

*'Residual diversity estimates' do not  
correct for sampling bias in  
palaeodiversity data*

Article

Accepted Version

Sakamoto, M., Venditti, C. ORCID: <https://orcid.org/0000-0002-6776-2355> and Benton, M. J. (2017) 'Residual diversity estimates' do not correct for sampling bias in palaeodiversity data. *Methods in Ecology and Evolution*, 8 (4). pp. 453-459. ISSN 2041-210X doi: [10.1111/2041-210X.12666](https://doi.org/10.1111/2041-210X.12666) Available at <https://centaur.reading.ac.uk/67779/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

Published version at: <http://dx.doi.org/10.1111/2041-210X.12666>

To link to this article DOI: <http://dx.doi.org/10.1111/2041-210X.12666>

Publisher: Wiley-Blackwell

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

[www.reading.ac.uk/centaur](http://www.reading.ac.uk/centaur)

**CentAUR**

Central Archive at the University of Reading

Reading's research outputs online

1 **'Residual diversity estimates' do not correct for sampling bias in**  
2 **palaeodiversity data**

3

4 SHORT TITLE: Do not use residuals method

5

6 WORD COUNT: 4,739

7

8 Manabu Sakamoto<sup>1</sup>, Chris Venditti<sup>1</sup> and Michael J. Benton<sup>2</sup>

9

10 <sup>1</sup>School of Biological Sciences, University of Reading, Reading, RG6 6AJ, UK

11 <sup>2</sup>School of Earth Sciences, University of Bristol, Bristol, BS8 1RJ, UK

12

13 EMAIL: m.sakamoto@reading.ac.uk

14

15

16 **ABSTRACT**

- 17 1. It is widely accepted that the fossil record suffers from various sampling  
18 biases – diversity signals through time may partly or largely reflect the  
19 rock record – and many methods have been devised to deal with this  
20 problem. One widely used method, the ‘residual diversity’ method, uses  
21 residuals from a modelled relationship between palaeodiversity and  
22 sampling (sampling-driven diversity model) as ‘corrected’ diversity  
23 estimates, but the unorthodox way in which these residuals are generated  
24 presents serious statistical problems; the response and predictor  
25 variables are decoupled through independent sorting, rendering the new  
26 bivariate relationship meaningless.
- 27 2. Here, we use simple simulations to demonstrate the detrimental  
28 consequences of independent sorting, through assessing error rates and  
29 biases in regression model coefficients.
- 30 3. Regression models based on independently sorted data result in  
31 unacceptably high rates of incorrect and systematically, directionally  
32 biased estimates, when the true parameter values are known. The large  
33 number of recent papers that used the method are likely to have  
34 produced misleading results and their implications should be reassessed.
- 35 4. We note that the ‘residuals’ approach based on the sampling-driven  
36 diversity model cannot be used to ‘correct’ for sampling bias, and instead  
37 advocate the use of phylogenetic multiple regression models that can  
38 include various confounding factors, including sampling bias, while  
39 simultaneously accounting for statistical non-independence owing to  
40 shared ancestry. Evolutionary dynamics such as speciation are inherently

41 a phylogenetic process, and only an explicitly phylogenetic approach will  
42 correctly model this process.

43 **KEY WORDS**

44 Palaeodiversity; residuals; modeling; sampling bias; fossil record; independent  
45 sorting

46 **INTRODUCTION**

47 It has been well known since the time of Darwin that the fossil record is largely  
48 incomplete (Darwin 1859), prompting generations of macroevolutionary  
49 researchers to take a cautious approach when interpreting patterns of  
50 palaeodiversity through time (Raup 1972; Raup 1976; Raup 1991; Prothero  
51 1999; Smith & McGowan 2007; Alroy 2010b). There have been many attempts to  
52 account for this sampling bias (Raup 1972; Raup 1976; Smith & McGowan 2007;  
53 Alroy 2010b), but one approach in particular, often referred to as the ‘residual  
54 diversity’ method, devised by Smith and McGowan (2007) (and modified by  
55 Lloyd (2012)), has been widely used (citation count ~215 to Aug 2016; Google-  
56 Scholar).

57

58 Using regression residuals as data ‘corrected’ for confounding factors is a widely  
59 used method in biology, social sciences, economics (King 1986; Freckleton  
60 2002), and even in palaeodiversity studies (Raup 1976). However, Smith and  
61 McGowan’s (2007) approach differs from these classical residuals approaches in  
62 one key way: the ‘residuals’ are generated not as regression residuals ( $\varepsilon = y - \hat{y}$ )  
63 from a simple regression of diversity ( $y$ ) on a proxy of sampling ( $x$ ), but from “*a*  
64 *model in which rock area at outcrop was a perfect predictor of sampled diversity*”  
65 (Smith & McGowan 2007), here referred to as the sampling-driven diversity  
66 model (SDDM). The SDDM is constructed as a regression model between  $y$  sorted  
67 from low to high values ( $y'$ ) and  $x$  sorted from low to high values ( $x'$ ), where the  
68 relationship between these two independently sorted variables  $y'$  and  $x'$  is  
69 assumed to represent the SDD generating process – though there is no reason to  
70 assume as such. ‘Residuals’ are obtained as the difference between the SDDM

71 predictions  $\hat{y}$ ' and the observed values  $y$ , which are then treated as the 'residual  
72 diversity estimates' (figure 1).  
73  
74 However, independently sorting  $y$  and  $x$  as outlined above decouples a paired,  
75 bivariate dataset, and is obviously problematic in statistics. Model fitting on  
76 decoupled data (e.g.  $y'$  and  $x'$ ) will lead to spurious predictions and 'residuals' as  
77 the estimated regression coefficients will be based on a forced (false) linear  
78 relationship (figure 1b). However, owing to continued wide use of the SDDM as a  
79 preferred method for identifying supposedly 'true' palaeodiversity signals (as  
80 recently as (Grossnickle & Newham 2016)), it appears that this basic statistical  
81 concept is somehow overlooked. While it has been suggested that the use of  
82 formation counts (the number of fossiliferous geological formations – a  
83 mappable unit of rock that represents a particular time and set of environments  
84 in a particular location – in a given time interval (Benton *et al.* 2011)) to 'correct'  
85 palaeodiversity time series data is unlikely to be meaningful because of  
86 substantial redundancy of the two metrics (Benton *et al.* 2011; Benton 2015),  
87 and a recent study has scrutinized the performance of SDDM residuals in  
88 accurately predicting true simulated biodiversity signals (Brocklehurst 2015),  
89 the performance of the SDDM itself has never been assessed. Here, we  
90 demonstrate the detrimental effects of decoupling data in regression modelling  
91 using simple simulations.

