Synthese (2021) 198 (Suppl 24):S5919–S5940
https://doi.org/10.1007/s11229-018-1820-x

S.I.: ABSTRACTION AND IDEALIZATION IN SCIENTIFIC MODELLING

Using models to correct data: paleodiversity and the fossil record

Alisa Bokulich[1]

Received: 14 August 2017 / Accepted: 17 May 2018 / Published online: 29 May 2018

Abstract Despite an enormous philosophical literature on models in science, surprisingly little has been written about data models and how they are constructed. In this paper, I examine the case of how paleodiversity data models are constructed from the fossil data. In particular, I show how paleontologists are using various model-based techniques to correct the data. Drawing on this research, I argue for the following related theses: first, the 'purity' of a data model is not a measure of its epistemic reliability. Instead it is the fidelity of the data that matters. Second, the fidelity of a data model in capturing the signal of interest is a matter of degree. Third, the fidelity of a data model can be improved 'vicariously', such as through the use of post hoc model-based correction techniques. And, fourth, data models, like theoretical models, should be assessed as adequate (or inadequate) for particular purposes.

Keywords Paleontology · Paleobiology · Evolution · Data · Model · Suppes · Fossil · Biodiversity · Representation · Simulations · Climate science · Sepkoski · Data models

1 Introduction

One of the most influential early papers in thinking about the philosophy of data is Patrick Suppes's (1962) paper "Models of Data." In this paper, Suppes introduces the seminal notion of a 'data model' and the related concept of a hierarchy of data models. He challenges the simplistic view that there are just two things: 'theory' and 'data', which are directly compared with one another, and argues that "one of the besetting

✉ Alisa Bokulich
abokulic@bu.edu

[1] Department of Philosophy, Boston University, Boston, USA

sins of philosophers of science is to overly simplify the structure of science…. a whole hierarchy of models stands between the model of the basic theory and the complete experimental experience" (Suppes 1962, p. 260). Rather than the "raw" data, what scientists are primarily interested in is a *model of the data*—a processed and abstracted version of the data that has been subjected to various statistical and other analyses.[1]

In this era of 'big data' there has been a renewed philosophical interest in understanding the nature of data in science. Leonelli (2016), in her excellent book *Data-Centric Biology*, identifies a number of key characteristics of data, the most important of which for our project here is the recognition that "despite their scientific value as 'given,' data are clearly made. They are the results of complex processes of interaction between researchers and the world" (2016, p. 71). How exactly data are made in this complex interaction between researchers and the world, and precisely what sorts of manipulations go into the construction of the various data models in Suppes's hierarchy, are questions that have remained surprisingly undertheorized in the philosophy of science.[2]

My aim in this paper is to shed further light on the nature of data models by focusing on the example of how paleodiversity data models are constructed from the fossil record. This methodologically rich case is instructive because it highlights a practice that I suspect is quite widespread in the sciences, despite not having received much philosophical attention—namely, the use of models to correct data. The idea that scientists use models to correct data might prima facie strike one as counterintuitive, if not downright problematic. The intuition here might be that any "model-tampered" data is in fact "corrupted" data. In what follows I argue that this intuition is mistaken. It is not the 'givenness' of data that makes it epistemically privileged, but rather its degree of fidelity, and the fidelity of data can be improved by removing artefactual elements and reducing noise. As we will see in detail in the case of paleodiversity data, modeling is a central means by which this is done. Indeed, models are used not just for correcting the data, but also for testing the adequacy of these data correction methods, by means of computer simulations involving what is called "synthetic" data.

So it is not the 'purity', but rather the *fidelity* of the data that matters. However, it is also important to remember that in assessing fidelity, what counts as signal and what counts as noise depends on the particular uses to which the data set will be put (i.e., what hypotheses the data will be used to provide evidence for or against). Moreover, the fidelity of data in capturing the signal of interest is not all or nothing, but rather is a matter of degree. Hence, rather than speaking of fidelity-full-stop, I will argue that we should instead be thinking of fidelity-for-a-purpose. Just as Parker (2010) cogently argues that *theoretical* models should be evaluated as adequate-for-purpose, so too should we evaluate *data* models as adequate or inadequate for particular purposes.

[1] What has often been overlooked in many discussions of data models is that Suppes's view of data models is tied to the Tarskian 'instantial' view of models. Elsewhere it is argued that the notion of data models should be disentangled from this instantial view, and that data models, like other models in science, should be understood as representations. This move is important not only philosophically for avoiding what van Fraassen (2008) calls the "loss of reality objection," but also for making adequate sense of scientific practice. See Parker and Bokulich (in preparation) for further discussion.

[2] For example, the mammoth *Springer Handbook of Model-Based Science* (Magnani and Bertolotti 2017), though covering many excellent topics in its 53 chapters, fails to have an entry on data models.

This is particularly important in the case of paleontology, where despite great progress in coming to understand—and finding ways to correct for—the many biases, gaps, and noise in the paleodiversity data, the possibility of a perfectly accurate depiction of past life is simply not in the offing. Nonetheless I will show how paleontologists are able to determine a range of purposes for which the various model-corrected paleodiversity data sets are adequate.

In philosophical discussions about scientific methodology, it is important to remain grounded in scientific practice; hence, in the next two sections I examine the historical emergence—and then current state of the art—of these model- and simulation-based data correction methods. In Sect. 2, I briefly trace the history of attempts to read the history of paleodiversity from the fossil record. From the beginning it was recognized that the data from the fossil record are a highly biased and incomplete representation of the history of life. Drawing on the work of historian David Sepkoski, I show how two important threads emerge from this history that are important for our philosophical discussion: First, we see how paleontologists were able to develop an increasingly quantitative understanding of the many different kinds of biases in the fossil record and determine the direction and magnitude of their impact on our picture of paleodiversity. Second, they were further able to make progress in determining how one could begin to mitigate the effects of those biases through the introduction of new computer simulation models and other model-based correction techniques. These two themes came to define what Sepkoski calls the 'generalized'—or what I prefer to call the 'corrected'—approach to reading the fossil record.

With this historical background in place, I turn in Sect. 3 to an examination of how this 'corrected' approach to reading the fossil record has been developed to a high degree of sophistication in contemporary paleontology. In particular, I examine three ways in which models are being used to correct the fossil data: the subsampling model approach, the residuals model approach, and the phylogenetic model approach. I show how scientists then test the reliability and robustness of these various model-based correction methods through computer simulations of hypothetical paleodiversities using synthetic data. In this research, models play a central role in the construction, correction, and testing of data models; hence, we see that models permeate the data-production process, or, as Edwards argues in the context of climate science, "without models there are no data" (2010, p. xiii).[3]

In Sect. 4, I use this case study of model-based data-correction techniques to argue for the following four philosophical theses: First, it is not the purity of the data that matters for epistemic privilege, but rather fidelity. Second, fidelity is a matter of degree. Third, the fidelity of one's data can be improved not just by introducing various forms of *physical* control (e.g., shielding, isolating, purifying) during data collection, but also through various forms of *vicarious* control (Norton and Suppe 2001, p. 72) after the data is collected. Model-based data correction techniques are an example of just this sort of vicarious control. Fourth, fidelity is a function of context; that is, it depends on the uses to which the data model will be put. Data can travel and be repurposed for different projects. As Leonelli explains, data journeys are "the material, social,

[3] A fuller discussion of some of the interesting parallels between data in paleontology and data in climate science is taken up in Parker and Bokulich (in preparation).

and institutional circumstances by which data are repackaged and transported across research situations, so as to function as evidence for a variety of knowledge claims" (2016, p. 5). Hence, it does not make sense to discuss 'fidelity (full stop)', but rather 'fidelity-for-a-purpose'.

