

Ordination analysis in sedimentology, geochemistry and palaeoenvironment—Background, current trends and recommendations

Or M. Bialik¹  | Emilia Jarochovska²  | Michal Grossowicz³

¹Marine Geology and Seafloor Surveying, Department of Geosciences, University of Malta, Msida, Malta

²GeoZentrum Nordbayern, Friedrich Alexander University of Erlangen-Nürnberg, Erlangen, Germany

³GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

Correspondence

Or M. Bialik, Marine Geology and Seafloor Surveying, Department of Geosciences, University of Malta, Msida MSD, Malta.
Email: or.bialik@um.edu.mt

Emilia Jarochovska, GeoZentrum Nordbayern, Friedrich Alexander University of Erlangen-Nürnberg, Erlangen, Germany.
Email: emilia.jarochovska@fau.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: Ja 2718/3-1; H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 101003394

Abstract

Ordination is the name given to a group of methods used to analyse multiple variables without preceding hypotheses. Over the last few decades, the use of these methods in Earth science in general, and notably in analyses of sedimentary sources, has dramatically increased. However, with limited resources oriented towards Earth scientists on the topic, the application of ordination analysis is at times suboptimal and misuse by authors can occur. This text was written for researchers with little to no experience with ordination with the aim of exposing them to the utility and the pitfalls of this branch of exploratory statistics. To do so, a detailed review of three ordination methods is offered: principal component analysis, non-metric multidimensional scaling and detrended correspondence analysis. A survey of 163 publications in Earth science is presented, in which these ordination methods were used together with a summary of how, why and on what type of data ordination was used. With common mistakes outlined and misuses in those publications identified. Notably, issues were found with reproducibility, documentation, data set dimensions and transformations. Based on this survey, a recommended workflow is offered for Earth scientists who wish to apply ordination. Additionally, this article is accompanied by highly annotated R scripts for novice users to use these methods.

KEYWORDS

clustering, DCA, dimension reduction, exploratory statistics, multivariate statistics, NMDS, PCA

1 | INTRODUCTION

Geology, as a science, combines descriptive (qualitative and semi-quantitative) and quantitative approaches. As the discipline evolved, the lexical richness of descriptive tools increased, technological advances have provided geoscientists with a wide range of quantitative tools. These advances allows modern geoscientists to generate an incredible richness

of data, often with high levels of complexity. For example, a study may now include textural description, elemental chemistry, isotopes and mineralogy from tens or even hundreds of samples with both spatial and temporal distribution. To deal with these progressively larger data sets, researchers have been turning to multivariate analysis methods, and particularly to ordination.

Geoscientists can approach these data in an exploratory fashion, a hypothesis testing one or a combination of both

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *The Depositional Record* published by John Wiley & Sons Ltd on behalf of International Association of Sedimentologists.

approaches. However, even when testing hypotheses, researchers usually look for positive evidence (Devezer et al., 2019). In that sense, geology and its sister disciplines do not adhere to the Popper school of testing scientific theories through falsifiability (Popper, 1934). Birks (1985) argued that geology was/is still in the empirical-descriptive or narrative phase of its development, and would develop onwards to an analytical phase governed by the hypothetico-deductive approach. While geosciences have indeed developed in this direction, it is a field that still has a very strong exploratory component. This, in turn, requires a different statistical process than that implemented in disciplines driven more by falsifiability (Mulaik, 1985).

Ordination has been used in Earth sciences for many years, with the earliest mention of it dating back to the 1940s (Griffiths, 1947), but its use has proliferated over the last 30 years as personal computing and statistical programs became more prevalent. Unfortunately, the most popular and widely accessible resources on multivariate statistics are oriented towards statisticians, ecologists or social scientists, and not geoscientists. For example, at the time of writing, of the 118 books on multivariate statistics available in the Springer catalogue only two are specified for geoscientists (Brown, 1998; Wackernagel, 2003), and only one of them discusses ordination methods. A similar underrepresentation is also observed in the Elsevier and Blackwell catalogues. This highly limits an Earth science student or scientist interested in applying these methods as the translation of terms and examples from these fields into geoscience might often be non-trivial. As a result, multivariate statistics in general, and ordination in particular, are not only less accessible to Earth scientists, they are also further removed from their intuitive toolbox. Most ordination techniques have been introduced by plant ecologists and many software packages implementing them are explicitly designed to handle species abundance or occurrence matrices (Hammer & Harper, 2007; Oksanen et al., 2019), which require different transformation techniques and choice of parameters. This results in useful statistical tools not being considered by Earth scientists when conducting research and compiling manuscripts. Moreover, as discussed later, when these methods are implemented, mistakes and misuses probably to occur.

