies on an incomplete understanding of how the likelihood calculation in these models works.

(Part 2) Although the peculiar behaviour of DEC+J in the 2-area case is confirmed (*j* maximizing near 3, resulting in jumps at every node being the preferred inference), the example is brittle. The odd "maximum jumps" behaviour breaks down with more areas, more taxa, or other data configurations. This shows the dangers of using specific human-constructed data patterns to make general judgments about models that have already been tested by extensive simulation-inference experiments. In addition, using AIC$_c$ to compare the models, as is recommended for such small datasets, actually shows that DEC+J would not be favoured over DEC unless the dataset has 9 or more species. Finally, inferences of *j* approaching 3 are extremely rare in empirical datasets.

(Part 3) There is nothing unfair about the DEC/DEC+J comparison, as both models make use of valid conditional likelihoods, as proven when likelihoods are summed across all possible datasets. Any particular model/parameter combination favours some data patterns and not others, but that is precisely the point of probabilistic models. The interesting question is which data patterns we tend to observe empirically, and if those patterns tend to be better fit by a particular model, that is a scientific result, not an unfair advantage.

(Part 4) I demonstrate the true source of the common advantage of DEC+J over DEC, namely that typical DEC inferences imply long residence times for widespread ranges, which predicts many widespread ranges will be observed at the tips. Such models fail model adequacy tests in cases (common in empirical datasets) where most or all species occupy single areas.

(Part 5) The claim that it is problematic for DEC+J to model dispersal as cladogenetic without including the continuous-time probability of speciation is falsified by adding to the likelihood calculation the probability of the phylogenetic tree under a Yule process; as the Yule tree likelihood is constant between DEC and DEC+J models, it makes no difference for model comparison. The Critique recommends ClaSSE for modelling jump dispersal, but I demonstrate that statistical comparison of DEC and DEC+J is identical to statistical comparison between two special cases of ClaSSE; the resultant likelihood differences are identical. Thus, on the Critique's own premises, statistical comparison of DEC and DEC+J is valid.

I conclude by pointing out the real flaws in DEC and DEC+J, namely that the Yule-process assumption means that both models ignore lineage extinction. I discuss the prospects, and substantial computational challenges, for using ClaSSE-like models to overcome these flaws in the future. Even if ClaSSE becomes computationally feasible for routine use in biogeography, the issues raised in the DEC/DEC+J debate regarding state spaces and residence times will remain crucial considerations that are ignored at modellers' peril. However, any such discussions must be premised on the notion that one of the purposes of science is to test our models by statistically comparing how well they fit data, rather than accepting a model based on intuition, popularity, or authority. The methods of statistical model comparison, which have become routine in dozens of other fields, are valid in biogeography as well, and should become routine.

Part 1. Untangling the Critique's likelihood calculations

The Critique's first quantitative claims involve a worked example: a 2-taxon tree with one species occupying area A, and the second occupying area B. R&S maximize the likelihood of this dataset under DEC and DEC+J. The Critique's depiction of the weights used in DEC and DEC+J is given in their Table 1, replicated here as Table 1a. For DEC, the ML parameter estimates for d and e are 0, and the log-likelihood is -2.890. For DEC+J, the ML occurs with d=0, e=0, and j=3, with a log-likelihood -0.693. The authors note that the difference in log-likelihood provided by the two models is greater than 2 units, approximate the p<0.05 significance cutoff for a LRT, with one degree of freedom for adding one free parameter. They ask, "should the mere observation of 2 sister species in different areas be interpreted as evidence in favour of ancestral jump dispersal over vicariance?"

Failure to replicate reported log-likelihoods with available programs. The problems begin here, as the claimed log-likelihoods are easily tested by running a 2-taxon tree in *Lagrange* (for DEC) or *BioGeoBEARS* (for DEC or DEC+J). I implemented the authors' worked example in Python *Lagrange* (version 20130526), C++ *Lagrange* (Smith 2010), and *BioGeoBEARS* (version 1.1.1, https://github.com/nmatzke/BioGeoBEARS). I assumed that two branches had lengths of 1.0, although the absolute timescale is irrelevant in this example. The Newick string used was: "(sp1:1,sp2:1);". For DEC, all three programs report a maximized log-likelihood of -1.792, with *d* and *e* approximately 0. For DEC+J, the BioGeoBEARS ML inference is *d*=0, *e*=0, with *j*=2.9999, and with a log-likelihood of 0.154. These disagree with what the Critique reports.

In the case of DEC, the reason for the difference is that, as programmed in Lagrange and BioGeoBEARS, DEC does not include state frequencies (SFs) at the root in the likelihood calculation (Matzke 2014; for further explanation, see Supp. Text). With 2 areas (A and B), there are *k*=4 states (null, A, B, and AB), but the null state is an impossible ancestor. Therefore, the log-likelihood under DEC+SFs is $-1.792 + \log\left(\frac{1}{k-1}\right) = -1.792 + \log\left(\frac{1}{3}\right) = -2.890$.

Table 1. The depiction of DEC and DEC+J weights in Table 1 of Ree & Sanmartin (2018), compared to the actual weights used in Lagrange (for DEC) and BioGeoBEARS (for DEC and DEC+J). Later columns show the translation of the weights into conditional probabilities, and the subsequent calculation of the total log-likelihood of the data for a 2-taxon tree with ranges of A on the left tip, and B on the right tip. Finally, the ancestral range probability calculation (valid for the root only) is shown.

1a. After Table 1 from Ree & Sanmartin (2018).

ance-stor	cladogenetic event	left	right	DEC	DEC+J
A	*in situ* speciation	A	A	$\frac{1}{3}$	$\frac{1}{3}\cdot\frac{3-j}{3}$
A	jump dispersal	A	B		$\frac{1}{12}\cdot j$
A	jump dispersal	B	A		$\frac{1}{12}\cdot j$
B	*in situ* speciation	B	B	$\frac{1}{3}$	$\frac{1}{3}\cdot\frac{3-j}{3}$
B	jump dispersal	B	A		$\frac{1}{12}\cdot j$
B	jump dispersal	A	B		$\frac{1}{12}\cdot j$
AB	vicariance	A	B	$\frac{1}{18}$	$\frac{1}{3}\cdot\frac{3-j}{3}$
AB	vicariance	B	A	$\frac{1}{18}$	$\frac{1}{3}\cdot\frac{3-j}{3}$
AB	peripheral isolate in A	A	AB	$\frac{1}{18}$	$\frac{1}{3}\cdot\frac{3-j}{3}$
AB	peripheral isolate in A	AB	A	$\frac{1}{18}$	$\frac{1}{3}\cdot\frac{3-j}{3}$
AB	peripheral isolate in B	B	AB	$\frac{1}{18}$	$\frac{1}{3}\cdot\frac{3-j}{3}$
AB	peripheral isolate in B	AB	B	$\frac{1}{18}$	$\frac{1}{3}\cdot\frac{3-j}{3}$

1b. DEC calculations in BioGeoBEARS, 2-taxon tree with *d*=*e*=0, data=(left:A, right:B).

	weights		sum of weights		$P(\text{event}\mid\text{ancest.}) = \frac{\text{weight}}{\sum\text{weights}}$		likelihood= $P(\text{event}\mid\text{ancestor})\cdot P(\text{data}\mid\text{event})$			Ancestral range probabilities (at root only)	
Matzke (2013) description	symb-olic	num-eric	symb-olic	num-eric	symb-olic	num-eric	P(data| event)	calcul-ation	result	formula	calculation
s**y**mpatry (narrow)	y	1	y	1	$\frac{y}{y}$	1	0	$1\cdot 0$	0	$\frac{P(\text{data}\mid\text{anc. in A})}{P(\text{data}\mid\text{all anc.})}$	$\frac{0}{0.167}=0$
founder/ **j**ump											
founder/ **j**ump											
s**y**mpatry (narrow)	y	1	y	1	$\frac{y}{y}$	1	0	$1\cdot 0$	0	$\frac{P(\text{data}\mid\text{anc. in B})}{P(\text{data}\mid\text{all anc.})}$	$\frac{0}{0.167}=0$
founder/ **j**ump											
founder/ **j**ump											
vicariance (narrow)	v	1	$2v+4s$	6	$\frac{v}{2v+4s}$	$\frac{1}{6}$	1	$\frac{1}{6}\cdot 1$	0.167	$\frac{P(\text{data}\mid\text{anc. AB})}{P(\text{data}\mid\text{all anc.})}$	$\frac{0.167}{0.167}=1$
vicariance (narrow)	v	1			$\frac{v}{2v+4s}$	$\frac{1}{6}$	0	$\frac{1}{6}\cdot 0$	0		
sympatry (**s**ubset)	s	1			$\frac{s}{2v+4s}$	$\frac{1}{6}$	0	$\frac{1}{6}\cdot 0$	0		
sympatry (**s**ubset)	s	1			$\frac{s}{2v+4s}$	$\frac{1}{6}$	0	$\frac{1}{6}\cdot 0$	0		
sympatry (**s**ubset)	s	1			$\frac{s}{2v+4s}$	$\frac{1}{6}$	0	$\frac{1}{6}\cdot 0$	0		
sympatry (**s**ubset)	s	1			$\frac{s}{2v+4s}$	$\frac{1}{6}$	0	$\frac{1}{6}\cdot 0$	0		
							Total likelihood	$=\frac{1}{6}$	0.167		
							Total log-likelihood	$=\log\left(\frac{1}{6}\right)$	-1.792		

1c. DEC+J calculations in BioGeoBEARS, 2-taxon tree with *d*=*e*=0, *j*=0.3, data=(left:A, right:B).

	weights			sum of weights		$P(\text{event}\mid\text{ancest.}) = \frac{\text{weight}}{\sum\text{weights}}$			likelihood= $P(\text{event}\mid\text{ancestor})\cdot P(\text{data}\mid\text{event})$			Ancestral range probabilities (at root only)	
event	symb-olic	num-eric	weight	symb-olic	num-eric	symb-olic	num-eric	P(event| anc.)	P(data| event)	calculation	result	formula	calculation
A->A,A	$y=1-\frac{j}{3}$	$1-\frac{0.3}{3}$	0.9	$y+2j$	1.5	$\frac{y}{y+2j}$	$\frac{0.9}{1.5}$	0.6	0	$0.6\cdot 0$	0	$\frac{P(\text{data}\mid\text{anc. in A})}{P(\text{data}\mid\text{all anc.})}$	$\frac{0.2}{0.5667}$ $=0.353$
A->A,B	j	0.3				$\frac{j}{y+2j}$	$\frac{0.3}{1.5}$	0.2	1	$0.2\cdot 1$	0.2		
A->B,A	j	0.3				$\frac{j}{y+2j}$	$\frac{0.3}{1.5}$	0.2	0	$0.2\cdot 0$	0		
B->B,B	$y=1-\frac{j}{3}$	$1-\frac{0.3}{3}$	0.9	$y+2j$	1.5	$\frac{y}{y+2j}$	$\frac{0.9}{1.5}$	0.6	0	$0.6\cdot 0$	0	$\frac{P(\text{data}\mid\text{anc. in B})}{P(\text{data}\mid\text{all anc.})}$	$\frac{0.2}{0.5667}$ $=0.353$
B->B,A	j	0.3				$\frac{j}{y+2j}$	$\frac{0.3}{1.5}$	0.2	0	$0.2\cdot 0$	0		
B->A,B	j	0.3				$\frac{j}{y+2j}$	$\frac{0.3}{1.5}$	0.2	1	$0.2\cdot 1$	0.2		
AB->A,B	$v=1-\frac{j}{3}$	$1-\frac{0.3}{3}$	0.9	$2v+4s$	5.4	$\frac{v}{2v+4s}$	$\frac{0.9}{5.4}$	$\frac{1}{6}$	1	$\frac{1}{6}\cdot 1$	0.1667	$\frac{P(\text{data}\mid\text{anc. AB})}{P(\text{data}\mid\text{all anc.})}$	$\frac{0.1667}{0.5667}$ $=0.295$
AB->B,A	$v=1-\frac{j}{3}$	$1-\frac{0.3}{3}$	0.9			$\frac{v}{2v+4s}$	$\frac{0.9}{5.4}$	$\frac{1}{6}$	0	$\frac{1}{6}\cdot 0$	0		
AB->A,AB	$s=1-\frac{j}{3}$	$1-\frac{0.3}{3}$	0.9			$\frac{s}{2v+4s}$	$\frac{0.9}{5.4}$	$\frac{1}{6}$	0	$\frac{1}{6}\cdot 0$	0		
AB->AB,A	$s=1-\frac{j}{3}$	$1-\frac{0.3}{3}$	0.9			$\frac{s}{2v+4s}$	$\frac{0.9}{5.4}$	$\frac{1}{6}$	0	$\frac{1}{6}\cdot 0$	0		
AB->B,AB	$s=1-\frac{j}{3}$	$1-\frac{0.3}{3}$	0.9			$\frac{s}{2v+4s}$	$\frac{0.9}{5.4}$	$\frac{1}{6}$	0	$\frac{1}{6}\cdot 0$	0		
AB->AB,B	$s=1-\frac{j}{3}$	$1-\frac{0.3}{3}$	0.9			$\frac{s}{2v+4s}$	$\frac{0.9}{5.4}$	$\frac{1}{6}$	0	$\frac{1}{6}\cdot 0$	0		
									Total likelihood	$=0.4+\frac{1}{6}$	0.5667		
									Total log-likelihood	$=\log(0.5667)$	-0.568		

For the difference in the DEC+J results, the situation is more complex. First, *BioGeoBEARS* +J models limit *j* to a maximum of 2.99999 rather than literally 3.0, because the conversion of cladogenetic event weights to cladogenetic event probabilities involves dividing the weight of an individual cladogenetic event by the summed weight of all possible events. When the ancestor is AB, having j=3.0 creates a 0/0 error in this calculation. Second, the authors apply SFs differently between their DEC and DEC+J worked examples. See Supplemental Text for a detailed explanation.

