

Statistical comparison of DEC and DEC+J is statistically valid;

...and, a path towards more
statistical sophistication and realism
in phylogenetic biogeography



Nicholas J. Matzke, Senior Lecturer, School of
Biological Sciences, University of Auckland
n.matzke@auckland.ac.nz



Wallis Bland, Ph.D. Candidate, School of Biological
Sciences, University of Auckland

Apologies for recording...

It's 1:50 am in New Zealand and if I wake these up I'm in trouble...

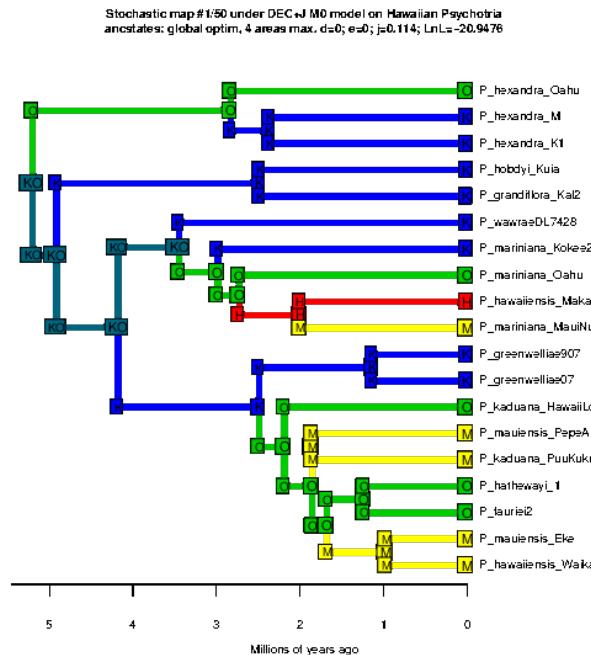
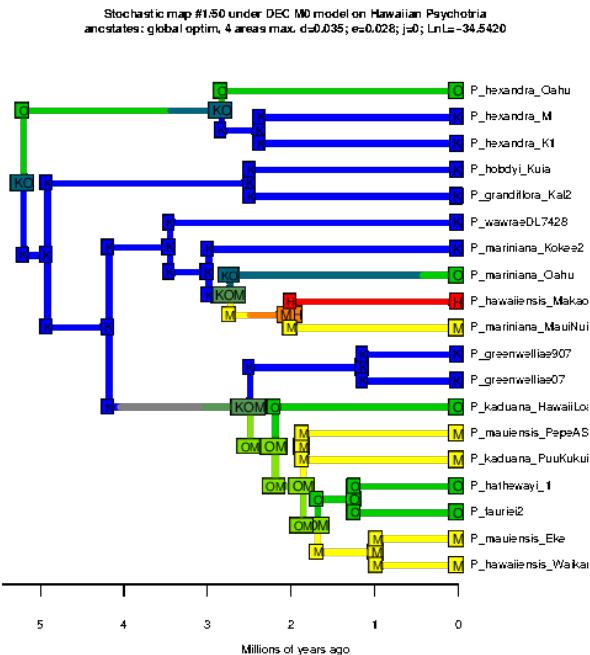


If someone can please relay questions to me / save them, I can answer online sometime this week.

Example BioGeoBEARS analysis: Hawaiian *Psychotria*

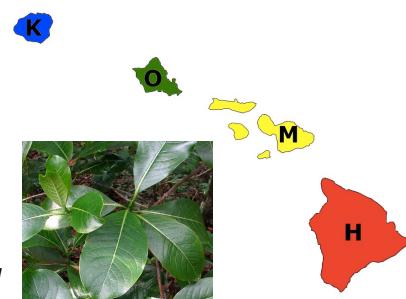
DEC lnL = -34.5

DEC+J lnL = -20.9



**This is about a
300,000x
improvement
in model fit**

**possible
histories
(DEC model)**

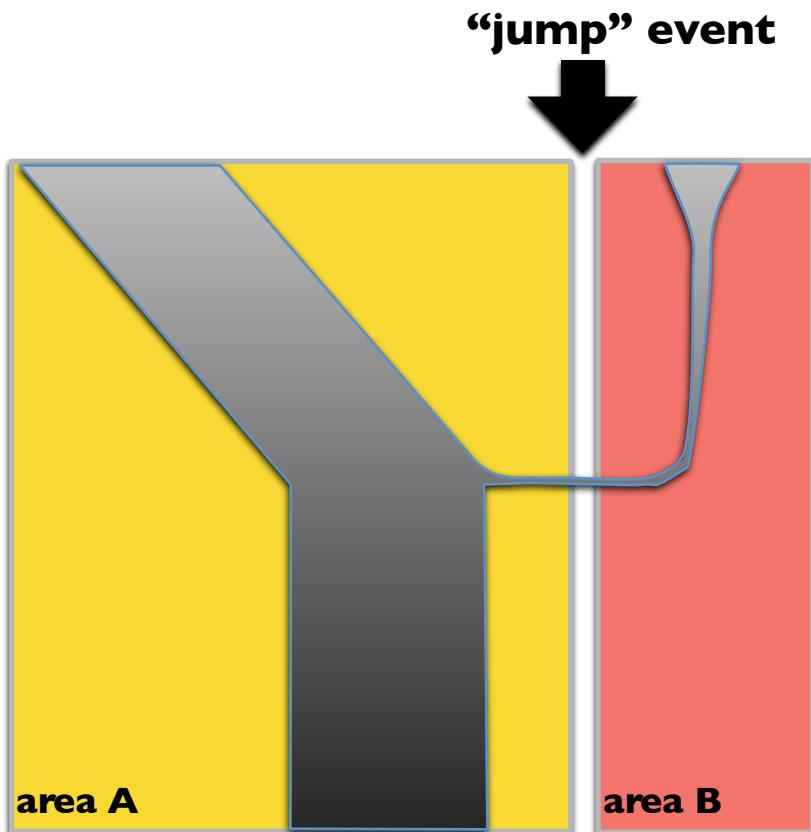


**possible
histories
(DEC+J model)**

Thanks for support



Founder-event speciation



- In founder-event speciation, a rare dispersal event “instantaneously” establishes a geographically isolated new lineage with one or a few individuals
- Widely discussed in population genetics, speciation literature, island biogeography
- However, ignored in traditional historical biogeography computer models

**I think we should use
statistical model choice
in biogeography**

E.g. comparison of two models:

- 1. model without founder-event speciation**
- 2. model with founder-event speciation**

	Process	Ranges Before	Ranges After	Character mapping	DIVA	DEC (LAGRANGE)	BayArea, BBM (RASP)
Anagenetic	Dispersal				✓	✓	✓
	Extinction				✓	✓	✓
	Range-switching				✓		
Cladogenetic	Sympatry (narrow)				✓	✓	✓
	Sympatry (widespread)						✓
	Sympatry (subset)					✓	
Cladogenetic	Vicariance (narrow)				✓	✓	
	Vicariance (widespread)				✓		
	Founder						

Which model should we use?

What about founder-event speciation?

