Evolutionary Ecology Research, 2017, **18**: 477–502

A survey of computational methods for fossil data analysis

Indrė Žliobaitė[1,2], Kai Puolamäki[3], Jussi T. Eronen[1,4] and Mikael Fortelius[1,5]

[1]*Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland,* [2]*Department of Computer Science, University of Helsinki, Helsinki, Finland,* [3]*Finnish Institute of Occupational Health, Helsinki, Finland,* [4]*BIOS Research Unit, Helsinki, Finland and* [5]*Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo, Norway*

ABSTRACT

Aim: (1) Survey and organize computational approaches to fossil data analysis into a methodological framework. (2) Highlight the kinds of research questions about evolutionary and environmental change that can be answered by applying computational algorithms to mammal fossil data to better understand past ecosystems and climates.

Questions: What models have been used for what research questions? What is their scope of application? What are their potential limitations?

Search methods: Our search of the literature was based on personal knowledge in combination with keyword-based searches. Papers were considered relevant if data-driven computational methods were used to analyse relationships between organisms and their environments at evolutionary time scales.

Conclusions: We demonstrate that different research questions may be answered with the same computational algorithm, and different algorithms may be needed to answer the same question in different contexts. We believe that in order to move forward, we need to match knowledge of methods with knowledge of the fossil record in a research question-driven way. Figure 2 presents a proposed workflow. Following this framework, we survey existing work and highlight what research questions can potentially be answered with which methods, some of which may not have been reported in the evolutionary palaeontology literature to date. The outcome of this survey is a proposal for a research agenda in computational fossil data analysis.

Keywords: big data, computational fossil data analysis, data mining, ecometrics, evolutionary palaeontology, machine learning, mammals.

INTRODUCTION

The relationship between environmental change and biotic processes as drivers of evolutionary change is a classic theme in investigations of the past, and a number of hypotheses have been formulated to capture it. An increasing emphasis on the computational analysis

Correspondence: I. Žliobaitė, Department of Geosciences and Geography, University of Helsinki, Helsinki 00014, Finland. e-mail: indre.zliobaite@helsinki.fi

of large datasets has revitalized the field, but also brings major challenges requiring an understanding of the underlying biology, the computational methodology, and how to link the two.

A major trend in evolutionary palaeontology research in the past decades has been the rapid augmentation of the use of large and complex datasets (see, for example, Brewer *et al.*, 2012) and increasingly complex algorithmic approaches. While ready-made computing toolboxes and the availability of global dataset compilations present an excellent opportunity to broaden the scale and scope of research on fossil data, this also results in the tendency to report algorithmic outputs in an increasingly mechanistic way, where results are judged based on faith in algorithms, with progressively little biological insight and interpretation. Indeed, method-driven rather than research question-driven studies are becoming common, relying exclusively as they do on outputs of complex and entangled computational methods. Such studies are hard to assess for editors, reviewers, and readers alike without placing those findings in both biological and methodological contexts.

This survey is intended as a methodological guideline for what kind of research questions about relations between organisms and their environments, as well as their evolution, can potentially be answered with what kinds of data-driven computational techniques. It is not our intention to advocate any particular technique or techniques – quite the opposite. Our main message is that the most popular or most recent algorithm does not provide answers automatically. How well an algorithm will help to answer a research question in evolutionary palaeontology will depend on how well biological concepts are translated into computing proxies, and on selecting meaningful model optimization criteria.

Although there are comprehensive methodological guides that focus on modern-day ecology (e.g. Legendre and Legendre, 1998; Quinn and Keough, 2002; Gotelli and Ellison, 2013), including that of Hammer and Harper (2006), with its focus on analysing a single taxon or a single locality, there is no source that addresses the polymorphic methodology of mammalian evolutionary palaeontology, particularly at a global scale and in a data-driven way. The main focus of this survey is on understanding the relationships between organisms and their environments over evolutionary time scales. Many of our examples are from mammals, because the mammal fossil record is denser and better resolved than other fossil groups. Moreover, having previously worked with mammal data ourselves, we understand the function better; but we hope that the principles are equally applicable to other fossil groups, including fossil vegetation data.

COMPUTATIONAL METHODS IN EVOLUTIONARY PALAEONTOLOGY

Computational methods involve developing and using data-driven analysis techniques beyond standard statistical hypothesis testing – such as machine learning, data mining methods, and overall big data analysis methods – for understanding fossil data at an appropriate scale (subcontinental to global scale, multi-site data) to extract knowledge about evolutionary processes. Computational approaches to evolutionary palaeontology are situated at the intersection of biology, earth sciences, ecology, and computing.

Such approaches to analysis of the fossil record are not covered by classical palaeontology, which typically focuses on a single site or taxa, and does not cover analysis at scale. Neither is it covered by geoinformatics alone, because taxon abundances, species traits, evolutionary relations, and other biological essentials need to be taken into account (Brewer *et al.*, 2012). Nor is this research covered by ecology alone, which focuses on analysis of communities, their

interactions, discovering community structures, spatial patterns, and ecological time series analysis.

A related term, palaeoinformatics, has mostly been used to refer to developing fossil databases (Brewer *et al.*, 2012). The advent of such databases in the last decade and before (Uhen *et al.*, 2013) has indeed drawn more research attention to computational techniques for analysing these rich datasets. Fossil data analysis at large scales poses specific challenges, such as temporal and spatial uncertainty, imprecise and incomplete assemblages due to fossilization and collection biases, and uncertainties in species identification. Thus, tailored computational methods are typically needed to accommodate these uncertainties, and the underlying assumptions behind the methods need to be well understood in order to draw meaningful conclusions from the patterns identified.

Traditional methods in statistics typically assume a known model about the physical phenomena, and fit parameters to that model based on observed data. Such methods typically assume that data come from a known physical or biological process with well-known and accurately captured properties, and which represent the population accurately. Such a setting allows one to verify a theoretical understanding of physical or biological processes. The result is reliable if the initial process model is good, and there is enough data of sufficient quality, and the fit is unbiased. Such methods are invaluable for studying phenomena at small scales (not necessarily small datasets); however, they do not necessarily apply equally well to large, noisy, and heterogeneously complex real-world data, where the underlying limiting assumptions of classical statistical methods are easily violated.

Of the statistical models designed to observe large-scale events, the most classic example is thermodynamics. In a sense, models that track, for example, animal population sizes, such as the neutral theory of biodiversity, are of a similar vein (Rosindell *et al.*, 2011). A typical feature of these models is that the details of small-scale interactions do not really matter for the large-scale results (as long as there is sufficient interaction and some quantities that are preserved and/or large-scale statistics), hence it is sufficient to use simple small-scale dynamics to model complex phenomena. However, one does not necessarily use the small-scale models to present large-scale phenomena. For instance, in understanding species diversity, a neutral theory could have advantages over a data-driven model, since the former by itself is a scientific hypothesis of how the world works, while extracting a process description from a data-driven model of such a process might not be that obvious.

Yet in some circumstances, modelling many processes operating at a global scale is intractable, as realistic physical or biological process models require a high degree of complexity. In such cases, data-driven approaches, including machine learning and data mining, are becoming increasingly useful, since instead of requiring a readily available model form, models can be constructed iteratively from the data.

Challenges of fossil data for computational methods

The fossil record of mammals bears on evolutionary history and environmental change in multiple ways. Most fossil datasets consist of occurrences or counts of taxa or morphotypes within a fossil assemblage that represents a particular place and time (Brewer *et al.*, 2012). In its basic form, fossil data records specimens with accompanying age, location, taxonomical information, and traits (either assigned per taxon, or measured per specimen). Through biostratigraphy, the fossil record provides a chronology of change to be used when other temporal data are either lacking or inferior in some way (e.g. Steininger, 1999; Agusti *et al.*, 2001; Wang

et al., 2013). Furthermore, mammalian fossil assemblages provide proxy data for reconstructing the environment via stable isotope analysis (West *et al.*, 2006; Passey *et al.*, 2009; Cerling *et al.*, 2010), species richness (Janis *et al.*, 2004), or trait distribution analysis (Eronen *et al.*, 2010a, 2010b; Polly, 2010).

Fossil data compilations are to a high degree interpretation based. There is a high degree of subjectivity and personal expertise encoded in the data, even if the best efforts are made to make interpretation uniform across databases (Uhen *et al.*, 2013). Moreover, since fossil datasets are compiled from different sources, the data are not just noisy, but of varying quality within single datasets. There are specific uncertainties and biases to be kept in mind when conducting computational data analysis and interpreting the results.

Fossil data record the presence of taxa, but absence from the fossil record does not necessarily indicate absence in reality. Thus, for instance, the actual times of a species' appearance and extinction are uncertain due to the uncertainty associated with their absence in the fossil record.

The dating of fossil localities is also uncertain. Many localities lack geological evidence of superposition or geochronologically datable materials (see, for example, Woodburne, 2004), as is the case for much of the terrestrial European Neogene (NOW Community, 2017). Traditionally, the solution has been to create more or less arbitrary biochronological entities, into which such fossil occurrences can be grouped.

Fossil data do not provide a uniform sample of organisms that were present millions of years ago, because not everything is equally likely to turn into fossils. One of the subfields of palaeontology – taphonomy – is dedicated to studying how organisms decay over time, how they become fossilized, and what biases affect the fossilization process (see, for example, Behrensmeyer *et al.*, 2000). In addition to fossilization biases, collection biases occur, since not everything that is fossilized is exposed or is equally accessible (e.g. fossils that happen to be at the bottom of a lake or in a remote place are less likely to be found), and not everything that is found in the field makes it into collections and databases.

Validation of models for fossil data can only be indirect, because no ground truth (what the biosphere really looked like millions of years ago) will ever become available. In many other domains, modelling results can usually be validated on a small sample of data where the true outcomes are known. For example, when developing prediction models for the effectiveness of a medical treatment, one might have access to a sample dataset where it is known whether those patients have recovered. Therefore, it is essential to critically consider the assumptions behind the chosen computational methods, in order to be able to judge to what extent the results are likely to be capturing a real phenomenon, and to what extent the outcomes may potentially be due to artefacts in the data, specific properties of methods, or both.