An additional concern regarding the use of statistical tools has to do with reproducibility (Nüst & Pebesma, 2020). Over the last few years, there has been a growing concern in the scientific community regarding the reproducibility of scientific results (Cooper, 2018; Nissen et al., 2016). Implementation affects reproducibility as different statistical packages and even slightly different workflows can result in different outcomes. Proper reporting of the assumptions, what analytical methods were used and why, and their output is considered crucial for comparability between studies (Amrhein et al., 2019). This raises the need for better reporting and

documentation of the statistical methods used and deeper understanding of their underlying algorithm.

There is a need to discuss the implementation of statistical tools so that they are more suitable for sedimentology, geochemistry, palaeoenvironmental studies *et hoc genus omne* (Latin: and everything else of this kind). Since, as noted, these fields often have a strong exploratory nature, this study set out to examine the use of ordination in these fields. Presented here is a review of the literature on ordination, with three selected popular methods described in detail: principal component analysis (PCA), non-metric multidimensional scaling (NMDS) and detrended correspondence analysis (DCA). The mathematics and equations are not detailed here, instead carefully selected references are provided on where to find them in an approachable form. In this work, a roadmap to sources is presented which the authors find particularly useful for geoscientists. This study presents a survey of the use of these ordination methods in Earth sciences, focusing on work done on rock, water, sediment and fossils. From this review, several common mistakes and issues were identified with the application of ordination methods and the interpretation of results. This text concludes by suggesting several recommended workflows and ways to avoid the identified pitfalls.

1.1 | What is ordination?

Ordination is the name for a family of multivariate analysis methods for exploratory data analysis (Gower, 1987). The common thread of these techniques is that they all order multivariate objects in a fashion that places similar objects near each other with dissimilar objects further away. The term exploratory data analysis here refers not so much to dealing with unknown settings for the first time but rather an approach to analysing data sets by summarising their main characteristics and leading to formulation of hypotheses rather than testing them (Chatfield, 1995). This approach falls into the broader category of machine learning without supervision (Hastie et al., 2009). As such, ordination methods were principally developed to allow a researcher to examine a data set *tabula rasa* and identify the relationships in it.

In the survey presented here, ordination was used for three principal goals: (a) to understand or identify the relationships between variables; (b) to differentiate or cluster data points; and (c) generate indices. The first two are clear derivatives of the exploratory nature of ordination methods. The third is a utilisation of the property of most ordination methods—dimension reduction. Ordination methods represent a higher number of dimensions by a smaller number of dimensions called components (Gauch & Whittaker, 1972; Syms, 2008). This can allow the representation of multiple variables as a much smaller number (one to three) of variables (indices). These indices

should then represent the highest proportion of the variability that can be shown in this reduced number of dimensions. Sedimentological data sets often consist of proxies, for example for the redox conditions, for distance from sediment source etc., where multiple variables are driven by a common underlying process. Despite that, the signal is recorded with different sensitivities and measured with different error terms. In dealing with proxy data, reduction of dimensions typically aims to reconstruct this underlying process through a combination of co-varying proxy variables approximating this process (Abdelhady & Fürsich, 2014; Bitušik et al., 2018; Gupta & Thomas, 1999; Koutsodendris et al., 2020; see Table S2).

There are multiple ordination methods developed over the course of the 20th century, from the classical canonical correlation (Austin, 1968) and polar ordination (Bray & Curtis, 1957) to more modern methods such as canonical correspondence analysis (ter Braak, 1986) and redundancy analysis (ter Braak & Prentice, 1988). This text focuses on and discusses three of them: PCA, NMDS and DCA, summarised in Table 1. These methods are in common usage across multiple disciplines, including Earth sciences, implemented in all major statistical packages and described in approachable textbooks. It is important to note that each ordination method is a class of algorithms in its own right, grouped together based on a common approach. This means that the same method might produce different outputs between two software packages due to differences in the underlying algorithms they use.

1.2 | Similarity Indices

An underlying aspect of ordination is the measurement of how similar or dissimilar data points are from one another.