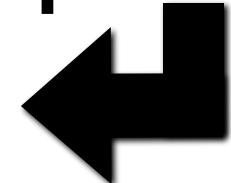


Figure 1, Matzke 2013, *Frontiers of Biogeography*

ORIGINAL
ARTICLE



Bayesian estimation of the global biogeographical history of the Solanaceae

Julia Dupin^{1*}, Nicholas J. Matzke², Tiina Särkinen³, Sandra Knapp⁴,
Richard G. Olmstead⁵, Lynn Bohs⁶ and Stacey D. Smith¹





ORIGINAL
ARTICLE



Bayesian estimation of the global biogeographical history of the Solanaceae

Julia Dupin^{1*}, Nicholas J. Matzke², Tiina Särkinen³, Sandra Knapp⁴,
Richard G. Olmstead⁵, Lynn Bohs⁶ and Stacey D. Smith¹

50 to 24 Ma



	SAm	CAm	Car	NAm	AF	EU	OZ
SAm	1	0.4	0.4	0.2	0.1	0.1	0.1
CAm	0.4	1	0.7	0.4	0.1	0.1	0.1
Car	0.4	0.7	1	0.4	0.1	0.1	0.1
NAm	0.2	0.4	0.4	1	0.7	1	0.3
AF	0.1	0.1	0.1	0.7	1	0.7	0.3
EU	0.1	0.1	0.1	0.8	0.7	1	0.3
OZ	0.1	0.1	0.1	0.3	0.3	0.3	1

24 to 10 Ma



	SAm	CAm	Car	NAm	AF	EU	OZ
SAm	1	0.6	0.5	0.5	0.1	0.2	0.1
CAm	0.6	1	0.7	0.6	0.1	0.1	0.1
Car	0.5	0.7	1	0.5	0.1	0.1	0.1
NAm	0.5	0.6	0.5	1	0.4	0.5	0.1
AF	0.1	0.1	0.1	0.4	1	1	0.3
EU	0.2	0.1	0.1	0.5	1	1	0.5
OZ	0.1	0.1	0.1	0.1	0.3	0.5	1

10 Ma to Present



	SAm	CAm	Car	NAm	AF	EU	OZ
SAm	1	1	0.75	1	0.1	0.2	0.1
CAm	1	1	0.7	1	0.1	0.1	0.1
Car	0.75	0.7	1	0.8	0.1	0.1	0.1
NAm	1	1	0.8	1	0.1	0.2	0.1
AF	0.1	0.1	0.1	0.1	1	1	0.7
EU	0.2	0.1	0.1	0.2	1	1	0.7
OZ	0.1	0.1	0.1	0.1	0.7	0.7	1

T
i
m
e



Global Solanaceae: distance, Biogeographical Stochastic Mapping

#	Models	Dispersal multipliers	Free parameters					Log-Likelihood	AIC	Δ AIC	AIC weights
			Number	$d^{(1)}$	$e^{(2)}$	$j^{(3)}$	$w^{(4)}$				
1	Basic Models										
1	DEC_NonTS	Non-TS	2	0.011	0.000	0.000	1.000	-1341.413	2686.826	317.604	0.000
2	DIVALIKE_NonTS	Non-TS	2	0.013	0.000	0.000	1.000	-1401.936	2807.871	438.648	0.000
3	BayAreaLIKE_NonTS	Non-TS	2	0.009	0.049	0.000	1.000	-1528.828	3061.656	692.433	0.000
	Time-Stratified Models										
4	DEC_TS	TS	2	0.025	0.000	0.000	1.000	-1210.832	2425.663	56.441	0.000
5	DIVALIKE_TS	TS	2	0.029	0.000	0.000	1.000	-1252.831	2509.661	140.438	0.000
6	BayAreaLIKE_TS	TS	2	0.023	0.044	0.000	1.000	-1430.614	2865.227	496.004	0.000
	+j Models										
7	DEC_NonTS_j	Non-TS	3	0.010	0.000	0.003	1.000	-1324.525	2655.050	285.828	0.000
8	DIVALIKE_NonTS_j	Non-TS	3	0.011	0.000	0.003	1.000	-1387.768	2781.536	412.313	0.000
9	BayAreaLIKE_NonTS_j	Non-TS	3	0.007	0.004	0.009	1.000	-1335.728	2677.457	308.234	0.000
10	DEC_TS_j	TS	3	0.022	0.000	0.008	1.000	-1191.907	2389.813	20.590	0.000
11	DIVALIKE_TS_j	TS	3	0.025	0.000	0.007	1.000	-1239.123	2484.246	115.024	0.000
12	BayAreaLIKE_TS_j	TS	3	0.015	0.001	0.020	1.000	-1213.992	2433.984	64.762	0.000
	+w Models										
13	DEC_TS_w	TS	3	0.026	0.000	0.000	1.114	-1204.980	2415.960	46.738	0.000
14	DIVALIKE_TS_w	TS	3	0.032	0.000	0.000	1.201	-1250.615	2507.231	138.008	0.000
15	BayAreaLIKE_TS_w	TS	3	0.026	0.045	0.000	1.307	-1425.518	2857.036	487.813	0.000
	+j +w Models										
16	DEC_TS_j_w	TS	4	0.029	0.000	0.009	1.887	-1180.611	2369.223	0.000	0.999
17	DIVALIKE_TS_j_w	TS	4	0.027	0.000	0.007	1.344	-1221.864	2451.729	82.506	0.000
18	BayAreaLIKE_TS_j_w	TS	4	0.016	0.001	0.021	1.117	-1210.247	2428.494	59.271	0.000

Global Solanaceae: distance, Biogeographical Stochastic Mapping

#	Models	Dispersal multipliers	Free parameters					Log-Likelihood	AIC	Δ AIC	AIC weights
			Number	$d^{(1)}$	$e^{(2)}$	$j^{(3)}$	$w^{(4)}$				
1	DEC_NonTS	Non-TS	2	0.011	0.000	0.000	1.000	-1341.413	2686.826	317.604	0.000
2	DIVALIKE_NonTS	Non-TS	2	0.013	0.000	0.000	1.000	-1401.936	2807.871	438.648	0.000
3	BayAreaLIKE_NonTS	Non-TS	2	0.009	0.049	0.000	1.000	-1528.828	3061.656	692.433	0.000
Time-Stratified Models											
4	DEC_TS	TS	2	0.025	0.000	0.000	1.000	-1210.832	2425.663	56.441	0.000
5	DIVALIKE_TS	TS	2	0.029	0.000	0.000	1.000	-1252.831	2509.661	140.438	0.000
6	BayAreaLIKE_TS	TS	2	0.023	0.044	0.000	1.000	-1430.614	2865.227	496.004	0.000
$+j$ Models											
7	DEC_NonTS_j	Non-TS	3	0.010	0.000	0.003	1.000	-1324.525	2655.050	285.828	0.000
8	DIVALIKE_NonTS_j	Non-TS	3	0.011	0.000	0.003	1.000	-1387.768	2781.536	412.313	0.000
9	BayAreaLIKE_NonTS_j	Non-TS	3	0.007	0.004	0.009	1.000	-1335.728	2677.457	308.234	0.000
10	DEC_TS_j	TS	3	0.022	0.000	0.008	1.000	-1191.907	2389.813	20.590	0.000
11	DIVALIKE_TS_j	TS	3	0.025	0.000	0.007	1.000	-1239.123	2484.246	115.024	0.000
12	BayAreaLIKE_TS_j	TS	3	0.015	0.001	0.020	1.000	-1213.992	2433.984	64.762	0.000
$+w$ Models											
13	DEC_TS_w	TS	3	0.026	0.000	0.000	1.114	-1204.980	2415.960	46.738	0.000
14	DIVALIKE_TS_w	TS	3	0.032	0.000	0.000	1.201	-1250.615	2507.231	138.008	0.000
15	BayAreaLIKE_TS_w	TS	3	0.025	0.045	0.000	1.107	-1255.518	2557.026	187.813	0.000
$+j+w$ Models											
16	DEC_TS_j_w	TS	4	0.029	0.000	0.009	1.887	-1180.611	2369.223	0.000	0.999
17	---	---	---	---	---	---	---	---	---	---	---
18	BayAreaLIKE_TS_j_w	TS	4	0.016	0.001	0.021	1.117	-1210.247	2428.494	59.271	0.000