With these challenges in mind, computational methods for fossil data analysis need to be particularly robust, stable with respect to small variations in data, tolerant to noise, and capable of explicitly handling uncertainties and estimating the confidence of the outputs.

A workflow of the computational approach

It is not unusual to start a study by following an impulse to apply a computational method (e.g. decision trees, Bayesian inference, support vector machines) that was recently successfully applied in a different context to one's own data. Yet the use of a previously successful method by itself would not guarantee good results, because a computational method (or an algorithm) is no more than an engine, and its success depends on the design of the computational task to answer a particular research question.

As an illustration, take the following example. A laboratory has built a boat using a Hybrid Toyota engine, and subsequently another lab decides to use a Hybrid Toyota engine when building an airplane. Like an algorithm, a particular engine may or may not work well for a particular purpose, such as flying an airplane. There are many system design aspects to consider prior to selecting what engine to use, and, in principle, different engines should be interchangeable. Thus, the choice of the algorithm should not outweigh the design of the whole computational research task for answering a particular research question.

Ideally, the process of designing a computational research task should cover the following steps.

1. Formulate the research question
2. Design the computational task setting
3. Define the performance criteria
4. Select which existing algorithm to use, or develop a new algorithm.

In the airplane example, the research question (1) is how to get across the ocean. The computational task setting (2) is the decision to build a plane, and designing how the plane should operate. The performance criteria (3) include how to judge whether the plane is working well, which could be a test flight to see if it crashes or not. Only then would it make sense to decide upon implementation details (4), such as what kind of engine to use (algorithm selection). In this sense, decision trees or Bayesian inferences are like engines – the choice to use decision trees or Bayesian inference arises only in Step 4. A decision tree would not solve the problem by itself if the first three steps were ignored. Their success in answering research questions in evolutionary palaeontology primarily depends on the intelligent design of the whole computational research task, rather than just the particular algorithm used to solve the task. Next, we summarize the main types of computational tasks, and provide examples for possible applications to fossil data analysis.

Overview of computational research tasks

Computational tasks can be grouped into those for finding structures and those for finding relations, as depicted in Fig. 1. Finding structures relates to extracting summaries or prototypes from data, such as finding common characteristics of species that often occur together. Finding relations relates to producing models in a functional form that can map designated input variables to the target variables, such as estimating body mass based on the size of mammal teeth.

Finding structures can be further divided into exploratory data analysis, which typically considers data presentation and visualization for manual analysis, and finding patterns, which includes obtaining data summaries, extracting common patterns, and detecting anomalies. Finding relations can be further divided into descriptive modelling, where the main interest is to obtain and analyse a model of relations between variables, and predictive modelling, where the main interest is to use the model to make predictions or estimates.

In terms of what types of research questions are to be answered, computational tasks can be divided into interpolation and extrapolation tasks. Interpolation tasks are about exploring and understanding what the data at hand contains. In contrast, extrapolation tasks are aimed at generalizing to new situations, different localities or different times. More generally, extrapolation tasks have the aim of understanding the causal processes behind the

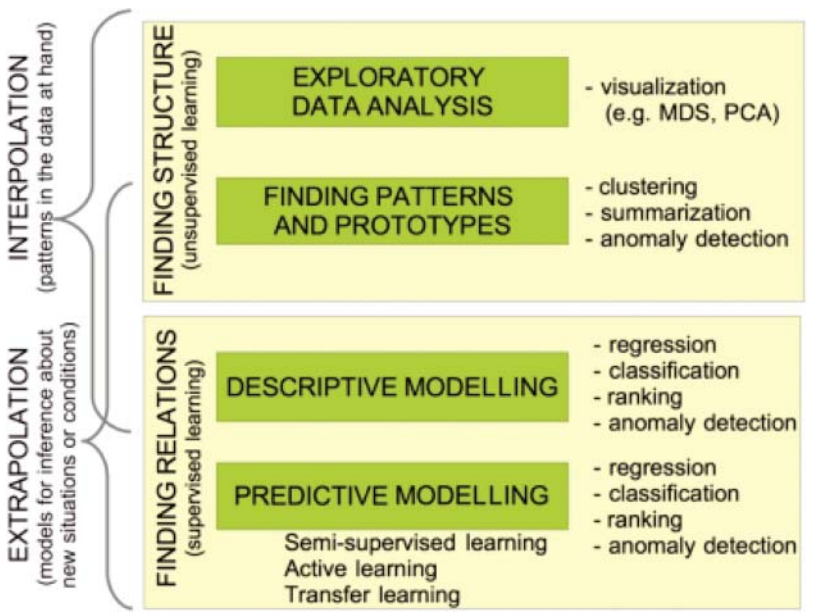

Fig. 1. Categorization of computational research tasks.

relation. For example, one can model the relation between the morphology of the teeth of large living mammals and the productivity of their environments. One can then make a generalizing assumption that the same functional relation applied in the past, where no direct measurements of the productivity of the environment can be made, but dental traits are available via the fossil record. In this way, by using a model built on living mammals, one can make estimates about past environmental conditions.

The transparency and interpretability of models are important criteria for selecting which algorithmic technique to use. There are so-called *black box* models, which are oriented towards achieving the best accuracy possible. In such models, extracting the exact reasoning for the prediction is difficult, but is possible to some degree (see, for example, Henelius *et al.*, 2014, 2015, 2017). Examples of black box models include feed-forward neural networks, support vector machines, Bayesian networks, and random forests. Less complex methods, such as linear regression, single decision trees, and nearest-neighbour approaches, are more readily accessible in terms of understanding the reasoning behind their predictions, and thus are more transparent. The choice of a method depends on the desired trade-offs between transparency, interpretability, potential accuracy, the complexity of the patterns to be captured, and the robustness of the outcomes.

There are many algorithms to consider for modelling relations. Generalized linear models (linear regression) are well understood and widely used in the natural sciences. Decision trees and Bayesian inference are gaining in popularity. A detailed tutorial on algorithmic techniques is beyond the scope of this survey; there are many comprehensive textbooks readers can consult. For a somewhat less technical treatment, we recommend Hand *et al.* (2001) and Berthold *et al.* (2010); more technical classics include Duda *et al.* (2000), Bishop (2006), and Hastie *et al.* (2009), with the latter being freely available online.

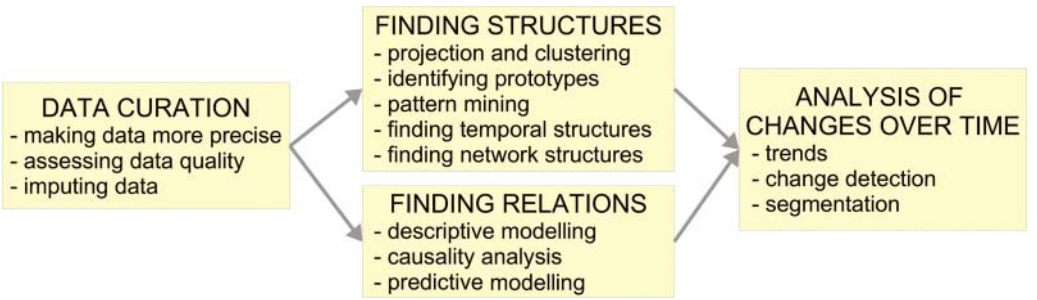

Fig. 2. Proposed computational task workflow in evolutionary palaeontology.

The present survey focuses primarily on computational approaches that have been applied to analyse the fossil record (including modern-day occurrences as an extension of the fossil record) in order to achieve an understanding of the relations between organisms and their environment, and how organisms along with their environment change over time.

We propose organizing the computational tasks in evolutionary palaeontology into a workflow based on four categories, as depicted in Fig. 2. A focused study could be in any one of those four categories, while a comprehensive study would follow the complete process from data curation, through identifying patterns and relations, to analysing changes in those patterns over time.

COMPUTATIONAL DATA CURATION

The fossil record is limited by what has been preserved and what has been found. Collecting more data on demand is laborious, time-consuming, and often not feasible because all the material of interest, which was exposed at a certain time and place, has already been collected. Moreover, collecting more data would not help counteract the persistent and systematic biases of the fossil record. Computational methods can be used to infer common patterns from large fossil databases and then using those patterns for identifying anomalies, uncertainties or unexpected patterns, and fixing them.

Making data more precise

Reconstructing age

Reconstructing age has been a subject of many computational studies, since it is essential to know the age and the place of origin of fossils in order to be able to interpret environmental estimates reconstructed from the fossil record.

Information relating to age is primarily available from stratigraphic methods, including biostratigraphy (Steininger 1999; Agusti *et al.*, 2001; Wang *et al.*, 2013), but can be refined computationally. Computational methods could potentially make the time estimates more precise, or time bands narrower, using species occurrence information. For example, the age range of a locality can be made more precise computationally by comparing nearby localities (Kallio *et al.*, 2011).

Seriation refers to computational methods for reconstructing age by establishing an order of succession for sites based on the co-occurrence of species from those sites. The main

principle involved is to find an order of localities from the oldest to the youngest, such that taxa are as consistent as possible in the timing of their first occurrence, co-occurrence, and extinction. Computational approaches include the graph-theoretical methods (Guex and Davaud, 1984), parsimony analysis (Hooker, 1996), Bayesian methods (Halekoh and Vach, 2004), correspondence analysis (Van de Velden *et al.*, 2009), and event ordination approaches, which are based on maximizing the fit of hypothesized time ranges to independent stratigraphic information, as summarized in Alroy (2000). Other approaches are based on partial orders (Ukkonen *et al.*, 2005), probabilistic modelling with Markov Chain Monte Carlo integration (Puolamäki *et al.*, 2006), and spectral methods (Fortelius *et al.*, 2006).