92

93

94 **MATERIAL AND METHODS**

95 We first generated random deviates,  $x$ , sampling from a normal distribution ( $\mu =$   
96  $0, \sigma = 1$ ), at a sample size  $n = 100$  (see SI for other sample sizes  $n = 30$  and  $1000$ ).  
97 We then calculated  $y$  using a linear relationship in the form of  $y = a + bx + e$ ,  
98 where  $a$  is the intercept,  $b$  is the slope and  $e$  is Gaussian noise. For simplicity, we  
99 fixed  $a = 0.4$  and  $b = 0.6$ , while varying  $e$  ( $\mu_e = 0, \sigma_e = 0.05, 0.1, 0.25, 0.5$ ) – other  
100 values of  $a$  and  $b$  should return similar if not identical results (though,  $b = 1$   
101 would be meaningless). Following Smith and McGowan (2007), we sorted  $y$  and  $x$   
102 independently of each other to generate  $y'$  and  $x'$ , and fitted an ordinary least  
103 squares (OLS) regression model to  $y'$  on  $x'$  (SDDM). For comparison, we fitted an  
104 OLS regression model to  $y$  on  $x$  in their original paired bivariate relationship (the  
105 standard regression model, SRM), the performance of which serves as a  
106 benchmark.

107

108 To test Smith and McGowan's (2007) assertion that the SDDM is indeed "*a model*  
109 *in which rock area at outcrop was a perfect predictor of sampled diversity*", we  
110 evaluated whether the estimated regression coefficients  $\alpha$  and  $\beta$  significantly  
111 differed from the true regression parameters,  $a$  and  $b$ , using a  $t$ -test. We repeated  
112 the procedure over 5000 simulations and calculated the percentage of times the  
113 estimated coefficients differed significantly from the true parameters. We would  
114 expect about 5% of the simulations to result in regression coefficients  
115 significantly different from the true parameters by chance alone; anything  
116 substantially above this threshold would indicate that the model has  
117 unacceptably high Type I error rates or falsely rejecting a true null hypothesis,  
118 where our null hypothesis is that the SDDM can correctly estimate the 'true'  
119 model parameters.

120

121 In addition, we tested for bias in the estimated regression slopes, i.e. whether the  
122 estimates systematically deviated from the simulation parameter  $b = 0.6$ . The  
123 mean of the 5000 slopes was subjected to a  $t$ -test against a fixed value of 0.6. If  
124 deviations were random, then we would not expect to find any significant  
125 differences between the mean slope and the theoretical value, with all slopes  
126 randomly distributed around it.

127

128

## 129 **RESULTS**

130 SRM coefficients were significantly different from the true model parameters in  
131 only  $\sim 5\%$  of the 5000 iterations across  $\sigma_e$  (figure 2a; table 1; SI), within  
132 acceptable levels of randomly detecting a statistical significance. Variation in  
133 regression lines across 5000 iterations are distributed randomly about the  
134 simulated line (figure 3a), with no significant difference between the mean  
135 regression slope and the simulation parameter  $b=0.6$  (table 2; SI). In contrast,  
136 SDDM coefficients were significantly different from the true parameters (figure  
137 2b) at a rate much higher than the conventionally accepted 5% (table 1; SI). The  
138 mean slope of the regression models significantly differed from the simulation  
139 parameter  $b$ , in a systematic and directional manner (figure 3b; table 2; SI) –  
140 SDDM regression coefficients are not only incorrect but grossly misleading. This  
141 systematic bias increases with increased noise in the data (table 2) – the more  
142 noise there is in the data, the more positive the relationship between  $y'$  and  $x'$   
143 becomes.

144

145

146 **DISCUSSION**

147 By establishing “*a model in which rock area at outcrop was a perfect predictor of*  
148 *sampled diversity*”, Smith and McGowan (2007) attempted to create a sampling-  
149 driven diversity model. However, their SDDM is not based on any hypothesized  
150 or empirical relationship between diversity and sampling, or formulated from  
151 first principles. This is in contrast to other well-formulated biological models  
152 such as various scaling models where the parameter of interest (i.e. scaling  
153 coefficient or the slope of the bivariate relationship) is founded on first-principle  
154 theories, e.g. the 2/3 rule for the scaling of area with mass. Rather, the SDDM is  
155 based on the assumption that  $y'$  and  $x'$  ( $y$  and  $x$  sorted independently of each  
156 other) form the expected theoretical bivariate relationship between  $y$  and  $x$ ,  
157 which this study shows to be incorrect (figures 2, 3), as one would expect since  
158 there is no reason to assume such a thing.

159

160 A further and perhaps more serious problem with using a forced pairing of  $y'$  and  
161  $x'$  is that each data point (pair of  $y'_i$  and  $x'_i$ ) does not represent a natural pairing  
162 and has no meaning; the new pairing is actually  $y_i$  and  $x_j$ , where the  $i$ th and  $j$ th  
163 orders are independent of each other. For instance, using the marine generic  
164 diversity and rock area data of Smith and McGowan (2007) (figure 4), the lowest  
165 marine generic diversity is in the Cambrian, Tommotian Stage (529 – 521 million  
166 years ago [Ma]; genus count = 309), while the smallest marine rock outcrop area  
167 (after removing 0 valued data (Smith & McGowan 2007)) is from the Early  
168 Permian, Asselian/Sakmarian Stage (299 – 290 Ma; rock area = 1). Similarly, the  
169 highest diversity is recorded for the Pliocene (5.3 – 2.58 Ma; genus count = 3911)

170 while the largest rock area is found in the Cenomanian (100 – 94 Ma; rock area =  
171 373). These two extreme points alone demonstrate that the paired diversity and  
172 rock area values are millions of years apart, and are independent of each other  
173 (figure 4).