Global Solanaceae: distance, Biogeographical Stochastic Mapping



Global Solanaceae: distance, Biogeographical Stochastic Mapping



Bayesian estimation of Solanaceae biogeography

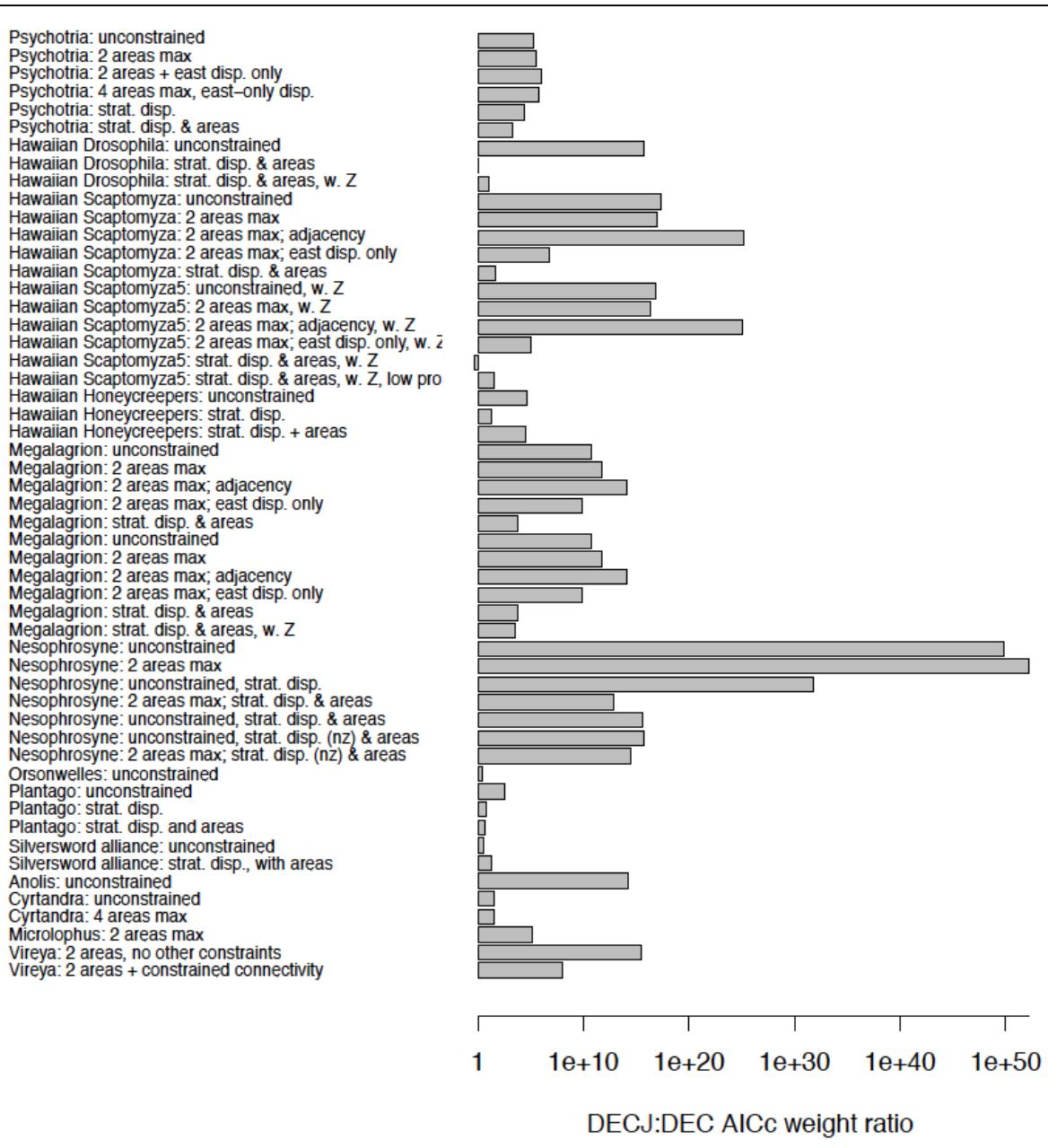
(a) Summary of dispersal events counts (and standard deviations)

	SAm	CAm	Car	NAm	AF	EU	OZ	
SAm	-	48.6 (2.63)	15.13 (1.62)	41.65 (2.83)	5.67 (0.89)	5.03 (1.14)	3.58 (0.77)	119.66 47%
CAm	11.35 (2.32)	-	10.12 (1.41)	23.26 (4.21)	0.22 (0.19)	0.35 (0.34)	0.32 (0.31)	45.62 18%
Car	1.21 (0.94)	4.20 (1.14)	-	3.09 (1.02)	0.22 (0.27)	0.11 (0.17)	0.14 (0.08)	8.97 4%
NAm	9.12 (1.51)	32.44 (4.01)	7.39 (1.20)	-	0.59 (0.29)	2.33 (0.78)	1.34 (0.52)	53.21 20%
AF	0.03 (0.21)	0.04 (0.22)	0.07 (0.33)	0.13 (0.31)	-	10.15 (1.39)	3.89 (0.83)	14.31 6%
EU	0.09 (0.22)	0.08 (0.32)	0.11 (0.21)	1.17 (0.32)	8.28 (1.32)	-	1.78 (0.37)	11.51 4%
OZ	0.06 (0.12)	0.04 (0.21)	0.04 (0.11)	0.12 (0.34)	1.39 (0.71)	0.88 (0.27)	-	2.53 1%
	21.86 9%	85.4 33%	32.86 13%	69.42 27%	16.37 6%	18.85 8%	11.05 4%	255.81 100%

DEC+J advantage is *very* common...

13 island clades,
many under
multiple geology
models

Ratio of
model
weights:
DEC+J:DEC



Matzke (2013),
*Frontiers in
Biogeography*

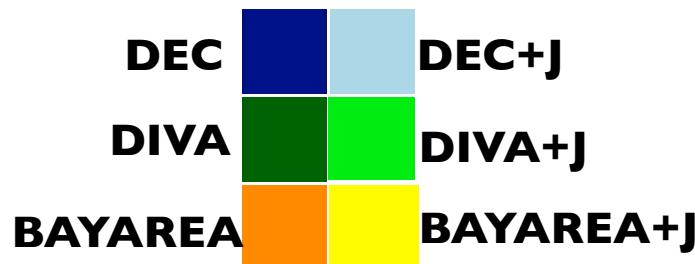
Matzke (2014),
*Systematic
Biology*

AICc model weights

DEC vs. DIVA vs. BayArea

...versus...