The use of similarity indices (also called similarity measures or similarity functions) is a means to evaluate this between samples. Different ordination methods use as input distance matrices, which describe the distances between each point and all the other points in the data set. However, similarity indices can also be used on their own. There are a large number of similarity indices, Sneath and Sokal (1973) and Johnston (1976) listed over 30 different indices and offer an extensive review of them. Some indices summarise the values of all variables into a single index (community structure measures sensu Pinkham & Pearson, 1976), whereas others pair attributes.

An important point to remember with regard to similarity indices is their underlying assumptions and methods. For example, while the input data to the index might be in absolutes, the index might treat them as proportions. These changes effectively turn an open data set into a closed one (Hammer & Harper, 2007, also see below regarding open and closed data sets), that is to say defined as part of a sum, which has a knock-on effect on the distributions of the variables. Many indices implemented in popular packages have been explicitly developed to handle biological diversity aspects and are derived from common ways of measuring diversity (see Jost et al., 2011, for a detailed analysis of the properties of common indices). As a result, many of them will not be relevant for sedimentological data. An overview of available implementations, general or dedicated to specific types of data, for R Software is available as Task View (<https://cran.r-project.org/view=Multivariate>).

Three indices that should be mentioned here, as they are relevant to the text, are Jaccard, Bray–Curtis and Gower. The Jaccard index (Jost et al., 2011) is one of the oldest indices still in use. This method was specifically designed for binary (0 or

TABLE 1 Summary description of the properties of PCA, NMDS and DCA

	PCA	NMDS	DCA
Brief description of operation	Orthogonal combination of n -dimensional variables	Lower dimensional optimisation of dissimilarity between data points	Consecutive dimensional warping of n -dimensional variables
Measurement of dimensional optimisation	Variance	Stress	Correspondence
Variables	Quantitative data, linear relationships	Quantitative, semi-quantitative, qualitative, or mixed	Non-negative, on a 0 to 1 range, quantitative or binary data
Assumes symmetrical distribution	Yes	No	No
Requires pre-treatments	Sometimes (normalising transformation)	Usually no, but this depends on the similarity index used	Sometimes (standardisation to [0, 1] range)
Consistent between runs	Yes	No	Yes
Preserves original dimensionality information	Yes	No	Yes
Relationships between variables	Axis are a composite of the vectors, vectors can be represented by eigenvalues	Axis are a qualitative representation of the total effect in the data set	Axis are a qualitative representation of the total effect but vectors can still be represented in this space

1, also called presence/absence or PA) variable sets by using the size of the intersection divided by the size of the union of the sample sets. Another commonly used index is the Bray–Curtis dissimilarity index (Bray & Curtis, 1957). This index has been specifically named as one of the more robust indices for communities (Bloom, 1981). This index is designed for compositional data sets which follow a consistent sampling protocol (e.g. the same number of points is counted in multiple thin sections; the same number of foraminifera counted in each sample), and is computed based on the ratio of lesser values to the sum of the vectors. Nonetheless, data sets need to be prepared in consideration of the properties of this index. The index approaches 0 when sample sizes are very different, regardless of how similar their composition is. In such cases, samples need to be standardised to be used with this index, for example if the number of counts in a sample can be standardised to correspond to the proportions in a sample; or if a different index may be more appropriate. For a detailed illustration of the limitations of the Bray–Curtis index and possible solutions, see Jost et al. (2011) and Chao et al. (2006). Gower's similarity coefficient (Gower, 1971; Podani, 1999) is a notably important index which will be expanded on further in the text as it is one of the more suitable for mixed data (van de Velden et al., 2019), that is to say a combination of compositional and non-compositional variables. This index uses a weight function to compute the similarity and eliminate objects equal to zero.

1.3 | Closed and open data sets, compositional and non-compositional data

The mathematical meaning of open and closed sets are complex topological terms defining sets of points meeting specific conditions in a topological space. Here a simpler case is used (Reyment & Savazzi, 1999) where the data matrix is constrained (or closed) or unconstrained (open). Closed data sets are also known as compositional, Egozcue (2009) provided a handy definition for these as ‘compositional data thus quantitatively expresses relative contributions of variables under consideration of a certain whole, which carry relative information between the components’. This is sometimes defined in simpler terms as ‘compositional data being a positive value multivariate data that sum up to a constant’. The underlying notion is that a change in one value of a given variable will permeate changes across the entire variable set. This interdependence is very well illustrated in data given in percent. If in an assemblage there is a set number of elements summed and divided to percent of the total, that total will be 100%. Then if on subsequent analysis any of the values from this assemblage are removed or changed—all other values will have to change so that the sum of the fractions will remain 100%.