174

175 This may be obvious, but independently sorting  $y$  and  $x$  has serious statistical  
176 consequences. For instance, in Smith and McGowan's (2007) data,  $\log_{10}$  marine  
177 generic diversity has no significant relationship with  $\log_{10}$  rock area in their  
178 original paired bivariate data (figure 4;  $r^2 = 0.0398$ ;  $p = 0.0979$ ), but once sorted,  
179 has a significantly strong positive relationship with  $\log_{10}$  rock area sorted  
180 independently of  $\log_{10}$  diversity (figure 4;  $r^2 = 0.903$ ;  $p < 0.001$ ). This general  
181 pattern is true in at least two more datasets (Benson *et al.* 2010; Benson &  
182 Upchurch 2013) (figures S1 and S2). The independent sorting procedure has  
183 forced a strong but false linear relationship between two variables that  
184 otherwise do not show any significant (or if significant, a very weak)  
185 relationship. In fact, two randomly generated deviates (e.g. sampled from a  
186 normal distribution) that have no relationship with each other (figure 5a), once  
187 sorted independently from lowest to highest will inevitably have a significant  
188 and strong relationship ( $r^2 = \sim 1$ ; figure 5b). Perhaps more detrimental, is the fact  
189 that the independently sorted bivariate relationship will always be strongly  
190 positive – a simulated negative relationship between  $x$  and  $y$  (figure 5c) will have  
191 a strong and positive relationship once they are sorted independently (figure  
192 5d).

193

194 In some clades (namely Mesozoic dinosaurs), diversity measures can have very  
195 strongly positive relationships with some sampling metrics, such as geological  
196 formation counts ( $\beta = 0.868$ ;  $r^2 = 0.85$ ;  $p < 0.001$  (Barrett, McGowan & Page  
197 2009)) or fossil collection counts ( $\beta = 0.865$ ;  $r^2 = 0.79$ ;  $p < 0.001$  (Butler *et al.*  
198 2011)), which would justify correcting for such confounding factors, if the  
199 sampling metrics were indeed non-redundant with diversity (Benton *et al.* 2011;  
200 Benton *et al.* 2013). However, even in such cases, it does not change the fact that  
201 the modelled relationship obtained from the SDDM will still be systematically  
202 biased (figure 3), and alternative methods should be considered.

203

204 It is problematic to stipulate that this forced relationship is the 'true' relationship  
205 between sampled palaeodiversity and the rock record. Our simulations show  
206 that regression models fitted on independently sorted data have unacceptably  
207 high Type I error rates when the data generation processes are known, meaning  
208 that Smith and McGowan's (2007) approach is not statistically viable. In  
209 particular, that the slopes are incorrectly estimated at very high rates (~100%  
210 when  $\sigma_e = 0.5$ ) has severe consequences in that SDDM predictions are  
211 systematically biased (figures 2b, 3b), leading to erroneous 'residuals'.

212 Inferences made from such problematic 'residuals' (Smith & McGowan 2007;  
213 Barrett, McGowan & Page 2009; Benson *et al.* 2010; Butler *et al.* 2011; Benson &  
214 Upchurch 2013) will inevitably be misleading (Brocklehurst 2015), lacking any  
215 biological or geological meaning.

216

217 Given our simulations, we strongly recommend against using the SDDM  
218 approach in modelling the relationship between palaeodiversity and rock record

219 data; the standard regression using unsorted data is a sensible option. However,  
220 using the residuals of a regression model as data for subsequent analyses has  
221 also long been known to introduce biased statistical estimates (King 1986;  
222 Freckleton 2002). Successive series of modelling removes variance and degrees  
223 of freedom from subsequent model parameter estimation, so the final models  
224 and statistical analyses do not account for the removed errors appropriately  
225 (King 1986). Instead, one can directly model the confounding effects along with  
226 effects of interest (e.g. environment, climate, etc) through multiple regressions  
227 (OLS, GLMs or generalized least squares [GLS]). In the context of palaeodiversity  
228 studies, one can fit a multiple regression model using some diversity metric as  
229 the response variable and sampling proxy as a confounding covariate, alongside  
230 additional predictor variables such as sea level, temperature, etc. The resulting  
231 model coefficients for the environmental predictors would be the effects of  
232 interest after accounting for the undesired effects of rock availability. Since  
233 diversity measures are frequently taken as counts, it is advisable to use models  
234 that appropriately account for errors in count data, such as the Poisson or  
235 negative binomial models (O'Hara & Kotze 2010). Whether or not to include time  
236 series terms (e.g. autoregressive [AR] terms) depends on the level of serial  
237 autocorrelation in the time series data and on sample size; palaeontological time  
238 series tend to be short, with 30 time bins or fewer being fairly typical (Mesozoic  
239 dinosaurs only span a maximum of 26 geological stages (Butler *et al.* 2011;  
240 Benson & Mannion 2012)), in which case complex models face the risks of over-  
241 parameterisation. Model selection procedures using the Akaike Information  
242 Criterion (Akaike 1973) or similar indices can help make this decision (Burnham  
243 & Anderson 2002). However, we do not lightly advocate the use of time series

244 modelling, especially if the dependent variable, sampled diversity, is in the form  
245 of counts, in which case appropriate time series methods are severely under-  
246 developed (but see generalised linear autoregressive moving average [GLARMA]  
247 models (Dunsmuir & Scott 2015) or Poisson exponentially weighted moving  
248 average [PEWMA] models (Brandt *et al.* 2000)), but more importantly since  
249 there are more appropriate alternative methods, i.e. phylogenetic approaches  
250 (Sakamoto, Benton & Venditti 2016).