DEC+J vs. DIVA+J vs. BayArea+J

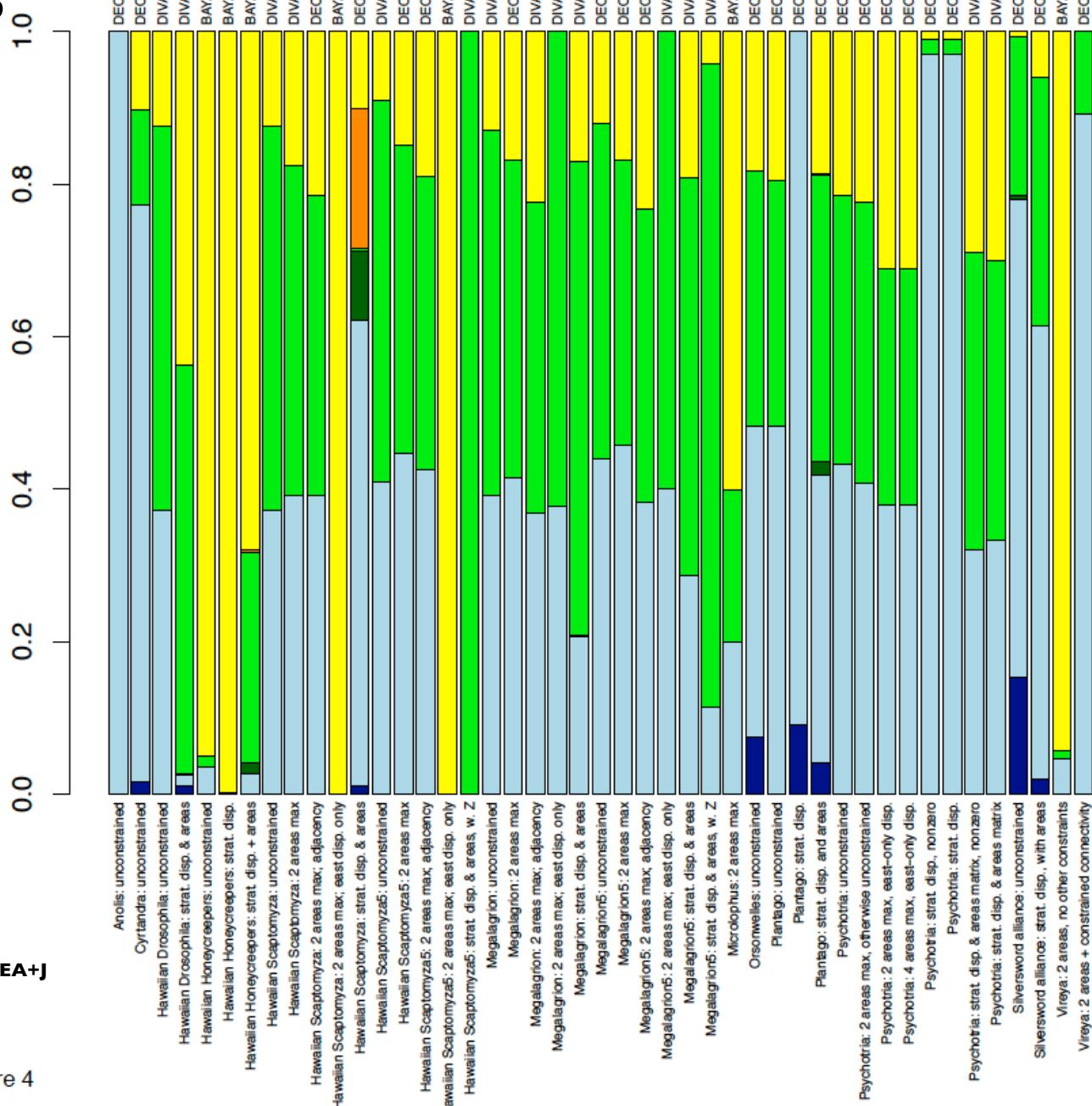


Islands

Figure 4



model probability



...but, *not* universal:



BioGeoBEARS Beer Challenge:

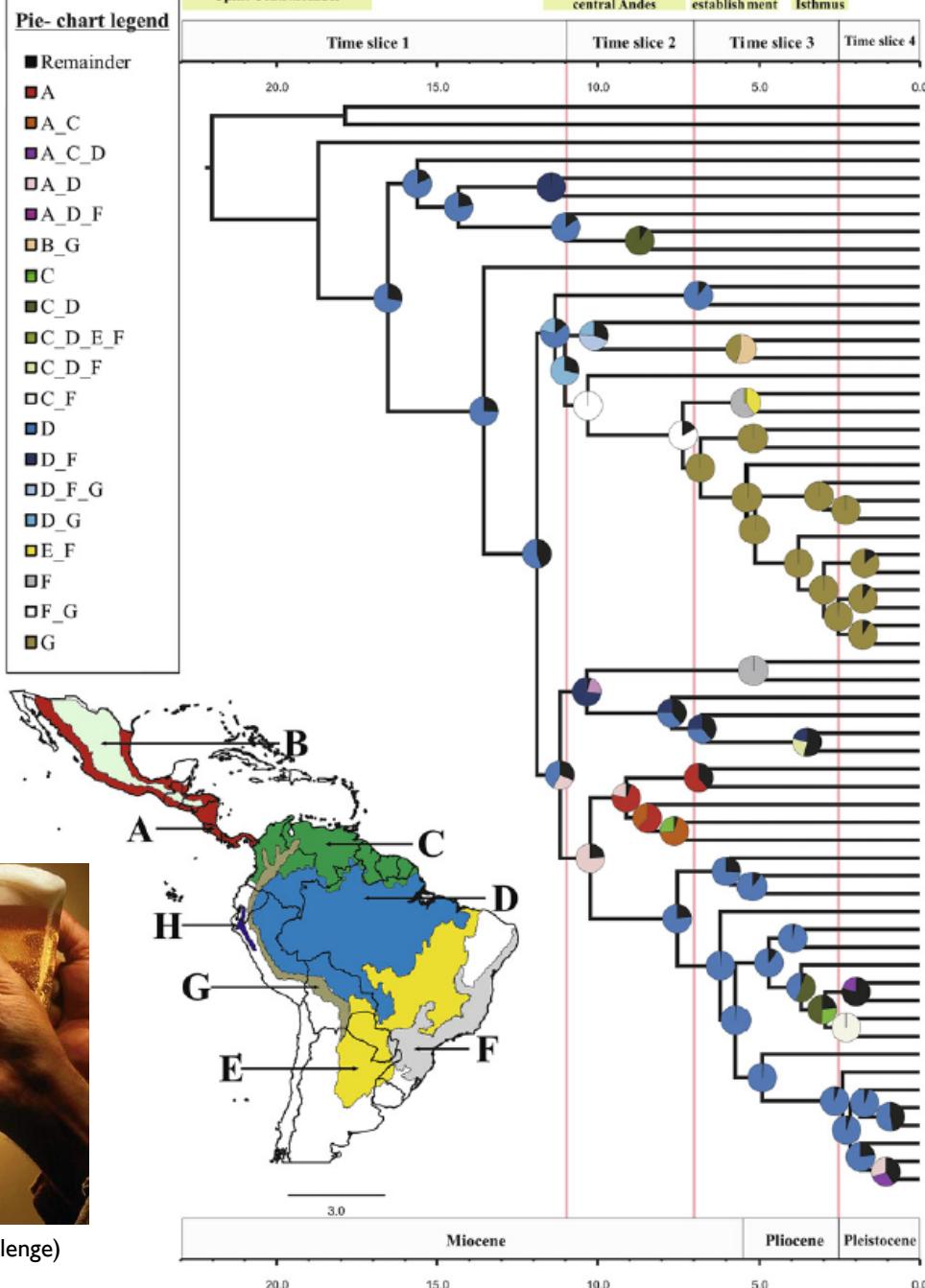
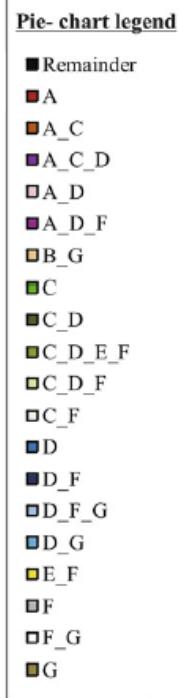
Find an empirical dataset that does *not* reject DEC