The main limitation of such data-driven approaches is that they rely on the certainty of absence. The idea is to minimize cases where taxa originate, disappear for a while, and then reappear. In reality, the absence of taxa is an uncertain parameter. Therefore, we can almost never establish a unique ordering, and at best we can estimate probabilities for different orderings. Different approaches try to account for this uncertainty in different ways: Alroy (2000) gives a best guess estimate for the ages of the find sites, while Ukkonen *et al.* (2005) express uncertainty using partial orders, and Puolamäki *et al.* (2006) estimate probabilities that site A predates site B. Gionis *et al.* (2006) and Ukkonen *et al.* (2009) place sites into buckets ordered in time, but the sites within a bucket are not ordered.

Existing computational approaches typically assign discrete time bands to sites. One promising direction for future research is the computational estimation of age as a point-in-time rather than a time interval, which could be done by considering the spatial location of the fossil specimens in relation to the stratigraphic boundaries. Point-in-time estimates would open up new possibilities for richer analysis of biospheric and environmental changes over time, because one would have many more data points for analysis over time from the same fossil record.

Estimating relative abundances

Estimating relative abundances is the task of computationally making the fossil record more informative. Current models for reconstructing environmental conditions from fossil data typically use taxon occurrence (see, for example, Fortelius *et al.*, 2014). Abundances may possess a complementary taxon occurrence data signal about ecological conditions and interactions between taxa.

The estimation of relative abundances in the fossil record has been based on counting actual specimens (Badgley, 1986) and statistically correcting for sampling effects (Moore *et al.*, 2007). It is well established in ecology that abundances are strongly related to occupancy (Gaston *et al.*, 2000), and similar reasoning has been used in the analysis of fossil data (Jernvall and Fortelius, 2004).

New computational approaches for estimating relative abundances from the fossil record could be developed using a similar principle as modern-day ecology niche modelling (Wal *et al.*, 2009). The main principle would be to model how suitable the environment is for a particular taxon, based on its morphological traits. A combination of probabilities of occurrence across communities would translate into relative abundances.

Assessing data quality

Assessing data quality is the second task in data curation that addresses quantifying biases, completeness, and uncertainties in the fossil record. Biases in the fossil record are widely

acknowledged (Behrensmeyer *et al.*, 2000), and often are considered to be unavoidable. In addition to purely statistical reasoning (Foote, 2016), research efforts have been made to identify and quantify biases computationally in relation to different depositional environments (Mitchell, 2015), or computationally compare fossil communities to other communities (Turvey and Blackburn, 2011). Such bias identification requires reference data that typically come from the modern day. A promising direction might be to computationally compare fossil communities with other fossil communities in terms of the distribution of their functional traits, and reason about completeness of the record from there. The main principle would be to computationally infer what is a normal or expected pattern, and identify deviations from that normal pattern.

Assessing data completeness

Assessing data completeness (Foote and Raup, 1996; Alba *et al.*, 2001) is typically based on fitting models to the survivorship curves of taxa. In addition, computational simulations have been conducted to assess the effects of data incompleteness on the outcomes of fossil data analyses, including the effects of extinction, extirpation, and exotic species on ecometric correlations (Polly and Sarwar, 2014), the effects of spatial sampling on the accuracy of trait distribution estimates (Saarinen *et al.*, 2010), and computationally distinguishing between false absences and true absences in the fossil record (Bingham *et al.*, 2007). All these methods are based on the continuous presence of taxa from origination to extinction. An interesting extension of this line of research would be to account for typically unimodal patterns of the rise and decline in a taxon's occupancy (Jernvall and Fortelius, 2004; Foote, 2007; Liow and Stenseth, 2007).

In addition to assessing data quality and completeness based on taxon survivorship, one could build upon anomaly detection methods in data mining and machine learning (Chandola *et al.*, 2009). The idea is to model the normal, then apply the model to identify which records fall outside the normal model. The task would be to identify data records that are not necessarily extreme, but that are in some ways out of their context. This is different from the statistical definition of outliers, which is about finding extreme values that are far from the centre of the data distribution.

Imputing data

The previous two sections concern curating the fossil record or a dataset as a whole. Imputing data refers to the analysis of individual records.

Resolving uncertainties in taxonomic identification

Resolving uncertainties in taxonomic identification is a challenging task, primarily reserved for human experts (Lieberman, 1999; MacLeod *et al.*, 2010; Culverhouse *et al.*, 2013; MacLeod and Steart, 2015). Automation has been used for resolving uncertainties in taxonomic records by looking for misspelling or string mismatches (Zermoglio *et al.*, 2016). Machine learning methods could be employed to extract patterns from metadata associated with a specimen (occurrence, patterns, timing, traits, co-occurrence context), which could, for instance, try to identify functionally similar species that might have been named differently in different geographic areas. In a similar way, machine learning could be deployed to narrow down options for taxonomic identification based on the context of their occurrence, such as which other species have been found in the vicinity. Taxon identification could be assigned probabilistically for difficult-to-identify specimens, and human experts could make the final decision.

Missing data imputation

Well-established statistical methods exist for missing data imputation (Allison, 2009). The main challenge with fossil data is that data are typically missing not at random, but with dependencies deeply linked to inherent biases in the fossil record. The main focus in palaeontology and archaeology with respect to missing data imputation has been to analyse the trade-off between the amount of missing data and the accuracy of imputation (Couette and White, 2010; Strauss *et al.*, 2003); a recent study has estimated the uncertainties of different imputation methods (Clavel *et al.*, 2014). Similar to taxonomic identification, machine learning could potentially be used in developing tailored missing data imputation methods for fossil data, by learning a model from the contextual information associated with each fossil specimen (e.g. age, placement, co-occurrence) and predicting the probability of occurrence at unknown sites. Machine learning methods from recommender systems (Koren *et al.*, 2009; Ricci *et al.*, 2011) could also be used. For example, methods for movie recommendations are well established, and the task setting is conceptually quite similar to that of the fossil record.

A typical recommended system dataset would record a number of users, and for each user it would record which movies that user has watched (and perhaps ratings of the movies given by the user). The system only knows about the movies watched, and the task is to infer whether a given user would like a movie that he or she has not yet watched. The models generalize from the idea that users who liked movie A also liked movie B. In the fossil setting, localities would be like users, species would be like movies (relative abundances could be like movie ratings), and the result would be to predict the probabilities of occurrence for unobserved species.

FINDING STRUCTURES IN FOSSIL DATA

Finding structures is about describing and summarizing data, identifying characteristic groups of individuals, types or segments, and describing each group, for example by designating a prototype individual, identifying common or exceptional patterns within the dataset, or constructing networks of interactions. Computationally inferring a food web would be an example of finding network structures. Another example task would be computationally identifying groups of plant-eating mammals that are similar to each other with respect to their diet.

Projection and clustering

Clustering organizes data in such a way that objects within the same cluster are similar to each other, but different from each other across clusters. Partitional clustering forms groups, and assigns objects into those groups. Hierarchical clustering produces a tree-like hierarchy of objects.

Clustering has been used to identify functionally similar groups of organisms with respect to their environment. For instance, Hempson *et al.* (2015) identified herbivore functional types in Africa for the last 1000 years, which they then analysed for associations with fire prevalence, soil nutrient status, and rainfall. Heikinheimo *et al.* (2007) used clustering to describe the occurrence of mammals in Europe in modern times; the resulting geographical pattern of clusters was analysed in relation to climate variables, geomorphology, and soil characteristics.

Projection techniques, such as principal component analysis (PCA), are often used for projecting data into lower dimensional space where similar objects appear close to each other. For example, Pineda-Munoz *et al.* (2016) used PCA to analyse the distribution of body masses of modern-day terrestrial mammals in relation to their diet. Meloro and Kovarovic (2013) used projection for identifying communities of large mammal species based on ecometric traits, and aligning those communities with the spatial distribution of species. Principal component analysis is conceptually closely related to clustering; the difference is that PCA solutions are continuous, while clustering assigns discrete cluster membership.

Projection techniques can be used to reduce the dimensionality of data before regression modelling. This is effective when the number of potential input variables is high, and input variables carry potentially overlapping information (i.e. input variables are correlated with each other). In this case, projection is used for data preprocessing, and is not related to finding structures or clustering. For example, Fernandez and Vrba (2006) used PCA to project data into a lower dimensional space before applying regression modelling for estimating Plio-Pleistocene climatic change in the Turkana Basin (East Africa) from the occurrence of taxa. Indeed, the number of taxa was high in relation to the number of site observations available for modelling, and projection helped to eliminate noise and build a more precise model. One drawback of projection preprocessing is that it masks the contributions of individual variables, and thus the resulting models lose some interpretability.

Clustering and projection rely on measuring pairwise similarity between objects as the first step of the analysis. Similarity between objects could also work as a stand-alone analysis. Kallio *et al.* (2011) investigated how to incorporate expert knowledge based constraints in order to obtain more meaningful similarity metrics. The key challenge is to define what kind of similarity patterns – for instance, between species assemblages – are expected across the globe, and then find ways to ignore trivial or obvious patterns when measuring similarity, in order to discover non-trivial or surprising patterns. Conceptually, this approach relates to pattern mining (see p. 488), where the key is to define constraints in order to discover patterns that are surprising in a statistical sense.

From the computational point of view, clustering and projection techniques require two mechanisms: (1) a distance function for measuring how similar any two objects are, and (2) a partitioning or projection mechanism to convert pairwise distances into a group structure. While a standard regression model can flexibly take a set of potential explanatory variables and assign importance weights automatically, in clustering one has to carefully craft a set of inputs that are judged to be relevant for similarity between individuals. One interesting future research direction that would go beyond standard clustering or projection algorithms would be to explore how to define meaningful distance measures incorporating human expertise in, for instance, biology or ecology. For example, if one wants to group the mammals of Europe, one needs to decide which are the relevant characteristics to measure (e.g. body mass, dental traits, colour), whether there are any important weights to consider, and whether it would make sense to incorporate the time dimension into measuring similarity between two individuals.