Compositional (closed) and non-compositional (open) data sets do not behave in the same way and have different geometrical properties (Aitchison, 1982; Filzmoser et al., 2009; Greenacre, 2018). As a result, the distributions of compositional data sets are different from those of non-compositional data sets (where the variables do not have interdependence). This requires that compositional data sets will be pre-treated if one wishes to apply methodologies designed for non-compositional data sets or that make assumptions about the underlying probability distribution (parametric methods), assumed for non-compositional data sets.

2 | ORDINATION METHODS

2.1 | Principal component analysis

Principal component analysis is an approach for finding variables (referred to as components) that account for the maximum amount of the variance in a multidimensional data set (Duneman, 1989; Hotelling, 1933). These components are linear combinations of the original variables. By summing the total variation of all components and defining the variation of each component by that number, one obtains the percent of explained variance on that component. The variances of each component are often presented in a form called a ‘scree plot’. The required input must be metric, that is to say, the input cannot be Boolean (true/false or presence/absence) nor can it be made of distinct categories (which are often used to describe sediments and rocks, e.g. facies and lithology) nor can it be ranked (see Section 4.2 for a classification of levels of measurement). To refrain from repeating the description of mathematical operations behind PCA, interested readers can find approachable explanations in the textbooks such as Rencher (2003) or Gauch (1982).

In a bare-bones algorithm, based only on variance, units are very important for two reasons. First, the size of numbers in each variable will affect the calculation of the sum of squared distances, emphasising large numbers. For example, if all variables are given in the same unit, and there are orders of magnitude of difference between variables (e.g. micrite in a point count of thin sections from a mudstone), the larger numbers will dominate the axis of maximal variance (referred to as PC1). PCA, by design, is optimised to represent the maximum variability and not nuances of small sample variance. Second, any unit that is given as a fraction of a whole, for example percent weight, mg/kg, ppm, etc. may lead to a closed-sum problem, which can compromise linear relationships between variables due to a poor representation of the dimensional centre and distribution around it (Filzmoser et al., 2009; Tolosana-Delgado, 2012).

Principal component analysis relies on the matrix of covariance between each pair of variables in the data set.

Covariances are, however, not meaningful if variables are measured in different units. A more complex situation, but one not unimaginable in sedimentology, would be if the variables are measured in the same units, but are not measured on a ratio scale. An example would be frequencies in bins, especially if bins are defined differently for two variables, or for ratio variables, which are unitless. Sometimes such cases may be hard to identify, but as a rule of thumb, the recommendation here is that if variables are not clearly measured in the same units, the matrix of covariance should not be used. Instead, PCA can be calculated using the matrix of correlations between variables. This is not the default option in most packages and needs to be consciously chosen by the user.

Principal component analysis was originally designed with data sampled from a multivariate normal distribution in mind. This means that, ideally, the variables should have a multinormal distribution or be normalised (Legendre & Legendre, 2012). Performance of PCA does not require variables to be normalised (see Section 2.5) or have a normal distribution, and deviation from normality might not bias the analysis (Ibanez, 1971). However, strong deviations from symmetry—which is a property of the normal distribution—will affect the performance of PCA negatively (Hammer & Harper, 2007). In the absence of symmetry in the distributions, the resulting ordination might be very sensitive to small changes. A common approach is to apply a monotone transformation that reduces the skewness (asymmetry) of the distribution. The topic of transformation is discussed further along the text (see Section 2.5); for more details, both Legendre and Legendre (2012) and Filzmoser et al. (2009) offer overviews of available transformation methods, the latter being more oriented towards Earth sciences. Principal component analysis also suffers from several other sensitivity issues related to the data distribution (Shi, 1993), some of which DCA and NMDS were specifically designed to overcome.

