

‘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 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

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 Sakamoto¹, Chris Venditti¹ and Michael J. Benton²

9

10 ¹School of Biological Sciences, University of Reading, Reading, RG6 6AJ, UK

11 ²School 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

predictions $\hat{y}$ ' and the observed values y , which are then treated as the 'residual diversity estimates' (figure 1).

However, independently sorting y and x as outlined above decouples a paired, bivariate dataset, and is obviously problematic in statistics. Model fitting on decoupled data (e.g. y' and x') will lead to spurious predictions and 'residuals' as the estimated regression coefficients will be based on a forced (false) linear relationship (figure 1b). However, owing to continued wide use of the SDDM as a preferred method for identifying supposedly 'true' palaeodiversity signals (as recently as (Grossnickle & Newham 2016)), it appears that this basic statistical concept is somehow overlooked. While it has been suggested that the use of formation counts (the number of fossiliferous geological formations – a mappable unit of rock that represents a particular time and set of environments in a particular location – in a given time interval (Benton *et al.* 2011)) to 'correct' palaeodiversity time series data is unlikely to be meaningful because of substantial redundancy of the two metrics (Benton *et al.* 2011; Benton 2015), and a recent study has scrutinized the performance of SDDM residuals in accurately predicting true simulated biodiversity signals (Brocklehurst 2015), the performance of the SDDM itself has never been assessed. Here, we demonstrate the detrimental effects of decoupling data in regression modelling using simple simulations.

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 α and β 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.

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

RESULTS

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

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)

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

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

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

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

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

Given our simulations, we strongly recommend against using the SDDM 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

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

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

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

ACKNOWLEDGEMENTS

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

DATA ACCESSIBILITY

This manuscript does not include data.

FUNDING

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- Hadfield, J.D. (2010) MCMC methods for multi-response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, **33**, 1-22.
- Hannisdal, B., Haaga, K.A., Reitan, T., Diego, D. & Liow, L.H. (2016) Common species link global ecosystems to climate change. *bioRxiv*, 043729.
- Harvey, P.H. & Pagel, M.D. (1991) *The comparative method in evolutionary biology*. Oxford University Press.
- King, G. (1986) How Not to Lie with Statistics - Avoiding Common Mistakes in Quantitative Political-Science. *American Journal of Political Science*, **30**, 666-687.
- Lloyd, G.T. (2012) A refined modelling approach to assess the influence of sampling on palaeobiodiversity curves: new support for declining Cretaceous dinosaur richness. *Biology Letters*, **8**, 123-126.
- Lloyd, G.T., Davis, K.E., Pisani, D., Tarver, J.E., Ruta, M., Sakamoto, M., Hone, D.W.E., Jennings, R. & Benton, M.J. (2008) Dinosaurs and the Cretaceous Terrestrial Revolution. *Proceedings Of The Royal Society B-Biological Sciences*, **275**, 2483-2490.
- O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology and Evolution*, **1**, 118-122.
- Prothero, D. (1999) Fossil record. *Encyclopedia of paleontology* (ed. R. Singer). Fitzroy Dearbon Publishers, Chicago, USA.
- Raup, D.M. (1972) Taxonomic Diversity during the Phanerozoic. *Science*, **177**, 1065-1071.
- Raup, D.M. (1976) Species Diversity in the Phanerozoic: An Interpretation. *PALEOBIOLOGY*, **2**, 289-297.
- Raup, D.M. (1991) *Extinction: bad genes or bad luck?* W. W. Norton, New York.
- Sakamoto, M., Benton, M.J. & Venditti, C. (2016) Dinosaurs in decline tens of millions of years before their final extinction. *Proceedings of the National Academy of Sciences*, **113**, 5036-5040.
- Smith, A.B. & McGowan, A.J. (2007) The shape of the phanerozoic marine palaeodiversity curve: How much can be predicted from the sedimentary rock record of western Europe? *Palaeontology*, **50**, 765-774.
- Stadler, T. (2013) Recovering speciation and extinction dynamics based on phylogenies. *Journal Of Evolutionary Biology*, **26**, 1203-1219.
- Stadler, T., Kuhnert, D., Bonhoeffer, S. & Drummond, A.J. (2013) Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings Of The National Academy Of Sciences Of The United States Of America*, **110**, 228-233.