In clustering, the definition of distance measure is central. A distance measure quantifies the difference between two sets of characteristics. A basic distance measure considers all the characteristics to be of equal importance. A more advanced measure may weight individual characteristics based on their variance, or even use computationally determined importance weights. One could computationally find, for example, that in measuring how similar two mammals are in terms of their morphological characteristics, body size should be twice as

important as body colour. One interesting direction would be to combine constraints based on expert knowledge (as in Kallio *et al.*, 2011) and supervised learning-based distance measures [which have been successfully used, for example, in bioinformatics (Qu and Xu, 2004)].

Identifying prototypes

A group of objects can be characterized by designating one or several representative objects from that group as a prototype. Prototypes can be real objects from the dataset at hand, or they can be artificially constructed. For example, suppose one would like to describe Bovids as a group. One can designate cattle as a prototype, and refer to Bovids as something similar to cattle. Or, one can construct a prototype defining typical characteristics, for example, body mass 223 kg, has horns, eats grass.

The identification of prototypes, following either clustering or projection, has been used to study relations between diet and habitat (Hempson *et al.*, 2015; Pineda-Munoz *et al.*, 2016), using off-the-shelf computational techniques.

An interesting future research direction relating to the identification of prototypes is the computational identification of chronofaunas. This involves identifying a set of species that are representative of a particular time; that is, a geographically restricted, natural assemblage of interacting animal populations that has maintained its basic structure over a geologically significant period of time (Olson, 1952). The Pikermian large mammal assemblage (Eronen *et al.*, 2009) is an example of prototype chronofauna identified manually and analysed computationally. Bingham and Mannila (2014) have attempted the data-driven identification of chronofaunas. Follow-up ideas include taking into computational consideration the time and location of the sites, incorporating taxonomical hierarchies, and also considering ecological characteristics as computational constraints.

Pattern mining

Clustering and projection is about relations of objects to each other, while pattern mining is about the characteristics of objects (e.g. traits of species or characteristics of sites). The aim of pattern mining is to discover sets of characteristics that commonly occur together – for example, it could potentially extract from data the observation that hypsodont teeth commonly occur with cement. Two datasets of living and fossil mammals that have become popular in the pattern mining community are now a sort of benchmark for testing pattern mining algorithms: the Mammals dataset (Heikinheimo *et al.*, 2007) and the Paleo dataset derived from the NOW database (NOW Community, 2017). Representative works using the Mammals include: diverse subgroup discovery (van Leeuwen and Knobbe, 2012), exceptional model mining (van Leeuwen, 2010; Duivesteijn *et al.*, 2015), summarizing data with surprising itemsets (Mampaey *et al.*, 2011, 2012), identifying components, and summarizing (van Leeuwen *et al.*, 2009). Representative studies using the Paleo dataset include: summarizing data with surprising itemsets (Mampaey *et al.*, 2011), with small sets of patterns (Lijffijt *et al.*, 2014), by compression (Bonchi *et al.*, 2011), as feature selection (Garriga *et al.*, 2008), finding tiles (Tatti and Vreeken, 2012b) or partial orders (Ukkonen *et al.*, 2009) (this also belongs to seriation), measuring the quality and significance of itemsets (Tatti, 2008, 2010; Hanhijärvi, 2011), measuring the difference between different data summaries (Tatti and Vreeken, 2012a), and assessing discovered structures through randomization (Gionis *et al.*, 2007; Hanhijärvi *et al.*, 2009). In pattern mining, fossil data have been used extensively for validating the development of algorithms, however very little has been done as regards analysing the

results from an evolutionary palaeontology perspective. These studies have been targeted at and published in data mining venues, where the primary interest is in theoretical properties and the behaviour of the methods. An interesting research direction would be to investigate the findings of these studies from the evolutionary palaeontology perspective, analyse whether the extracted patterns match the current knowledge, and, more interestingly, what new insights they might provide.

Finding temporal structures

Finding temporal structures relates to data curation tasks for reconstructing age (see p. 483) and analysis over time (see p. 492). In principle, many algorithmic techniques could be used in all three contexts – the difference is in the research question or task. The main goal of data curation is to make data more representative. In analysis over time, the main goal is to track changes or developments in concepts of interest over time. Finding temporal structures is primarily aimed at identifying the structures themselves. One direction for finding temporal structures in fossil data could be to identify time periods that are in some ways similar to each other. Similar time periods would then define similar kinds of environment. Conceptually, the task would be similar to clustering, but instead of grouping objects one would group time periods. The cut-off points to find homogeneous time periods would also be defined computationally. Computational methods for tasks of this kind are available, and are known as motif discovery (Das and Dai, 2007; Mueen *et al.*, 2009). Such methods have been widely used, for instance, in bioinformatics for analysing gene sequences. As far as we are aware, they have yet to be applied in evolutionary palaeontology contexts, but they could, for instance, treat occurrences of traits over time as a continuous sequence.

Finding network structures

Finding network structures is about community detection or finding modules in a given network [for a survey of algorithmic techniques, see Fortunato (2010)]. Within the scope of our survey, network analysis has to date been primarily applied to trophic group detection in food webs (Gauzens *et al.*, 2015), i.e. the detection of groups of species that are functionally similar. A computational approach is an interesting alternative to expert identification of trophic groups, helpful for simplifying and understanding large and complex food webs that may contain thousands of links.

A growing means of discovering network structures is analysing how a network changes over time (Rozenshtein *et al.*, 2016), including event detection (Rozenshtein *et al.*, 2014) and motif discovery (Paranjape *et al.*, 2016). Network structures that are changing over time would be interesting to explore in fossil data analysis, in particular how the network structure is changing in relation to environmental changes.

FINDING RELATIONS IN FOSSIL DATA

Finding relations refers to producing models that can relate proxies derived from the data in such a way that some target characteristics can be computed from proxy data. For example, given occurrence patterns of taxa, the goal may be to numerically estimate climatic characteristics of localities, such as temperature or precipitation. A standard linear regression is an example of finding relations, but there are many algorithmic techniques beyond

regression, such as decision trees, support vector machines, nearest-neighbour approaches, Bayesian models, and neural networks. In machine learning, this setting is referred to as supervised learning, because there is a designated target characteristic.

Three types of tasks can be distinguished within finding relations: (1) descriptive modelling, (2) predictive modelling, and (3) causality analysis.

Descriptive and predictive modelling

The difference between a descriptive and predictive model is that in descriptive modelling, the model itself is the main interest for the purpose of understanding how proxies relate to each other. In predictive modelling, the main goal is to produce models that would accurately extrapolate outside the modelling data – to new situations, new areas, or unseen time intervals. Descriptive and predictive modelling are often used in the same study, or it could even be that the same models are analysed as descriptive and then used as predictive.

There are two main tasks for finding relations at the interplay between organisms and their environment. Environmental niche modelling or species distribution modelling relate the observed presence of a species to environmental conditions. It is an active research area, both from the ecology and computing perspectives, and computational techniques have been widely made use of, including nearest-neighbour-based techniques (envelope models) (Pearson and Dawson, 2003), statistical and machine learning techniques (Elith *et al.*, 2006), and, for example, maximum entropy modelling (Phillips *et al.*, 2006). Niche modelling mainly applies to the present day or the recent past, where environmental observations are available as measurements, there is relatively little uncertainty as regards taxonomic identification, and taxa of interest are available for all times of the analyses. Species distribution modelling mainly focuses on characterizing species in terms of their environments. Some studies have analysed how species relate to their environments by focusing on traits rather than taxonomic indicators, and modelling abundance as a function of traits (Shipley *et al.*, 2006; Brown *et al.*, 2014; Warton *et al.*, 2015). This is an interesting line of approach for applying the method to longer time scales, where species largely differ across different time periods but functional traits remain directly comparable.

Knowing how present-day communities relate to their environments, one can build models capturing patterns of this relation, which can then be extrapolated to the fossil communities to analyse past processes as drivers of evolutionary change. The reason that this approach would work is that even though communities change as taxa evolve over time, the principles of how communities relate to their environments and how they survive in those environments persist over time. That is, evolutionary processes leave patterns. If the processes in the present are similar to the processes in the past, they will leave similar patterns (Reed, 2013). Most of the taxa in the present are likely to be different from those in fossil assemblages. Yet the functional traits of taxa, governed by the laws of physics, chemistry, and physiology, are likely to be similar in the present and in the past. For example, animals that run tend to leave the same pattern of skeletal architecture, such as long limbs (Reed, 2013). Adaptations to certain habitats will likely be similar in the present and in the past. From the analytical perspective, similar patterns of adaptation within communities would indicate similar habitats.

In deep time analysis, the task of finding relations is aimed at answering different questions than in modern-day ecology. In the deep time setting, the objective is to estimate environmental conditions from observed fossil finds. The two main types of approaches,

similar to those used in ecology, are taxon-based (Atkinson *et al.*, 1987; Mosbrugger and Utescher, 1997; Kuhl *et al.*, 2002; Fernandez and Vrba, 2006; Birks *et al.*, 2010; Utescher *et al.*, 2014; Myers *et al.*, 2015) and trait-based, also known as ecometric approaches (Wolfe, 1995; van Dam, 2006; Eronen *et al.*, 2010b; Polly *et al.*, 2011; Liu *et al.*, 2012; Fortelius *et al.*, 2016; Vermillion *et al.*, 2017). The principal difference between the two approaches is similar to that of modern-day niche modelling. Taxon-based approaches are mainly concerned with describing how, as they operate via assumed links between the nearest living relatives of past species, and do not consider the functional relationships between organisms and the environment. Trait-based approaches are concerned with explaining why, via explicitly taking into account relations between traits of organisms and their environment. Traits such as leaf shape and tooth structure mediate interactions between organisms and their surroundings, and thus determine the place and conditions in which the organism can live most productively.