251

252 Fundamentally, macroevolutionary studies aim to increase our understanding of  
253 evolutionary processes (speciation and extinction through time), rather than the  
254 resulting patterns or phenomena (sampled diversity, e.g. richness). Thus, we  
255 should seek to characterize the process using biologically meaningful and  
256 interpretable models instead of describing the patterns. Further, simply  
257 exploring error in the fossil record in itself seems rather fruitless because  
258 uncertainty depends on the questions being posed; palaeontological studies of  
259 macroevolution should be no different than other statistical approaches in the  
260 natural sciences in that uncertainty is assessed while exploring the phenomena  
261 of interest (Benton 2015). Explicitly phylogenetic approaches (e.g. (Lloyd *et al.*  
262 2008; Didier, Royer-Carenzi & Laurin 2012; Stadler 2013; Stadler *et al.* 2013;  
263 Sakamoto, Benton & Venditti 2016) offer the best and most appropriate means to  
264 tackle questions of evolutionary processes. Especially when extrinsic causal  
265 mechanisms for changes in biodiversity are tested using regression models,  
266 ignoring phylogeny is in serious violation of statistical independence  
267 (Felsenstein 1985; Harvey & Pagel 1991). It is also worth noting that  
268 subsampling approaches (e.g. Alroy's SQS (Alroy 2010a; Alroy 2010b; Alroy

269 2010c)) are gaining wide popularity as modern methods to account for sampling  
270 bias, they are not without problems (Hannisdal *et al.* 2016), and certainly do not  
271 take shared ancestry described by phylogeny into account, thus also suffering  
272 statistical non-independence (Felsenstein 1985; Harvey & Pagel 1991), and can  
273 frequently result in incorrect interpretation of the data. For instance, while  
274 recent studies using binned time series approaches (including SDDM and SQS)  
275 have led to mixed conclusions regarding the long-term demise of dinosaurs  
276 before their final extinction at the Cretaceous-Paleogene (K-Pg) boundary 66  
277 million years ago (Ma) (Barrett, McGowan & Page 2009; Lloyd 2012; Brusatte *et*  
278 *al.* 2015), an explicitly phylogenetic Bayesian analysis has strongly suggested  
279 that dinosaurs were indeed in a long-term decline tens of millions of years prior  
280 to the K-Pg mass extinction event, in which speciation rate was exceeded by  
281 extinction rate and dinosaurs were increasingly incapable of replacing extinct  
282 taxa with new ones (Sakamoto, Benton & Venditti 2016). Such evolutionary  
283 dynamics cannot be identified using time-binned (tabulated) data. Phylogenetic  
284 mixed modelling approaches (Hadfield 2010) further allow the incorporation of  
285 confounding variables such as sampling but also environmental effects  
286 (Sakamoto, Benton & Venditti 2016). Therefore, in order to advance our  
287 understanding of the evolutionary dynamics of biodiversity, speciation and  
288 extinction through time (or the underlying process generating the observed  
289 patterns in sampled diversity, e.g. taxon richness), while accounting for sampling  
290 and phylogenetic non-independence, it is imperative that we have an abundance  
291 of large-scale comprehensive phylogenetic trees of fossil (and extant) taxa.

292

293

294 **ACKNOWLEDGEMENTS**

295 We thank Jo Baker, Ciara O'Donovan and Henry Ferguson-Gow for discussion  
296 and insightful comments. We also thank Neil Brocklehurst and Michel Laurin for  
297 reviewing this manuscript and providing helpful commentary. We have no  
298 conflicts of interest.

299

300

301 **DATA ACCESSIBILITY**

302 This manuscript does not include data.

303

304 **FUNDING**

305 MS and CV are funded by Leverhulme Trust Research Project Grant RPG-2013-  
306 185 (awarded to CV). MJB is funded by Natural Environment Research Council  
307 Standard Grant NE/I027630/1.

308

309

310 **REFERENCES**

- 311 Akaike, H. (1973) Information theory and an extension of the maximum  
312 likelihood principle. *2nd International Symposium on Information Theory*  
313 (eds B.N. Petrov & F. Csaki), pp. 267–281. Akademiai Kiado, Budapest.
- 314 Alroy, J. (2010a) Fair sampling of taxonomic richness and unbiased estimation of  
315 origination and extinction rates. *Quantitative methods in paleobiology.*  
316 *Paleontological Society Papers*, **16**, 55-80.
- 317 Alroy, J. (2010b) Geographical, Environmental and Intrinsic Biotic Controls on  
318 Phanerozoic Marine Diversification. *Palaeontology*, **53**, 1211-1235.
- 319 Alroy, J. (2010c) The Shifting Balance of Diversity Among Major Marine Animal  
320 Groups. *Science*, **329**, 1191-1194.
- 321 Barrett, P.M., McGowan, A.J. & Page, V. (2009) Dinosaur diversity and the rock  
322 record. *Proceedings Of The Royal Society B-Biological Sciences*, **276**, 2667-  
323 2674.
- 324 Benson, R.B.J., Butler, R.J., Lindgren, J. & Smith, A.S. (2010) Mesozoic marine  
325 tetrapod diversity: mass extinctions and temporal heterogeneity in

326 geological megabiases affecting vertebrates. *Proceedings Of The Royal*  
327 *Society B-Biological Sciences*, **277**, 829-834.

328 Benson, R.B.J. & Mannion, P.D. (2012) Multi-variate models are essential for  
329 understanding vertebrate diversification in deep time. *Biology Letters*, **8**,  
330 127-130.

331 Benson, R.B.J. & Upchurch, P. (2013) Diversity trends in the establishment of  
332 terrestrial vertebrate ecosystems: Interactions between spatial and  
333 temporal sampling biases. *Geology*, **41**, 43-46.

334 Benton, M.J. (2015) Palaeodiversity and formation counts: redundancy or bias?  
335 *Palaeontology*, **58**, 1003-1029.

336 Benton, M.J., Dunhill, A.M., Lloyd, G.T. & Marx, F.G. (2011) Assessing the quality of  
337 the fossil record: insights from vertebrates. *Comparing the Geological and*  
338 *Fossil Records: Implications for Biodiversity Studies*, **358**, 63-94.