SUPPORTING INFORMATION

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

TABLES

Table 1. Type I error rates (%) for SRM (Standard Regression Model) and SDDM (Sampling-Driven Diversity Model) estimates (intercept α and slope β) across residual error (σ_e).

σ_e	SRM		SDDM	
	α	β	α	β
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

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 (σ_e).
 428

σ_e	SRM			SDDM		
	mean-slope	t-value	p-value	mean-slope	t-value	p-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

FIGURES

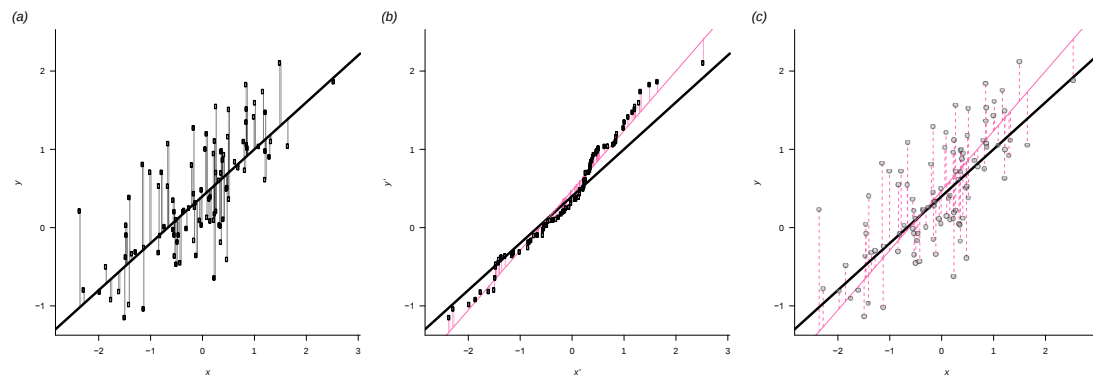


Figure 1. Procedure for generating 'residuals' from a sampling-driven diversity model. (a) A paired, bivariate dataset x (sampling proxy) and y (sampled diversity) was simulated so that x is randomly drawn from a normal distribution ($\mu = 0, \sigma = 1$) and y is calculated as $y = a + bx + e$ where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

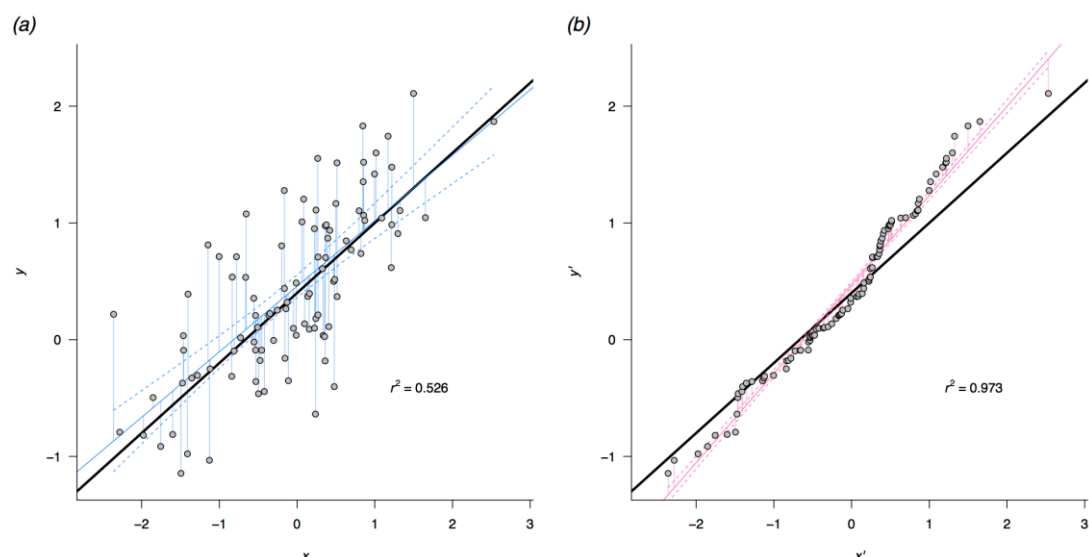
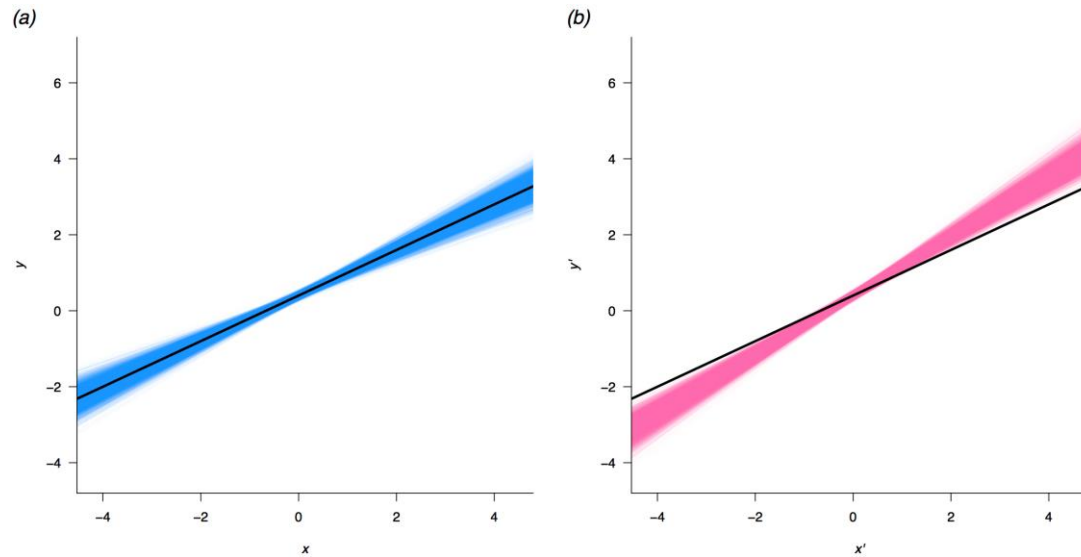


