Associate Editor

Comments to the Author:

I am now in receipt from of two referee reports for "Etienne et al. How reliably can we infer diversity-dependent diversification from phylogenies?". Both referees were extremely positive about the ms and only have minor issues to be addressed. I think this is a valuable contribution to MEE and encourage the authors to consider the referees' comments carefully in a revised version.

Thank you for these nice words. We have revised the ms in the light of these minor comments. In addition, we have rerun our analyses to check for a potential numerical inaccuracy. This changed our results quantitatively in some cases, but our qualitative conclusions remain practically unaltered.

Reviewer(s)' Comments to Author:

Reviewer: 1

This is a nice technical paper on a topic of considerable interest in macroevolutionary modeling. I am unsurprised by the general findings that Type I error rates can be high – we tried to address this via simulation in our 2008 Proc B paper – but the general approach here is obviously a substantial improvement. I have one major comment that I'd lead to see the authors address. Given that all of these models are approximations of an underlying reality, and that we really have no clue how diversity-dependence might act, what is the benefit of using a formal diversity-dependent model? Would there be analytical advantages to going with an exponential decay process for speciation rate as function of time, which can at least do an acceptable job (in some areas of parameter space) at estimating rates under diversity-dependent processes? In BAMM, for example, I use an exponential approximation but it can still estimate mean rates of the process and branch-specific rates with some accuracy, even if it's the "wrong" model. I expect that there are some areas where the BAMM approximation would perform poorly – especially for clades that have been at equilibrium for a long tim – but perhaps not.

We appreciate this comment. In fact, last year we already started a separate project studying whether a time-dependent model can be used as a proxy for the computationally more demanding diversity-dependent model. This turns out to be a very tricky issue that goes far beyond the focus of this ms which deals primarily with the question whether the presence or absence of diversity-dependence can be reliably detected. However, we certainly take the reviewer's

point that this merits some discussion. We have therefore added a paragraph in the Discussion section.

The flip side of this issue is that true secular changes in diversification, for reasons that have nothing to do with diversity-dependence, will presumably lead to high Type I error rates using the procedure here. The DD model is going to soak up variation in rates through time, so rates through time – for any reason – can lead to spurious conclusions about diversity-dependence if assessed in a CR vs DD framework.

We agree that spurious conclusions could be drawn about diversity-dependence in a CR vs DD framework. But this is a general philosophical issue: if one model is preferred over others, it only means that this model performs best in explaining the data, but it does not imply that the mechanism underlying the model is necessarily active. There are infinitely many other models that could explain the data. The appropriate test of time-dependence vs diversity-dependence is of course a TD vs DD framework, which is a study in its own right as we explained above.

I do wonder whether the processes that lead to declining apparent speciation in phylogenies are separable and think it's still an open question. And even if you found support for (say), a linear time-dependent decline in speciation over the DD model, could we really reject diversity-dependence, given that the true process of diversity-dependence might not be well-approximated by the functional relationship in the logistic formulation?

This is exactly the issue that we have already started to explore in the separate project explained above. In this project we aim to find a time-dependent function that matches the diversity-dependent relationship as much as possible, so that any difference found between their performances on data can really be attributed to time vs diversity-dependence. This is a very interesting point, but beyond the focus of this ms.

Some expanded discussion of these important conceptual issues could help frame some key questions for future work in this area and for limits to our interpretation of whatever it is that is coming out of formal DD models.

We have added a discussion of these important points.

L177 - 184: Does this approach for conditioning – integrating backwards to -Inf before the crown age – yield appreciable differences when compared to simply conditioning on the crown age itself (if not, the latter could be preferable in the interest of computational tractability).

Yes, it certainly does, as our results in Figure 1 show.

L202 - 204: I don't follow why you can't do the forward simulations to fixed tree size. Does the Hartmann et al (Syst Biol, doi: 10.1093/sysbio/syq026) algorithm get around these issues?

This method works IF one can easily simulate until extinction or a point where the process is not likely to get back to the predefined tree size. For diversity-dependent diversification extinction is the only possibility because the system is drawn towards the equilibrium when far away from equilibrium. And for the same reason, simulation until extinction may take a very long time. Of course, it may also be short when starting with 2 lineages, but then extinction happens early on and the predefined tree size has not been reached. We have added some sentences to this extent in the ms.

L220 - 227: This is an interesting tradeoff in conditioning and I expect is not widely appreciated.

Indeed, we agree that this underappreciated.

L95: I have found that there are some communication issues when discussing these models, because people get hung up on the concept of "carrying capacity", even though it's being used phenomenologically here. I expect someone will read this and ask "what can we learn from this model if it's making this unrealistic assumption about the existence of a"carrying capacity" "? It might be useful to head off this possible misunderstanding by also noting that you can also view this model as a simple linear effect of species richness on speciation rate, basically the simplest model you can use to study how richness might affect diversification. Jim Mallet has an interesting paper addressing some of the ways that the K formulation of this model has led to misinterpretation in population ecology as well (Mallet, Evolutionary Ecology Research, 14:627, 2012).

OK, point taken. We have noted this in the ms.