In Sect. 5, I argue that we can see the importance of model-based correction techniques not just at the very abstract level of global paleodiversity data, but also much farther down the data-model hierarchy, at the level of the categorized and prepared fossil rocks themselves. Drawing on the work of Caitlin Wylie, I discuss how one level of the data-model hierarchy can be underdetermined by the data-model level below it. Here too we will see the importance of judging data models as adequate (or inadequate) for particular purposes. These themes are drawn together and reiterated in the concluding Sect. 6.

2 A brief history of using fossils to read the history of life

When it comes to studying the history and evolution of life, the fossil record is a unique and vitally important source of data.[4] Very early on, however, it was recognized that the fossil record is a highly incomplete and biased representation of that history. Hence, the actual history of life, and the waxing and waning of its diversity, may differ significantly from what is suggested by a literal reading of the "raw data". Sir Charles Lyell in his 1830 *Principles of Geology* notes,

> [W]e are bound to remember, whenever we infer the poverty of the flora or fauna of any given period of the past, from the small number of fossils occurring in ancient rocks, that it has been evidently no part of the plan of Nature to hand down to us a complete or systematic record of the former history of the animate world….[S]uch failure may have arisen, not because the population of the land or sea was scanty at that era, but because in general the preservation of any relics of the animals or plants of former times is the exception to a general rule. (Lyell 1830, pp. 145–146)

Given the dynamic nature of the Earth and its rocks, coupled with the vastness of time in geological history, the "general rule", as Lyell argues, is that nearly all evidence of past life would be destroyed and lost.

This issue became particularly acute for Charles Darwin who both wanted to use the fossil record to support his theory of evolution by natural selection and was keenly aware that the failure to find a continuous gradation of forms in the fossil record could be used by his critics as evidence against the theory. In his *Origin of Species* (1859), Darwin devotes an entire chapter (Chapter IX: "On the Imperfection of the Geological Record") to this problem, and it is a theme that reappears in several other chapters as well. Darwin rightly recognizes a number of important factors that bias the fossil record (which he summarizes, for example, in Darwin 1859, pp. 341–342).

[4] Of course, the fossil record is not just critical for understanding the processes of biological evolution, but also gives information about the history of the climate and the movements of tectonic plates. Thus, one must pay attention to the purpose for which the data is intended.

In paleobiology today these biasing factors are often referred to as “filters” through which the biological “signal” becomes distorted and partially lost (see, e.g., Benton and Harper 2009).

First, there are taphonomic filters or biases, relating to what types of organisms are likely to get preserved. Organisms with soft bodies are far less likely to be preserved than ones with bones or shells. Even for organisms with hard parts, the chemical conditions of the death site must be right for preservation and mineralization. Second, there are further biological and ecological biases due to whether the species is common, with many individuals and short lifespans, or rare; and its ecological location and migration behavior may be relevant as well.

Third, as both Lyell and Darwin note, there are many geological sources of bias as well. Only some environments are sites of sediment deposition; sites where there is rapid erosion will not be preserved. Even if a fossil is preserved initially, tectonic movements involve temperatures and pressures that can metamorphose the rock, destroying the fossil. Even if the fossil survives these tectonic movements, it needs to be uplifted to the surface where it can be found, and moreover be found before being destroyed through further erosion.

Finally, there are various anthropogenic biases, such as the unlikely event the fossil is actually found and identified. Geographical biases can arise from the collecting efforts of paleontologists: the majority of fossils today have been collected in Europe and North America, while other parts of the world are not as well explored. Additional anthropogenic biases may arise from the interests of collectors in certain “charismatic species,” and as Darwin notes, the fossil must be recorded in a museum collection (or today a computer database), and not just end up in someone’s private collection, in order for it to become a part of the scientific record. A detailed understanding of these many different biases in the data of the fossil record—and more importantly the development of sophisticated analytical techniques to correct for them—is thus critical for understanding the rise and fall of taxonomic diversity throughout history.

The field of paleobiology arguably came into its own in the 1970s, in what Sepkoski and Ruse (2009) have called the ‘paleobiological revolution,’ where there was a movement to not just collect and describe individual fossils, but to conduct large-scale quantitative analyses of patterns in the history of life (Sepkoski 2012a). The historian Sepkoski (2012a, b, 2013, 2016) recounts in detail how the paleobiological revolution can be traced to a small, influential group of paleontologists—including Stephen Jay Gould, Thomas Schopf, Dan Simberloff, and David Raup—who met at the Marine Biological Lab (MBL) in Woods Hole, Massachusetts, and sought to introduce new quantitative methods and the use of computer simulation models into a hitherto merely “idiographic” paleontology. A key outcome of this collaboration was a computer simulation model known as the MBL model,[5] which could be used to stochastically generate “synthetic” phylogenetic trees, with patterns of speciation and extinction. The MBL model, which was a minimal model largely devoid of biological assumptions, could then serve “as a ‘base level model’ or ‘criterion of subtraction’ for ascertaining what amount of apparent order requires no deterministic cause [and]… then seek standard

[5] For more on the MBL model see, for example, Huss (2009).

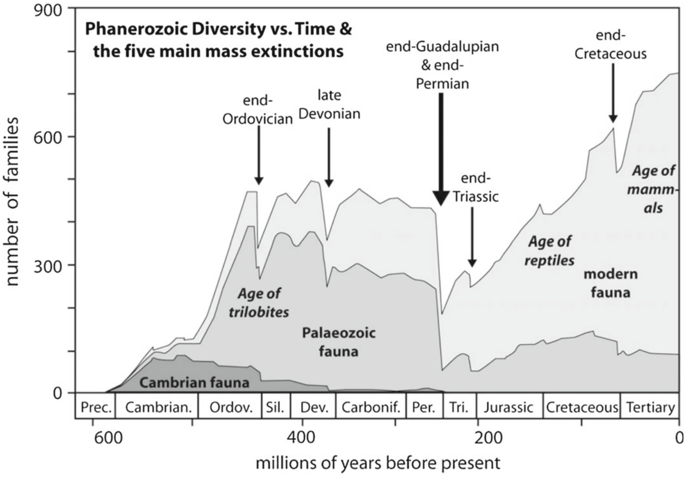

Fig. 1 Sepkoski's curve representing how marine fauna diversity has varied over time from the Cambrian through Tertiary Period, with the "big five" mass extinctions indicated. (Metcalfe and Isozaki 2009, Fig. 1, after Sepkoski 1984; with permission from Elsevier)

explanations for the residuum of order" (Gould et al. 1977, p. 24).[6] Although not listed as an author on the early MBL papers, Alroy (2010b, p. 70) recounts that the FORTRAN code used in the MBL model was written by John J. "Jack" Sepkoski, who was at the time Gould's graduate student assistant.[7]

Jack Sepkoski is best known for the key role he played in the other pillar of the paleobiological revolution, namely the construction of large-scale computer databases of global fossil data. While Sepkoski was a graduate student at Harvard in 1973, Gould set him on the mammoth task to "compile data on all orders within classes and as many families within orders and genera within families [over the past 600 million years] as [he] could obtain" (Sepkoski 1994, p. 135). This project involved 10 years of digging, not in the dirt, but in the library, and resulted in his *Compendium of Fossil Marine Families* (1982) and his famous paleodiversity curve, now referred to as the "Sepkoski curve" (see Fig. 1).

Although Sepkoski was well aware that his data on paleodiversity were highly imperfect, they nonetheless proved an adequate representation of the history of marine life for drawing some conclusions about large-scale patterns in paleodiversity, such as the discovery of three distinct marine faunas (the trilobite-rich Cambrian, the brachiopod-rich Paleozoic, and the bivalve-gastropod-rich Modern faunas) and the discovery of the "big five" mass extinctions.[8]

[6] Such subtraction models play an important role not only in current paleontological research (e.g., Smith and McGowan 2007, "residuals method"), but also in current climate research, where they have been termed "intermediate models" (e.g., Edwards 2001, p. 61).