Problems with the likelihood formulae in the Critique. In *Lagrange* and *BioGeoBEARS*, for DEC and DEC+J, the conditional probability of any particular range-inheritance scenario at a cladogenesis event is just the weight of that individual event, divided by the summed weight of all possible events in that particular row of the cladogenetic transition matrix (Table 1; Table 2AB; for the 3-area case, see Supplemental Table 1). However, R&S provide formulae (Table 1) for the weights that do not correspond to either program; an additional issue is mixing the likelihood calculation with the calculation of ancestral range probabilities. See Supp. Text for detailed discussion, and Table 1 for a comparison of the incorrect and correct calculations. It is easy to run small example datasets in *Lagrange* and *BioGeoBEARS* (the scripts and datasets are provided in Supp. Data), so it is surprising R&S did not check their likelihood calculations against the relevant programs.

Failure to replicate the 4-taxon worked example likelihoods. The critique's reported likelihoods for the 4-taxon example were checked against *Lagrange* (for DEC) and *BioGeoBEARS* (for DEC and DEC+J), and did not match (Table 3). The complete set of input files and scripts for all runs are given in Supplemental Data (FigShare: 10.6084/m9.figshare.19166393). The Critique reports a maximized logL under DEC of -8.50. However, Python *Lagrange*, C++ *Lagrange*, and *BioGeoBEARS* all report a logL of -4.481. For DEC+J, the Critique reports -3.47, while *BioGeoBEARS* DEC+J reports -1.171. Given the several difficulties encountered in trying to match up the Critique's 2-taxon worked example with the actual likelihood calculations done by *Lagrange* and *BioGeoBEARS*, I did not attempt to reverse-engineer the calculation of the Critique's incorrect likelihoods.

Table 2. Parameterization of the weights of cladogenetic range-change events for various biogeographical models in *BioGeoBEARS*. Here, 2 areas are assumed, meaning there are 3 possible ranges just before speciation, and 9 possible (left, right) pairs of descendants just after speciation. The conditional probability of each pair of ranges is calculated by taking the individual weight of the specific event, divided by sum of weights for all events in that that row. Blank cells indicate 0 weight/conditional probability.

2A. Parameterization of the DEC cladogenesis model for 2 areas. The letters specify the weights of the range-inheritance events of narrow-range sympatry (*y*), subset sympatry (*s*), and vicariance (*v*). Under DEC, each of these events has a weight of 1, that is, *y*=*s*=*v*=1.

Ranges of descendant pairs

	Left	A	A	A	B	B	B	AB	AB	AB
	Right	A	B	AB	A	B	AB	A	B	AB
Ancest. ranges	A	*y*								
	B					*y*				
	AB		*v*	*s*	*v*		*s*	*s*	*s*	

2B. Parameterization of the DEC+J cladogenesis model for 3 areas. The added letter, *j*, gives the weight of jump-dispersal events. As *j* increases, y, s, and v take the value 1-(*j*/3).

Ranges of descendant pairs

	Left	A	A	A	B	B	B	AB	AB	AB
	Right	A	B	AB	A	B	AB	A	B	AB
Ancest. ranges	A	*y*	*j*		*j*					
	B		*j*		*j*	*y*				
	AB		*v*	*s*	*v*		*s*	*s*	*s*	

2C. Parameterization of the DIVALIKE cladogenesis model for 3 areas. In DIVALIKE, following the assumptions of the parsimony DIVA program of Ronquist (1997), subset speciation (*s*) is given 0 weight, and widespread vicariance is allowed (unlike DEC), although only becomes relevant with 4 or more areas (and thus a scenario like ABCD->AB,CD becomes possible). The *j* parameter can be added, in the same locations in the table as DEC+J, to create DIVALIKE+J (not shown). In DIVALIKE+J, *j* ranges from 0 to 2, and *y* and *v* are reduced by *j*/2.

Ranges of descendant pairs

	Left	A	A	A	B	B	B	AB	AB	AB
	Right	A	B	AB	A	B	AB	A	B	AB
Ancest. ranges	A	*y*								
	B					*y*				
	AB		*v*		*v*					

2D. Parameterization of the BAYAREALIKE cladogenesis model for 3 areas. In BAYAREALIKE, following the assumptions of the Bayesian BayArea program of Landis et al. (2013), both subset speciation (*s*) and vicariance (*v*) are given 0 weight, and widespread sympatry is allowed (unlike DEC). The *j* parameter can be added, in the same locations in the table as DEC+J, to create BAYAREALIKE+J (not shown). In BAYAREALIKE+J, *j* ranges from 0 to 1, and *y* is reduced by *j*.

Ranges of descendant pairs

	Left	A	A	A	B	B	B	AB	AB	AB
	Right	A	B	AB	A	B	AB	A	B	AB
Ancest. ranges	A	*y*								
	B					*y*				
	AB									*y*

2E. For completeness, the implied parameterization of the cladogenesis process for an M*k* model (Markov model with *k*=2 states) for 2 areas is shown below. The anagenetic rate matrix for M*k* is, of course, different from the above models, as it allows "range-switching" along branches (e.g. sudden transitions from A to B), rather than relying on range expansion and range contraction as in the models above). The M*k* model also disallows ranges with a range size greater than 1 area. For Mk (and any similar purely-anagenetic model), the only allowed cladogenetic "events" are those that copy the ancestral value to both descendants, e.g. A->A,A; B->B,B.

Ranges of descendant pairs

	Left	A	A	A	B	B	B	AB	AB	AB
	Right	A	B	AB	A	B	AB	A	B	AB
Ancest. ranges	A	*y*								
	B					*y*				
	AB									

Part 2. The “all jump” inferences on tiny datasets are not that surprising, and they disappear with larger example datasets

Degeneracy and all-jump (or mostly-jump) inferences. Despite the several problems with the Critique’s calculation of likelihoods, I can verify that running the two-taxon and four-taxon examples in the *BioGeoBEARS* DEC+J model produces ML estimates of d=0, e=0, and j=2.99999 (Table 3). This parameter inference results in scenarios where jump dispersal is the most probable scenario at each node (but not the only scenario, due to the difference between DEC+J and DEC+J$_{\mathrm{mod1}}$, discussed in Supp. Text).

Table 3. Lagrange and BioGeoBEARS results of running DEC and DEC+J on ultrametric, pectinate trees, range "A" in the most derived position, and range "B" at all other tips.

							Root state probabilities				
Program and model	**Number of species**	**Number of free parameters**	**maximized lnL**	***d***	***e***	***j***	**A**	**B**	**AB**	**AIC$_c$**	**AIC$_c$ model weight**
"DEC", R&S 2-taxon example	2	2	-2.890	0	0	0			1	N/A[1]	
DEC, Python Lagrange	2	2	-1.792	4.29E-09	3.29E-09	0			1		
DEC, C++ Lagrange	2	2	-1.79176	1.13E-06	0.0001389	0			1		
DEC, BioGeoBEARS	2	2	-1.791759	7.68E-09	6.40E-09	0	1.48E-16	1.48E-16	1		
DEC+J, BioGeoBEARS	2	3	0.1541502	1.00E-12	1.00E-12	2.99999	0.4285714	0.4285714	0.1428572		
"DEC+J", R&S 2-taxon example	2	3	-0.693	0	0	3	0.5	0.5	0		
DEC, Python Lagrange	3	2	-3.472	1.79E-01	5.05E-07	0	0	0.3216	0.6784	N/A[1]	
DEC, C++ Lagrange	3	2	-3.47196	0.08971	7.07E-08	0	0	0.321649	0.678351		
DEC, BioGeoBEARS	3	2	-3.471964	0.08959	1.00E-12	0	6.12E-14	0.3213795	0.6786205		
DEC+J, BioGeoBEARS	3	3	-0.492477	1.00E-12	1.00E-12	2.99999	0.4090907	4.09E-01	0.1818182		
"DEC", R&S 4-taxon example	4	2	-8.50	0.060	0	0				N/A[1]	
DEC, Python Lagrange	4	2	-4.481	0.1011	4.29E-09	0	0	0.7578	0.2209		
DEC, C++ Lagrange	4	2	-4.48106	0.10183	8.14E-06	0	0	0.758828	0.241172		
DEC, BioGeoBEARS	4	2	-4.481012	0.10106	1.00E-12	0	6.43E-14	0.757828	0.2421718		
DEC+J, BioGeoBEARS	4	3	-1.170587	1.00E-12	1.00E-12	2.99999	0.4029848	0.402985	0.1940299		
"DEC+J", R&S 4-taxon example	4	3	-3.47	0.00	0.00	3					
DEC, Python Lagrange	5	2	-5.11514	0.07287	3.40E-07	0		0.929327			
DEC, C++ Lagrange	5	2	-5.11514	0.07287	3.40E-07	0		0.929327			
DEC, BioGeoBEARS	5	2	-5.115132	0.07286	1.00E-12	0	1.79E-14	0.9293209	0.0706791	20.2	99.88%
DEC+J, BioGeoBEARS[2]	5	3	-1.858772	1.00E-12	1.00E-12	2.99999	0.4009899	0.4009903	0.1980198	33.7	0.12%
DEC, Python Lagrange	6	2	-5.542	0.05183	4.29E-09	0		0.9826			
DEC, C++ Lagrange	6	2	-5.54188	0.05178	1.88E-06	0		0.98255			
DEC, BioGeoBEARS	6	2	-5.541854	0.05181	1.00E-12	0	3.56E-15	0.9825586	0.0174414	19.1	88.17%
DEC+J, BioGeoBEARS[2]	6	3	-2.550271	1.00E-12	1.00E-12	2.99999	0.4003292	0.4003297	0.1993411	23.1	11.83%
DEC, Python Lagrange	7	2	-5.867	0.03825	4.29E-09	0		0.9961			
DEC, C++ Lagrange	7	2	-5.86716	0.03829	4.97E-07	0		0.996078			
DEC, BioGeoBEARS	7	2	-5.86715	0.03825	1.00E-12	0	6.34E-16	0.9960760	0.0039240	18.7	70.59%
DEC+J, BioGeoBEARS[2]	7	3	-3.242869	1.00E-12	1.00E-12	2.99999	0.4001095	0.4001100	0.1997805	20.5	29.41%
DEC, Python Lagrange	8	2	-6.138	0.02933	4.29E-09	0		0.9992			
DEC, C++ Lagrange	8	2	-6.13769	0.02935	1.70E-06	0		0.999159			
DEC, BioGeoBEARS	8	2	-6.137646	0.02933	1.00E-12	0	1.09E-16	9.99E-01	0.0008382	18.7	53.34%
DEC+J, BioGeoBEARS	8	3	-3.471416	1.00E-12	1.00E-12	0.090724	6.08E-07	0.9997825	0.0002169	18.9	46.66%
DEC, Python Lagrange	9	2	-6.373	0.02321	4.29E-09	0		0.9998			
DEC, C++ Lagrange	9	2	-6.37269	0.02342	7.77E-07	0		0.999826			
DEC, BioGeoBEARS	9	2	-6.372617	0.0232	1.00E-12	0	1.84E-17	0.9998268	0.0001732	18.7	41.51%
DEC+J, BioGeoBEARS	9	3	-3.62951	1.00E-12	1.00E-12	0.076489	1.69E-08	0.9999622	3.78E-05	18.1	58.49%

[1] AIC$_c$ is not defined unless the number of data exceeds the number of free parameters by 2 or more.