Taygetis
butterflies
favor
DEC



(winner of Nick Matzke beer challenge)



The Critique:

PERSPECTIVE

WILEY

Journal of
Biogeography



Conceptual and statistical problems with the DEC+J model of founder-event speciation and its comparison with DEC via model selection

Richard H. Ree¹

Isabel Sanmartín²

¹Life Sciences Section, Integrative Research Center, The Field Museum,
Chicago, IL, USA

²Real Jardín Botánico, RJB-CSIC, Madrid, Spain

Correspondence

Richard H. Ree, Life Sciences Section, Integrative Research Center, The
Field Museum, Chicago, IL, USA.

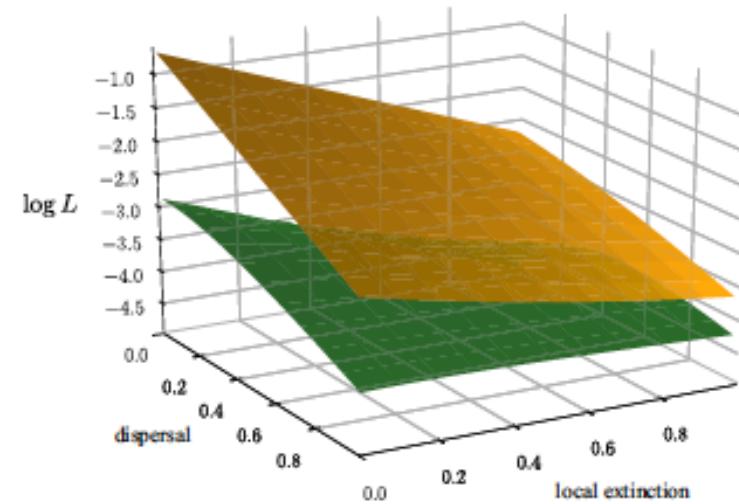
Email: rree@fieldmuseum.org

2.4 | Cladogenetic range evolution is not time dependent: a conceptual flaw, exacerbated in DEC+J

2.5 | Degenerate likelihoods in DEC and DEC+J: A worked example

2.6 | Pathological consequences of degeneracy in DEC+J

DEC+J differs from DEC in expanding the scope and influence of cladogenetic range evolution, both by allowing jump dispersal scenarios and assigning a free parameter to their partial likelihood



The paper has been influential...

Conceptual and statistical problems with the DEC+ J model of founder-event speciation and its comparison with DEC via model selection

RH Ree, I Sanmartín - Journal of Biogeography, 2018 - Wiley Online Library

Phylogenetic studies of geographic range evolution are increasingly using statistical model selection methods to choose among variants of the dispersal-extinction-cladogenesis (DEC) model, especially between DEC and DEC+ J, a variant that emphasizes "jump ...

☆ 99 Cited by 252 Related articles All 6 versions

The authors are distinguished leaders in the field, editors at the *Journal of Biogeography*, and major authorities.

The paper is now very often cited in papers to justify just running DEC and ignoring DEC+J, no matter what the difference is in log-likelihood or AIC weights

The citations are almost always uncritical.

Unfortunately...the Critique make many mistakes, large and small

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. Matzke¹

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

n.matzke@auckland.ac.nz

<https://osf.io/vqm7r>

1. InLs are wrong (disagree with BioGeoBEARS and Lagrange)
2. Examples are tiny (2, 4 species) and brittle
3. BioGeoBEARS likelihoods properly add up, there is no anti-DEC bias
4. True source of DEC+J advantage is model adequacy
5. ClaSSE actually is directly comparable, basically, just add Yule

But first, let's back up!

The Critique, stripped down to its essence, basically says: let's just ignore likelihood and AIC, and choose models another way.

For simple inference of ancestral ranges on a fixed phylogeny, a DEC-based model may be defensible if statistical model selection is not used to justify the choice, and it is understood that inferences about cladogenetic range inheritance lack any relation to time, normally a fundamental axis of evolutionary models.

THINGS THAT MAKE YOU GO

...if taken seriously, this advice would make biogeography unique in the world of phylogenetic methods, and unique in inferential/statistical science.



Likelihood & AIC are everywhere...is biogeography really a magical exception where these principles don't apply?

Practical use of the information-theoretic approach

[KP Burnham, DR Anderson - Model selection and inference, 1998 - Springer](#)

Abstract Model building and data analysis in the biological sciences somewhat presupposes that the person has some advanced education in the quantitative sciences, and statistics in particular. This requirement also implies that a person has substantial knowledge of ...

★ 99 Cited by 53396 Related articles All 8 versions

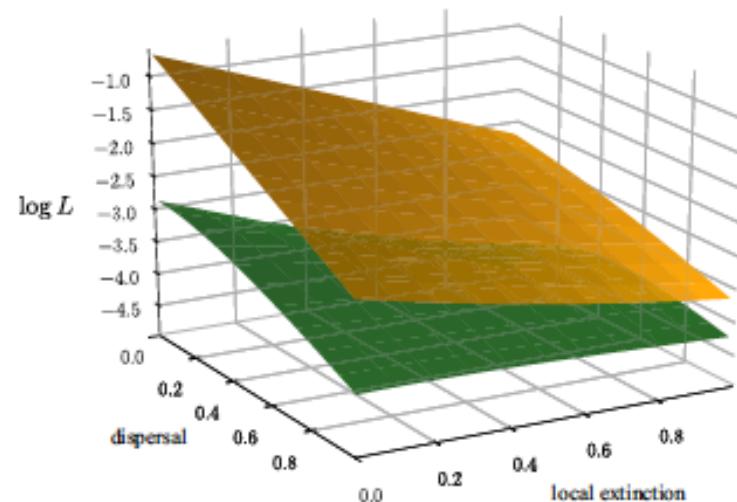
Used in e.g.: phylogenetic comparative methods, DNA model selection, ecological model comparison, paleobiology, etc. etc.