[7] The historian David Sepkoski is the son of the paleontologist Jack Sepkoski.

[8] This issue of the adequacy of a data model for a purpose will be discussed further below.

In the early work of these MBL collaborators, historian David Sepkoski shows that one can see three different approaches to “reading the fossil record”: an optimistic (or what he calls “literal”) reading; a highly abstract, idealized reading, largely detached from the historical data; and what Sepkoski calls a “generalized,” or, as I prefer to call it, “corrected” rereading of the fossil record, which uses simulation models not to replace, but rather to correct the historical data.[9]

The first “optimistic” reading can be seen in the most influential paper that appeared in the proceedings of a 1971 symposium on models in paleobiology organized by Schopf: Niles Eldredge and Stephen Jay Gould’s now famous paper on “punctuated equilibrium.” This paper, following the conclusions of Eldredge’s dissertation work on the fossil record of Devonian trilobites, argues that evolution proceeds not through a constant gradualism, but rather is characterized by long periods of stasis, in which species appear stable and do not undergo any cumulative change, that are then interrupted by short periods of rapid evolutionary change, effected through the geographical isolation of a much smaller population. If this is the dominant mode by which evolution takes place, then one would *not* expect to find the continuous gradation of forms between species that Darwin worried so much about being largely absent from the fossil record.

Eldredge and Gould’s conclusion in this paper is that paleontologists have been misled by an excessive pessimism about biases in the fossil record. They conclude,

> [M]any breaks in the fossil record are real; they express the way in which evolution occurs, not the fragments of an imperfect record…. Acceptance of this point would release us from a self-imposed status of inferiority among the evolutionary sciences. The paleontologist’s gut-reaction is to view almost any anomaly as an artifact imposed by… an imperfect fossil record…. We suspect this record is much better… than tradition dictates. (Eldredge and Gould 1972, pp. 96–97)

While Eldredge and Gould were right to suggest that paleontologists were too quick to dismiss unexpected patterns in the fossil record as “noise” rather than a genuine “biological signal”, the well-documented biases in the fossil record, which were increasingly being understood in quantitative detail, precluded a wholesale reading of the fossil data at face value.[10]

An alternative approach, championed by David Raup, is to construct a “corrected” reading of the fossil record. In a 1972 paper, Raup, like Darwin, notes that “systematic biases exist in the raw data such that the actual diversity picture may be quite different from that afforded by a direct reading of the raw data” (p. 1065). Before data can be corrected, however, the relevant sources of bias—and an understanding of the concrete effects or artifacts that those biases produce on the data—need to be identified. Raup discusses seven biases that affect the diversity counts. Among these are the fact that the durations of geological time units are not all the same (a long time interval will show higher diversity than a short one) and the “Lagerstätten effect”.

[9] Due to limited space, I will only very briefly discuss the first, skip the second, and focus primarily on the third "corrected" approach to reading the fossil record.

[10] For an excellent philosophical discussion of punctuated equilibrium in connection with paleontology see Turner (2011).

Lagerstätten are geological sites, such as the famous Burgess shale, where a (typically anoxic, rapid sedimentary) environment led to exceptionally good fossil preservation, including soft tissue records. As Raup notes, the distribution of Lagerstätten through time is not uniform, hence time periods that have a Lagerstätte preservation site will lead to increased diversity estimates over those time periods without Lagerstätten.[11] Raup further identifies a cluster of biasing factors that is referred to as the "Pull of the Recent:" For example, not only are younger (more recent) rocks likely to have better preservation of fossils and have a broader geographic representation today, but various taxonomic practices can also contribute to the Pull of the Recent. The point of enumerating these problems, however, is not just to lament the biases in the fossil record, but to determine the direction and magnitude of their effects on the observed diversity curve, and ultimately to find ways to "correct" the data by appropriately adjusting the diversity curve in light of these biases.

A particularly noteworthy innovation in Raup's (1972) paper is his new proposed methodology for how this data correction research program can be carried out. His proposal is to use the newly developed simulation model to generate an idealized "synthetic" (or hypothetical) initial diversity distribution (i.e., before fossilization), then add into the simulation model various "biases" that would delete various portions of the record, and finally compute the resulting diversity curves. Raup concludes,

> The simulation demonstrates that diversity patterns such as are observed in the fossil record can be produced by the application of known biases to quite different diversity data. The simulation does not of course prove the alternative model for Phranerozoic diversity because of our present ignorance of the actual impact of the biases. (Raup 1972, p. 1071)

Raup recognizes that there is an underdetermination problem here in that multiple combinations of initial diversity curves plus biases could reproduce the observed data, and thus he sets paleobiology with the following task:

> There are undoubtedly other plausible models as well, depending on the weight given to each of the biases. Future research should therefore be concentrated on a quantitative assessment of the biases so that a corrected diversity pattern can be calculated from the fossil data. (Raup 1972, p. 1071)

An enumeration of the various biases in the data and a quantitative understanding of their effects on that data are thus essential to the project of correctly reading the history of paleodiversity from the fossil record.

In addition to introducing the use of computer simulations for fixing biases in the fossil data, Raup (1975) also introduces a second important tool for constructing corrected data models, known as rarefaction or subsampling.[12] As Raup explains,

[11] As an example, Raup notes that the observed diversity of insects during the Cretaceous is essentially zero, not because the actual diversity was zero, but because of the absence of Lagerstätten of this time period to record them.

[12] This method was first developed by the Woods Hole benthic ecologist Howard Sanders. While ecologists tend to use the term 'rarefaction', paleontologists typically prefer the term 'subsampling' (see Alroy 2010b, p. 61 for a discussion of the terminology).

“rarefaction is basically an interpolation technique making it possible to estimate how many species would have been found had the sample been smaller than it actually was” (Raup 1975, p. 333). Paleobiology in the twenty-first century has pursued with great advantage these two correction methods, and in what follows we examine both the current state of the art of this “corrected” approach to reading the fossil record and the philosophical lessons it can teach us about data modeling more generally.

3 Paleodiversity and correcting the fossil record: three approaches

Simply counting the number of taxa (e.g., species, genera, families) that appear in the fossils from each successive geological time interval provides what is called the “raw taxic diversity,” but as we saw in the last section, scientists from the beginning have recognized that this highly biased data should not be accepted at face value. There are currently three broad methods for correcting the fossil data, which will be discussed in turn: (1) Subsampling approaches, (2) Residuals approaches, and (3) Phylogenetic approaches. All three of these approaches involve the use of models in some way.

The first approach to correcting paleodiversity data is rarefaction or subsampling. The aim of subsampling methods is to correct those biases in the fossil data that arise from differences in the sample size. Although a complete or comprehensive sample is not possible in paleontology, the aim is to correct the data so that it is at least a “fair” sample. However, what does it mean to have a “fair” sample? In what is now referred to as the “classical rarefaction” method introduced into paleontology by Raup (1975), it was assumed that a fair sample was one that was uniform—that all the samples had roughly the same number of individuals (either specimens or more often in paleontology “occurrences”, which is the number of taxa in a collection of specimens). In a series of papers published in 2010, John Alroy argues that the classical rarefaction method is not in fact adequate for correcting these sampling biases. Intuitively, the concern is that when diversity (or “richness”) is low and a species is very common, you don’t need to sample much to find out what there is. When diversity is high and any given species is more rare, you need to sample harder to get an accurate picture of what there is.