[2] For these runs, because the likelihood profile for *j* had 2 peaks, the optimizer had to be changed from the default BioGeoBEARS_run_object$use_optimx =TRUE, to BioGeoBEARS_run_object$use_optimx = "GenSA", in order to find the ML solution at *j*=2.99999 during optimization.

The result is indeed surprising to intuition, and the Critique’s central arguments hang on this result, the claim that it is “degenerate,” the claim that such “degenerate” results are problematic, and the claim that similarly “degenerate” results are common on bigger, real-world datasets. The reason for R&S's choice of the word "degenerate" is obscure; it does not correspond to standard usages of the term; this relatively unimportant terminological issue is discussed in Supp. Text.

Is the 4-taxon DEC+J result "pathological"? The differences in log-likelihoods between DEC and DEC+J (Δ logL) produced by the two models are not as large as the Critique alleges -- the Critique reports ΔlogLs of 2.197 and 5.030 for the 2-taxon and 4-taxon case, but they are actually 1.946 and 3.310 (as can be calculated from Table 3). Nevertheless, it does surprise an intuition that many phylogeneticists and biogeographers would probably share, namely that a 4-taxon, rooted tree with the range pattern (((A,B),B),B) should infer the most probable root state as B, with a single transition to A happening high up in the tree. Whether or not the observed result *should* surprise a well-informed common sense is another matter.

I suspect that the intuition is derived from biologists' extensive experience with parsimony reconstructions (typically "Fitch" parsimony, Clark et al., 2009) and, within model-based frameworks, Markov-*k* and DNA substitution models. However, intuitions are not always reliable, especially when they are being transferred from a simple model to a more complex one with additional processes that are not included in the simple model. This is the case with biogeographical models, where the anagenetic process involves range expansions and range contractions rather than direct transitions between A and B, and a cladogenetic process has also been added.

In the case of the 4-taxon example with *j*=2.99999, the jump dispersal process is contributing most of the likelihood. There are 3 internal nodes, one of them being the root. If the root is A, then the tip data are explained by 3 jumps to B (3 jump dispersals of the form A->A,B), each having a conditional probability of $\frac{1}{2}$ of being appropriate to explain the data. The likelihood of the data under this scenario is $\left(\frac{1}{2}\right)^3 = \frac{1}{8}$. If the root is B, three jump dispersals can also explain the data (B->A,B at the root, A->A,B at the higher nodes), contributing $\frac{1}{8}$ to the likelihood. If the root state is AB, there are three scenarios that can explain the tip data. First, vicariance at the root (AB->A,B) followed by two jump dispersals (A->A,B), with a probability of $\frac{1}{6} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{24}$. Second, subset sympatry at the root

(AB->AB,B), followed by another subset sympatry (AB->AB,B), followed by vicariance (AB->A,B), with a probability of $\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{216}$. Third, subset sympatry at the root (AB->AB,B), followed by vicariance at the middle node (AB->A,B), followed by jump dispersal (A->A,B), with a probability of $\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{2}=\frac{1}{72}$. The total likelihood across all possible ancestral ranges A, B, and AB is therefore $\frac{1}{8}+\frac{1}{8}+\frac{1}{24}\cdot\frac{1}{216}\cdot\frac{1}{72}=0.3102$, logL=-1.1706 (Table 3). Note that 4-taxon datasets where A and B alternate at the tips have identical log-likelihood (=-1.1706) under the d=e=0, j=2.99999 model (see example "4species_ABalternate_ML_DECvDECj_BGB" in Supplemental Data), suggesting that all that is happening is that coincidence is a relatively good explanation of the data when there are only 2 areas and 4 tips.

Our intuition against jump dispersals in the Critique's 4-taxon example stems from the assumption that within-area speciation, e.g. B->B,B, should be a more probable explanation when several sister species have ranges in B. It turns out that this intuition is true in likelihood terms as well, but *only* if there are a sufficient number of internal nodes to make a series of high-probability jump dispersals conferring likelihood $\frac{1}{2}$ a lower likelihood contribution than a series of high-probability within-area speciation events followed by a single, lower probability jump dispersal event. While our intuition can easily propose a scenario like B->B,B at the root, followed by B->B,B and then B->A,B, a probabilistic model has to assign the same set of conditional probabilities across all of the nodes. Raising the probability of within-area scenarios like B->B,B thus necessarily lowers the probability of jump dispersal scenarios like B->A,B, and vice versa. The parameter j describes this tradeoff. In the 2-area, 4-taxon example, if $j=\frac{3}{7}$=0.4286, then the probabilities of the three possible scenarios at a particular node with range B are B->B,B=0.5, B->A,B=0.25, B->B,A=0.25 (Supplemental Excel File 1). This model might more closely reflect our intuition for what we think the best-fit process "should" be in the 4-taxon case. However, the history that contributes the highest likelihood to the data in this case is B->B,B at the root and at the middle node, followed by B->A,B at the top node, for a total probability of $\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{4}=\frac{1}{16}$, less than the $\frac{1}{8}$ conferred by *each* of the two all-jump-dispersal scenarios when j=2.99999. Numerous other scenarios are possible under DEC+J when d=0, e=*0*, $j=\frac{3}{7}$, but all scenarios together add up to a total likelihood of only 0.147 (Supplemental Excel File 1).

The (almost) "all jumps" result breaks down with increasing dataset size. How many internal nodes would it take, in the case of a 2-area system, with one area found only in the most derived position, to successfully inform a DEC+J model that the probability of within-area speciation is greater than 0, and thus the ML estimate of *j* cannot be 2.99999? I investigated this by creating datasets with taxa numbers ranging from 2 to 10, as well as 15 and 20 species. Following the Critique's two-taxon example, the trees had minimum branch lengths of 1 unit, and each additional sister branch was added with a common ancestor 1 unit below the previous node. All other branchlengths extend up to the present, to produce an ultrametric tree. For each dataset, I ran ML analyses under DEC and DEC+J. I also constructed curves describing how the log-likelihood varies with *j* (with *d* and *e* fixed to 0) by calculating the log-likelihood of each dataset at each value of *j* from 0 to 2.99999 at increments of 0.01. These calculations were done with the function `bears_optim_run` with `skip_optim` set to `TRUE`.

The ML results for <10 species are in Table 3, and the likelihood profile curves are in Figure 1A. Dashed lines indicate cases where the likelihood is maximized at *j*=2.99999, and solid lines indicate ML solutions at some other value. At *n*=4 taxa, confirming R&S's observation, the likelihood rises monotonically with *j*. However, at *n*=5 taxa, a second likelihood peak begins to be seen, representing scenarios where most nodes experience within-area speciation events, followed by a jump dispersal at the highest node. At *n*=6 or 7, the peak becomes more distinct, and at *n*=8, the likelihood peak at *j*=0.091 becomes the ML solution. As the number of taxa increases further, the peak at small *j* soon dominates. These results suggest that the jumps-at-every-node result observed in the Critique, while counterintuitive, is partially a product of tiny dataset size. Maximum likelihood inference, and many of the key results used in ML inference (for example, the Likelihood Ratio Test and AIC) rely on asymptotic results that describe the expected behaviour of the inference (consistency, unbiasedness, etc.) as the size of the sample becomes large (Anisimova et al., 2001; Burnham & Anderson, 2002). There is no guarantee that ML will perform well on very small datasets. We should, however, be reassured by the fact that intuitively reasonable inference begins to occur as the dataset size increases -- keeping in mind that, as we are inferring 3 free parameters, even *n*=8 is a small dataset. Judging an inference method with 3 free parameters on a dataset of size *n*=2 or *n*=4 is not a reliable strategy.

Figure 1:

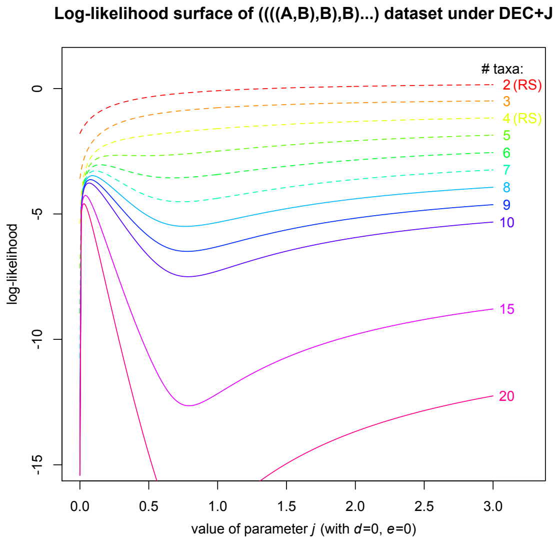


Figure 1A. Likelihood profiles for datasets following the pattern in the Critique's worked examples, namely, pectinate, ultrametric trees, with all tips inhabiting single areas, and with only one tip in the most-nested clade inhabiting a different area.

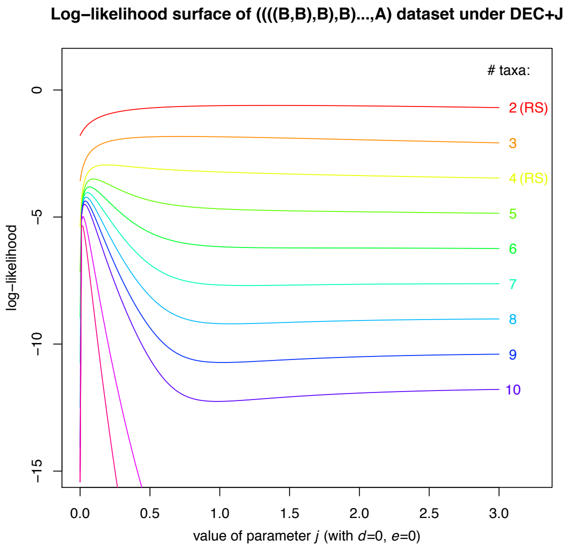


Figure 1B. As in 1A, but with the maximum number of areas set to 3 instead of 2.

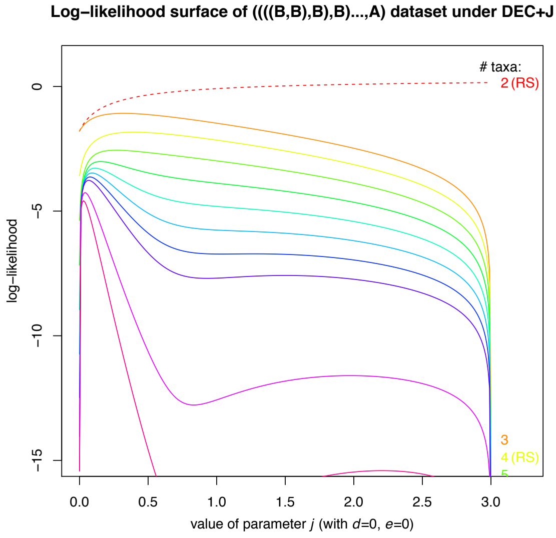


Figure 1C. As in 1A, but with the unique tip range being sister to the most-derived group.

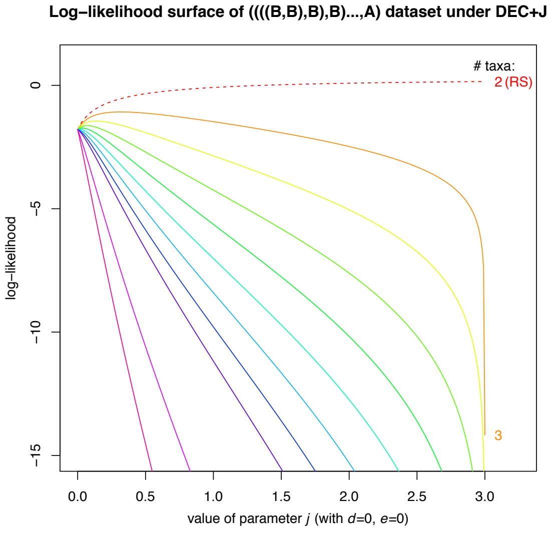


Figure 1D. As in 1A, but with the unique tip range moved to the outgroup.