But, maybe the Critique has found a problem that means likelihood/AIC are flawed in this specific case?

Critique claims “degeneracy” in the DEC+J likelihood optimizations

2.6 | Pathological consequences of degeneracy in DEC+J

DEC+J differs from DEC in expanding the scope and influence of cladogenetic range evolution, both by allowing jump dispersal scenarios and assigning a free parameter to their partial likelihood



But: it is the Critique's likelihoods that are wrong

Table 3. Lagrange and BioGeoBEARS results of running DEC and DEC+J on ultrametric, pectinate trees, range "A" in the Critique.

Program and model	Number of species	Number of free parameters	maximized lnL	d	e	j
"DEC", R&S 2-taxon example	2	2	-2.890	0	0	0
DEC, Python Lagrange	2	2	-1.792	4.29E-09	3.29E-09	0
DEC, C++ Lagrange	2	2	-1.79176	1.13E-06	0.0001389	0
DEC, BioGeoBEARS	2	2	-1.791759	7.68E-09	6.40E-09	0
"DEC", R&S 4-taxon example	4	2	-8.50	0.060	0	0
DEC, Python Lagrange	4	2	-4.481	0.1011	4.29E-09	0
DEC, C++ Lagrange	4	2	-4.48106	0.10183	8.14E-06	0
DEC, BioGeoBEARS	4	2	-4.481012	0.10106	1.00E-12	0

Running the Critique's "worked examples" for DEC in *BioGeoBEARS* shows that their likelihood calculations cannot be replicated.

But, *BioGeoBEARS* DEC likelihood agree with Lagrange, which was written by Critique author Ree!

Anyone can check this by running the examples themselves (files in Supp. Mat).

The “degenerate” examples are extremely brittle

Figure 1:

Log-likelihood surface of (((A,B),B),B,...) dataset under DEC+J

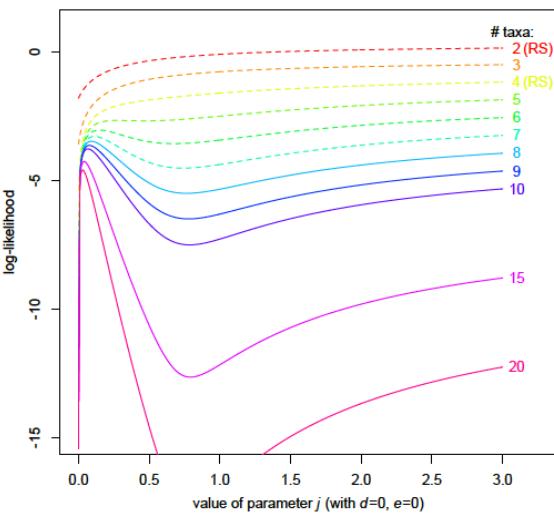


Figure 1A. Likelihood profiles for datasets following the pattern in the Critique’s worked examples, namely, pectinate, ultrametric trees, with all tips inhabiting single areas, and with only one tip in the most-nested clade inhabiting a different area.

Log-likelihood surface of (((B,B),B),B,...,A) dataset under DEC+J

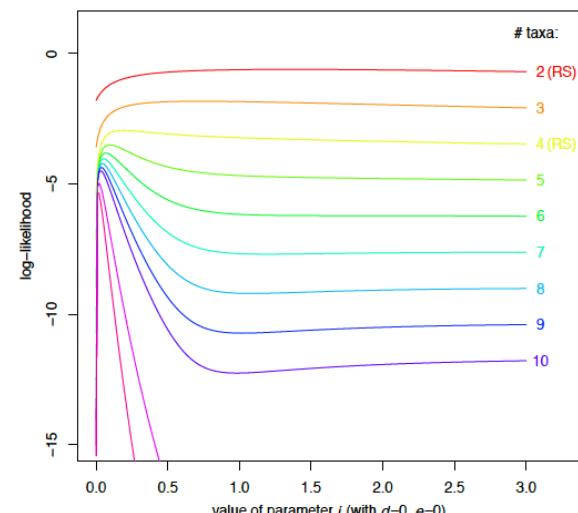


Figure 1B. As in 1A, but with the maximum number of areas set to 3 instead of 2.

Log-likelihood surface of (((B,B),B),B,...,A) dataset under DEC+J

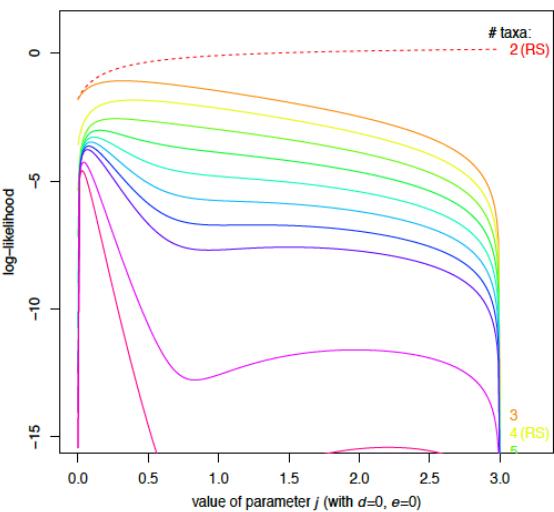


Figure 1C. As in 1A, but with the unique tip range being sister to the most-derived group.

Log-likelihood surface of (((B,B),B),B,...,A) dataset under DEC+J

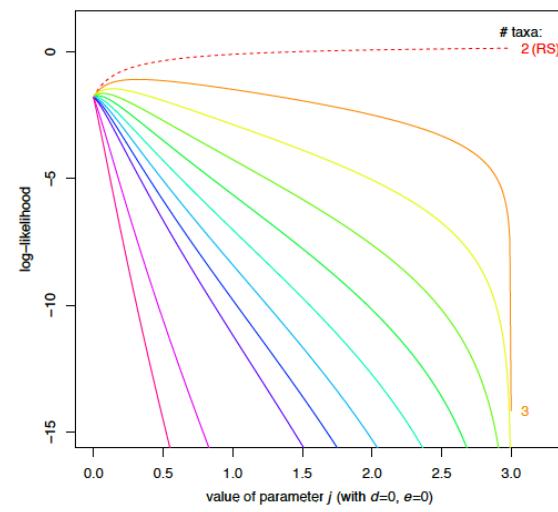


Figure 1D. As in 1A, but with the unique tip range moved to the outgroup.

Human-chosen datasets of 2 & 4 taxa are a bad way
to test a method anyway

Maximum likelihood and AIC, as methods, need
reasonable amounts of data, compared to the number of
parameters

AICc should have been used on these tiny datasets; if it is
used, the Critique's claims of DEC+J advantage on these
tiny datasets disappear

The real test, the industry standard, is simulation/
inference studies. Thousands of simulations were done in
Matzke (2014), *Systematic Biology*, verifying the
inference. These were totally ignored in the Critique.