Figure 2. Regression modelling on a decoupled bivariate dataset fails to estimate the simulation slope parameter. (a) A bivariate dataset (y and x) was generated so as to follow a theoretical relationship (thick line) with intercept $a = 0.4$, slope $b = 0.6$ and noise (e [$\mu_e = 0$, $\sigma_e = 0.5$]).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



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

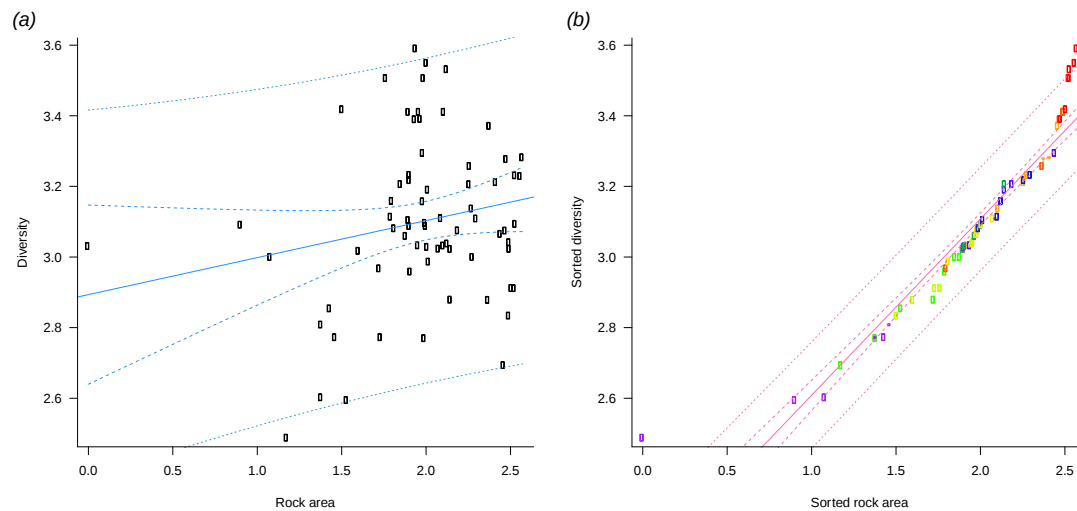
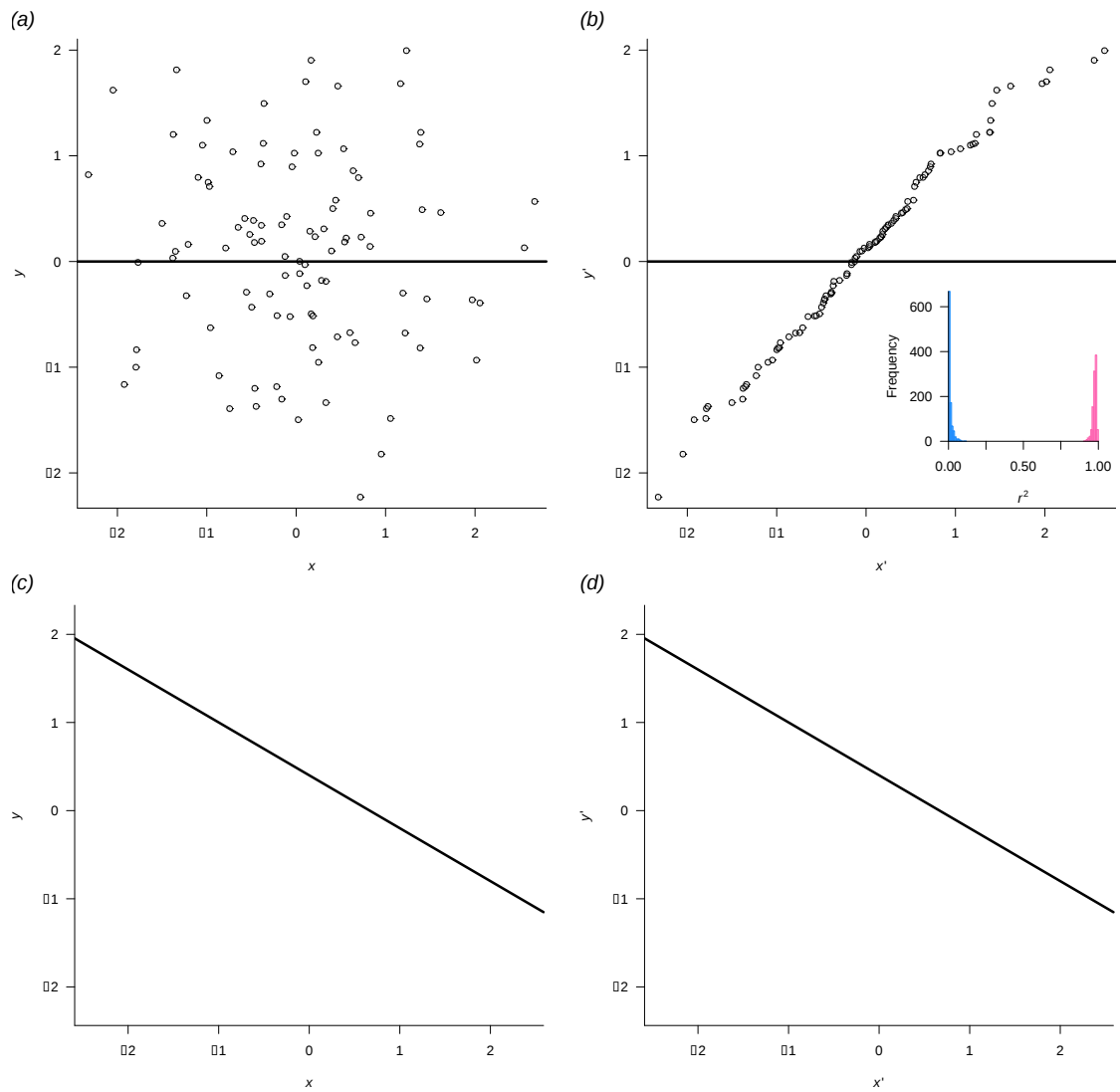


Figure 4. The difference between the original paired, bivariate relationship (a) and the forced, false relationship (b) shown using the data from Smith and McGowan (2007). Log-transformed marine generic diversity has a non-significant and weak relationship with log-transformed rock area ($\beta = 0.105$;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



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 ($r^2 \approx 1$; inset pink) despite the unrelated nature of the
 498 original datasets ($r^2 \approx 0$; 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