From the computational perspective, both types of approach can in principle use the same predictive modelling algorithms (either nearest-neighbour-based approaches or the explicit modelling of functional relationships via regression, decision trees, or the like), although the actual modelling choices have varied in the literature. The main difference between the taxon-based and trait-based approaches is in computational task setting; that is, how the data are transformed and aggregated for feeding into predictive modelling algorithms. In taxon-based approaches, a fossil site is described by a vector of indicator species taking values of present or absent. Determining the modern-day match for model calibration is accomplished via assigning the nearest living relative for each fossil species. Then, either virtual modern-day sites are created to exactly match the fossil sites in terms of species occurrence (co-existence approaches), or existing modern-day sites are used for inferring a relationship between species occurrence and climate variables (assemblage and calibration function approaches). With trait-based (ecometric) approaches, a vector of the distribution of functional traits (e.g. mean hypsodonty) is used to describe a fossil site, while modern-day sites are described in the same way. Instead of focusing on individual organisms, trait-based approaches deal with the functional composition of communities, assuming that trait variables are sufficiently general to accurately represent the functional relation of extinct taxa to the environment.

The main advantage of trait-based approaches in biological interpretation is that they can – to some extent – model the properties and interactions of unseen or extinct taxa as well, when the nearest living relative may be substantially different from the past species. The assumption here is that the relation (e.g. between the environment and tooth morphology) is preserved even if the other properties of the species and biome were very different.

Open computational challenges relate mainly to trait-based approaches. One interesting direction for future research would be the data-driven extraction of trait summaries, going beyond simple mean over site or considering projections of traits, for instance, in the spirit of principal component analysis. Another interesting direction would be to make models context-aware, in the sense that they could, for instance, take into account the state of evolution (what organisms and what traits were possible and feasible) at different times.

Causality analysis

Causality modelling (Pearl, 2009) is aimed at inferring causal relationships between proxies. It is well known that correlation does not generally imply causation; that is, the fact that

variables relate to each other does not mean that one causes the other, as there may be some hidden variable that causes both. As a caricature example, suppose one finds a relation between high hypsodonty and lack of precipitation. Removing some hypsodont horses from the area will not result in more rainfall.

Causality modelling is a very challenging task to accomplish without randomized trials, which are not possible by nature for the fossil data, yet some computational possibilities exist. The main principle is to input preliminary assumed causation and then computationally infer the strength and the direction of the causal links. Structural equation modelling is one type of such approaches (Bollen and Pearl, 2013). Structural equation modelling has been used in ecology (Menendez *et al.*, 2007; Grace *et al.*, 2016), and could potentially be extended to fossil data analysis. Hannisdal (2011; Hannisdal *et al.*, 2017) has experimented with information-theoretic approaches for inferring causality by taking into account the time order and time lag between observations of stable isotope ratios. The causality inference is based on the assumption that if one event occurs after the other, and there is a relation between the two events, then the direction of causality (if any) can only be forwards, not backwards. Other than this attempt, the application of causality analysis is lacking in evolutionary palaeontology, although it presents a promising direction for future research due to its potential for understanding and explaining evolution as a process.

ANALYSIS OF CHANGES OVER TIME

Analysis over time is about characterizing how communities and their environments change over time. Analysis can be based on raw fossil data, or on outputs produced via finding structures and finding relations. For instance, relations between the dental traits of plant-eating mammals and climate variables can be encoded using regression models, and then climate estimates produced by those models can be analysed over time. Analysis of changes over time is often performed manually via visual inspection of time series plots, but there are interesting computational approaches for analysis of changes over time that could be applied in similar settings. The most pressing computational time series analysis tasks for the analysis of fossil data over time would be: (1) aggregating for trends, (2) change and anomaly detection, (3) segmentation, and (4) sequential pattern mining.

Aggregating for trends

Most methods of classical time series analysis (see, for example, Shumway and Stoffer, 2011) fall under this category. The idea is to aggregate data into a single time series, which should then be easier to analyse by visual inspection. The most common approaches include spatial and temporal averaging, smoothing, identifying trends, and seasonality components via time series models.

Within the confines of this survey, aggregation and trend analysis has been used in the spatio-temporal analysis of faunal similarities to a reference (Pikermian) fauna (Eronen *et al.*, 2009), and in smoothing precipitation estimates over computational localities over time (Fortelius *et al.*, 2016).

One interesting possibility lies in determining point-in-time age estimates of fossil localities, or even individual fossils, thus extending the currently prevailing assignment of time ranges. Such a development in fossil dating would greatly increase the time granularity of the current fossil record, which would open up possibilities for applying advanced

computational analysis of such time series, including, for instance, the analysis of long-term cycles and reoccurring trends.

Segmentation

Segmentation of temporal data refers to partitioning data into bins. Conceptually, segmentation is similar to clustering, except that data points, which are close in time, have to be assigned to one group, and the task is to determine where the boundaries of groups are located. [For an overview of computational techniques, see Bingham (2010).]

Currently, time bins describing the ages of fossil records are determined based on domain expertise. In fossil data analysis, segmentation could potentially be used for identifying time bins computationally.

Sequential pattern mining

Sequential pattern mining is conceptually similar to pattern mining, except that it operates on data over time. Pattern mining is about finding items (e.g. species) that often occur together. Sequential pattern mining if about finding events that often occur one after another. [For an overview of computational techniques, see Mooney and Roddick (2013).] In fossil data analysis, sequential pattern mining could potentially be used for event detection and analysis, for example of extraordinary extinction events.

Change and anomaly detection

Change and anomaly detection is about identifying exceptional events or alterations in regimes. [Basseville and Nikiforov (1993) and Chandola *et al.* (2009) provide good overviews of common techniques used.] We are not aware of such studies in fossil data analysis, but if the time granularity of the fossil record increases due to finer dating, this would open interesting possibilities for identifying and characterizing events in a biosphere via computational change and anomaly detection.

METHODOLOGICAL CHALLENGES

The increasing availability of software tools makes advanced computational techniques more accessible for non-specialists, but at the same time easy access to 'black-box' implementations increases the risk of a misleading interpretation of the outcomes. The assumptions behind operations differ from algorithm to algorithm, but there are several generic aspects to take into consideration when analysing the results of computationally intensive studies.

There is no one best algorithm

A great number of computational algorithms are available, many of which have been applied to numerous problems in various fields, including evolutionary palaeontology. Algorithms differ in their assumptions – what form of relation between variables is assumed, what kind of data generation process is assumed, and how the model parameters are fit. The choice of more advanced algorithms (such as support vector machines, neural

networks, and Bayesian networks) over simple baseline methods is hampered by two factors. First, there is the issue of diminishing returns: the baseline methods often already have good estimation accuracy. While more advanced methods can improve the baseline accuracy, the improvements offered are marginal, and come at the cost of higher complexity. Second, while the use of more complex algorithms can be theoretically justified, the process of learning the theoretical framework may fail to take into account important aspects of the real problems associated with it. Even if one algorithm performs better than another on present-day data, for instance, the situation may be reversed on fossil data due to systematic differences in the fossil world compared with the present-day environment. Often, simpler methods are more robust for studying such changes, because they make fewer hidden assumptions. A complex model may uncover functional relations that may result in better performance on present-day data, but it may perform sub-optimally on past data where the same assumptions may not apply. These issues are elegantly discussed by Hand (2006) in the context of learning theory and supervised classification.

Moreover, relative performance will depend on the experience the person making the comparison has in using the methods. A researcher may find that his or her favourite method is best, merely because he or she understands the underlying assumptions behind the method well and has experience in how to squeeze the best performance from that method. The default parameters given in standard textbooks or software tools are not necessarily the best set for more specific tasks. For example, the fact that one is more familiar with decision trees is a good justification for choosing to employ decision trees, since it is more likely that they will be used following good practice.

Reliable validation

Reliable validation is critical for interpreting results. Statistical hypothesis testing offers a standard way of testing the results where appropriate, but just as important is understanding the relations and results, and verifying them with independent evidence. The methodology for statistical significance testing is well established in standard cases, such as comparing the means of two samples. In more complex cases, out-of-sample testing, such as cross-validation, may be an easier and safer option, especially when fitting predictive models for extrapolation purposes. Cross-validation is a procedure for the iterative testing of models. A dataset is divided into, for example, ten subsets at random. Nine subsets are used for model calibration and the tenth, which was not used for model calibration, is used for model testing. When model performance has been measured and recorded, the testing subset is put back into the modelling data, and another subset is removed as a testing subset. This procedure is repeated ten times, and the results are averaged. This is repeated until all the data have been used for validation (see, for example, Duda *et al.*, 2000). This procedure simulates out-of-sample testing.

Models can be characterized by their complexity, which essentially means model flexibility. For example, a regression model corresponding to a polynomial of a higher degree is more complex than a polynomial with a lower degree. The more flexible the model is, the more susceptible it is to capturing one-off deviations from the data rather than generic underlying patterns, especially if the calibration sample size is small. In such cases, the fit to the calibration data will be very good, but the model would fail to generalize on unseen data. In order to find the right balance between the data and model complexity, one should always validate on unseen data, for example by using a cross-validation procedure.

In computational tasks of finding patterns, the need for statistical significance testing is well known and widely applied, but there are two important issues to be kept in mind. First, a statistical test is meaningful as long as the null model is chosen correctly and the correct questions are asked. However, consider computing the co-occurrence of two species that both live in Africa. If we compute the global correlation of their occurrence, the test statistics will show a highly significant result; however, this only tells us that both species live in Africa, which we could have discovered more easily simply by looking at a map. To determine whether these species interact, one must choose the null model, by taking into account only the intersection of the species' areas of occurrence (see, for example, Kallio *et al.*, 2011).