Alroy argues that to correct the data for sampling biases one should “track not the number of items that are drawn but the ‘coverage’ of the data set represented by the species that have been drawn…. The coverage of any one species is its relative frequency” (Alroy 2010a, p. 1216). This approach makes use of a method developed by Alan Turing and his co-worker at Bletchley Park, Jack Good, to estimate the total population frequencies of species represented in a sample when little is known about the underlying population (Good 1953). Alroy calls this correction method shareholder quorum subsampling (SQS), though it more frequently referred to today as coverage-based rarefaction.

The SQS method is a significant improvement over classical rarefaction in correcting for sampling biases, though it does not, of course, address the problem of unknown taxa (which may render the coverage of the entire frequency incomplete) and it depends on the idealized assumption of random sampling, which does not hold in the case of real fossil data. Subsampling methods also require large databases of fossil information to be effective, such as the continually growing Paleobiology Database (PaleoBioDB), where the SQS method is offered as an analysis algorithm through the Fossilworks.org gateway to PaleoBioDB. For some taxa, however, the fossil data are simply too sparse to use subsampling methods.

When there are multiple subsampling methods available (e.g., classical rarefaction versus SQS), the question becomes which—if any—is a reliable method for correcting the fossil data? While there are certainly relevant theoretical and conceptual considerations, one can also assess the adequacy of these correction methods empirically, even without having access to the true historical diversity curve with which to compare it. This is done by means of a computer simulation of a hypothesized initial diversity (i.e., using synthetic data) against which the adequacy of various subsampling methods can be tested (e.g., Collins and Simberloff 2009; Alroy 2010a, p. 1218).

The second broad approach for correcting biases in the fossil data is the residuals method. The central idea behind this method is to see the "raw"[13] taxic diversity curve from the fossil data as a combination of biological and geological (as well as anthropogenic and other) signals.[14] If one can model the effects of the geological signal alone, then one could "subtract" it from the raw diversity curve, leaving behind the desired predominantly biological signal. The geological signal is understood as "the amount of sedimentary rock preserved at outcrop"—sedimentary rock, because that is the type of rock in which fossils are formed and preserved, and 'at outcrop', because tectonic and erosional processes need to bring the sedimentary rock from that time period up to the surface of the Earth where it can be found by paleontologists. The problem for constructing paleodiversity data curves over time is that the amount of sedimentary rock available at outcrop from different geological time periods is highly variable.

The data correction method of Smith and McGowan (2007) involves constructing a model in which rock outcrop area is taken to be a perfect predictor of sampled diversity and then using this as a "subtraction model" to obtain the hitherto masked biological signal. The intuition is that the remaining "residual" part of the paleodiversity signal, which is unexplained by the rock outcrop area, can be attributed to the genuine biological signal (by a sort of Mill's method of residues). Smith and McGowan's approach initiated a whole family of residual model approaches. Again the adequacy of these various methods can be tested by means of simulation models, where one starts with the synthetic data of a hypothetical initial diversity, then introduces various sampling biases to produce the biased "observed paleodiversity", and then evaluates how well the data correction methods are able to recover the initial diversity.

[13] Note that the raw taxic diversity estimate is not really "raw," insofar as it already involves substantial theoretical categorization, cleaning up, and processing. Paleontologists often seem to use the term 'raw' to refer to the level of data model below the data-correction techniques they are investigating; hence it is a term that can shift with context.

[14] My use of the notion of "signal" here bears some affinity to Turner's (2007) informational interpretation of traces (e.g., 18–20). More recently Currie (2018, Chapter 3) has argued that a strictly ontological notion of trace, such as the informational view, should be replaced with an epistemic notion of trace that builds in the notion of evidential relevance. A discussion of these interesting issues is outside the scope of this work.

Such a simulation-based study of the effectiveness of various residual model data correction methods was recently carried out by Neil Brocklehurst. Comparing a corrected data model for paleodiversity against the raw paleodiversity he shows that the

> optimum implementation of the residual diversity estimate consistently outperforms the raw, taxic diversity estimate…. This method is indeed an appropriate method to correct for sampling and can provide a better representation of the true history of a clade than the raw data. (Brocklehurst 2015, p. 10)[15]

In other words, the data that have been corrected via the residuals "subtraction model" method are a more accurate, more reliable representation of the history of diversity (as tested and shown by simulation modeling[16]), and hence, are better data to use in testing macroevolutionary or other hypotheses. For our philosophical project here, it is important to note the representational language being used: the raw data are a *representation* of the history of biodiversity, albeit an imperfect one. The concern is to develop data-correction methods that will produce a *better* representation of the history of diversity; however, if one is not careful in adequately developing and testing these data corrections methods, then one can end up with a data model that is a *worse* representation of this history.

In saying that simulation tests indicate that some residuals corrected data are better than the raw, one does not mean that they are a perfectly accurate depiction of the history. Paleontologists are not under any illusion that there is such a thing as a perfect data model that is indistinguishable from the history of biodiversity. There is a whole continuum of data models of varying accuracy. As will be discussed more later, the relevant question is whether the data are adequate for the uses to which they are being put. For different purposes, different correction methods and data models may be more or less appropriate.

So far we have examined two different approaches to correcting the data from the fossil record: subsampling model methods and residual model methods. A third prominent approach to correcting the fossil data is known as the phylogenetic model method (Norwell 1993; Smith 1994). This method makes use of cladistic analysis and phylogenetic tree models to correct gaps in fossil data. Cladistics is a method for inferring ancestral relationships among taxa using 'characters,' which are typically morphological (e.g., anatomical) or genetic traits. On the basis of the similarities and differences between those traits, one then constructs a cladogram (by, e.g., using parsimony or maximum likelihood). In the cladistic approach, it is assumed that a group of organisms is related by descent from a common ancestor, and that when a lineage splits it divides into exactly two 'sister' taxa, which appear at the same time.

[15] It should be noted that there are many different ways to implement residual diversity model corrections (involving, for example, different choice of proxies); hence, Brocklehurst's conclusion here only applies to the "optimal" implementation of the method. Significant problems have been raised with other widely-used implementations of the residuals method, especially those that use the more restricted clade-bearing formations as the proxy (see Sakamoto et al. 2017 for a discussion). I thank Mike Benton (personal communication) for underscoring this point.

[16] These tests are of course fallible, depending on the reliability of the assumptions made in the simulation; however, this is arguably no different than elsewhere science, which is understood to be an iterative, ongoing process.

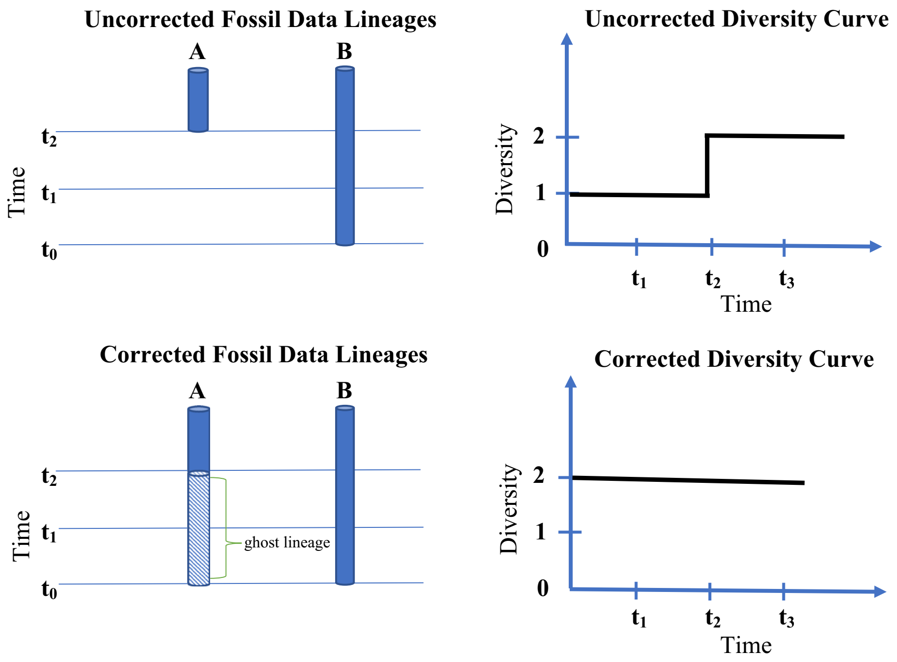

Fig. 2 Phylogenetic model corrected data with ghost lineage added for taxon A. Note the differences between the corrected and uncorrected paleodiversity curves. (Redrawn after Upchurch and Barrett 2005)

This assumption is critical to the phylogenetic correction method in that it licenses the inference that any taxon is as old as its sister.