Human-constructed datasets are not simulation results. Although judging methods from small datasets is problematic, it is even more problematic to make judgments about a model's performance in general, across thousands of observed and simulated datasets, on the basis of one particular data configuration and problem setup. The problem with a human-constructed data pattern is that it is not actually the product of a probabilistic process – that is, there is no sense of how commonly the human-constructed pattern would actually emerge from whatever intuitive process is being assumed. For example, in the case of the 4-taxon example and its derivatives, consider the intuitive model, where it is assumed that within-area speciation is high probability, and jump-dispersal is low probability. A simulation under such a model would not usually produce a pectinate tree with a single tip in the most-derived position inhabiting a different area than the other tips. The most common result on a small tree would be to have all tips in the same area (within-area speciation at each node). Less common, but not rare, would be tips other than the most-derived position inhabiting the different area. A likelihood analysis is taking all of these possibilities into consideration, and if a human-constructed data pattern is actually rare under the model that human intuition infers, it opens up the possibility that some non-intuitive model will be a better fit.

The (almost) "all jumps" result is only preferred if standard practices in statistical model comparison are ignored. Above, we have been relying purely on the likelihoods returned by DEC and DEC+J when applied to various constructed datasets. The argument has been about whether or not the inferred values of *j* and the observed likelihood improvements are reasonable or unreasonable. However, the entire discussion thus far has ignored basic principles of statistical model comparison that should be applied especially in the case of small datasets. As noted above, the theory behind the Likelihood Ratio Test uses approximations that are reliable only asymptotically as dataset size becomes large. The Akaike Information Criterion (AIC; Akaike, 1974; Burnham & Anderson, 2002) is another approach to correct for the bias in comparing models via the log-likelihood only, penalizing models by the number of free parameters they have. However, the AIC itself is known to be biased in the situation where the dataset size is small compared to the number of free parameters. In this situation, the sample-size corrected AIC, or AIC$_c$, is recommended (Burnham & Anderson, 2002).

AIC$_c$ increases the penalty for free parameters when the dataset size is small; as dataset size increases, AIC$_c$ converges on traditional AIC. For comparing models with 2 and 3 parameters, AIC$_c$ is likely to make little difference in model comparison when the dataset size is above ~20. However, it can make a large difference for small datasets (Van Dam & Matzke, 2016). I calculated AIC$_c$ and AIC$_c$ model weight (Franklin et al., 2001) for all of the DEC/DEC+J model comparisons in Table 3. Note that AIC$_c$ is not even defined when the dataset size exceeds the number of free parameters by less than 2; therefore, the Critique's worked examples are too small to even apply AIC$_c$. However, even with phylogenies of 5-8 species, DEC+J fails to obtain higher model weight than DEC. Only at 9 species does DEC+J begin to outperform DEC (58.5% to 41.5%).

The brittleness of conclusions from narrow, constructed examples. Another problem with making general judgments based on narrow constructed examples is that slight changes to the problem setup might reveal that the unintuitive result is actually quite brittle to the specific setup of the problem. To check this, I explored a few variations on the Critique's example setup, by modifying the likelihood-profile analysis described above. The variations were (Figure 1B) setting the maximum number of areas to 3 instead of 2, but not otherwise changing the tree or data; (1C) moving the "unique" single-area range from a most-derived position to the branch sister to the most-derived clade; (1D) moving the unique range to the "outgroup" position in the tree; and (1E) moving the unique range to a branch that diverges from the node above the root node.

In none of these slightly-modified example datasets does the phenomenon of ML optimizing *j* at 2.99999 occur (Figures 1B-1E). Instead, *j* is estimated to have some low but positive value in all cases, agreeing with intuition.

The reason for this radically different behaviour has to do with the likelihood conferred on the data by the all-jump-dispersal scenarios under these modified setups. In Figure 1B, when the maximum number of areas is 3, the following jump-dispersal scenarios are possible at a node, conditional on the range being A just before speciation: A,B; B,A; A,C; C,A. Each of these has a conditional probability of 0.25. When the maximum number of areas was 2, the possible results of jump dispersal starting from A were only A,B and B,A, each with a probability of 0.5. Thus, the simple change from 2 areas to 3 results in the likelihood of the data under the all-jumps scenario decreasing by half at each internal node. As a result, the all-jumps scenarios cease to have the highest probability, even for small datasets.

For Fig. 1C-1E, merely moving the location of the unique range removes the competitiveness of the all-jumps scenarios. In 1C, moving the unique range to the branch sister to the most-derived clade results in the derived clade having a range of (B,B). Such a data pattern cannot be explained by a jump dispersal event, so there must be some weight on *y* to explain a within-area speciation event, therefore *j* does not optimize near 3.

In Fig. 1D, having the unique range in the outgroup instead of the most derived position produces a data pattern that can be explained by vicariance of AB at the root, followed by within-area speciation at every higher node. The within-area speciation events will be more probable the closer *j* is to 0. Therefore, DEC+J never achieves a log-likelihood more than about 0.7 units higher than DEC on this dataset, and the advantage rapidly declines to 0 (as does the estimate of *j*) as the dataset size increases. The situation is similar in 1E. These are examples of data patterns where DEC+J does not outperform DEC. (Many such patterns are possible, belying any claims that DEC+J has an "unfair" advantage. Whether or not DEC-favouring patterns are often observed in real empirical data is a different question.)

In summary, the surprising behavior noted in the Critique's worked examples, where ML inference favoured the maximum value of *j* and preferred (almost) all-jump dispersal histories, is in fact the result of the specific structure of that problem setup. The decision to use a 2-area system interacts with the fact that the phylogenies in use are binary trees (all nodes have 2 descendant lineages) to create a situation where the "right" jumps to explain the dataset have high probability.

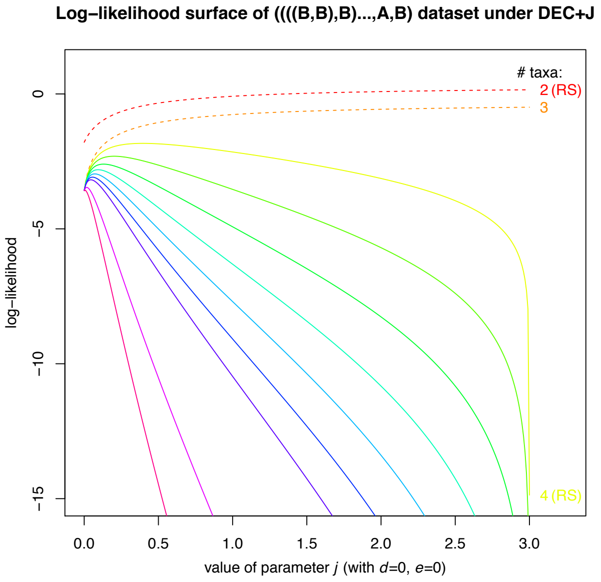


Figure 1. Likelihood profiles for various datasets under DEC+J with *d*=0, *e*=0, and *j* varying on the x-axis. The likelihood was calculated for $0 \leq j < 3$, incrementing *j* by 0.01.

Figure 1E. As in 1A, but with the unique tip range moved to the first tip inside the outgroup.

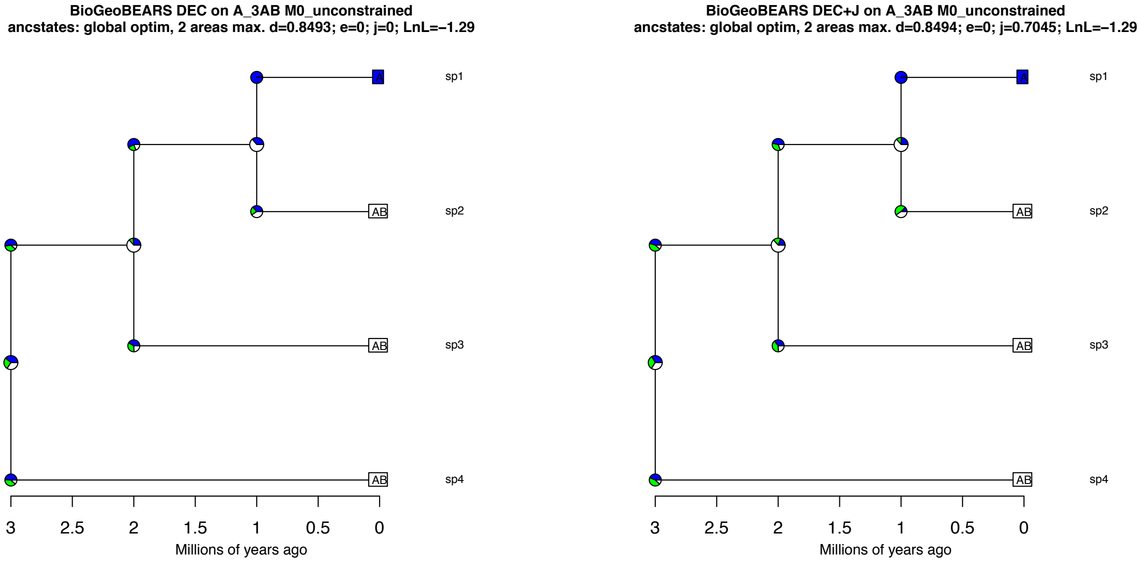

BioGeoBEARS DEC on A_3AB M0_unconstrained
ancstates: global optim, 2 areas max. d=0.8493; e=0; j=0; LnL=−1.29
A
AB
AB
AB
sp1
sp2
sp3
sp4
3
2.5
2
1.5
1
0.5
0
Millions of years ago
BioGeoBEARS DEC+J on A_3AB M0_unconstrained
ancstates: global optim, 2 areas max. d=0.8494; e=0; j=0.7045; LnL=−1.29
A
AB
AB
AB
sp1
sp2
sp3
sp4
3
2.5
2
1.5
1
0.5
0
Millions of years ago

Figure 2. An example data pattern where DEC+J (right) has no advantage over DEC (left).

Some constructed dataset patterns always prefer DEC. It is trivial to construct datasets where DEC+J has no advantage over DEC. For example, if we take the pectinate trees from the previous examples, and the replace the range "A" with "AB" at the tips, then DEC+J has no log-likelihood advantage, despite the extra free parameter (Supplemental Table 2), no matter the dataset size. Instead, DEC+J infers a dominant role for range-expansion dispersal, inferring virtually identical (and large) values of the parameter *d*, and the same ancestral state probabilities (Figure 2). The fact that a particular data pattern favours a particular model is not evidence of unfairness; instead, it reinforces the point that inferences made on individual data patterns are not a reliable basis on which to judge the general behaviour of an inference model. Instead, different data patterns will be better fit by different models.

Part 3. Testing for fairness

The Critique alleges that DEC+J has an "unfair" advantage over DEC in explaining datasets. Of course, there are ways that an unfair advantage *could* exist in likelihood-based model comparison. For example, because some implementations of models include or exclude constant multipliers to the likelihood, such as due to equal state frequencies (see Stadler, 2013 for examples of different implementations of tree models), one would not want to compare a model that included a constant multiplier with another model that left it out. Another way that a likelihood-based comparison could be unfair would be if one of the likelihood calculations had a mistake that artificially raised or lowered the likelihood difference between models.

However, there is a way to check for fairness. Likelihoods are conditional probabilities, and for data that are discrete, the conditional probabilities must add up to 1 when summed across all possible datasets (Felsenstein, 1992). This condition applies as long as no constant multipliers to the likelihood have been left out; if they have been left out, then the likelihoods will add up to 1/(the constant).

Summing likelihoods across all possible datasets. I performed this experiment for the DEC and DEC+J models, for two-taxon trees with 2-area and 3-area problems. For the 2-area problem, there are 4 possible ranges at each tip, and therefore 4x4=16 possible datasets for a 2-taxon tree. For the 3-area problem, there are 8 possible ranges, and therefore 8x8=64 possible datasets for a 2-taxon tree. For each dataset collection, I calculated the likelihood under four model/parameter combinations: (1) DEC, with cladogenetic range-change only (d=e≅0); (2) DEC with both anagenetic and cladogenetic range-change processes (d=0.2, e=0.1); (3) DEC+J, using the same d and e as for combination #2, but adding j=0.15; (4) DEC+J, dominated by jump-dispersal (d=e≅0, j=2.9). The scripts for these runs are included in Supplemental Data, so that researchers may repeat the experiment with other choices of the parameter values.