2.2 | Non-metric multidimensional scaling

Non-metric multidimensional scaling (here NMDS, sometimes nMDS, NMS or MDS) is an indirect gradient analysis approach which produces an ordination (Kruskal, 1964). However, rather than using some metric of distance, NMDS substitutes the original distance data with ranks. The use of ranks omits some of the issues associated with using absolute distance (e.g. sensitivity to transformation), and as a result, is a much more flexible technique that accepts a variety of types of data (Field et al., 1982), it is also one of the best general predictive ordination methods (Wildi, 2018). Non-metric multidimensional scaling operates as an iterative process: first, it generates a dissimilarity matrix for every pair of samples, for example by using the

Bray–Curtis dissimilarity index. The NMDS algorithm then finds an optimal monotonic transformation of the similarities in order to obtain optimally scaled data, and tests that versus the similarity matrix. It then rearranges the configurations in order to minimise the stress—the difference between the original ranked distances and those in the transformed output. The algorithm repeats these steps until the stress is reduced to a pre-defined level. Ideally, the stress should be less than 0.05 although stress of 0.05–0.10 is still good, despite the risk of false inferences. That said, a very large number of samples could lead to high stress despite reasonable inference (Dexter et al., 2018).

This iterative nature of NMDS makes it computationally heavy (Alotaibi et al., 2011) which could be an issue with very large data sets, although more modern algorithms have improved computation time (Taguchi & Oono, 2005). Additionally, this means that NMDS is expected to return the similar but not identical result on different runs (Hammer & Harper, 2007). This last property is particularly important to note when using NMDS to generate indices. Even if data are not changed, different permutations could result in slightly different results, which in turn could decrease the signal-to-noise ratio. This can be circumvented by seeding the random number generator by a fixed number (see Section 3.3.2.4).

2.3 | Detrended correspondence analysis

Detrended correspondence analysis is a derivative of the correspondence analysis (CA) method (Hill, 1979). While CA operates very similarly to PCA, it determines the component position by maximising the correspondence between variables and data points rather than the variance. Detrended correspondence analysis adds another layer of operation aimed at neutralising the ‘arch effect’ CA suffers from, where the points forming a gradient reconstructed along the second axis are distributed along an arch relative to the first axis. This arch results in misrepresentation of the gradient orthogonal to the first gradient, but is a conflation of both gradients. This detrending can be done in two methods, either by expressing each subsequent axis as a polynomial function of the prior axis or by a segmentation method in which segments of each axis are centred to have a zero-mean relative to the subsequent axis. Following this step, a non-linear rescaling is implemented to shift sample scores along each axis so that the average width would become 1. These distortions of the axis, especially with the segmentation method, is a core criticism of DCA (Wartenberg et al., 1987). Detractors argue that this masks, through mathematical manipulations, the data's real curvilinear structure and thus hinders the understanding of the real data and identification of the causes of the observed distribution.

2.4 | Clustering and cluster testing

Clustering is a different subset of exploratory techniques for identifying groups and subgroups in a multivariate data set, in other words—identify which objects in a data set are similar to each other and to what level (Romesburg, 2004). Ordination methods are capable of detecting groups (clusters) if they are present in the data set, but they are not clustering methods. Paired with their relationship to the variables, it is possible to formulate hypotheses about distribution patterns in the data. The clusters might be a priori, defined for example by lithology or temporal position and reinforced by the ordination. Alternatively, the clusters might be defined a posteriori (learning without supervision) to the ordination based on the inferred distribution of samples or variables across the main components. In either case, apparent clustering does not necessarily mean that the clusters are indeed different in a statistically significant manner. This is where statistical tests come into play to ensure that the clusters are indeed dissimilar, given the size of the data set and the variation of the variables. As ordination methods are applied to multivariate data, so must be the test, and matched to the type of data and cluster analysed. One of the most common multivariate tests (Warne et al., 2012) is the multivariate analysis of variance (MANOVA), a multivariate extrapolation of analysis of variance (ANOVA). Similar to PCA, MANOVA uses the variance of the population, additionally, it assumes a multi-normal distribution (Olson, 1974). Therefore, it is acceptable to use this statistical test when clustering post PCA. Another problem with applying MANOVA is its internal assumption that the variables are independent, which is inherently problematic with closed set data (Finch, 2005). Non-parametric variations of MANOVA (NPMANOVA) have been developed to mitigate some of these issues (Puri & Sen, 1971). The non-parametric analysis of similarities (ANOSIM) is a good fit to test groups inferred specifically from NMDS as both use ranking (Buttigieg & Ramette, 2014; Clarke, 1993). Essentially, only one assumption is made by ANOSIM, that the ranges of ranked dissimilarities within groups are equal, or at least very similar.