The phylogenetic method allows one to correct the fossil data by filling in certain gaps as follows:[17] Consider two taxa A and B that cladistic analysis has determined are sister taxa. The first appearance of A in the fossil record is at time t_2 while the first appearance of B in the fossil record is at an earlier time t_0 (see Fig. 2).

Since A and B descended from a common ancestor that existed prior to B at t_0, there must be a lineage linking A from t_2 back in time to its first appearance at t_0. Because A is not actually observed in the fossil data as existing in the stratigraphic interval from t_0 to t_2, but is only inferred, it is called a ‘ghost lineage.’ Note that this corrected phylogenetic diversity estimate (PDE) will be different from the raw taxic diversity estimate (TDE), because A will be added to the diversity count for that earlier time period, even though no fossils of A were found in that time period. The diversity curves will likewise be different: “TDE suggests that only one taxon (B) is present during time t_0, so that the appearance of A at t_2 would be interpreted as an increase in diversity. The PDE, in contrast, suggests that diversity has remained constant during t_0–t_2” (Upchurch and Barrett 2005, p. 108).

A second way that phylogenetic methods correct fossil data is by using what are known as ‘Lazarus’ taxa. A Lazarus taxon is a taxon that disappears from the fossil

[17] This example follows Upchurch and Barrett (2005, p. 108).

record for a long period of time, suggesting that it has gone extinct, but then a representative appears again millions of years later or even as a living specimen in the present.[18] A famous example of a Lazarus taxon of the latter sort is the coelacanth, which is a lobe-finned fish. Although coelacanths have a long fossil record from the Devonian to the Cretaceous, they do not appear in the fossil record after the Cretaceous and were thought to be extinct until a live specimen was caught off the coast of South Africa. Given our theoretical understanding of evolution, there must be a continuous lineage that connects the Cretaceous population to the present population, and hence a ghost lineage is added to taxon counts in the intervening 80 million years, even though no fossils of coelacanths appear in that stratigraphic interval.[19] Hence the phylogenetic method corrects the fossil data by filling in gaps in the stratigraphic (temporal) range of a taxon on the basis of what can be theoretically inferred from cladistic analysis. Phylogenetic correction methods are of course only as good as the cladograms or phylogenetic trees on which they are based, and these in turn can be revised in light of new data or analyses.

Ghost lineages can extend the range of a taxon either forward in time (as in the case of the coelacanth) or backwards in time (as in the previous example of the ‘A’ taxon), though the latter is far more common. Even when one does not find a Lazarus taxon, it is still possible that there is a portion of the lineage after the last appearance of a taxon that is simply unsampled in the fossil record—what is sometimes called a ‘zombie lineage.’[20] These zombie lineages cannot, however, be inferred on the basis of phylogenetic methods. There is thus an asymmetry in the phylogenetic correction method in that, while origination times are frequently extended backward, extinction times are less likely to be extended forward (see, e.g., Foote 1996).

As with the residual model correction method, the reliability of phylogenetic-model corrected data methods can be tested by means of computer simulations involving synthetic data. In a study initiated by Jack Sepkoski and Christine Janis (published after Sepkoski’s death by Lane et al. 2005) a computer simulation known as GHOSTRANGE was used to test two central problems with phylogenetic methods: the asymmetry of the corrections backward in time (but not forward) and the problem of incorrect phylogenies. They summarize the results of their simulation analyses as follows:

> [W]e show here that in the majority of the diversification scenarios simulated the phylogenetic method of estimating diversity [PDE] is superior to the taxic [TDE]…. However, the expected backward skew in diversity predicted by the biased nature of only correcting the first appearance times of taxa… is apparent in many other circumstances. These include time intervals leading up to an ‘event horizon’ such as a mass extinction event, the termination of a clade, or end of an analysis time period. (Lane et al. 2005, p. 30)

[18] Lazarus taxa, which are genuine descendants, must be carefully distinguished from ‘Elvis taxa’, which are not actually descendants of the original taxon, but merely appear to be, due to a similar morphology resulting from convergent evolution (Erwin and Droser 1993).

[19] The story of the coelacanth along with a clear illustrations of ghost lineages can be found at http://www.ucmp.berkeley.edu/taxa/verts/archosaurs/ghost_lineages.php.

[20] Lane et al. (2005) propose the term ‘zombie lineage’ for the unsampled terminal (as opposed to initial) portion of a taxon’s range (pp. 22–23), though some authors use ‘ghost lineage’ for both.

In other words, their simulation studies not only show that the phylogenetically corrected data is a better representation of the "true" simulated paleodiversity than the raw taxic data under most scenarios, but also specifies those scenarios where PDE breaks down and becomes unreliable. In those latter scenarios where it breaks down, they show how PDE contributes to what is known as the Signor–Lipps effect (Signor and Lipps 1982), whereby a number of biasing factors (related to reduced sample size and artificial range truncation) will cause diversity to appear to decline gradually prior to a mass extinction event. Biasing effects on paleodiversity data curves such as these play a central role in the high-profile debate about whether or not the nonavian dinosaurs were in a long-term decline prior to the Chicxulub asteroid impact at the K-Pg (formerly K-T) boundary that led to their extinction.

Although the performance of these data correction methods within the context of a computer simulation is not a perfect indicator of their performance when it comes to real-world data, it is important information to take into account, and arguably provides minimum constraints on the adequacy of any method.[21] While simulation studies seem to clearly show that model-corrected data using any one of these correction methods typically outperform the raw taxic diversity data in providing a better representation of paleodiversity, it is not clear that they can show that one of these data correction methods is always better than the others. Which method is more reliable in any given context is likely going to depend on which types of organisms one is looking at.[22] For example, when it comes to terrestrial vertebrates (such as the dinosaurs), despite the highly incomplete and biased data, one can work out fairly reliable phylogenies because vertebrate remains give many diagnostic characters for cladistic analysis. Hence, the phylogenetic-model correction method is likely to be a reliable tool for correcting terrestrial vertebrate data. On the other hand, when it comes to marine invertebrates, despite a much more complete fossil record, phylogenetic correction methods are less likely to be as reliable. This is because shell geometry, for example, gives very few diagnostic characters to use in phylogenetic reconstruction. Hence different data correction methods may work better for different groups. For these sorts of reasons, paleontologists typically argue that multiple correction methods should be used in coordination (e.g., Foote 1996). Indeed the more one can learn about the strengths and weakness of various correction methods, the better one can guard against the biases they may introduce, and the more effectively they can be deployed.

4 Model corrected data: not purity, but fidelity-for-a-purpose

The process of collecting fossil data together to paint a picture of how biodiversity has changed across the globe from the Cambrian explosion 541 million years ago until the present is an example of what Paul Edwards, in the context of climate modeling, calls *making data global*. He defines this as "building complete, coherent, and consis-

[21] As Brocklehurst notes, a method that cannot even perform well in the simplified simulation scenario is unlikely to perform better under the more complicated conditions found in the real world (2015, p. 12).