The results for the 3-area problem are shown in Table 4. It can be seen that, regardless of the model and parameters, the likelihoods under a particular model, added up across all possible data patterns, add up to 7.0 (ignoring rounding errors). If state frequencies were applied, with a 0 base frequency for the null range, and $\frac{1}{7}$ for the other ranges, these likelihoods would add up to 1.0. For the 2-area problem (Supplemental Table 3), the likelihoods add up to 3.0 (or 1.0, if SFs of (0, 1/3, 1/3, 1/3) were to be applied). It is crucial to recognize that this result is not just luck or coincidence; it is the mathematical consequence of having implemented models such that they rely on valid conditional probabilities.

Table 4. Log-likelihoods and likelihoods of all 64 possible data patterns for a 2-taxon tree (unit branchlengths) with three geographic areas, under four different sets of DEC/DEC+J parameters. (a) A DEC "cladogenesis-only" model. (b) A DEC model with both anagenetic and cladogenetic processes. (c) A DEC+J model with both anagenetic and cladogenetic processes, with a significant weight on cladogenetic jump dispersal. (d) A DEC+J model with no anagenetic processes, and with jump dispersal dominanting the cladogenetic process. Summing the likelihoods of all possible data patterns shows that, regardless of the model and parameters, the total of all likelihoods is 7. This demonstrates that no models have "unfair" advantages, rather, each data pattern is best fit (grey shading) by one of the four given parameters sets. (Note that none of the four model-parameter sets given here is likely to be the maximum likelihood (ML) inference for any of the 64 data patterns; the purpose is just to demonstrate that any particular model & parameters will prefer some data patterns over others, and vice versa.)Summing the likelihoods of all possible data patterns shows that, regardless of the model and parameters, the total of all likelihoods is 7. This demonstrates that no models have "unfair" advantages, rather, each data pattern is best fit (grey shading) by one of the four given parameters sets. (Note that none of the four model-parameter sets given here is likely to be the maximum likelihood (ML) inference for any of the 64 data patterns; the purpose is just to demonstrate that any particular model & parameters will prefer some data patterns over others, and vice versa.)

Data (left, right)	(a) DEC, cladogenetic range-change only: *d*=0.00001, *e*=0.00001, *j*=0		(b) DEC, mix of anagenetic and cladogenetic range-change: *d*=0.2, *e*=0.1, *j*=0		(c) DEC+J, mix of anagenetic and cladogenetic range-change, with jump dispersal added: *d*=0.2, *e*=0.1, *j*=0.15		(c) DEC+J, only cladogenetic range-change, dominated by jump-dispersal: *d*=0.0001, *e*=0.0001, *j*=2.9	
	Data lnL	Data likelihood	Data lnL	Data likelihood	Data lnL	Data likelihood	Data lnL	Data likelihood
000,000	-21.6396	0	-3.6198	0.027	-3.6307	0.026	-21.9162	0
000,100	-11.2253	0	-2.6187	0.073	-2.6286	0.072	-11.5019	0
000,010	-11.2253	0	-2.6187	0.073	-2.6286	0.072	-11.5019	0
000,001	-11.2253	0	-2.6187	0.073	-2.6286	0.072	-11.5019	0
000,110	-12.3884	0	-3.0451	0.048	-3.043	0.048	-12.0615	0
000,011	-12.3884	0	-3.0451	0.048	-3.043	0.048	-12.0615	0
000,101	-12.3884	0	-3.0451	0.048	-3.043	0.048	-12.0615	0
000,111	-12.8992	0	-2.837	0.059	-2.8313	0.059	-12.8991	0
100,000	-11.2253	0	-2.6187	0.073	-2.6286	0.072	-11.5019	0
100,100	-0.0001	1	-0.8903	0.411	-1.3316	0.264	-5.8551	0.003
100,010	-1.7918	0.167	-2.3853	0.092	-1.8164	0.163	-0.685	0.504
100,001	-1.7918	0.167	-2.3853	0.092	-1.8164	0.163	-0.685	0.504
100,110	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
100,011	-2.4849	0.083	-2.6711	0.069	-2.261	0.104	-0.568	0.567
100,101	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
100,111	-2.4849	0.083	-1.8572	0.156	-1.8514	0.157	-2.4848	0.083
010,000	-11.2253	0	-2.6187	0.073	-2.6286	0.072	-11.5019	0
010,100	-1.7918	0.167	-2.3853	0.092	-1.8164	0.163	-0.685	0.504
010,010	-0.0001	1	-0.8903	0.411	-1.3316	0.264	-5.8551	0.003
010,001	-1.7918	0.167	-2.3853	0.092	-1.8164	0.163	-0.685	0.504
010,110	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
010,011	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
010,101	-2.4849	0.083	-2.6711	0.069	-2.261	0.104	-0.568	0.567
010,111	-2.4849	0.083	-1.8572	0.156	-1.8514	0.157	-2.4848	0.083
001,000	-11.2253	0	-2.6187	0.073	-2.6286	0.072	-11.5019	0
001,100	-1.7918	0.167	-2.3853	0.092	-1.8164	0.163	-0.685	0.504
001,010	-1.7918	0.167	-2.3853	0.092	-1.8164	0.163	-0.685	0.504
001,001	-0.0001	1	-0.8903	0.411	-1.3316	0.264	-5.8551	0.003
001,110	-2.4849	0.083	-2.6711	0.069	-2.261	0.104	-0.568	0.567
001,011	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
001,101	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
001,111	-2.4849	0.083	-1.8572	0.156	-1.8514	0.157	-2.4848	0.083
110,000	-12.3884	0	-3.0451	0.048	-3.043	0.048	-12.0615	0
110,100	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
110,010	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
110,001	-2.4849	0.083	-2.6711	0.069	-2.261	0.104	-0.568	0.567
110,110	-11.9184	0	-2.4889	0.083	-2.5834	0.076	-15.3191	0
110,011	-12.2061	0	-2.795	0.061	-2.7243	0.066	-11.3781	0
110,101	-12.2061	0	-2.795	0.061	-2.7243	0.066	-11.3781	0
110,111	-13.3046	0	-2.6314	0.072	-2.6234	0.073	-13.3045	0
011,000	-12.3884	0	-3.0451	0.048	-3.043	0.048	-12.0615	0
011,100	-2.4849	0.083	-2.6711	0.069	-2.261	0.104	-0.568	0.567
011,010	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
011,001	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
011,110	-12.2061	0	-2.795	0.061	-2.7243	0.066	-11.3781	0
011,011	-11.9184	0	-2.4889	0.083	-2.5834	0.076	-15.3191	0
011,101	-12.2061	0	-2.795	0.061	-2.7243	0.066	-11.3781	0
011,111	-13.3046	0	-2.6314	0.072	-2.6234	0.073	-13.3045	0
101,000	-12.3884	0	-3.0451	0.048	-3.043	0.048	-12.0615	0
101,100	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
101,010	-2.4849	0.083	-2.6711	0.069	-2.261	0.104	-0.568	0.567
101,001	-1.7918	0.167	-1.8506	0.157	-1.9656	0.14	-5.192	0.006
101,110	-12.2061	0	-2.795	0.061	-2.7243	0.066	-11.3781	0
101,011	-12.2061	0	-2.795	0.061	-2.7243	0.066	-11.3781	0
101,101	-11.9184	0	-2.4889	0.083	-2.5834	0.076	-15.3191	0
101,111	-13.3046	0	-2.6314	0.072	-2.6234	0.073	-13.3045	0
111,000	-12.8992	0	-2.837	0.059	-2.8313	0.059	-12.8991	0
111,100	-2.4849	0.083	-1.8572	0.156	-1.8514	0.157	-2.4848	0.083
111,010	-2.4849	0.083	-1.8572	0.156	-1.8514	0.157	-2.4848	0.083
111,001	-2.4849	0.083	-1.8572	0.156	-1.8514	0.157	-2.4848	0.083
111,110	-13.3046	0	-2.6314	0.072	-2.6234	0.073	-13.3045	0
111,011	-13.3046	0	-2.6314	0.072	-2.6234	0.073	-13.3045	0
111,101	-13.3046	0	-2.6314	0.072	-2.6234	0.073	-13.3045	0
111,111	-23.0258	0	-2.7457	0.064	-2.7373	0.065	-23.0258	0
Totals		**7.002**		**7.001**		**7.007**		**7.005**

Note: Deviations from integer 7 are due to rounding errors.

Particular models favour particular data patterns; this is not "unfair." In Table 4, for each data pattern, the model that confers the highest likelihood on that data pattern is highlighted in grey. From this, it is apparent that there is no such thing as a universal or "unfair" advantage for one model over another, when both models rely on comparable conditional probabilities. Instead, each model produces some data patterns with higher frequency and others with lower frequency. It may well be the case that certain data patterns are more commonly observed in real-life empirical datasets, and it may be that the kinds of data patterns most commonly observed in real-life happen to be fit better by one model than another. But this is not bias or unfairness. Instead, this is reality impinging on our models. If a particular model tends to fit empirical datasets more poorly than another, this is important information, not something that should ignored or explained away (for discussion of the issue of null ranges at tips, see Supp. Text).

Part 4. Model adequacy: A way to see the actual advantage of DEC+J on certain datasets

A common strategy in assessing phylogenetic comparative methods is to test for model adequacy (Ripplinger & Sullivan, 2010; Pennell et al., 2015), recently applied in biogeography (Tejero-Cicuéndez et al., 2021). The motivation for model adequacy is that in most cases we know that the available inference models are much simpler than the "true" model operating in the real world. Methods such as AIC and Bayes factors can be used to compare the relative fit of models to the data, but showing that one model has relatively better fit than another does not prove that the better model is capturing key features of the data. The basic method to assess model adequacy involves (a) running inference under each model of interest, and inferring parameters (and perhaps ancestral state probabilities); (b) taking these estimates and simulating new datasets under them (posterior predictive simulations); (c) describing the results of these simulations graphically or statistically, for example with summary statistics that capture key features of the data; (d) comparing the summary statistics to the same summary statistics for the original dataset. If the observed data fall well within the distributions produced by the simulations, the model can be judged to be "adequate" in that respect. If not, then it is "inadequate" at modeling that feature of the data. What "key features" of the data should be used to make these judgments is admittedly a subjective decision by the researcher, and presumably with sufficient data and effort any computationally feasible model could be judged to be an inadequate description of ultra-complex reality to some degree, but model adequacy at least provides a method to identify large, obvious discrepancies between inferred models and the data they are trying to fit.

Previous publications (Matzke, 2014; Massana et al., 2015) have suggested the importance of "narrow" ranges (single-area ranges) versus "widespread" ranges (ranges occupying 2 or more areas) for determining the relative fit of DEC and DEC+J models, taking the view that particular data patterns favour each model. The Critique, on the other hand, alleges that the key advantage of DEC+J is an "unfair" advantage due to the exclusion of time-dependent probabilities from the jump-dispersal process. We can test this hypothesis with a model adequacy experiment.

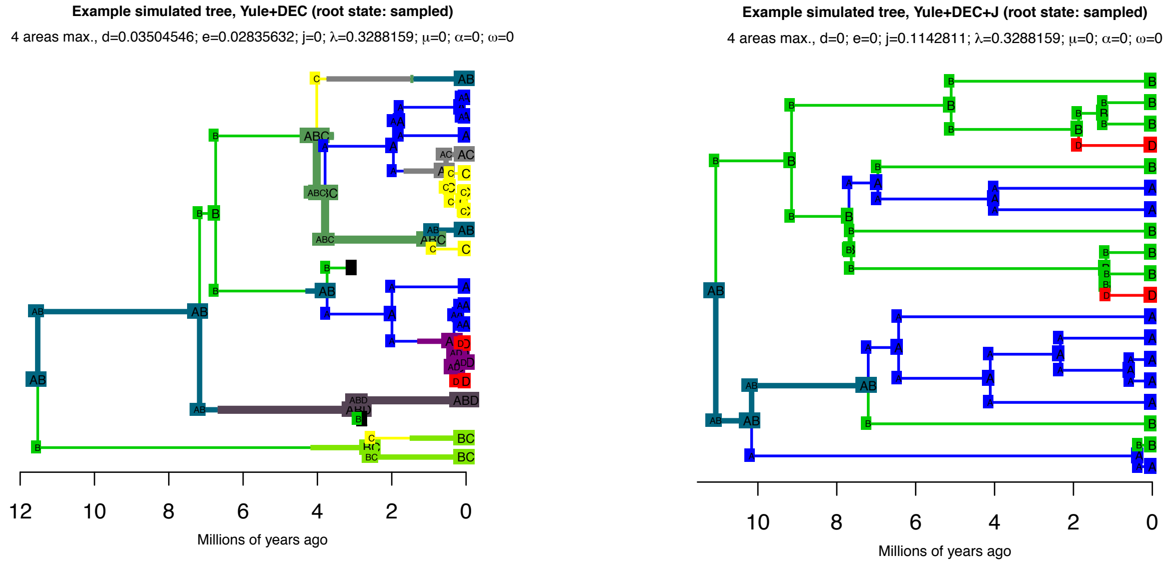
Model adequacy simulation experiment. I took the phylogeny and range data from Hawaiian *Psychotria*, Ree and Smith's (2008) example dataset used in *Lagrange* and *BioGeoBEARS*. First, I inferred parameters and ancestral ranges under DEC and DEC+J, as done in the example *BioGeoBEARS* script at http://phylo.wikidot.com/biogeobears#script. I also inferred the speciation rate using Maximum Likelihood with the APE function `yule`. The result was that the birthrate $\lambda = 0.3289132$. The APE function `birthdeath` returns the same result, and also estimates the extinction rate as 0.0. Then, I ran 8 sets of 100 simulations each, using the simulation code slightly modified from Matzke (2014) (available in Supplemental Material), using the inferred *d*, *e*, and *j* parameters, and a Yule process for the phylogeny, using the inferred λ. Each simulation records the complete simulated history of ranges along branches, as well as the states at nodes, so I calculated the proportion of branchlength spent in widespread ranges. Dividing by the total branchlength in each tree provides the fraction of the tree occupied by widespread ranges. The means and 2.5-97.5% percentiles of this statistic across each batch of 100 trees are shown in Table 5. I also tabulated the mean fraction of tips that are widespread for each simulated tree.