Thus ANOSIM is an example of a permutation test, which overcomes various limitations of parametric methods (Kowalewski & Novack-Gottshall, 2010; chapter 1.2 in Legendre & Legendre, 2012). The permutation concept allows sampled data to be compared against a distribution drawn from random simulations rather than from a theoretical distribution, like the Gaussian or the Poisson, and therefore to analyse data that does not satisfy the statistical assumptions underlying traditional parametric tests (Collingridge, 2013). It is also recommended to use a permutation test when sample sizes are small, or if the degrees of freedom are low. Since each data set can draw a distribution using permutations, every possible statistical test can be used on

the alternative distribution, thereby creating a permutation test. Therefore, this concept can be used also in multivariate analysis. For example, ANOSIM is a permutation test, as is PerMANOVA, the permutation alternative for the above mentioned MANOVA. Another statistical test, ADONIS, is suitable for mixed data sets as it relies on similarity indices, so the choice of a suitable similarity index makes it possible to compare variables which otherwise would not be comparable (Anderson, 2001).

2.5 | Transformations and standardisation

Variables used as input for ordination may need pre-treatment. The character of the pre-treatment depends on the requirements of a particular method. Sections 4.1.1 and 4.1.3, as well as Bialik et al. (2021), offer an example where the same data set needs different pre-treatment for PCA and for DCA. Furthermore, some pre-treatment operations are specific to particular types of variables, such as in the case of a spectrogram (Abudulla et al., 2013) or ecological community composition. For the latter, pre-treatment considerations are explained in detail by Legendre and Gallagher (2001). Owing to the breadth of possibilities, the most important situations requiring pre-treatment directly are identified when explaining aspects of the respective ordination algorithm that necessitate it (see Section 2.1 here for PCA requirements). Furthermore, some pre-treatments, such as standardisation to the same variance (see below), depend on the type of variable (explained in Section 4.2), as variance cannot be meaningfully compared between variables measured on different scales. It is recommended to iterate using pre-treatments and test their impact on the final output.

Pre-treatment of variables to fulfil the requirements of the chosen statistical method can be divided into two groups: transformation, which changes the shape of the distribution of a variable (e.g. from a log-normal to normal), and standardisation, which does not change the shape of the distribution, but changes its position (Figure 1). A common case of transformation is the normalising transformation, or normalisation (Legendre & Legendre, 2012), which is the adjustment of the variable's distribution to resemble the Gaussian (normal) distribution. In cases of positive skewness (long right tail) root transformations are useful, whereas in cases of negative skewness, log-transformations can be used. In addition to these general cases, there are many specific ones, for example bimodal distributions, where case-by-case choice may be necessary. A normal distribution will have skewness equal to zero, but for non-normal distributions other measures of symmetry are available (e.g. QuAsy by Hohmann & Jarochovska, 2019). In either case, skewness or a different measure of symmetry should be as close to zero as possible. The exact value depends not only on the departure from

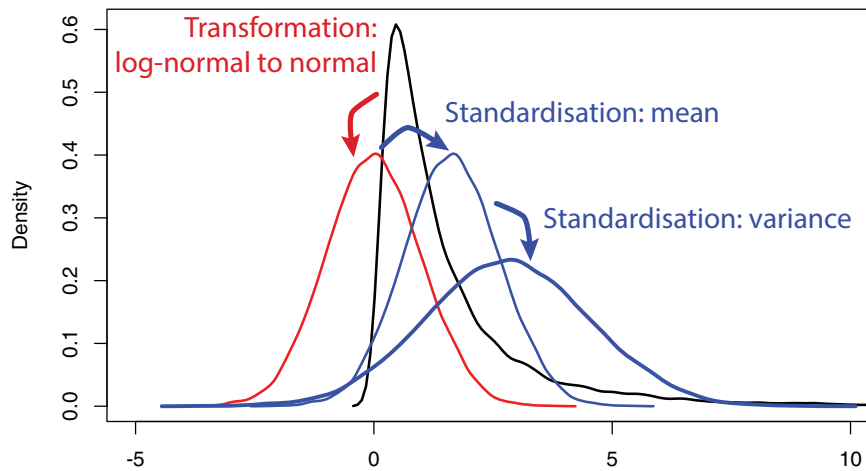


FIGURE 1 Example of a common transformation from a log-normal distribution with mean equal 0 and standard deviation equal 1 to normal distribution. Log-transformation causes the mean and standard deviation of the generated normal distribution to be different from the original distribution, therefore two common standardisation operations are shown: shifting (translation) of the distribution so that its mean equals zero and expanding it so that its standard deviation equals one. This combination allows the distribution to be used with methods requiring normality, while also preserving the original mean and variance. X represents any real-valued random variable

symmetry, but also on the type of underlying distribution, therefore no rule of thumb can be given and documentation of variables becomes even more important.