[22] I am very grateful to an anonymous referee for calling my attention to this important point and the following examples.

tent global data sets from incomplete, inconsistent, and heterogeneous data sources" (Edwards 2010, p. 251). It involves not only mammoth compilation and standardization projects, such as that undertaken by Sepkoski (1982) and the PaleoBio Database, but also involves the various modeling methods described in the last section, whereby sophisticated interpolation, correction, and subsampling techniques are applied to correct for biased and gappy data.

As we saw in detail in the previous section, the construction of paleodiversity data models involves the use of various other models to construct, correct, and test the data at almost every step. In the case of subsampling approaches to creating a corrected data model of the fossil record, computer models are used both to carry out the random subsampling algorithm and to test, via simulation studies, the ability of these methods to correct for the sampling biases in the "raw" data, without introducing further biases of their own. In the residuals approach, subtraction models that represent the biasing effect of the geological record are constructed and then used to filter out this geological signal from the raw fossil data, leaving behind a more accurate biological signal of the paleodiversity. The reliability of these methods too were tested using further simulation models. Finally the third data correction approach uses cladistic models of phylogenetic relationships to interpolate (i.e., fill in) some of the data missing from the extant fossil record. As with the other two approaches to correcting the fossil data, the reliability and robustness (e.g., under ignorance of true phylogenies) of these methods were further tested via simulation models.

Traditionally it is assumed that the "purer" or less processed the data is, the more epistemically reliable it is. In the case of paleodiversity and the fossil record examined here, we saw just the opposite. As simulation studies showed, both the optimal residuals-model-corrected fossil data and the phylogenetically-corrected data did a better job tracking the "true" paleodiversity than the raw fossil data did. The purity of the data is not a measure of its epistemic reliability. Indeed the epistemic reliability of data at any level in the data-model hierarchy is something to be assessed and not assumed. As Edwards notes,

> Instead the question is how well scientists succeed in controlling for the presence of artifactual elements in both theory and observation—and this is exactly how the iterative cycle of improving data… proceeds. (Edwards 2010, p. 282)

In other words, it is not the purity but rather the *fidelity* of the data that matters.

A central part of empirical research is the continual development of new techniques to improve the fidelity of data by learning to identify and then control, shield, or compensate for various sources of distortion in the data. Norton and Suppe have introduced the helpful distinction between *physical control* and *vicarious control* (2001, p. 72). Physical control is what we are all familiar with in the context of experimentation: one tries to isolate the variable we are interested in measuring by physically removing (e.g., by reducing friction or purifying a sample) or shielding from (e.g., the Earth's magnetic field, air currents, or radiation) other factors that can come into influence the result of our measurement in unwanted ways. In the context of laboratory-based science one typically tries to accomplish this through a well-designed experimental setup. In many cases, however, (both inside and outside the context of laboratory-based science) there can be sources of noise or error that are hard to control by physical means. The notion of vicarious control describes the removal of unwanted effects *after* the experiment is conducted by measuring (or estimating) their influence and then removing them (e.g., mathematically) during data reduction.[23] Learning what all the sources of error are, and how to most effectively control or compensate for them—both physically and vicariously—is something science seeks to continually improve through further research in an iterative cycle of data improvement.

It is important to recognize that the fidelity of one's data in representing some facet of the world need not be all or nothing, but rather is a matter of degree. The key question is not whether the data model is a perfectly accurate depiction, but rather whether it is a representation that is adequate for the purposes to which the data model will be put. In other words, the adequacy of a data model depends on what sort of theoretical claims it is intended to provide evidence for or against. In the more general context of theoretical models, Wendy Parker cogently argues that "what we can sensibly aim to test or confirm are not scientific models themselves, but their adequacy for particular purposes" (Parker 2010, p. 291). Model evaluation should, thus, be understood as an activity to determine the set of purposes for which a model is adequate. I want to explicitly extend this notion of adequacy-for-purpose to *data models* as well.[24]

One can see this issue of the adequacy of a data model for a purpose in the case of paleodiversity data in paleontology. As we saw in Sect. 2, the raw taxic diversity data models were sufficient to provide evidence that the tempo and mode of evolution did not always proceed by gradualism, but rather, as Eldredge and Gould (1972) argued, could proceed through a process of punctuated equilibrium. However, in Sect. 3, we saw that the raw taxic diversity data models were not adequate for the purpose of resolving whether the nonavian dinosaurs were in a long term decline prior to the Chicxulub impact. To provide adequate evidence for or against this hypothesis, a phylogenetic-corrected data model of the fossil record is required (see, e.g., Sakamoto et al. 2016, 2017).

In their article "Assessing the Quality of the Fossil Record", Michael Benton and colleagues detail the range of studies for which current representations of the fossil record, despite the many known biases, are still adequate:

> [T]he fossil record, error-ridden and incomplete as it is, is *adequate for many purposes*, although *none of these provides evidence that error in the fossil record is negligible*: (1) the order of fossils in the rocks generally matches closely the order of nodes in morphological or molecular trees;… (2) at coarse scales of observation (families and stratigraphic stages), there is no evidence that this matching becomes worse deeper in time;… (3) macroevolutionary patterns, including posited mass extinctions and diversifications, are largely immune to changes in palaeontological knowledge;…(4) congruence between stratigraphy and phylogeny has also been largely stable through the twentieth century, despite an order-of-magnitude increase in the number of fossils;… (5) new fossil finds,

[23] Data reduction is just another term for the process by which raw data is turned into a scientifically useful data model by being cleaned up, ordered, and corrected.

[24] This notion of the adequacy of a data model for a purpose is elaborated in greater detail in Parker and Bokulich (in preparation).

> even of reputedly poor sampled groups such as primates and humans, do not always alter perceptions of evolutionary patterns;… and (6) new post-Cambrian Lagerstätten rarely add new families to existing knowledge, just new species and genera. (Benton et al. 2011, p. 67; emphasis added)

There are two important points in the above passage worth highlighting for our philosophical project: First, rather than evaluating data models as accurate (or inaccurate), they should instead be evaluated as adequate (or inadequate)-for-a-particular-purpose. And, second, saying that a data model is adequate-for-purpose does not mean that it is a data model free of all errors and biases. Hence, in the context of the data of the fossil record, the relevant question is not whether all the biases in the fossil record have been removed such that it is a perfect depiction of paleodiversity over time, but rather whether those biases render the data model inadequate for testing the particular hypotheses the scientist is interested in. There are many hypotheses in science for which even an incomplete and biased data model is still adequate. Whether it is adequate or inadequate in any particular context, however, is something that needs to be scientifically investigated and assessed. Moreover, as we've seen in detail, in some cases one can improve the adequacy of a data model for a purpose by using various data-correction techniques.

5 Corrected data models (almost) all the way down

So far I have focused on the role of models in correcting data at the relatively abstract level of global paleodiversity data. However, one can arguably see the role of corrected data models and the importance of assessing fidelity-for-a-purpose at every level of the data-model hierarchy in paleontology—including at the level of the prepared fossil rocks themselves.

At the bottom of the data-model hierarchy are the fossil rocks, which can be thought of as a physical data model.[25] The fossils in this context are taken as a representation of past life on Earth.[26] It is an imperfect representation of those past life forms in that it is a static, often 2-dimensional projection, where only certain parts of the organism are represented (e.g., typically not the soft-bodied parts). The fossil rock representation of the organism is constructed through natural (e.g., chemical and geological) and often anthropogenic processes, the latter of which went largely unnoticed by the philosophical community until the work of STS scholar Caitlin Wylie.