Multiple hypothesis testing carries a risk of finding relations that might only be due to random fluctuations. For example, if we take 20 random correlations, normally we would accept a finding as statistically significant at the 5% confidence level. In a perfect world, we could form our hypothesis based on one set of data and test the hypothesis on another set of independent data. Statistical techniques, such as Bonferroni or Holm-Bonferroni correction, do exist to take into account multiple testing, but it is not a straightforward process to apply them fairly. It is easy to apply correction if, for instance, many species are tested for correlation in the same manner, but it is not so easy if many computational methods and their parameter settings are explored iteratively. Counting everything would perhaps be too conservative, and eventually deem any results insignificant, but not taking into account multiple testing carries a risk of arriving at misleading conclusions. Finding a balance is challenging. [For a review of multiple hypothesis testing, see Dudoit *et al.* (2003) or Hanhijärvi (2011).]

Statistical significance does not mean practical significance – or vice versa. Specifically, the absence of statistical significance does not necessarily mean that there is no relation; it may just mean that the sample is small. Given a sufficiently large sample, even minor differences can be found to be statistically significant; however, one should always consider whether it matters in practice, especially if the difference is very small even though it is statistically significant. Therefore, human judgement should always be used alongside statistical tests.

Circularity of reasoning

Circular reasoning may occur when information that would normally be unavailable before the conclusions are reached is used for inference. For example, one may model a species habitat from the climate variables, and then try to model climate from the estimated species habitat. In data mining and machine learning, a similar phenomenon is known as data leakage (Kaufman *et al.*, 2011), which is the unintentional introduction of predictive information about the target by the data collection, aggregation, and preparation process, which otherwise would not have been available. The main protection against circular reasoning is the critical judgement of the researchers involved.

The dangers of circular reasoning are well acknowledged in the community, but this awareness sometimes comes back as a 'double-edged sword', inducing an excessive fear of circularity. Using data from the same source to form different proxies within the same study may be methodologically justified, as long as different known processes operate behind those proxies, and there is a different scientific reasoning behind them. While circular reasoning should be avoided, circularity or crossing feeds of data does not necessarily have to be avoided, and indeed often cannot be avoided.

CONCLUDING REMARKS

The increasing availability of global-scale fossil data, powerful computing tools and interfaces, and the development of advanced analysis methods for 'big data', present new opportunities for large-scope research in evolutionary palaeontology. Easier access to data and tools has revitalized the field and facilitated many excellent studies. However, this has also resulted in a tendency to produce more and more mechanistic studies that are algorithm driven, and which focus on reporting algorithmic outputs with relatively few biological insights and interpretations. We have argued for matching knowledge of methods, datasets, and biological concepts in a research question-driven way.

We believe that the main potential of data-driven analysis is not in making human experts work faster, more broadly or more efficiently. Computational analysis can capture more complex patterns than the human eye or standard statistical tests. It can be viewed as an extension of our eyes, but not a substitute for human reasoning. Expert knowledge will not be replaced, but it will be translated into computing proxies and incorporated into analysis. The machine outputs will then need to be translated back to the humans. It is the process of summarizing existing scientific knowledge, defining computational tasks and concepts, and interpreting the outcomes that requires the major research and collaborative effort. The main research challenge is not in matching fossil data with an algorithm, but in matching knowledge of biology, knowledge of fossil data, and knowledge of algorithmic methods to define new meaningful research questions that can be answered in a computational way.

We hope that this survey serves as a step forward, offering a methodological guideline for what kinds of research questions in evolutionary palaeontology can potentially be answered with which kinds of algorithmic techniques.

ACKNOWLEDGEMENTS

This work received support from the Finnish Centre of Excellence for Algorithmic Data Analysis Research ALGODAN (K.P.), the Academy of Finland (decision 288814) (K.P.), the ECHOES Project funded by the Academy of Finland (M.F., I.Z.), and the Kone Foundation (J.T.E.). M.F. acknowledges generous support from the Alexander von Humboldt Foundation. This is a contribution to the ICCB (Integrative Climate Change Biology) program.

REFERENCES

Agusti, J., Cabrera, L., Garces, M., Krijgsman, W., Oms, O. and Pares, J.M. 2001. A calibrated mammal scale for the Neogene of western Europe: state of the art. *Earth-Sci. Rev.*, **52**: 247–260.

Alba, D.M., Agusti, J. and Moya-Sola, S. 2001. Completeness of the mammalian fossil record in the Iberian Neogene. *Paleobiology*, **27**: 79–83.

Allison, P.D. 2009. Missing data. In *The Sage Handbook of Quantitative Methods in Psychology* (R.E. Millsap and A. Maydeu-Olivares, eds.), pp. 72–89. Thousand Oaks, CA: Sage.

Alroy, J. 2000. New methods for quantifying macroevolutionary patterns and processes. *Paleobiology*, **26**: 707–733.

Atkinson, T.C., Briffa, K.R. and Coope, G.R. 1987. Seasonal temperatures in Britain during the past 22,000 years, reconstructed using beetle remains. *Nature*, **325**: 587–592.

Badgley, C. 1986. Counting individuals in mammalian fossil assemblages from fluvial environments. *Palaios*, **1**: 328–338.

Basseville, M. and Nikiforov, I.V. 1993. *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall [available at: http://people.irisa.fr/Michele.Basseville/kniga/].

Behrensmeyer, A.K., Kidwell, S.M. and Gastaldo, R.A. 2000. Taphonomy and paleobiology. *Paleobiology*, **26**: 103–147.

Berthold, M.R., Borgelt, C., Hoeppner, F. and Klawonn, F. 2010. *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*. Dordrecht: Springer.

Bingham, E. 2010. Finding segmentations of sequences. In *Inductive Databases and Constraint-Based Data Mining* (S. Dzeroski, B. Goethals and P. Panov, eds.), pp. 177–197. Dordrecht: Springer.

Bingham, E. and Mannila, H. 2014. Towards computational techniques for identifying candidate chronofaunas. *Ann. Zool. Fenn.*, **51**: 43–48.

Bingham, E., Kaban, A. and Fortelius, M. 2007. The aspect Bernoulli model: multiple causes of presences and absences. *Pattern Anal. Appl.*, **12**: 55–78.

Birks, H.J.B., Heiri, O., Seppa, H. and Bjune, A.E. 2010. Strengths and weaknesses of quantitative climate reconstructions based on Late-Quaternary. *Open Ecol. J.*, **3**: 68–110.

Bishop, C.M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Bollen, K.A. and Pearl, J. 2013. Eight myths about causality and structural equation models. In *Handbook of Causal Analysis for Social Research* (S.L. Morgan, ed.), pp. 301–328. Dordrecht: Springer.

Bonchi, F., van Leeuwen, M. and Ukkonen, A. 2011. Characterizing uncertain data using compression. In *Proceedings of the 11th SIAM International Conference on Data Mining*, pp. 534–545. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Brewer, S., Jackson, S.T. and Williams, J.W. 2012. Paleoecoinformatics: applying geohistorical data to ecological questions. *Trends Ecol. Evol.*, **27**: 104–112.

Brown, A.M., Warton, D.I., Andrew, N.R., Binns, M., Cassis, G. and Gibb, H. 2014. The fourth-corner solution using predictive models to understand how species traits interact with the environment. *Meth. Ecol. Evol.*, **5**: 344–352.

Cerling, T.E., Harris, J.M., Leakey, M.G., Passey, B.H. and Levin. N.E. 2010. Stable carbon and oxygen isotopes in East African mammals: modern and fossil. In *Cenozoic Mammals of Africa* (L. Werdelin and W.J. Sanders, eds.), pp. 949–960. Berkeley, CA: University of California Press.

Chandola, V., Banerjee, A. and Kumar, V. 2009. Anomaly detection: a survey. *ACM Comput. Surv.*, **41**: 15:1–15:58.

Clavel, J., Merceron, G. and Escarguel, G. 2014. Missing data estimation in morphometrics: how much is too much? *Syst. Biol.*, **63**: 203–218.

Couette, S. and White, J. 2010. 3D geometric morphometrics and missing-data: can extant taxa give clues for the analysis of fossil primates? *C.R. Palevol.*, **9**: 423–433.

Culverhouse, P.F., Macleod, N., Williams, R., Benfield, M.C., Lopes, R.M. and Picheral, M. 2013. An empirical assessment of the consistency of taxonomic identifications. *Mar. Biol. Res.*, **10**: 73–84.

Das, M.K. and Dai, H.K. 2007. A survey of DNA motif finding algorithms. *BMC Bioinform.*, **8** (suppl. 7): S21.

Duda, R.O., Hart, P.E. and Stork, D.G. 2000. *Pattern Classification*, 2nd edn. New York: Wiley-Interscience.

Dudoit, S., Popper Shaffer, J. and Boldrick, J.C. 2003. Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, **18**: 71–103.

Duivesteijn, W., Feelders, A.J. and Knobbe, A. 2015. Exceptional model mining. *Data Min. Knowl. Discov.*, **30**: 47–98.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A. *et al.* 2006. Novel methods improve prediction of species distributions from occurrence data. *Ecogeography*, **29**: 129–151.

Eronen, J.T., Ataabadi, M.M., Micheels, A., Karme, A., Bernor, R.L. and Fortelius, M. 2009. Distribution history and climatic controls of the late Miocene Pikermian chronofauna. *Proc. Natl. Acad. Scci. USA*, **106**: 11867–11871.

Eronen, J.T., Puolaṁaki, K., Liu, L., Lintulaakso, K., Damuth, J., Janis, C. *et al.* 2010a. Precipitation and large herbivorous mammals I: estimates from present-day communities. *Evol. Ecol. Res.*, **12**: 217–233.

Eronen, J.T., Puolaṁaki, K., Liu, L., Lintulaakso, K., Damuth, J., Janis, C. *et al.* 2010b. Precipitation and large herbivorous mammals II: application to fossil data. *Evol. Ecol. Res.*, **12**: 235–248.