Table 5. Summary statistics from model adequacy study, simulating from the ML-inferred parameters for the *Psychotria* dataset. All DEC simulations were run with the ML-inferred patterns for the *Psychotria* dataset under the DEC model (λ=0.3288159, μ=0, d=0.03504546, e=0.02835632, j=0). Similarly, the Yule+DEC+J simulations used the DEC+J-inferred parameters (λ=0.3288159, μ=0, d=0, e=0, j=0.1142811).

Simulation model	Starting range	Number of tips	average fraction of branchlength that is widespread	95% percentiles	mean fraction of tips that are widespread	95% percentiles
Yule+DEC	A	50	0.164	(0.0065, 0.382)	0.178	(0.0276, 0.341)
Yule+DEC+J	A	50	0.000	(0, 0)	0.000	(0, 0)
Yule+DEC	sampled	50	0.210	(0.0065, 0.463)	0.188	(0.0526, 0.368)
Yule+DEC+J	sampled	50	0.038	(0, 0.185)	0.0063	(0, 0.053)
Yule+DEC	A	19	0.182	(0.0828, 0.281)	0.192	(0.0715, 0.34)
Yule+DEC+J	A	19	0.000	(0, 0)	0.000	(0, 0)
Yule+DEC	sampled	19	0.198	(0.0896, 0.331)	0.195	(0.081, 0.34)
Yule+DEC+J	sampled	19	0.016	(0, 0.09)	0.0016	(0, 0.02)
Psychotria (empirical)		19			0 (all tips single-area)	

It is apparent that simulations from the DEC parameters, even though they were optimized to fit the empirical *Psychotria* dataset, nevertheless produce simulated datasets that do not resemble the empirical dataset with respect to range size. Depending on the simulation conditions, an average of 16-21% of tips are widespread, while the empirical dataset had 0 widespread tips. The 95% percentiles describing variation within a group of simulations also exclude 0 for all runs. Clearly, the DEC model is inadequate, at least for the key variable of range-size, for fitting a dataset like *Psychotria*, which has 0 tips in widespread ranges.

The reasons for this mismatch become clear when simulation histories are visualized. Figure 3 gives an example. The width of the branch corresponds to the number of areas occupied for each lineage at every point in the phylogeny. A DEC simulation (3A), using the parameters inferred from *Psychotria*, is able to spread lineages to different areas using a *d* parameter of 0.035, but it is not able produce all tips with size 1. DEC does have processes that reduce range size – the cladogenesis processes lower the average range size when they operate, and the range contraction parameter *e*, is 0.028, comparable to *d*. (The problem would be worse in the much more common situation where inferred $e \cong 0$.) But, even though the parameters controlling range expansion and range contraction are optimized to match the *Psychotria* dataset, the resulting simulation under the inferred process just isn't able to avoid producing some tips with widespread ranges.



Example simulated tree, Yule+DEC (root state: sampled)

4 areas max., d=0.03504546; e=0.02835632; j=0; λ=0.3288159; μ=0; α=0; ω=0

Example simulated tree, Yule+DEC+J (root state: sampled)

4 areas max., d=0; e=0; j=0.1142811; λ=0.3288159; μ=0; α=0; ω=0

Figure 3. Example simulation histories under DEC and DEC+J with parameters fit to the Hawaiian *Psychotria* dataset. 3A: DEC+Yule simulation; 3B: DEC+J+Yule simulation.

Explanation: residence times of widespread ranges under DEC. The key problem involves the concept of *residence time*. Stochastic models imply certain average residence times for every state in the state space, given a starting state, parameter values, and a time after the starting point. DEC fundamentally relies on a range-expansion process, a process that occurs rarely for typically-inferred values of *d*. For example, when *d*=0.035 in the *Psychotria* example, this suggests that range expansion events happen about once for every 28.6 million years of branch length (1/0.035). The simulated 19-species phylogenies have an average of about 60 million years of branchlength (95% CI: 32-88), suggesting that typically just a few range expansion events will occur. After the range has expanded, there will usually be a substantial waiting time until either a cladogenesis event or a range contraction event occurs. Thus, the lineages in the simulation will spend a substantial proportion of time residing in widespread ranges. The more time that lineages spend in widespread ranges, the greater the chance that one or more tips will inhabit widespread ranges when the simulation stops, and in fact, residence time in widespread ranges and the number of observed tips in widespread ranges are strongly correlated in the simulations (Figure 4).

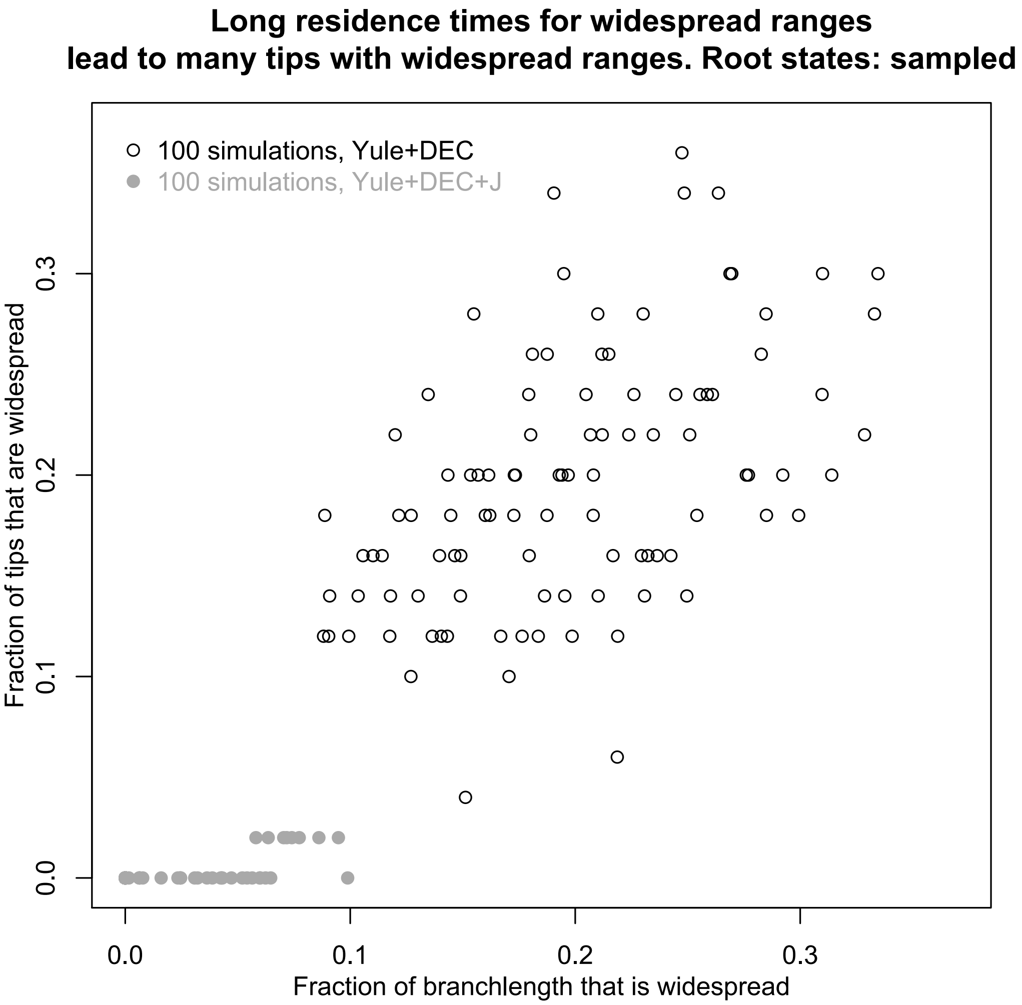


Figure 4. The correlation between the residence time spent in widespread ranges during a simulation, and the proportion of tips occupying widespread ranges at the end of the simulation. Black: DEC simulations. Grey: DEC+J simulations. The simulations correspond to rows 7 and 8 of Table 5.

On the other hand, simulations from the DEC+J model, fit to the same *Psychotria* dataset, have no difficulty producing datasets where most or all of the tips inhabit single areas (Table 5). Figure 3B shows an example simulation under this model. Even when the root states are sampled from inferred root state probabilities under DEC+J (which give some probability to the root inhabiting multiple areas), the model has no trouble producing few-to-no tips with widespread ranges.

In short, DEC+J is a better fit to datasets like Hawaiian *Psychotria* because it is more adequate – it puts high probability on states (single-area ranges) that are like the states observed in the empirical dataset (also single-area ranges). The DEC model is a poorer fit, because it puts a higher probability on producing tips with widespread ranges, which are not found in this dataset. Insofar as datasets with characteristics like the *Psychotria* dataset are common in historical biogeography, this explains the commonly, but not universally, observed advantage of the DEC+J model over the DEC model.

To put it more simply: a model that tends to produce states that are not commonly observed will tend to have a poorer fit than a model that produces states that are commonly observed. This is not "unfair" – rather, comparing how well different models explain datasets is the main point of statistics, and perhaps of science in general.

DEC's view of biogeographical evolution: slow and stately. Another crucial point may now be made. Biogeographers often plot ancestral range probabilities calculated at nodes on their phylogeny. However, they often do not think enough about what their inferred model is suggesting happens along branches. The DEC model, at least when *d* is greater than 0 and at typically-inferred values, and *e* is comparable or 0, suggests that range expansion dispersal events occur rarely in evolutionary time -- usually with average waiting times between range-expansion events of millions of years. Furthermore, it suggests that these widespread ranges then persist for, typically, millions more years, until either an independent cladogenesis event occurs that produces a daughter (or daughters) with smaller range sizes, or until a rare range-contraction event occurs. This is a slow, stately view of the evolution of geographic range.

A priori, it is perfectly possible that biogeographical evolution actually works in this slow-and-stately way. And it accords with a view of biogeography that was common in the heyday of vicariance biogeography and which presumably influenced the programming, and community reception, of DIVA and then DEC (the Critique argues that DEC is a modified version of the M*k* model, but the influence from DIVA is also strong). But with the decline of vicariance biogeography, it should also be recognized that it is perfectly possible that geographic ranges might not evolve in this slow-and-stately fashion, particularly at the coarse scales that most historical biogeography analyses are done. It might be quite rare, for instance, that a species can maintain, for millions of years, a geographic range spanning multiple continents or multiple isolated islands. The key point is that it should not be assumed that the DEC model is an "obvious" default choice for modeling biogeographical evolution. It is one possible choice among many possible models, most of which have not been programmed. It has become a near-default in the biogeographical community basically because it was the first such model available, and was relatively easy to use. But, if ten different biogeographers sat down in different rooms and individually programmed their own simulations representing their best guess at how biogeographical evolution works, would they all come up with DEC?

The importance of the state space. Once it is accepted that the residence time in different states is a crucial consideration, a variety of other interesting discussions would be worthwhile. To briefly mention them: it is clear that the choice of the states and the state space is a critical consideration for statistical model comparison, as choices that create states that are not observed in the data can have a major impact on model fit (FitzJohn, 2012; Massana et al., 2015). On the other hand, the DEC model's range-expansion/range-contraction process clearly implies the possibility of various ranges, even if they are not observed. It would be desirable to subject the state space itself to statistical model comparison, but philosophical difficulties ensue (Supp. Text).