[25] More precisely, I have in mind those fossil rocks that have been collected, prepared, and categorized. I will not engage the difficult question here of where exactly to draw the line between (raw) data and a data model. It may very well be that the distinction is one of degree with vague boundaries, rather than a difference of kind (though as with other vague categories, that does not mean there are no important differences); and where the line is drawn may further be context dependent. My inclination here is to say that if a fossil rock has been collected, categorized, and/or prepared, that is sufficient for it to count as a data model.

[26] As noted before, fossil data can be taken to be a representation of more than just past life (e.g., they can also represent facts about the geological or paleoclimatological record).

Before a chunk of rock containing a fossil can be counted as a useful scientific specimen, it typically needs to 'prepared.'[27] This work is carried out not by the paleontologists themselves, but rather by skilled technicians known as fossil preparators, who remove what is called the matrix (the excess rock) from around the fossil. As Wylie shows, this is far from a trivial process:

> Because fossils often look similar to their matrices, preparators rely on geological knowledge of rock formations and mineral characteristics to distinguish a matrix nodule from an unusual bone growth, for example.... [They also need knowledge of anatomy and biology.] Knowing the location of important traits on a skull allows a preparator to search for them while removing matrix, and also to be careful when preparing near the structures' expected locations. (Wylie 2009, p. 6)

The fossil preparator can thus be understood as taking the "raw data" of the fossiliferous rock and constructing from it a physical data model that is in a form useful for scientific investigation and paleontological theorizing.

In recounting a joke heard in a museum fossil preparation lab about how an accidental slip of the instrument could lead to the "discovery" of a new species, Wylie notes that this highlights the sometimes difficult decisions preparators have to make in distinguishing what is signal from what is noise. She observantly remarks,

> Scientists recognize the underdetermination of knowledge by data: they know that multiple interpretations of data are possible, and that, as a result, their interpretations must be defended and will most likely be debated. But reminding them that specimens themselves are underdetermined by raw material—e.g., that specimens may take different forms and yield different data depending on how they are processed—is more dangerous, because it threatens the natural objects that are the foundation of empirical research. (Wylie 2016)

In addition to the traditional underdetermination of theory by data, Wylie is here calling attention to the underdetermination of data model by the data level below it. This arguably can happen at any pair of levels in the data-model hierarchy, and a central issue of scientific debate is often how this ladder of data models should be climbed.

One can see the importance of the notion of adequacy-for-purpose even at the level of the fossil specimen, insofar as how that specimen is prepared will often depend on the theoretical uses to which it will be put. Wylie explains,

> A major decision for the preparator is how and to what extent a specimen is prepared. Finney [a fossil preparator she interviewed] believes specimens should not be prepared unless needed for a researcher's specific study, and in that case preparation should be done as required for that researcher's question and no more. (Wylie 2009, p. 10)

[27] Although not always required, preparation is typically needed for vertebrate fossils, and sometimes needed for invertebrate fossils as well.

That is, a fossil specimen should be prepared only to the extent to which it is adequate to provide the requisite evidence for the paleontologist's specific theoretical questions.[28] Some theoretical questions will require more of the matrix—or even more of the fossil itself—to be removed in order for it to be an adequate data model to provide evidence for or against a particular hypothesis, while for other sorts of questions a minimal preparation may be adequate.

Once the fossil specimens are prepared, they are then categorized both taxonomically and chronologically—a process that requires substantial theoretical knowledge and inference. At almost every level of the data model hierarchy—from the datum of the individual prepared fossil specimen up to the most sophisticated phylogenetically-corrected global fossil data set—involves the use of models. There is thus what Edwards calls a *model-data symbiosis* (Edwards 2010, pp. 281–282), whereby models and data are in a mutually dependent and mutually beneficial relationship.[29] This is not to say that there are no distinctions between data and models, but rather is a call to recognize the complicated ways in which data and models depend on each other. Furthermore, as Edward's term implies, models need not be a corruption of data, but rather are the very means by which data become scientifically useful for testing and further theorizing.[30]

6 Conclusion

It has long been recognized that the data of the fossil record are both highly incomplete and strongly biased by a number of geological and other "filters." Nonetheless, paleontologists have developed a suite of data-correction techniques whereby some of these biases can be mitigated, and even some gaps filled. In particular, we examined three prominent data-correction techniques used in the construction of paleodiversity data models: the subsampling model approach, the residuals model approach, and the phylogenetic model approach. As we saw, models are being used not just in constructing and correcting these data models, but also in testing the reliability and robustness of the data-correction methods, by means of computer simulations involving synthetic data. These simulation studies indicate that the model-corrected data can provide a *better* representation of the history of biodiversity than the "raw" diversity data do. The importance of such data-correction techniques in constructing data models that are more useful for scientific theorizing was seen not just at the highly abstract level of global paleodiversity data, but also lower down in the data-model hierarchy, at the level of the prepared fossil rocks themselves.

[28] While most numerical data-model correction techniques are reversible, many physical data-model correction techniques are not, and hence call for more caution.

[29] A fuller discussion of this notion of model-data symbiosis and a taxonomy of the different ways that data can be model-filtered is provided in Bokulich (forthcoming).

[30] Of course not all model-corrected data will be better than the raw—it will depend on the particular concrete details of the scientific case. Data correction methods typically work best when there is a) a detailed, quantitative understanding of the biases and their effects on the data and b) robust, independent lines of evidence providing the grounds for the model-based corrections.

In this scientific case study we saw a number of important themes emerge for our philosophical understanding of data models: first, the purity of a data model is not a measure of its epistemic reliability. Rather, what is epistemically important is its *fidelity* in representing the relevant feature of the world. Second, the fidelity of a data model is a *matter of degree*. A paleodiversity data model can do a better or worse job of capturing the biological signal of interest. Third, the fidelity of a data model can be improved not just by means of physical control during data collection, but also through *vicarious control* after the data have been collected. This can be done by modeling various sources of distortion or noise in the data, and then removing them during data reduction. Fourth, because a data model can function as evidence for a variety of different knowledge claims, the fidelity of a data model must be judged *relative to a particular purpose*. As we saw in the case study, while there are some theoretical questions for which a given paleodiversity data model is adequate, there are others for which it is not. Hence data models, like theoretical models, should be judged as adequate-for-purpose.

One might think that without access to the true history of biodiversity, assessments of adequacy and attempts to correct data to bring it more in line with the true history would be hopeless. What is remarkable, however, is the ingenuity with which scientists have made these seemingly intractable questions tractable. In this regard we've seen how paleontologists have, first, come to understand in growing detail the contours of our ignorance about the history of biodiversity; second, developed a suite of methods for correcting the fossil data; and, third, found ways to test the reliability and robustness of these methods under our ignorance.

Acknowledgements I am grateful to Wendy Parker, Adrian Currie, Mike Benton, and two anonymous referees for helpful comments on an earlier version of this paper. I also thank Demetris Portides for first encouraging me to write this paper and for his patience seeing it through to completion. I gratefully acknowledge the support of the Institute of Advanced Study at Durham University, COFUND Senior Research Fellowship, under EU grant agreement number 609412.

References

Alroy, J. (2010a). Geographical, environmental, and intrinsic biotic controls on phanerozoic marine diversification. *Paleontology, 53*(6), 1211–1235.

Alroy, J. (2010b). Fair sampling of taxanomic richness and unbiased estimation of origination and extinction rates. In J. Alroy & G. Hunt (Eds.), *Quantitative methods in paleobiology* (pp. 55–80). Baltimore: The Paleontological Society.

Benton, M., Dunhill, A., Lloyd, G., & Marx, F. (2011). Assessing the quality of the fossil record: Insights from vertebrates. In A. McGowan & A. Smith (Eds.), *Comparing the geological and fossil records: Implications for biodiversity studies* (Vol. 358, pp. 63–94). London: Geological Society.