Fernandez, M.H. and Vrba, E.S. 2006. Plio-Pleistocene climatic change in the Turkana basin (East Africa): evidence from large mammal faunas. *J. Human Evol.*, **50**: 595–626.

Foote, M. 2007. Symmetric waxing and waning of marine animal genera. *Paleobiology*, **33**: 517–529.

Foote, M. 2016. On the measurement of occupancy in ecology and paleontology. *Paleobiology*, **42**: 707–729.

Foote, M. and Raup, D. 1996. Fossil preservation and the stratigraphic ranges of taxa. *Paleobiology*, **22**: 120–141.

Fortelius, M., Gionis, A., Jernvall, J. and Mannila, H. 2006. Spectral ordering and biochronology of European fossil mammals. *Paleobiology*, **32**: 206–214.

Fortelius, M., Eronen, J.T., Kaya, F., Tang, H., Raia, P. and Puolamäki, K. 2014. Evolution of Neogene mammals in Eurasia: environmental forcing and biotic interactions. *Annu. Rev. Earth Planet. Sci.*, **42**: 579–604.

Fortelius, M., Zliobaite, I., Kaya, F., Bibi, F., Bobe, R., Leakey, L. *et al.* 2016. An ecometric analysis of the fossil mammal record of the Turkana basin. *Phil. Trans. R. Soc. Lond. B: Biol. Sci.*, **371**: 0232.

Fortunato, S. 2010. Community detection in graphs. *Phys. Rep.*, **486** (3): 75–174.

Garriga, G.C., Ukkonen, A. and Mannila, H. 2008. Feature selection in taxonomies with applications to paleontology. In *Discovery Science*, Proceedings of the 11th International Conference, DS 2008, Budapest, Hungary, 13–16 October 2008 (J.-F. Boulicaut, M.R. Berthold and T. Horváth, eds.), pp. 112–123. Berlin: Springer.

Gaston, K.J., Blackburn, T.M., Greenwood, J.J.D., Gregory, R.D., Quinn, R.M. and Lawton, J.H. 2000. Abundance–occupancy relationships. *J. Appl. Ecol.*, **37** (suppl. 1): 39–59.

Gauzens, B., Thébault, E., Lacroix, G. and Legendre, S. 2015. Trophic groups and modules: two levels of group detection in food webs. *J. R. Soc. Interface*, **12** (106): 20141176.

Gionis, A., Mannila, H., Puolamäki, K. and Ukkonen, A. 2006. Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 561–566. New York: ACM.

Gionis, A., Mannila, H., Mielikainen, T. and Tsaparas, P. 2007. Assessing data mining results via swap randomization. *ACM Trans. Knowl. Discov. Data*, **1** (3): 14.

Gotelli, N.J. and Ellison, A.M. 2013. *A Primer of Ecological Statistics*, 2nd edn. Sunderland, MA: Sinauer Associates.

Grace, J.B., Anderson, T.M., Seabloom, E.W., Borer, E.T., Adler, P.B., Harpole, W.S. *et al.* 2016. Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature*, **529**: 390–393.

Guex, J. and Davaud, E. 1984. Unitary associations method: use of graph theory and computer algorithm. *Comput. Geosci.*, **10**: 69–96.

Halekoh, U. and Vach, W. 2004. A Bayesian approach to seriation problems in archaeology. *Comput. Stat. Data Anal.*, **45**: 651–673.

Hammer, O. and Harper, D.A.T. 2006. *Paleontological Data Analysis*. Malden, MA: Blackwell.

Hand, D.J. 2006. Classifier technology and the illusion of progress. *Stat. Sci.*, **21**: 1–14.

Hand, D.J., Smyth, P. and Mannila, H. 2001. *Principles of Data Mining*. Cambridge, MA: MIT Press.

Hanhijärvi, S. 2001. Multiple hypothesis testing in pattern discovery. In *Discovery Science*, Proceedings of the 14th International Conference, DS 2011, Espoo, Finland, 5–7 October 2011 (T. Elomaa, J. Hollmén and H. Mannila, eds.), pp. 122–134. Berlin: Springer.

Hanhijärvi, S., Ojala, M., Vuokko, N., Puolamäki, K., Tatti, N. and Mannila, H. 2009. Tell me something I don't know: randomization strategies for iterative data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 379–388. New York: ACM.

Hannisdal, B. 2011. Non-parametric inference of causal interactions from geological records. *Am. J. Sci.*, **311**: 315–334.

Hannisdal, B., Haaga, K.A., Reitan, T., Diego, D. and Liow, L.H. 2017. Common species link global ecosystems to climate change: dynamical evidence in the planktonic fossil record. *Proc. R. Soc. Lond. B: Biol. Sci.*, **284**: 20170722.

Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Dordrecht: Springer.

Heikinheimo, H., Fortelius, M., Eronen, J. and Mannila, H. 2007. Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. *J. Biogeogr.*, **34**: 1053–1064.

Hempson, G.P., Archibald, S. and Bond, W.J. 2015. A continent-wide assessment of the form and intensity of large mammal herbivory in Africa. *Science*, **350**: 1056–1061.

Henelius, A., Puolamäki, K., Boström, H., Asker, L. and Papapetrou, P. 2014. A peek into the black box: exploring classifiers by randomization. *Data Min. Knowl. Discov.*, **28**: 1503–1529.

Henelius, A., Puolamäki, K., Karlsson, I., Zhao, J., Asker, L., Boström, H. *et al.* 2015. GoldenEye++: a closer look into the black box. In *Statistical Learning and Data Sciences* (A. Gammerman, V. Vovk and H. Papadopoulos, eds.), pp. 96–105. Lecture Notes in Artificial Intelligence #9047. Dordrecht: Springer.

Henelius, A., Puolamäki, K. and Ukkonen, A. 2017. Interpreting classifiers through attribute interactions in datasets. In *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning* (WHI 2017), pp. 8–13 [proceedings available at: https://arxiv.org/abs/1708.02666; paper available at: http://arxiv.org/abs/1707.07576].

Hooker, J. 1996. Mammalian biostratigraphy across the Paleocene–Eocene boundary in the Paris, London and Belgian basins. In *Correlation of the Early Paleogene in Northwestern Europe* (R. Knox, R. Corfield and R. Dunay, eds.), pp. 205–218. London: Geological Society of London.

Janis, C.M., Damuth, J. and Theodor, J.M. 2004. The species richness of Miocene browsers, and implications for habitat type and primary productivity in the North American grassland biome. *Palaeogeogr., Palaeoclimatol., Palaeoecol.*, **207**: 371–398.

Jernvall, J. and Fortelius, M. 2004. Maintenance of trophic structure in fossil mammal communities: site occupancy and taxon resilience. *Am. Nat.*, **164**: 614–624.

Kallio, A., Puolamäki, K., Fortelius, M. and Mannila, H. 2011. Correlations and co-occurrences of taxa: the role of temporal, geographic, and taxonomic restrictions. *Palaeontol. Electron.*, **14** (1): 4A.

Kaufman, S., Rosset, S. and Perlich, C. 2011. Leakage in data mining: formulation, detection, and avoidance. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 556–563. New York: ACM.

Koren, Y., Bell, R. and Volinsky, Ch. 2009. Matrix factorization techniques for recommender systems. *Computer*, **42** (8): 42–49.

Kuhl, N., Gebhardt, G., Litt, T. and Hense, A. 2002. Probability density functions as botanical-climatological transfer functions for climate reconstruction. *Quaternary Res.*, **58**: 381–392.

Legendre, P. and Legendre, L. 1998. *Numerical Ecology*. Amsterdam: Elsevier.

Lieberman, D.E. 1999. Homology and hominid phylogeny: problems and potential solutions. *Evol. Anthropol.*, **7**: 142–151.

Lijffijt, J., Papapetrou, P. and Puolamäki, K. 2014. A statistical significance testing approach to mining the most informative set of patterns. *Data Min. Knowl. Discov.*, **28**: 238–263.

Liow, L.H. and Stenseth, N.C. 2007. The rise and fall of species: implications for macroevolutionary and macroecological studies. *Proc. R. Soc. Lond. B: Biol. Sci.*, **274**: 2745–2752.

Liu, L., Puolamäki, K., Eronen, J.T., Ataabadi, M.M., Hernesniemi, E. and Fortelius, M. 2012. Dental functional traits of mammals resolve productivity in terrestrial ecosystems past and present. *Proc. R. Soc. Lond. B: Biol. Sci.*, **279**: 2793–2799.

MacLeod, N. and Steart, D. 2015. Automated leaf physiognomic character identification from digital images. *Paleobiology*, **41**: 528–553.

MacLeod, N., Benfield, M. and Culverhouse, P. 2010. Time to automate identification. *Nature*, **467**: 154–155.

Mampaey, M., Vreeken, J. and Tatti, N. 2011. Tell me what I need to know: Succinctly summarizing data with itemsets. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 573–581. New York: ACM.

Mampaey, M., Vreeken, J. and Tatti, N. 2012. Summarizing data succinctly with the most informative itemsets. *ACM Trans. Knowl. Discov. Data*, **6** (4): 16:1–16:42.

Meloro, C. and Kovarovic, K. 2013. Spatial and ecometric analyses of the Plio-Pleistocene large mammal communities of the Italian peninsula. *J. Biogeogr.*, **40**: 1451–1462.

Menendez, R., Gonzalez-Megias, A., Collingham, Y., Fox, R., Roy, D.B., Ohlemuller, Rr. *et al.* 2007. Direct and indirect effects of climate and habitat factors on butterfly diversity. *Ecology*, **88**: 605–611.

Mitchell, J.S. 2015. Preservation is predictable: quantifying the effect of taphonomic biases on ecological disparity in birds. *Paleobiology*, **41**: 353–367.