The fairness of the DEC / DEC+J likelihood comparisons is easily verified

Table 5. 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 dominating 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+, only cladogenetic range-change, dominated by jump-dispersal: $d=0.0001$, $e=0.0001$, $j=2.9$				
	Data InL	Data likelihood	Data InL	Data likelihood	Data InL	Data likelihood	Data InL	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	-6.685	0.504
100,001	-1.7918	0.167	-2.3853	0.092	-1.8164	0.163	-6.685	0.504
100,110	-1.7918	0.167	-1.8506	0.157	-1.9566	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	-1.9566	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	-6.685	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	-6.685	0.504
010,110	-1.7918	0.167	-1.8506	0.157	-1.9566	0.14	-5.192	0.006
010,011	-1.7918	0.167	-1.8506	0.157	-1.9566	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	-6.685	0.504
001,010	-1.7918	0.167	-2.3853	0.092	-1.8164	0.163	-6.685	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	-1.9566	0.14	-5.192	0.006
001,101	-1.7918	0.167	-1.8506	0.157	-1.9566	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	-11.2253	0	-2.6187	0.073	-2.6286	0.072	-11.5019	0
110,100	-1.7918	0.167	-1.8506	0.157	-1.9566	0.14	-5.192	0.006
110,010	-1.7918	0.167	-1.8506	0.157	-1.9566	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	-1.9566	0.14	-5.192	0.006
011,001	-1.7918	0.167	-1.8506	0.157	-1.9566	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	-1.9566	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	-1.9566	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	-1.7918	0.167	-1.8506	0.157	-1.9566	0.14	-5.192	0.006
111,001	-1.7918	0.167	-1.8506	0.157	-1.9566	0.14	-5.192	0.006
111,110	-12.2061	0	-2.795	0.061	-2.7243	0.066	-11.3781	0
111,011	-12.2061	0	-2.795	0.061	-2.7243	0.066	-11.3781	0
111,101	-11.9184	0	-2.4889	0.083	-2.5834	0.076	-15.3191	0
111,111	-13.3046	0	-2.6314	0.072	-2.6234	0.073	-13.3045	0
111,100	-13.3046	0	-2.6314	0.072	-2.6234	0.073	-13.3045	0
111,010	-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.

This means some data favor DEC, some favor DEC+J. This is not “unfair”, this is data impinging on our models...which is the whole point of statistics, and science.

This is because: for 3 areas, there are $2^3 = 8$ possible ranges. And the null range is an impossible ancestor (it is a possible tip state, though). $8-1=7$.

Short version: For a 3-area problem, 2-area tree, if you add up likelihoods across all possible pairs of tip ranges, the likelihoods add up to...7.0.

True source of DEC+J advantage: the DEC model is statistically inadequate on many (not all) datasets

DEC needs widespread (2+ areas) ranges as intermediates. This predicts many tips should be widespread.

If most tips are single-area ranges, which is common (not universal), well, then a model that produces this, like DEC+J, will fit better.

Thus, the DEC model is statistically inadequate on many common datasets.

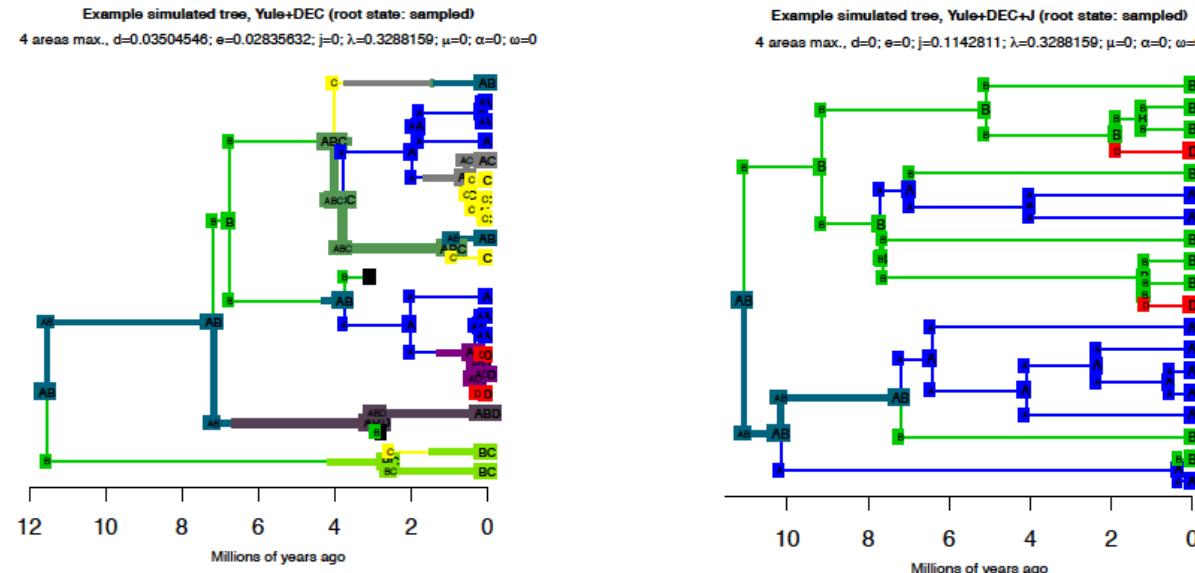


Figure 3. Example simulation histories under DEC and DEC+J with parameters fit to the Hawaiian *Psychotria* dataset. 3A: DEC+Yule simulation; 3B: DEC+J+Yule simulation.

True source of DEC+J advantage: the DEC model is statistically inadequate on many (not all) datasets

Table 6. 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 ($\lambda=0.3288159$, $\mu=0$, $d=0.03504546$, $e=0.02835632$, $j=0$). Similarly, the Yule+DEC+J simulations used the DEC+J-inferred parameters ($\lambda=0.3288159$, $\mu=0$, $d=0$, $e=0$, $j=0.1142811$).

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)	
<i>Psychotria</i> (empirical)					0 (all tips single-area)	

True source of DEC+J advantage: the DEC model is statistically inadequate on many (not all) datasets