339 Benton, M.J., Ruta, M., Dunhill, A.M. & Sakamoto, M. (2013) The first half of  
340 tetrapod evolution, sampling proxies, and fossil record quality.  
341 *Palaeogeography Palaeoclimatology Palaeoecology*, **372**, 18-41.

342 Brandt, P.T., Williams, J.T., Fordham, B.O. & Pollins, B. (2000) Dynamic modeling  
343 for persistent event-count time series. *American Journal of Political*  
344 *Science*, **44**, 823-843.

345 Brocklehurst, N. (2015) A simulation-based examination of residual diversity  
346 estimates as a method of correcting for sampling bias. *Palaeontologia*  
347 *Electronica*, **18**.

348 Brusatte, S.L., Butler, R.J., Barrett, P.M., Carrano, M.T., Evans, D.C., Lloyd, G.T.,  
349 Mannion, P.D., Norell, M.A., Peppe, D.J., Upchurch, P. & Williamson, T.E.  
350 (2015) The extinction of the dinosaurs. *Biological Reviews*, **90**, 628-642.

351 Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference:*  
352 *a practical information - theoretical approach*, 2nd edn. Springer, New  
353 York.

354 Butler, R.J., Benson, R.B.J., Carrano, M.T., Mannion, P.D. & Upchurch, P. (2011) Sea  
355 level, dinosaur diversity and sampling biases: investigating the 'common  
356 cause' hypothesis in the terrestrial realm. *Proceedings Of The Royal Society*  
357 *B-Biological Sciences*, **278**, 1165-1170.

358 Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection, or the*  
359 *Preservation of Favoured Races in the Struggle for Life*, First Edition edn.,  
360 London, UK.

361 Didier, G., Royer-Carenzi, M. & Laurin, M. (2012) The reconstructed evolutionary  
362 process with the fossil record. *Journal Of Theoretical Biology*, **315**, 26-37.

363 Dunsmuir, W.T.M. & Scott, D.J. (2015) The glarma Package for Observation-  
364 Driven Time Series Regression of Counts. *Journal of Statistical Software*,  
365 **67**, 1-36.

366 Felsenstein, J. (1985) Phylogenies and the Comparative Method. *American*  
367 *Naturalist*, **125**, 1-15.

368 Freckleton, R. (2002) On the misuse of residuals in ecology: regression of  
369 residuals vs. multiple regression. (vol 71, pg 542, 2002). *Journal of Animal*  
370 *Ecology*, **71**, 722-722.

371 Grossnickle, D.M. & Newham, E. (2016) Therian mammals experience an  
372 ecomorphological radiation during the Late Cretaceous and selective  
373 extinction at the K–Pg boundary. *Proceedings of the Royal Society of*  
374 *London B: Biological Sciences*, **283**.

375 Hadfield, J.D. (2010) MCMC methods for multi-response Generalized Linear  
376 Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*,  
377 **33**, 1-22.

378 Hannisdal, B., Haaga, K.A., Reitan, T., Diego, D. & Liow, L.H. (2016) Common  
379 species link global ecosystems to climate change. *bioRxiv*, 043729.

380 Harvey, P.H. & Pagel, M.D. (1991) *The comparative method in evolutionary*  
381 *biology*. Oxford University Press.

382 King, G. (1986) How Not to Lie with Statistics - Avoiding Common Mistakes in  
383 Quantitative Political-Science. *American Journal of Political Science*, **30**,  
384 666-687.

385 Lloyd, G.T. (2012) A refined modelling approach to assess the influence of  
386 sampling on palaeobiodiversity curves: new support for declining  
387 Cretaceous dinosaur richness. *Biology Letters*, **8**, 123-126.

388 Lloyd, G.T., Davis, K.E., Pisani, D., Tarver, J.E., Ruta, M., Sakamoto, M., Hone,  
389 D.W.E., Jennings, R. & Benton, M.J. (2008) Dinosaurs and the Cretaceous  
390 Terrestrial Revolution. *Proceedings Of The Royal Society B-Biological*  
391 *Sciences*, **275**, 2483-2490.

392 O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in*  
393 *Ecology and Evolution*, **1**, 118-122.

394 Prothero, D. (1999) Fossil record. *Encyclopedia of paleontology* (ed. R. Singer).  
395 Fitzroy Dearbon Publishers, Chicago, USA.

396 Raup, D.M. (1972) Taxonomic Diversity during the Phanerozoic. *Science*, **177**,  
397 1065-1071.

398 Raup, D.M. (1976) Species Diversity in the Phanerozoic: An Interpretation.  
399 *PALEOBIOLOGY*, **2**, 289-297.

400 Raup, D.M. (1991) *Extinction: bad genes or bad luck?* W. W. Norton, New York.

401 Sakamoto, M., Benton, M.J. & Venditti, C. (2016) Dinosaurs in decline tens of  
402 millions of years before their final extinction. *Proceedings of the National*  
403 *Academy of Sciences*, **113**, 5036-5040.

404 Smith, A.B. & McGowan, A.J. (2007) The shape of the phanerozoic marine  
405 palaeodiversity curve: How much can be predicted from the sedimentary  
406 rock record of western Europe? *Palaeontology*, **50**, 765-774.

407 Stadler, T. (2013) Recovering speciation and extinction dynamics based on  
408 phylogenies. *Journal Of Evolutionary Biology*, **26**, 1203-1219.

409 Stadler, T., Kuhnert, D., Bonhoeffer, S. & Drummond, A.J. (2013) Birth-death  
410 skyline plot reveals temporal changes of epidemic spread in HIV and  
411 hepatitis C virus (HCV). *Proceedings Of The National Academy Of Sciences*  
412 *Of The United States Of America*, **110**, 228-233.

413

414 **SUPPORTING INFORMATION**

415 **SI-text.** Supporting information and results pertaining to the effects of sample  
416 size (Tables S1 and S2) as well as examples of discrepancies between original  
417 paired bivariate relationship and the independently sorted relationship from the  
418 literature (Figs S1 and S2).