Mooney, C.H. and Roddick, J.F. 2013. Sequential pattern mining – approaches and algorithms. *ACM Comput. Surv.*, **45** (2): 19:1–19:39.

Moore, J.R., Norman, D.B. and Upchurch, P. 2007. Assessing relative abundances in fossil assemblages. *Palaeogeogr., Palaeoclimatol., Palaeoecol.*, **253**: 317–322.

Mosbrugger, V. and Utescher, T. 1997. The coexistence approach – a method for quantitative reconstructions of tertiary terrestrial palaeoclimate data using plant fossils. *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, **134**: 61–86.

Mueen, A., Keogh, E., Zhu, Q., Cash, S. and Westover, B. 2009. Exact discovery of time series motifs. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 473–484. Alexandria, VA: American Statistical Association.

Myers, C.E., Stigall, A.L. and Lieberman, B.S. 2015. PaleoENM: applying ecological niche modeling to the fossil record. *Paleobiology*, **41**: 226–244.

NOW Community, The. 2017. *New and Old Worlds Database of Fossil Mammals (NOW)*. Licensed under CC BY 4.0 [available at: http://www.helsinki.fi/science/now/].

Olson, E.C. 1952. The evolution of a Permian vertebrate chronofauna. *Evolution*, **6**: 181–196.

Paranjape, A., Benson, A.R. and Leskovec, J. 2016. Motifs in temporal networks [available at: https://arxiv.org/abs/1612.09259].

Passey, B.H., Ayliffe, L.K., Kaakinen, A., Zhang, Z.Q., Eronen, J.T., Zhue, Y. *et al.* 2009. Strengthened East Asian summer monsoons during a period of high-latitude warmth? Isotopic evidence from Mio-Pliocene fossil mammals and soil carbonates from northern China. *Earth. Planet. Sci. Lett.*, **277**: 443–452.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edn. Cambridge: Cambridge University Press.

Pearson, R.G. and Dawson, T.P. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecol. Biogeogr.*, **12**: 361–371.

Phillips, S.J., Anderson, R.P. and Schapire, R.E. 2006. Maximum entropy modeling of species geographic distributions. *Ecol. Model.*, **190**: 231–259.

Pineda-Munoz, S., Evans, A.R. and Alroy, J. 2016. The relationship between diet and body mass in terrestrial mammals. *Paleobiology*, **42**: 659–669.

Polly, D.P. 2010. Tiptoeing through the trophics: geographic variation in carnivoran locomotor ecomorphology in relation to environment. In *Carnivoran Evolution: New Views on Phylogeny,*

Form, and Function (A. Goswami and A. Friscia, eds.), pp. 374–410. Cambridge: Cambridge University Press.

Polly, D.P. and Sarwar, S. 2014. Extinction, extirpation, and exotics: effects on the correlation between traits and environment at the continental level. *Ann. Zool. Fenn.*, **51**: 209–226.

Polly, D.P., Eronen, J.T., Fred, M., Dietl, G.P., Mosbrugger, V., Scheidegger, C. *et al.* 2011. History matters: ecometrics and integrative climate change biology. *Proc. R. Soc. Lond. B: Biol. Sci.*, **278**: 1131–1140.

Puolamäki, K., Fortelius, M. and Mannila, H. 2006. Seriation in paleontological data using Markov chain Monte Carlo methods. *PLoS Comp. Biol.*, **2**: 62–70.

Qu, Y. and Xu, S. 2004. Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. *Bioinformatics*, **20**: 1905–1913.

Quinn, J.P. and Keough, M.J. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.

Reed, K. 2013. Multiproxy paleoecology: reconstructing evolutionary context in paleoanthropology. In *A Companion to Paleoanthropology* (R.D. Begun, ed.), pp. 204–225. Oxford: Wiley-Blackwell.

Ricci, F., Rokach, L., Shapira, B. and Kantor, P.B., eds. 2011. *Recommender Systems Handbook*. Dordrecht: Springer.

Rosindell, J., Hubbell, S.P. and Etienne, R.S. 2011. The unified neutral theory of biodiversity and biogeography at age ten. *Trends Ecol. Evol.*, **26**: 340–348.

Rozenshtein, P., Anagnostopoulos, A., Gionis, A. and Tatti, N. 2014. Event detection in activity networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1176–1185. New York: ACM.

Rozenshtein, P., Gionis, A., Prakash, B.A. and Vreeken, J. 2016. Reconstructing an Epidemic Over Time. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1835–1844. New York: ACM.

Saarinen, J., Oikarinen, E., Fortelius, M. and Mannila, H. 2010. The living and the fossilized: how well do unevenly distributed points capture the faunal information in a grid. *Evol. Ecol. Res.*, **12**: 363–376.

Shipley, B., Vile, D. and Garnier, E. 2006. From plant traits to plant communities: Aa statistical mechanistic approach to biodiversity. *Science*, **314**: 812–814.

Shumway, R.H. and Stoffer, D.S. 2011. *Time Series Analysis and Its Applications*, 3rd edn. Dordrecht: Springer.

Steininger, F.F. 1999. Chronostratigraphy, geochronology and biochronology of the Miocene European land mammal megazones (ELMMZ) and the Miocene mammal-zones. In *The Miocene Land Mammals of Europe* (G.E. Rossner and K. Heissig, eds.), pp. 9–24. Munich: Verlag Dr. Friedrich Pfeil.

Strauss, R., Atanassov, M. and Oliveira, J.D. 2003. Evaluation of the principal-component and expectation-maximization methods for estimating missing data in morphometric studies. *J. Vert. Paleontol.*, **23**: 284–296.

Tatti, N. 2008. Maximum entropy based significance of itemsets. *Knowl. Inf. Syst.*, **17**: 57–77.

Tatti, N. 2010. Probably the best itemsets. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 293–302. New York: ACM.

Tatti, N. and Vreeken, J. 2012a. Comparing apples and oranges: measuring differences between exploratory data mining results. *Data Min. Knowl. Discov.*, **25**: 173–207.

Tatti, N. and Vreeken, J. 2012b. Discovering descriptive tile trees by mining optimal geometric subtiles. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 9–24. Dordrecht: Springer.

Turvey, S.T. and Blackburn, T.M. 2011. Determinants of species abundance in the Quaternary vertebrate fossil record. *Paleobiology*, **37**: 537–546.

Uhen, M.D., Barnosky, A.D., Bills, B., Blois, J., Carrano, M.T., Carrasco, M.A. *et al.* 2013. From card catalogs to computers: databases in vertebrate paleontology. *J. Vert. Paleontol.*, **33**: 13–28.

Ukkonen, A., Fortelius, M. and Mannila, H. 2005. Finding partial orders from unordered 0–1 data. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 285–293. New York: ACM.

Ukkonen, A., Puolamäki, K., Gionis, A. and Mannila, H. 2009. A randomized approximation algorithm for computing bucket orders. *Inf. Process. Lett.*, **109**: 356–359.

Utescher, T., Bruch, A.A., Erdei, B., Francois, L., Ivanov, D., Jacques, F.M.B. *et al.* 2014. The coexistence approach – theoretical background and practical considerations of using plant fossils for climate quantification. *Palaeogeogr., Palaeoclimatol., Palaeoecol.*, **410**: 58–73.

van Dam, J.A. 2006. Geographic and temporal patterns in the late Neogene (123ma) aridification of Europe: the use of small mammals as paleoprecipitation proxies. *Palaeogeogr., Palaeoclimatol., Palaeoecol.*, **238**: 190–218.

van de Velden, M., Groenen, P. and Poblome, J. 2009. Seriation by constrained correspondence analysis: a simulation study. *Comput. Stat. Data Anal.*, **53**: 3129–3138.

van Leeuwen, M. 2010. Maximal exceptions with minimal descriptions. *Data Min. Knowl. Discov.*, **21**: 259–276.

van Leeuwen, M. and Knobbe, A. 2012. Diverse subgroup set discovery. *Data Min. Knowl. Discov.*, **25**: 208–242.

van Leeuwen, M., Vreeken, J. and Siebes, A. 2009. Identifying the components. *Data Min. Knowl. Discov.*, **19**: 176–193.

Vermillion, W.A., Head, J.J., Polly, D., Eronen, J.T. and Lawing, A.M. (accepted) 2017. Ecometrics: a trait-based approach to paleoclimate and paleoenvironmental reconstruction. In *Methods in Paleoecology: Reconstructing Cenozoic Terrestrial Environments and Ecological Communities* (D.A. Croft, S.W. Simpson and D.F. Su, eds.). Dordrecht: Springer.

Wal, J.V.D., Shoo, L.P., Johnson, C.N. and Williams, S.E. 2009. Abundance and the environmental niche: environmental suitability estimated from niche models predicts the upper limit of local abundance. *Am. Nat.*, **174**: 282–291.

Wang, X., Flynn, L. and Fortelius, M., eds. 2013. *Fossil Mammals of Asia: Neogene Terrestrial Biostratigraphy and Chronology*. New York: Columbia University Press.

Warton, D.I., Shipley, B. and Hastie, T. 2015. CATS regression – a model-based approach to studying trait-based community assembly. *Meth. Ecol. Evol.*, **6**: 389–398.

West, J.B., Bowen, G.J., Cerling, T.E. and Ehleringer, J.R. 2006. Stable isotopes as one of nature's ecological recorders. *Trends Ecol. Evol.*, **21**: 408–414.

Wolfe, J.A. 1995. Paleoclimatic estimates from tertiary leaf assemblages. *Annu. Rev. Earth Planet. Sci.*, **23**: 119–142.

Woodburne, M.O. 2004. *Late Cretaceous and Cenozoic Mammals of North America: Biostratigraphy and Geochronology*. New York: Columbia University Press.

Zermoglio, P.F., Guralnick, R.P. and Wieczorek, J.R. 2016. A standardized reference data set for vertebrate taxon name resolution. *PLoS One*, **11** (1): e0146894.