Reviewer: 2

I very much liked this manuscript for two reasons: In the first place because it provides a much needed cautionary note concerning macroevolutionary inferences from phylogenies. Second, because many years ago I submitted a technical comment checking how useful likelihood ratio tests were in selecting models of diversification, and the reviewers thought it was a waste of time to check the performance of the well-established likelihood ratio test. So, it gave me some satisfaction to read the current manuscript!

We believe that this reviewer was ahead of his time.

A critical issue in studies like the present is whether simulation and estimation algorithms are correct. (One may get biased estimates and error rates because of incorrect simulation and/or estimation algorithms/equations). It is almost impossible for reviewers to check this. Because estimates are approximately unbiased under dual conditioning, I suppose the algorithms, and hence the results and conclusions presented, are correct.

This ms was already an enormous computational effort for us, so one can hardly expect reviewers or readers to check this thoroughly. Nevertheless, we have made the code available in the R package DDD. We have also added the simulated trees for readers who would like to check our results or would like to additional analyses on them.

Therefore, (for once) I do not have any major comments: the paper is well written (apart from a few minor issues listed below) and of suitable lengths. The figures are very informative.

We thank the reviewer for his kind words.

While I have no major comments, I do have a suggestion the authors might find useful to think about: what about confidence intervals? The likelihood ratio comparing a CR model to a density-dependent model evidently doesn't follow a Chi-square distribution, but what about the ratio of likelihoods of the same model given different parameter values? In other words, could it be that confidence intervals around parameters are reliable even if point estimates are biased? For me, this leads to the question whether Bayesian estimation would perform any better here. For the nested models here you could perhaps calculate Bayes' factor as the Savage-Dickey ratio (assuming any K>>N equals infinity), and hence compare models without relying on Wilk's theorem.

The reviewer raises two points. First, a point about confidence intervals. We have addressed this issue already in the manuscript by suggesting parametric bootstrapping to obtain confidence intervals (and assessing bias). These are by construction reliable, because they tell us the spread in estimations from simulations with the maximum likelihood parameters. Second, the reviewer suggests that model selection using Bayes factor may perform better than a likelihood ratio test. Because the Bayes factor is simply the ratio of marginal likelihoods we do not expect that these perform better, and introduce another complication: one must specify priors. Using appropriate priors could result in better type I and II errors, but the appropriate priors may change from data set to data set. We have therefore chosen not to discuss this in the ms.

Line 38: You haven't written what K is, not even what model of density-dependence you use, so the symbol K comes out of the blue sky here.

Thank you for noting this. We have corrected this.

Line 65 This is incorrect: the branches /branching time intervals are typically longer towards the root. Consider inserting "than expected under a constant-rates speciation-extinction model"

We have rephrased this.

Line 84. Is"1)" and "2)" really necessary here?

We have removed these.

Line 99 remove "deep"

We have removed this.

Line 101 you haven't defined K' yet, or explained why it measures the number of available niches. You also don't explain that further down, so this needs fixing.

We have more clearly defined K and K'.

Line 132 bootstrapping

Done.

210-211 It isn't obvious what "this" model is. Perhaps "the density-dependent model" would be clearer.

We have rephrased this.

210-227: I think you mean Fig1 when you wrote Fig2, and vice versa, in this section.

Thank you. We have changed this.

250 You cite Tekle, but it might be appropriate to cite also Wilk's original paper. The reason why the chi-square approximation isn't appropriate could be non-independence inherent in time series, in addition to the points you raise.

We have inserted a citation to Wilk's paper.

227 What you say is that improper conditioning improves precision at the expense of accuracy. (Not unusual.) It might be helpful to summarize it that way.

We have incorporated this suggestion.

248 -272. I find the phrasing, especially the word "bootstrap" a bit too difficult here. The reason why the likelihood ratio test doesn't perform well is that the sampling distribution of the likelihood ratio is not the usual chi-square distribution. You don't need to mention regularity conditions here. What about the following: "The mismatch between our Type I error rate and the intended significance level occurs because the distribution of the likelihood ratio is not the expected Chi-square (for reasons we will discuss furher down). Therefore, we suggest to estimate the actual distribution of the likelihood ratio using simulations (Tekle et al. 2015)." Steps 1-4 describe this procedure. One could use kernel density estimation instead of using the fraction LR>LR0 to reduce the number of simulations, but if you use the fraction, then use it properly (doi: 10.1086/341527 or some equivalent paper). (I'm not asking to fix the values in the ms, with 500 simulations it wouldn't make a real difference. However, I do think you should replace that sentence about the fraction (262) with the proper method from the doi above.)

We have inserted this suggestion.

338 You cite Lambert and Stadler 2013, but I believe this was already demonstrated clearly by Maddison et al 2007: 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.

We have inserted this reference.

343 "The failure of the chi 2 -likelihood ratio test is due to..." I wonder how you can be so sure what this is due to. Probably the K=infinity issue plays a role, but Wilk's theorem also assumes that data are idd, which seems violated here.

We have rephrased this.

355 differs

Done.

I liked the information-dense figures, and have only a few comments on the legends:

Thank you for these kinds words.

501-502 Perhaps a more effective phrasing would be "Note the discrepancy between Type I error rates and significance levels."

We have rephrased this.

508 and further: you have logarithms of the likelihood ratio, not loglikelihood ratios.

The reviewer is right (of course). We have rephrased this.