419 **TABLES**

420 Table 1. Type I error rates (%) for SRM (Standard Regression Model) and SDDM  
 421 (Sampling-Driven Diversity Model) estimates (intercept  $\alpha$  and slope  $\beta$ ) across  
 422 residual error ( $\sigma_e$ ).  
 423

$\sigma_e$	SRM		SDDM	
	$\alpha$	$\beta$	$\alpha$	$\beta$
0.05	5.34	4.90	26.1	28.5
0.10	4.84	4.92	40.2	48.4
0.25	4.82	4.78	57.3	91.3
0.50	5.48	5.14	68.7	100.0

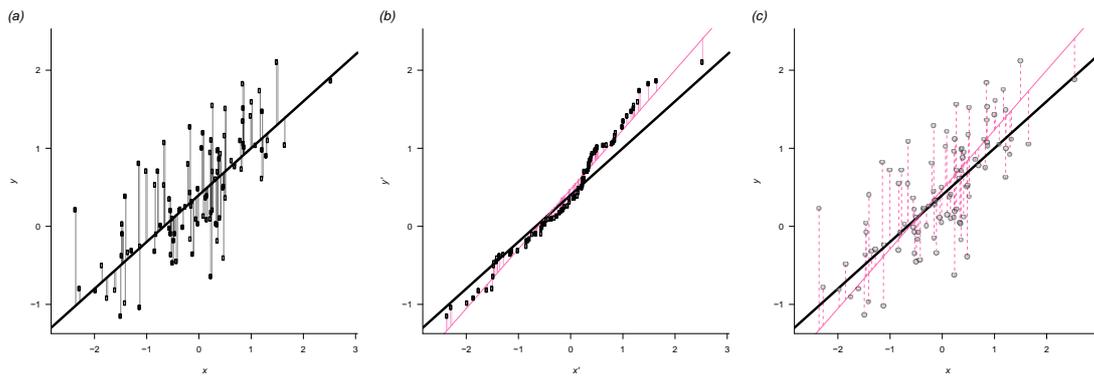
424

425 Table 2. *t*-test results between mean regression slopes of 5000 iterations and the  
 426 theoretical slope  $b = 0.6$ , for SRM (Standard Regression Model) and SDDM  
 427 (Sampling-Driven Diversity Model) across residual error ( $\sigma_e$ ).  
 428

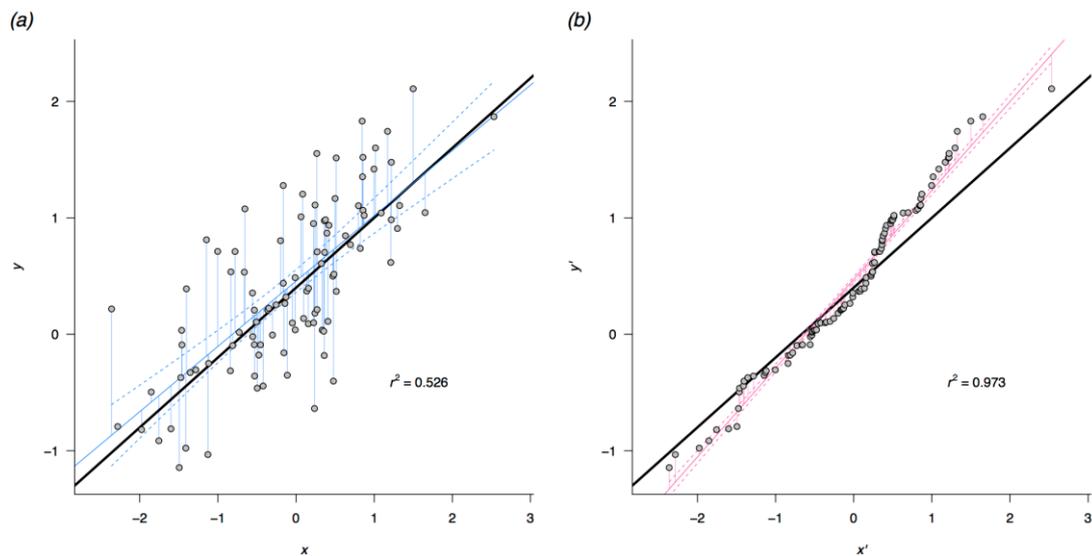
$\sigma_e$	SRM			SDDM		
	mean-slope	<i>t</i> -value	<i>p</i> -value	mean-slope	<i>t</i> -value	<i>p</i> -value
0.05	0.6	1.230	0.220	0.602	20.9	0
0.10	0.6	-1.790	0.073	0.607	46.0	0
0.25	0.6	-0.042	0.967	0.646	131.0	0
0.50	0.6	0.685	0.493	0.775	244.0	0