Further examples of transformations include transformations aiming to achieve linear relationships between variables. For example, when the data set contains variables with uniform and exponential distributions, those with exponential distributions can be log-transformed. Ordination methods are typically robust with respect to shapes of distributions, with the notable exception of PCA (see above for details). If variables used as input for PCA have substantially different shapes of distributions, a correlation matrix instead of a covariance matrix should be used for calculations.

Normalising transformation should not be confused with dividing elemental concentration by a reference, such as by PAAS (Australian Post-Archean Shale; Piper & Bau, 2013) for rare Earth elements, ‘conservative’ elements, or TOC (total organic carbon). This operation, confusingly, is also referred to as ‘normalisation’. The decision whether to apply this type of normalisation is a matter of the research question and not a matter of fulfilling the requirements of a statistical method. Either original variable or variable modified using this type of normalisation may be used in ordination analysis, but the incorporation of both at once should be avoided, as it will introduce spurious correlations. Spurious correlation may be also produced by division of one variable by another one (Van der Weijden, 2002) and the need to perform it, as well as the effect it has on the analysis should be considered individually, depending on the purpose of the analysis.

Standardisation denotes any method of changing the absolute position of a variable's distribution without

changing its shape. This may include translation, that is ‘shifting’ so that the mean equals zero, or expansion/contraction (multiplying/dividing by a factor) so that the range of values falls into the $[-1,1]$ range. ‘Standardisation’ is sometimes taken to mean only a specific case of general standardisation, that is standardising to ‘z-scores’, so that the variable's mean equals zero and its variance equals one. The choice of standardisation method is dictated primarily by the statistical method and not by the type of variable, for example PCA based on a correlation matrix typically does not require any standardisation, but some similarity indices used for NMDS may require standardisation and then the choice of standardisation method should follow the recommendations for the specific similarity index (see Jost et al., 2011; Legendre & Legendre, 2012).

3 | LITERATURE SURVEY OF ORDINATION USE IN GEOSCIENCES

3.1 | Approach and methods

3.1.1 | Data compilation

In order to evaluate the use of ordination methods, sedimentological, geochemical and general geoscience journals were searched using keywords “principal component analysis”, “PCA”, “non-metric multidimensional scaling”, “NMDS”, “detrended correspondence analysis”, “DCA”, and “ordination”. Several queries were carried out using the Publish or Perish software v7 (Harzing, 2019) for each term in selected journals in the field, followed by manual review

of individual manuscripts. Google Scholar was elected as the primary search venue due to its more inclusive definitions, the fact that it is not restricted by any pay wall and is a common resource used by many researchers. This resulted in partial export of the journal titles which had to be checked manually in several cases. The initial search yielded 5,710 results from 23 journals for all search terms. This was then cleaned to remove repetitions, yielding 4,681 records (tables S1 and S2 in Bialik et al., 2021) Numbers of articles published per year in each journal were obtained from Scimago Journal and Country Rank (<https://www.scimagojr.com/>). The time range was limited to focus only on the period of proliferation of personal computing and introduction of widespread communication between computers ('social computing', Sharma et al., 2016). From the narrower range of years, 117 were selected manually by the authors in no particular order (semi-randomly) from the list generated by the initial search with supplementary searches to explore additional journals from adjacent fields. In this manual phase, specific attention was given to newer studies (published in 2020) to evaluate the current state of use. An attempt was made to include no more than one article from the same author in order to obtain a wider spectrum of methodologies and minimise biases. Additionally, 100 manuscripts were selected fully randomly from the initial query using a random number generator. Each manuscript was manually reviewed to identify the methods used, workflow and software, type of data, what for and how the ordination methods were used, the number of data points, groups and variables for which the analysis was carried out and how the data were presented and curated (Table S3, Bialik et al., 2021). This compilation includes only studies where PCA, NMDS or DCA were used, all others were omitted. Primarily, this search focused on articles where fossils or sediment were analysed or that used methodologies that would also be relevant to the analysis thereof. Manuscripts included in the final survey were limited to only those implementing the methodology (i.e. not just mentioned, and no review papers). Studies where ordination methods were part of a data processing workflow (as is sometimes the case for X-ray absorption near edge structure (XANES), several palaeomagnetic and image processing methods among many others) and not the data analysis were not included. In total, 163 studies were included in the final survey with 174 individual analyses listed collected from 43 different journals in the fields of palaeontology, palaeoenvironment, sedimentology and general geology. Metadata for all the examined manuscripts included in the database is provided in the associated data repository (Table S2, Bialik et al., 2021). Out of respect for colleagues in the field, the database was anonymised and, in the following, only positive examples were named. Names and DOIs of the articles are listed separately and have been randomised.