Part 5. Time-dependent probabilities: just add Yule

The last major argument R&S offer is that the DEC+J model does not model biogeographical change as a time-dependent process. Time-dependence, the Critique claims, is a fundamental feature of all evolutionary models, and if DEC+J does not adhere to this feature, then it can and should be ignored in favour of DEC. If AIC and the usual practices of statistical model choice suggest that DEC is a much poorer fit for the data than DEC+J, then so much the worse for AIC and statistical model choice! Apparently, the fact that AIC and related methods are now ubiquitous in evolutionary biology and other quantitative sciences – the key work, Burnham and Anderson (2002), has over 50,000 citations -- can be discarded: biogeography should take a different path and prefer a traditional model regardless of the difference in statistical fit.

Connecting DEC/DEC+J to ClaSSE. There are a variety of ways to respond to the Critique's assertions about the importance of time-dependence, and its lack of time-dependence in cladogenetic models (Supp. Text). However, none of the above counterarguments need to be resolved, because it can be shown that statistical comparison of DEC and DEC+J is exactly equivalent to comparing two special cases of the ClaSSE model. The probabilities of each cladogenetic range-changing event allowed by these models, including jump dispersal, can easily be transformed into continuous-time rates by multiplying by the rate of speciation inferred by fitting a Yule process to the phylogeny. However, in a clear falsification of the Critique's argument, adding in the likelihood due to the continuous-time probability (density) of speciation events does nothing to change the log-likelihood difference between the two models, and thus does not affect the results of statistical model comparison.

I will begin by considering the ClaSSE model, the very model that the Critique endorses as an appropriate constructed model without the alleged flaws of DEC and (especially) DEC+J. ClaSSE is the most generic case of a series of models descending from the BiSSE (Binary State-dependent Speciation/Extinction) model Maddison et al., 2007. In BiSSE, there are six parameters: transition rates between two states, and speciation and extinction rates for each state. The MuSSE model (for multistate characters) is similar, but allows more than two states. The GeoSSE model (Goldberg & Igić, 2012) adds cladogenetic range change, where the various cladogenetic range-change processes allowed in DEC are given rate parameters. Thus, in a two-area system, there are three possible ranges/states (A,B,AB), each with a transition rate to the other two states (these could take the value of DEC's *d* and *e*, or other values), with an extinction rate, and with speciation rate(s). The states A and B just have a typical speciation rate, and the range-inheritance event simultaneous with speciation is just within-area speciation (A->A,A or B->B,B). However, the AB state can have multiple kinds of speciation (vicariance (A->A,B), subset sympatry (AB->A,AB), and subset sympatry (AB->B,AB)), which can have the same or different rates.

ClaSSE (Goldberg & Igić, 2012) takes this progression to the logical conclusion, where every state in the state space of size k has transition rates to every other state (k-1 transition rates per state, k*(k-1) total), an extinction rate (k extinction rates), and a speciation rate for every possible pair of states that could descend from a particular state (k*k possible types of speciation from each state, or k^3 speciation rates total). The number of parameters therefore rapidly explodes with larger state spaces, and attempting to make all parameters free and infer them from the data would be impossible except for the smallest problems. However, as noted by the Critique and previously (Matzke, 2014), numerous parameters can be fixed to zero or set to equal each other, allowing a diverse range of special case models with fewer free parameters to be built, for example to study on diversification is influenced by elevation zone (Condamine et al. 2018) or chromosomal evolution (Freyman & Höhna, 2018).

DEC and DEC+J as special cases of ClaSSE. DEC and DEC+J can be produced by setting ClaSSE parameters as follows. First, the anagenetic transition rates between states are set to the values they would take according to the DEC Q matrix, using d and e (or 2d, 3d, etc., in cases where multiple source areas are available). Second, the extinction rates for all states are set to 0. Third, the total speciation rate for each state is set to equal the Yule-estimated rate for the tree in question, λ_{Yule}. Therefore, if the possible ranges are A, B, and AB, then λ_{A}= λ_{B}= λ_{AB}=λ_{Yule}. Finally, the rates of each individual type of speciation, for example $\lambda_{\text{AB->A,B}}$, are set to be fractions of λ_{Yule}, according to the conditional probabilities that they would have DEC or the DEC+J model in question. As the conditional probabilities of cladogenetic range inheritance events in DEC and DEC+J sum to 1, $\lambda_{\text{AB->A,B}}$ equals λ_{Yule}*P(AB->A,B|AB). All of these steps are demonstrated using *diversitree* in example scripts in Supplemental Data (see also Supplemental Figure 1).

To complete the equivalency between ClaSSE comparisons and DEC/DEC+J comparisons, state frequencies at the root must be set to be equal (diversitree ROOT.EQUI option), or equal except with 0 probability assigned to the null state (ROOT.GIVEN option). Finally, the completeness of taxon sampling is set to 1, and the option to condition on tree survival is set to false.

Experiment: Comparing log-likelihoods calculated under DEC/DEC+J and ClaSSE. Employing this method, I calculated the likelihood of a constructed example dataset using *BioGeoBEARS* DEC and DEC+J, and the special-case ClaSSE equivalents in diversitree. The example dataset was a 4-taxon tree using the standard hominoid topology (((human,chimp),gorilla),orang), but with the shortest branches given length 1, and the other branches scaled to make the tree ultrametric. The ranges were set to B ("Asia") for orangs, and A ("Africa") for the others. Obviously, this is not meant to resemble the real phylogenetic dates or a real analysis, rather, the purpose is to compare likelihoods between independently programmed implementations. Likelihoods were calculated across a wide range of combinations of the *d*, *e*, and *j* parameters (*d*, *e*: 0, 0.01, 0.05, 0.1, 0.3, 1, 3, 5; *j*: 0, 0.01, 0.05, 0.1, 0.3, 1, 2.99).

The results are plotted in Figure 5. It is apparent that the likelihoods conferred by *BioGeoBEARS* DEC and DEC+J, and an implementation of these models in diversitree's ClaSSE, are extremely highly correlated across many orders of magnitude and many parameter combinations. The match is not quite perfect (slope 0.9951, R-squared=0.9991), but the strategy used to calculate the likelihood in SSE models (Maddison et al., 2007) is a numeric approximation and not analytically exact, so minor disagreements are expected. The Critique's statement, "Needless to say, ClaSSE likelihoods cannot be compared with DEC and DEC+J" is incorrect – DEC and DEC+J likelihoods differ from their respective ClaSSE submodel likelihoods only by terms that are constant across both models. Exclusion of constant terms in likelihood comparisons is a ubiquitous practice (e.g. Burnham & Anderson, 2002; Stadler, 2013).

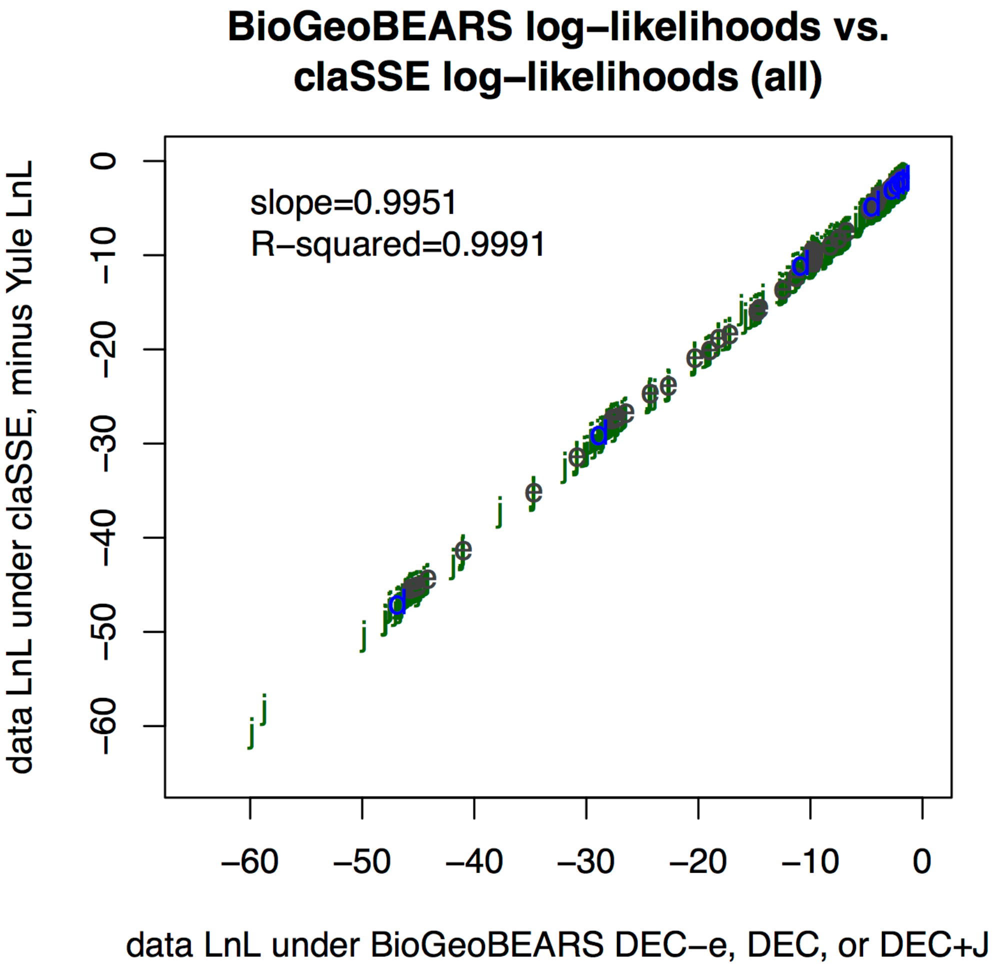


Figure 5. Log-likelihoods calculated under BioGeoBEARS DEC and DEC+J for a wide range of parameter values, regressed against log-likelihoods calculated under diversitree ClaSSE models set up to produce DEC and DEC+J behavior (as in Figure 5). The letter *d* represents the DEC model when *e*=0, *j*=0. The letter *e* represents the DEC model when *d*>0, *e*>0, *j*=0. The letter *j* represents the DEC+J model when *j*>0.

Implications of DEC and DEC+J being submodels of ClaSSE. The failure to recognize that the continuous-time rate of any particular DEC or DEC+J cladogenetic range-inheritance event is just the Yule speciation rate, times the conditional probability of that cladogenetic range-inheritance event, is the key missing conceptual step in the Critique's analysis of DEC and DEC+J. The consequences for the Critique's argument are severe. First, the claim that DEC and DEC+J are conceptually flawed because they invoke processes that do not act in continuous time is rendered untrue once the Yule log-likelihood is added. Second, the notion that this conceptual flaw would differentially advantage DEC+J over DEC is falsified, because the likelihood contribution of the Yule process is the same for both models – both models are using the same phylogeny, and they both have the same number and timing of cladogenetic events to work with. Under a Yule process, possibilities like extinct lineages or incomplete taxon sampling are ruled out by assumption, therefore there are no missing speciation events in the tree. (We will return to the dubious validity of that assumption momentarily, but Yule models are universally recognized to be valid evolutionary models; of course, valid models may not be correct; "all models are wrong" as famously stated by Box).

Finally, because the Critique itself endorses ClaSSE as an appropriate model for biogeography, the fact that the DEC/DEC+J comparison has been shown to be identical to comparing two special cases of ClaSSE introduces an irreconcilable self-contradiction into the Critique's core argument. Also, the Critique's claim that "DEC and DEC+J cannot be compared using standard statistical methods" must be considered false by anyone who recognizes that various submodels of ClaSSE can be statistically compared using AIC and related methods. As BiSSE, GeoSSE, etc. are submodels of ClaSSE, every paper that ever compared two or more SSE-derived models has done such statistical comparisons -- probably this includes every paper ever published using an SSE model.

Conclusion: legitimate criticisms, and the path forward in biogeography

The framing of DEC and DEC+J as special cases of ClaSSE has another advantage: it gives us a clearer view of what these models are assuming, and thus a better vantage point from which to critique these assumptions and construct a research program to test them.