429

430 **FIGURES**



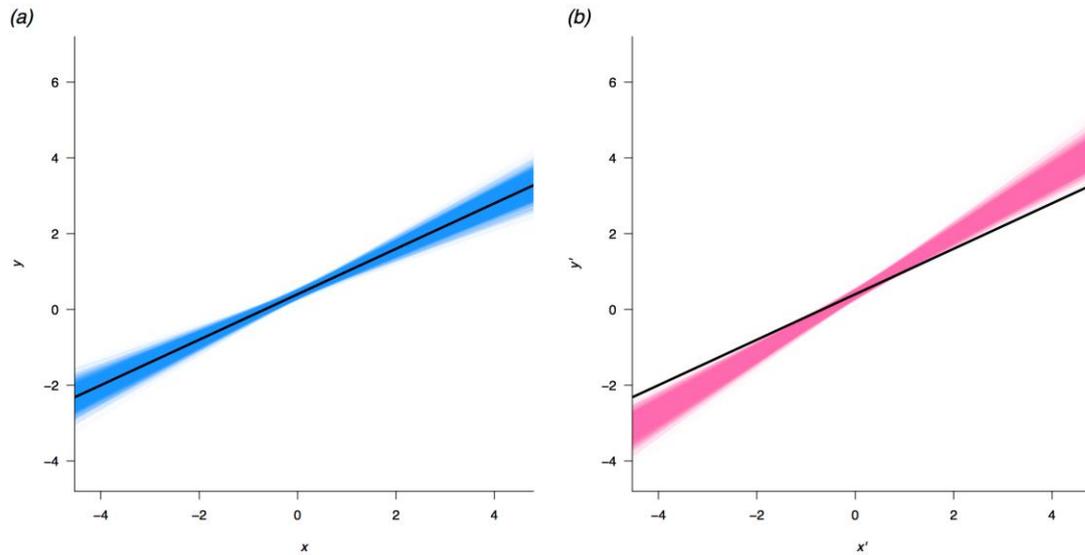
431  
432 Figure 1. Procedure for generating ‘residuals’ from a sampling-driven diversity  
433 model. (a) A paired, bivariate dataset  $x$  (sampling proxy) and  $y$  (sampled  
434 diversity) was simulated so that  $x$  is randomly drawn from a normal distribution  
435 ( $\mu = 0, \sigma = 1$ ) and  $y$  is calculated as  $y = a + bx + e$  where  $a = 0.4, b = 0.6$  and  $e$  is  
436 noise ( $\mu = 0, \sigma = 0.5$ ). The thick black line is the expected relationship  $Y = a + bx$ .  
437 Vertical lines represent the true residuals or deviations in  $y$  from the thick line.  
438 (b) Following Smith and McGowan (2007)  $x$  and  $y$  are sorted from low to high  
439 values independent of each other ( $x'$  and  $y'$  respectively), and an ordinary least  
440 squares (OLS) regression model (pink line) is fitted to  $y'$  on  $x'$ . Despite the pink  
441 line supposedly representing the data generating process, it is clear that it is not  
442 a good estimator of the true known generating process, the thick line. (c) The  
443 OLS model from (b) is used as the sampling-driven diversity model (SDDM) or  
444 the expected relationship between  $y$  and  $x$ , from which ‘residuals’ are computed  
445 as the deviations in  $y$  from the pink line (vertical pink dotted lines). It is  
446 immediately clear that there is a substantial difference between the true  
447 residuals (a) and the SDDM ‘residuals’ (c).  
448



449

450 Figure 2. Regression modelling on a decoupled bivariate dataset fails to estimate  
 451 the simulation slope parameter. (a) A bivariate dataset ( $y$  and  $x$ ) was generated  
 452 so as to follow a theoretical relationship (thick line) with intercept  $a = 0.4$ , slope  
 453  $b = 0.6$  and noise ( $e$  [ $\mu_e = 0$ ,  $\sigma_e = 0.5$ ]). The best-fit regression line (blue) is not  
 454 significantly different from the theoretical line (dashed 95% confidence intervals  
 455 encompass the thick line; see table 1 for Type I error rates over 5000  
 456 simulations), with  $y$  and  $x$  forming a moderately strong relationship ( $r^2 = 0.526$ )  
 457 appropriate for the degree of  $e$  modelled. Regression model residuals (vertical  
 458 lines) show no structure, as expected. (b) The bivariate data in (a) were sorted  
 459 independently of each other ( $y'$  and  $x'$ ), to which a regression model was fitted.  
 460 The best-fit sampling-driven diversity model (SDDM) regression line (pink)  
 461 deviates strongly from the theoretical relationship (dashed 95% confidence  
 462 intervals do not encompass the thick line; table 1), and  $y'$  and  $x'$  form a very  
 463 strong (but false) linear relationship ( $r^2 = 0.973$ ). Regression residuals (vertical  
 464 lines) show clear structure. One pair of model comparison out of 5000  
 465 simulations is shown.