Benton, M., & Harper, D. (2009). *Introduction to paleobiology and the fossil record*. Chichester: Wiley.

Bokulich, A. (forthcoming). Towards a taxonomy of the model-ladenness of data. In *Presentation in Symposium session: Exploring model-data symbiosis in the geosciences*. Philosophy of Science Association Biennial Meeting, November 2018, Seattle, WA.

Brocklehurst, N. (2015). A simulation-based examination of residual diversity estimates as a method of correcting for sampling bias. *Palaeontologia Electronica, 18.3.7T,* 1–15.

Collins, M., & Simberloff, D. (2009). Rarefaction and nonrandom spatial dispersion patterns. *Environmental and Ecological Statistics, 16,* 89–103.

Currie, A. (2018). *Rock, bone, and ruin: An optimist's guide to the historical sciences*. Cambridge, MA: The MIT Press.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray. Retrieved from https://en.wikisource.org/w/index.php?title=On_the_Origin_of_Species_(1859)&oldid=6512451.

Edwards, P. (2001). Representing the global atmosphere: Computer models, data, and knowledge about climate change. In C. Miller & P. Edwards (Eds.), *Changing the atmosphere: Expert knowledge and environmental governance* (pp. 31–65). Cambridge, MA: MIT Press.

Edwards, P. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge, MA: MIT Press.

Eldredge, N., & Gould, S. J. (1972). Punctuated equilibria: An alternative to phyletic gradualism. In T. Schopf (Ed.), *Models in paleobiology* (pp. 82–115). San Francisco: Freeman, Cooper, and Co.

Erwin, D., & Droser, M. (1993). Elvis taxa. *Palaios, 8,* 623–624.

Foote, M. (1996). Perspective: Evolutionary patterns in the fossil record. *Evolution, 50*(1), 1–11.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika Trust, 40*(3/4), 237–264.

Gould, S. J., Raup, D., Sepkoski, J., Jr., Schopf, T., & Simberloff, D. (1977). The shape of evolution: A comparison of real and random clades. *Paleobiology, 3,* 23–40.

Huss, J. (2009). The shape of evolution: The MBL model and clade shape. In D. Sepkoski & M. Ruse (Eds.), *The paleobiological revolution: Essays on the growth of modern paleontology*. Chicago: University of Chicago Press.

Lane, A., Janis, C., & Sepkoski, J. (2005). Estimating paleodiversities: A test of taxic and phylogenetic methods. *Paleobiology, 31*(1), 21–34.

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. Chicago: University of Chicago Press.

Lyell, C. (1830). *Principles of geology: Being an attempt to explain the former changes of the earth's surface, by references to causes now in operation*. London: John Murray. Retrieved from http://www.esp.org/books/lyell/principles/facsimile/contents/lyell-v1-aa-fm.pdf.

Magnani, L., & Bertolotti, T. (Eds.). (2017). *Springer handbook of model-based science*. Dordrecht: Springer.

Metcalfe, I., & Isozaki, Y. (2009). Current perspectives on the permian-triassic boundary and end-permian mass extinction: Preface. *Journal of Asian Earth Sciences, 36,* 407–412.

Norton, S., & Suppe, F. (2001). Why atmospheric modeling is good science. In C. Miller & P. Edwards (Eds.), *Changing the atmosphere: Expert knowledge and environmental governance* (pp. 67–105). Cambridge, MA: MIT Press.

Norwell, M. (1993). Tree-based approaches to understanding history: Comments on ranks, rules, and the quality of the fossil record. *American Journal of Science, 293,* 407–417.

Parker, W. (2010). Scientific models and adequacy for purpose. *The Modern Schoolman, 87,* 285–293.

Parker, W., & Bokulich, A. (in preparation). Data models, representation, and adequacy-for-purpose.

Raup, D. (1972). Taxonomic diversity during the phanerozoic. *Science, 177*(4054), 1065–1071.

Raup, D. (1975). Taxonomic diversity estimation using rarefaction. *Paleobiology, 1,* 333–342.

Sakamoto, M., Benton, M., & Venditti, C. (2016). Dinosaurs in decline tens of millions of years before their final extinction. *Proceedings of the National Academy of Science, 113*(18), 5036–5040.

Sakamoto, M., Venditti, C., & Benton, M. (2017). 'Residual diversity estimates' do not correct for sampling bias in palaeodiversity data. *Methods in Ecology and Evolution, 8,* 453–459.

Sepkoski, J. (1982). Compendium of fossil marine families. *Milwaukee Public Museum Contributions in Biology and Geology, 51,* 1–125.

Sepkoski, J. (1984). A kinetic model of phanerozoic taxonomic diversity. III. Post-paleozoic families and mass extinctions. *Paleobiology, 10*(2), 246–267.

Sepkoski, J. (1994). What I did with my research career: Or how research on biodiversity yielded data on extinction. In W. Glenn (Ed.), *Mass-extinction debates: How science works in a crisis*. Stanford, CA: Stanford University Press.

Sepkoski, D. (2012a). *Reading the fossil record: The growth of paleobiology as an evolutionary discipline*. Chicago: University of Chicago Press.

Sepkoski, D. (2012b). 'Replying life's tape': Simulations, metaphors, and historicity in Stephen Jay Gould's view of life. *Studies in History and Philosophy of Biological and Biomedical Sciences, 58,* 73–81.

Sepkoski, D. (2013). 'Towards a natural history of data': Evolving practices and epistemologies of data in paleontology, 1800–2000. *Journal of the History of Biology, 46,* 401–444.

Sepkoski, D. (2016). 'Replaying life's tape': Simulations, metaphors, and historicity in Stephen Jay Gould's view of life. *Studies in History and Philosophy of Biological and Biomedical Sciences, 58,* 73–81.

Sepkoski, D., & Ruse, M. (2009). *The paleobiological revolution: Essays on the growth of modern paleontology*. Chicago: University of Chicago Press.

Signor, P., III, & Lipps, J. (1982). Sampling bias, gradual extinction patterns and catastrophes in the fossil record. In L. Silver & P. Schultz (Eds.), *Geological implications of large asteroids and comets on the earth* (Vol. 190, pp. 291–296). Boulder: Geological Society of America.

Smith, A. (1994). *Systematics and the fossil record: Documenting evolutionary patterns*. Oxford: Blackwell Science Ltd.

Smith, A., & McGowan, A. (2007). The shape of the phanerozoic marine paleodiversity curve: How much can be predicted from the sedimentary rock record of Western Europe. *Palaeontology, 50*(4), 765–774.

Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science: Proceedings of the 1960 international congress* (pp. 252–261). Stanford: Stanford University Press.

Turner, D. (2007). *Making prehistory: Historical science and the scientific realism debate. Cambridge studies in philosophy and biology*. Cambridge: Cambridge University Press.

Turner, D. (2011). *Paleontology: A philosophical introduction*. Cambridge: Cambridge University Press.

Upchurch, P., & Barrett, P. (2005). Phylogenetic and taxic perspectives on sauropod diversity. In K. Rogers & J. Wilson (Eds.), *The sauropods: Evolution and paleobiology* (pp. 104–124). Berkeley: University of California Press.

van Fraassen, B. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Clarendon Press.

Wylie, C. (2009). Preparation in action: Paleontological skill and the role of the fossil preparator. In: M. Brown, J. Kane, & W. Parker (Eds.), *Methods in fossil preparation: Proceedings of the first annual fossil preparation and collections symposium* (pp. 3–12).

Wylie, C. (2016). "Overcoming underdetermination" on extinct: The philosophy of palaeontology blog (April 11, 2016). Retrieved August 5, 2017 from http://www.extinctblog.org/extinct/2016/4/11/overcoming-underdetermination.