Extinction and sampling. The biggest criticism that can be made of DEC and DEC+J is that they both make the Yule assumption of no lineage extinction, as previously mentioned in Matzke (2014). The no-extinction assumption has to be false for most clades (Marshall, 2017), and in addition, taxon experts often know that their phylogeny suffers from incomplete sampling of the total diversity of their study group (a problem similar to extinction). The flawed assumptions of DEC and DEC+J (and other similar models) could well have real impacts on inference of parameters and ancestral ranges, because the existence of extinct or unsampled lineages means that there are unobserved cladogenetic events in the tree, which is problematic if cladogenetic range-changing processes are operating. On the other hand, the simulations performed by Matzke (2014) included scenarios with lineage extinction and other violations of DEC/DEC+J assumptions, with detectable but quite limited negative impacts on parameter inference and model comparison. As noted in Matzke (2014), at least when extinction has a moderate rate and is geographically unbiased, ignoring it does not seem to cause fatal problems for the usage of DEC and DEC+J, and comparison of the same. Of course, if extinction rates have been large or biased, then the problems could well be severe. However, this is not much different than the situation with other phylogenetic comparative methods. For phylogenetic inference of ancestral body sizes and the rates of body size evolution, severely mistaken inference could occur if extinction or taxon sampling had biases correlated with body size.

State-dependence of speciation and extinction. Another major assumption of DEC and DEC+J is the state-*in*dependence of the overall speciation rate. The assumed speciation rate is the same for all states (except for the null range, where it is 0, although setting the speciation rate for the null range to other values does not change the likelihood, as the null range is an impossible ancestor at all points in the tree). There is state-dependence of speciation only in the limited sense that probabilities of specific range-inheritance scenarios, given speciation, are conditional on the range before speciation.

Advantages of framing DEC and DEC+J as ClaSSE submodels. The framing of DEC and DEC+J as special cases of ClaSSE is useful, in that it invites us to focus on the potential of statistically comparing these simple models with more complex ones in a ClaSSE framework. For example, it is clear that lineage extinction is a major process in evolutionary history, but its inference is fraught (Marshall, 2017). What dataset sizes and characteristics will retain sufficient information to favour a biogeographical model with lineage extinction? Similar questions could be asked about state-dependence versus state-independence for speciation and extinction, and for different types of cladogenetic range-changing processes (vicariance, jump-dispersal, etc.). Once the paradigm of statistical model comparison is routinely accepted in historical biogeography, all such questions become statistically accessible. Historical biogeography can thus join the larger world of macroevolutionary modeling, using rigorous statistical methods to measure the support that data lend to different models. This would be a marked improvement over the history of historical biogeography, which has been marked by contentious, non-statistical debates over assumptions.

Routine use of statistical model comparison would also be an improvement over some common practices in the historical biogeography literature of recent years, where the biogeographical portion of research papers would often follow one of the two following stereotyped strategies. The first common practice has been to run the three methods available in RASP -- parsimony DIVA, likelihood-based Lagrange DEC, and the Bayesian Binary Model (more recently BayArea). As these are three different inference frameworks, the results are not statistically comparable, and researchers were stuck either presenting all three, or picking one based on intuition or prior beliefs. The second common practice was to choose just one model, commonly DEC or DIVA, and treat the result as "the reconstruction" of biogeographical history, with little consideration of whether the assumptions of these models were good fits for the data at hand.

BioGeoBEARS, by providing likelihood-based implementations of all of these models, and adding the option for jump dispersal to each model, took a first step towards remedying the limitations of previous practice. However, all of these models still have obvious flaws, and the point of *BioGeoBEARS* and the DEC+J model was not to provide the One True Model – which science will never reach in any case – but to open up some of the assumptions made by DEC and other models to statistical testing. Once statistical model comparison has been accepted as a framework, the way is open for advancing our understanding of biogeographical process, arguably a more important scientific goal than merely inferring ancestral ranges on a phylogeny. This approach has been used to explore the role of distance, connectivity, and dispersal-modifying traits (Van Dam & Matzke, 2016; Dupin et al., 2017; Matos-Maraví et al., 2018; Klaus & Matzke, 2020; Garcia-R & Matzke, 2021).

Prospects for practical use of ClaSSE in biogeography. The process of model building and statistical testing will continue in historical biogeography. I therefore endorse R&S's recommendation to explore the use of ClaSSE for biogeographical problems. However, as noted in Matzke (2014), there are substantial technical challenges for doing so. Apart from the parameter explosion, computational speed is a major problem. The SSE models use numerical integration on a series of differential equations to approximate the likelihood, as analytic strategies such as matrix exponentiation (used in DEC-like models and innumerable other phylogenetic models) are not available. Many users of *Lagrange* or *BioGeoBEARS* users have already run up against the quite strict limits of matrix exponentiation as a strategy: once the state space grows beyond 1500 or 2000 possible ranges, the likelihood calculations become so slow that ML optimization runs will not complete in any reasonable amount of time. Multicore parallel processing, implemented in *BioGeoBEARS*, allows the state space to be pushed only a little further. The numeric integration method used in SSE models, on the other hand, is even slower, thus imposing even more strict limits on the state space. Even if the calculation issues are overcome, allowing ClaSSE to be run on normal biogeographical problems with 100-2000 possible ranges, there will be a fairly strict limit to the complexity of models that can be supported given existing or imaginable biogeographical datasets. DEC and DEC+J may well be overly simple, but any ClaSSE model feasible for inference will have to similarly use a small set of parameters to describe the most important processes. The goal should be to find simple ClaSSE models that also can pass basic model adequacy tests. The issue of the residence times of range sizes, in particular, will undoubtedly recur as soon as ClaSSE-derived models are employed in biogeography, and any models that fail to take this into account will fare poorly.

The advantage of statistical model comparison in biogeography is that it allows researchers with differing intuitions about how biogeography works to statistically test them by comparing model fits. Frequent model testing would be an indication that historical biogeography is a vibrant and advancing science. The Critique seems to suggest that the main point of probabilistic models in historical biogeography is to infer ancestral ranges, certainly a common view in the literature. However, most important use of probabilistic models in historical biogeography should be to learn about the processes that have produced the distribution of biodiversity around the globe. A beneficial by-product of comparing models is, of course, that better-fit models are likely to be more accurate for estimating ancestral ranges, as demonstrated by the simulation-inference studies in Matzke (2014). Better inference of ancestral ranges is good, but learning about the fundamental processes working in biogeography is a far more fundamental scientific goal than just inferring the history of individual clades. By comparing models on many clades, we can begin to understand the variation in the importance of different processes, for example depending on the physical geography, habitat, and ecological dispersal ability of different taxa. In order for this research program to advance, the one thing researchers must be prepared to do is revise their intuitions, if statistical evidence starts accumulating against them.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723. doi:10.1109/TAC.1974.1100705

Anisimova, M., Bielawski, J. P., & Yang, Z. (2001). Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution. *Molecular Biology and Evolution, 18*(8), 1585-1592. doi:10.1093/oxfordjournals.molbev.a003945

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach, second edition*. New York: Springer-Verlag.

Clark, J. R., Wagner, W. L., & Roalson, E. H. (2009). Patterns of diversification and ancestral range reconstruction in the southeast Asian–Pacific angiosperm lineage *Cyrtandra* (Gesneriaceae). *Molecular Phylogenetics and Evolution, 53*(3), 982-994. doi:10.1016/j.ympev.2009.09.002

Dupin, J., Matzke, N. J., Särkinen, T., Knapp, S., Olmstead, R. G., Bohs, L., & Smith, S. D. (2017). Bayesian estimation of the global biogeographical history of the Solanaceae. *Journal of Biogeography, 44*(4), 887-899. doi:10.1111/jbi.12898

Felsenstein, J. (1992). Phylogenies from restriction sites: a maximum-likelihood approach. *Evolution, 46*(1), 159-173. doi:10.2307/2409811

FitzJohn, R. G. (2012). Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution, 3*(6), 1084-1092. doi:https://doi.org/10.1111/j.2041-210X.2012.00234.x

Franklin, A. B., Shenk, T. M., Anderson, D. R., & Burnham, K. P. (2001). Statistical Model Selection: An Alternative to Null Hypothesis Testing. In T. M. Shenk & A. B. Franklin (Eds.), *Modeling in Natural Resource Management: Development, Interpretation, and Application* (pp. 75-90). Washington: Island Press.

Freyman, W. A., & Höhna, S. (2018). Cladogenetic and Anagenetic Models of Chromosome Number Evolution: A Bayesian Model Averaging Approach. *Systematic Biology, 67*(2), 195-215. doi:10.1093/sysbio/syx065

Garcia-R, J. C., & Matzke, N. J. (2021). Trait-dependent dispersal in rails (Aves: Rallidae): Historical biogeography of a cosmopolitan bird clade. *Molecular Phylogenetics and Evolution, 159*, 107106. doi:https://doi.org/10.1016/j.ympev.2021.107106

Goldberg, E. E., & Igić, B. (2012). Tempo and mode in plant breeding system evolution. *Evolution, 66*(12), 3701-3709. doi:10.1111/j.1558-5646.2012.01730.x

Klaus, K., & Matzke, N. J. (2020). Statistical Comparison of Trait-dependent Biogeographical Models indicates that Podocarpaceae Dispersal is influenced by both Seed Cone Traits and Geographical Distance. *Systematic Biology, 69*(1), 61-75. doi:10.1093/sysbio/syz034

Maddison, W. P., Midford, P. E., & Otto, S. P. (2007). Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology, 56*(5), 701-710. doi:10.1080/10635150701607033

Marshall, C. R. (2017). Five palaeobiological laws needed to understand the evolution of the living biota. *Nature Ecology & Evolution, 1*, 0165. doi:10.1038/s41559-017-0165

Massana, K. A., Beaulieu, J. M., Matzke, N. J., & Meara, B. C. (2015). Non-null Effects of the Null Range in Biogeographic Models: Exploring Parameter Estimation in the DEC Model. *bioRxiv*. doi:10.1101/026914

Matos-Maraví, P., Matzke, N. J., Larabee, F. J., Clouse, R. M., Wheeler, W. C., Sorger, D. M., . . . Janda, M. (2018). Taxon cycle predictions supported by model-based inference in Indo-Pacific trap-jaw ants (Hymenoptera: Formicidae: *Odontomachus*). *Molecular Ecology, 27*(20), 4090-4107. doi:10.1111/mec.14835

Matzke, N. J. (2013). Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. *Frontiers of Biogeography, 5*(4), 242-248.

Matzke, N. J. (2014). Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. *Systematic Biology, 63*(3), 951-970. doi:10.1093/sysbio/syu056

Pennell, M. W., FitzJohn, R. G., Cornwell, W. K., & Harmon, L. J. (2015). Model Adequacy and the Macroevolution of Angiosperm Functional Traits. *The American Naturalist, 186*(2), E33-E50. doi:10.1086/682022

Ree, R. H., & Sanmartín, I. (2018). Conceptual and statistical problems with the DEC+J model of founder-event speciation and its comparison with DEC via model selection. *Journal of Biogeography, 45*(4), 741-749. doi:https://doi.org/10.1111/jbi.13173

Ree, R. H., & Smith, S. A. (2008). Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology, 57*(1), 4-14. doi:10.1080/10635150701883881

Ripplinger, J., & Sullivan, J. (2010). Assessment of Substitution Model Adequacy Using Frequentist and Bayesian Methods. *Molecular Biology and Evolution, 27*(12), 2790-2803. doi:10.1093/molbev/msq168

Stadler, T. (2013). How Can We Improve Accuracy of Macroevolutionary Rate Estimates? *Systematic Biology, 62*(2), 321-329. doi:10.1093/sysbio/sys073

Tejero-Cicuéndez, H., Patton, A. H., Caetano, D. S., Šmíd, J., Harmon, L. J., & Carranza, S. (2021). Reconstructing Squamate Biogeography in Afro-Arabia Reveals the Influence of a Complex and Dynamic Geologic Past. *Systematic Biology*. doi:10.1093/sysbio/syab025

Van Dam, M. H., & Matzke, N. J. (2016). Evaluating the influence of connectivity and distance on biogeographical patterns in the south-western deserts of North America. *Journal of Biogeography, 43*(8), 1514-1532. doi:10.1111/jbi.12727

Conflict of Interest

The author declares no conflicts of interest.

Data Availability

All scripts and input/output files for analyses discussed in the paper are permanently available on FigShare at http://dx.doi.org/10.6084/m9.figshare.19166393 , or on GitHub at: https://github.com/nmatzke/M21 .

Biosketch

Nicholas J. Matzke is a Senior Lecturer in the School of Biological Sciences at the University of Auckland, New Zealand. He is the author of *BioGeoBEARS*.