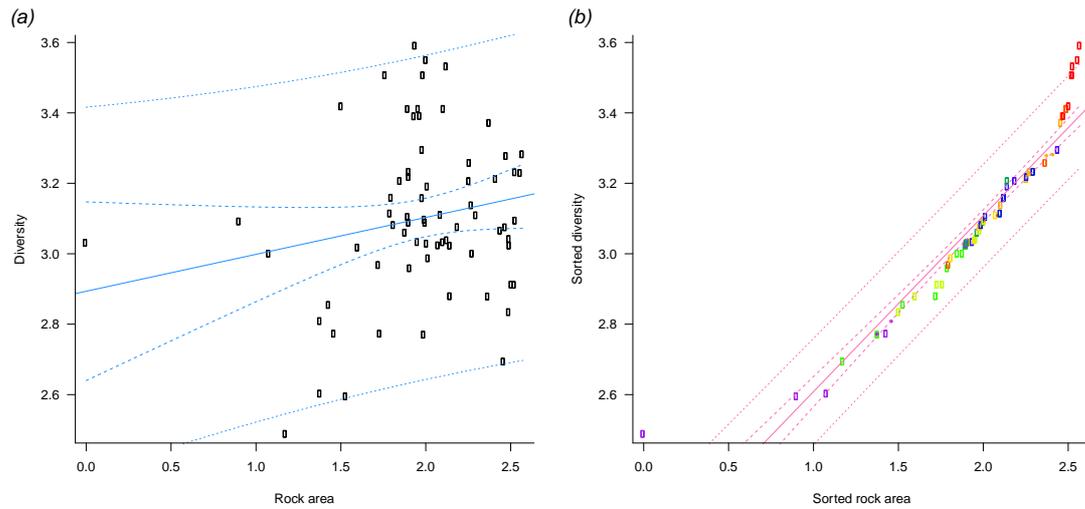
466



467

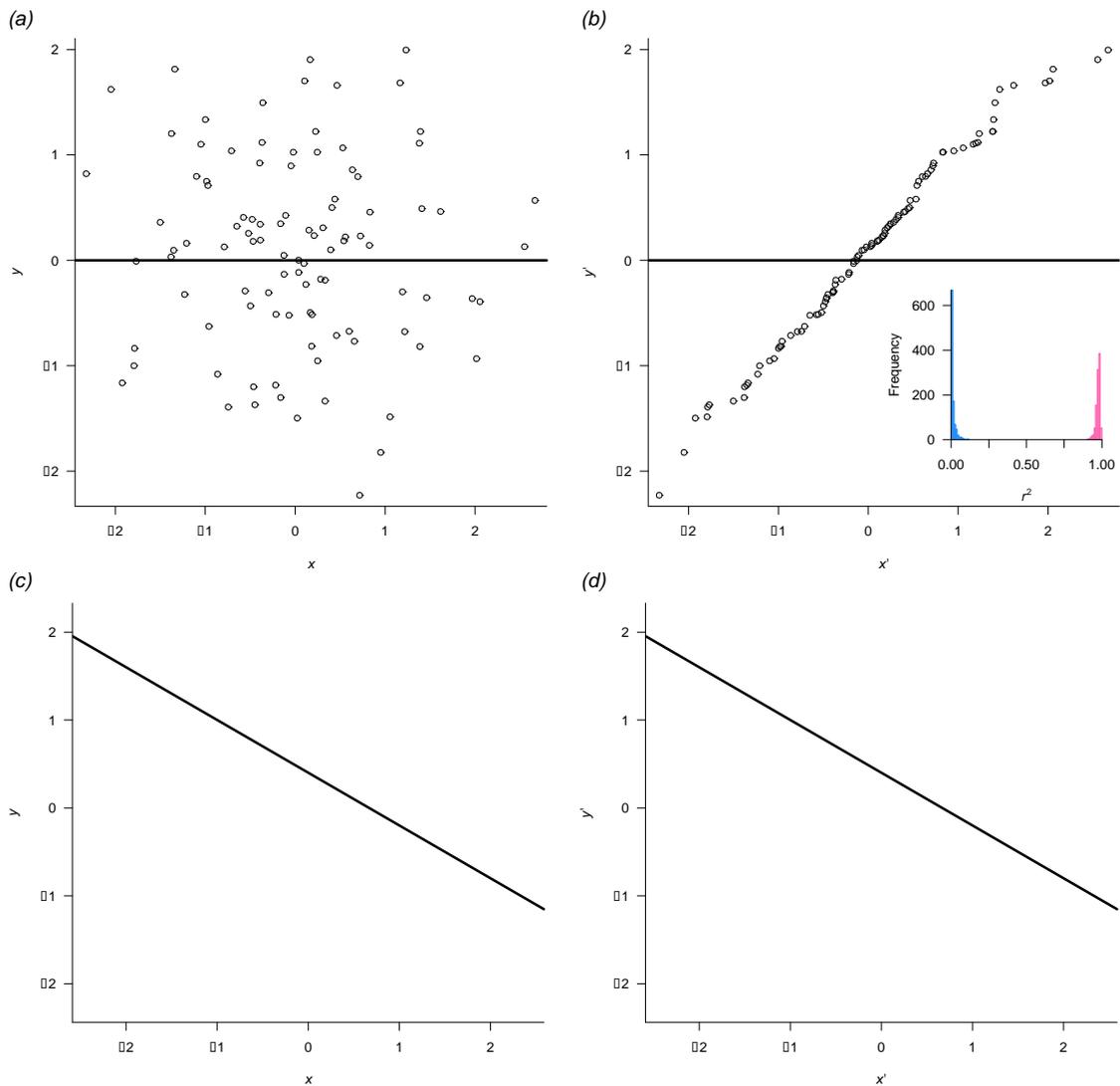
468 Figure 3. SDDM regression predictions are systematically biased. (a) Standard  
 469 regression lines (blue) for 5000 simulated datasets at  $\sigma_e = 0.5$  deviate randomly  
 470 around the theoretical relationship (thick line) with the mean slope showing no  
 471 significant difference from the theoretical slope  $b = 0.6$  (table 2). (b) SDDM  
 472 regression lines on decoupled datasets (pink) deviate systematically away from  
 473 the theoretical relationship (thick line), with a significant difference between the  
 474 mean regression slope and the theoretical slope (table 2).

475



476

477 Figure 4. The difference between the original paired, bivariate relationship (a)  
 478 and the forced, false relationship (b) shown using the data from Smith and  
 479 McGowan (2007). Log-transformed marine generic diversity has a non-  
 480 significant and weak relationship with log-transformed rock area ( $\beta = 0.105$ ;  $r^2 =$   
 481  $0.0398$ ;  $p = 0.0979$ ; a). However, once diversity and rock area are sorted  
 482 independently of each other following Smith and McGowan (2007), then the  
 483 relationship becomes significant and strong ( $\beta = 0.499$ ;  $r^2 = 0.903$ ;  $p < 0.001$ ; b).  
 484 Points are coloured according to their geological age with cooler colours on the  
 485 older and warmer colours on the younger ends of the time scale. Filled and  
 486 outline colours in (b) correspond to the ages of the rock record and diversity  
 487 respectively, and demonstrate visually the mismatch between  $y'$  and  $x'$ . Dashed  
 488 lines are confidence intervals, while dotted lines are prediction intervals.  
 489



490

491 Figure 5. Independently sorting any two variables results in a forced positive  
 492 relationship. (a) Two randomly generated variables  $y$  and  $x$  show no significant  
 493 relationships across 1000 simulations, with the slopes of the regression lines  
 494 (blue) distributed around the expected slope of zero. (b) When regression  
 495 models are fitted on independently sorted datasets ( $y'$  and  $x'$ ), estimated slopes  
 496 are significantly different from the expected value of zero, and result in a strong  
 497 positive relationship ( $r^2 = \sim 1$ ; inset pink) despite the unrelated nature of the  
 498 original datasets ( $r^2 = \sim 0$ ; inset blue). (c) A bivariate dataset ( $y$  and  $x$ ) was  
 499 generated so as to follow a theoretical relationship (thick line) with intercept  $a =$   
 500 0.4, slope  $b = -0.6$  and noise ( $e$  [ $\mu_e = 0$ ,  $\sigma_e = 0.5$ ]). Standard regression lines (blue)

501 deviate randomly around the theoretical relationship with the mean slope  
502 showing no significant difference from the theoretical slope  $b = -0.6$ . (d) However  
503 once sorted independently, regression lines (pink) deviate systematically away  
504 from the theoretical relationship, with all estimated slopes being positive. Thus  
505 SDDM slope estimates are systematically and directionally biased.  
506