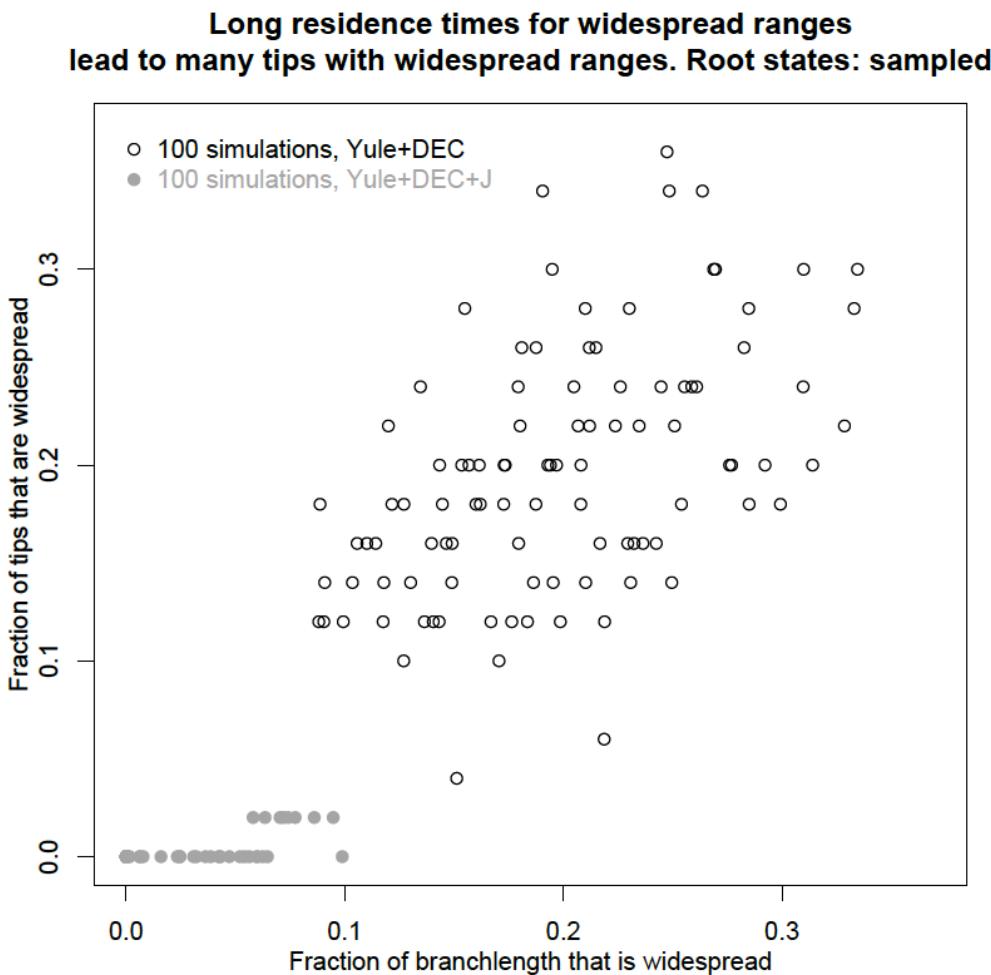


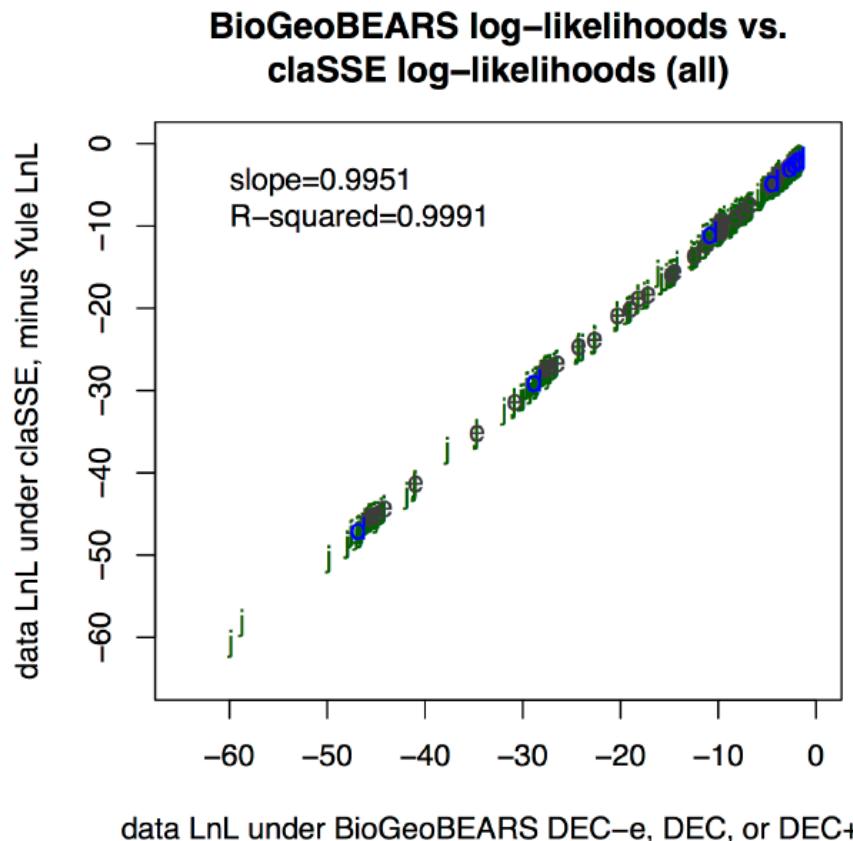
Figure 4. The correlation between the residence time spent in widespread ranges during a simulation, and the proportion of tips occupying widespread ranges at the end of the simulation. Black: DEC simulations. Grey: DEC+J simulations. The simulations correspond to rows 7 and 8 of Table 6.

True source of DEC+J advantage: the DEC model is statistically inadequate on many (not all) datasets

Also: “time-independence” has nothing to do with DEC+J’s advantage. The advantage is simple, and was just explained.

We can add in time-dependence by adding the likelihood of the tree under a Yule process. This adds the same lnL constant to DEC and DEC+J, so doesn’t effect model comparison.

Lastly: Comparison of DEC and DEC+J lnLs produces identical lnL differences to appropriately constructed ClaSSE submodels.



The Critique endorses
ClaSSE as appropriate, so
the Critique refutes itself.

See Matzke (2021) for
detailed explanation, and
the R scripts!

Figure 6. 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 $j>0$.

There are some legitimate critiques of DEC+J

1. The use of “relative weights” of cladogenetic events is not very intuitive. E.g., setting the range of j as (0.0, 2.99999) is arbitrary.

But these can easily be converted to rates, e.g.:

$$\text{speciation rate} * j / (\text{sum of weights})$$

2. DEC+J (and DEC) ignore extinction. This is a big limitation, which could be fixed by using ClasSE models.

a. Accurate inference of extinction is a problem across most of phylogenetics

b. ClasSE models have to iteratively solve systems of (4*number of states) Ordinary Differential Equations. SLOW WHEN YOU HAVE 1000+ states



B I U x^2 x_2 \leftarrow \rightarrow ∇ Σ Σ Zoom Insert

Sections

SYNTHESIS

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

Handling Editor: Hans-Peter Comes

AQ2 Nicholas J. Matzke ^{1,*} 

¹ School of Biological Sciences, University of Auckland, Auckland, New Zealand

Correspondence

Nicholas J. Matzke, School of Biological Sciences, University of Auckland, Auckland, New Zealand.

Email: n.matzke@auckland.ac.nz

Funding Information

Marsden Fund

University of Auckland : 3722433

Marsden Grants : 18-UOA-034 : 16-UOA-277

Australian National University

Australian Research Council : DE150101773

University of Tennessee, Knoxville

National Institute for Mathematical and Biological Synthesis

NSF National Science Foundation: DBI-1300426 : EFJ0832858 : 0919124

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 (*Journal of Biogeography*, 45, 741–749, 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 analyse the key claims made by the critique.

Out of time...for more, please see:

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. Matzke¹

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

n.matzke@auckland.ac.nz <https://osf.io/vqm7r>

Thought questions:

1. Why were none of these mistakes caught previously?
2. If 10 biogeographers invented their own biogeography models, would they all be DEC? How should we tell which models are better?
3. “All models are wrong, but some are useful.” Even ClaSSE models will have to be quite simple – what simplifications are acceptable?
4. Biogeography should be statistically sophisticated – it shouldn’t rely on tradition, verbal arguments, or authority. Look to the future!