3.1.2 | Data analysis

All analyses have been performed using R Software (R Core Team, 2020) and visualised using the packages *ggplot2* (Wickham, 2016) and *ggbiplot* (Vu, 2011). The R code and further analyses and figures are available in a human-readable RMarkdown format as S4 in Bialik et al. (2021).

3.2 | Results—Findings from the survey of published ordination analyses

The number of publications utilising ordination shows a consistent increase since the beginning of the 1990s, this pattern is visible both in the analysis of the total number of mentions and in the random selection sub-sample (Figure 2).

Three main uses for the ordination methods are identified in the data set (Figure 3): Clustering and differentiation between groups (42%), assessing the relationship between variables (33%) and generating indexes (21%), in a few cases the authors set out to do more than one of these in the same study. Additionally, a smaller number of analyses set out to use ordination to identify most significant parameters ($n = 5$), test the relationship between one parameter of interest to the other variables ($n = 2$) and assign to pre-defined groups ($n = 1$). In the sampled studies, ordination was mostly commonly applied to fossil assemblages (37%) and elemental chemistry (31%), with other types of data including grain type (8%), physical properties (7%), mineralogy (4%), organic molecules (3%), isotopes (3%) and others (Figure 4). Some 18% of the analysis reviewed used more than one type of data. Most of the data sets (65%) were compositional and another 17% were mixed. This wide swath of data types illustrates the versatility and power of ordination in geosciences, particularly for work with sedimentary material.

Of the three methods investigated (PCA, MNDS and DCA), the most ubiquitous is PCA, accounting for 84% of all results found in the initial search (Table S2, Bialik et al., 2021). The other two methods each account for around 12%

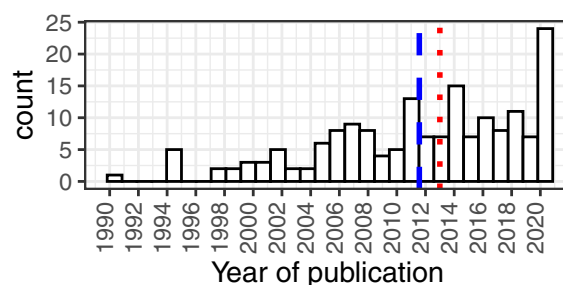


FIGURE 2 Distribution of geoscience articles mentioning keywords used in this survey (see Methods for the list of keywords) through years ($n = 174$). The blue dashed line indicates the mean year (2011) and the red dotted line the median (2013)

FIGURE 3 Distribution of objectives for ordination analyses used in this survey ($n = 174$)

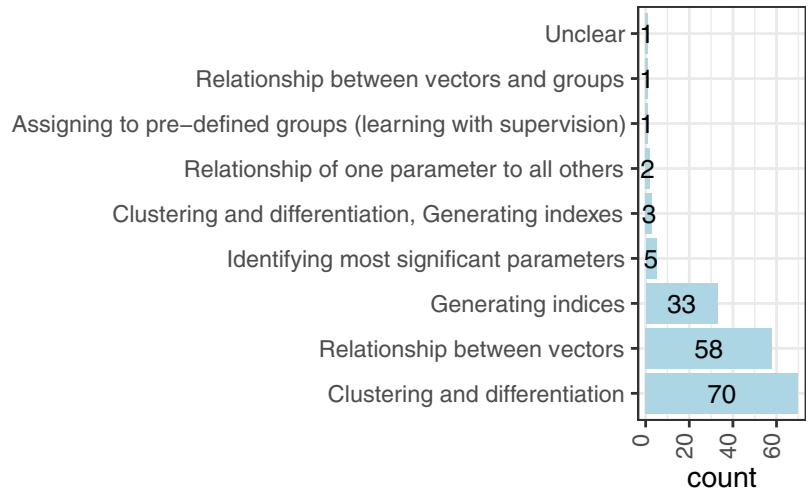


FIGURE 4 Types of data in the survey ($n = 174$)

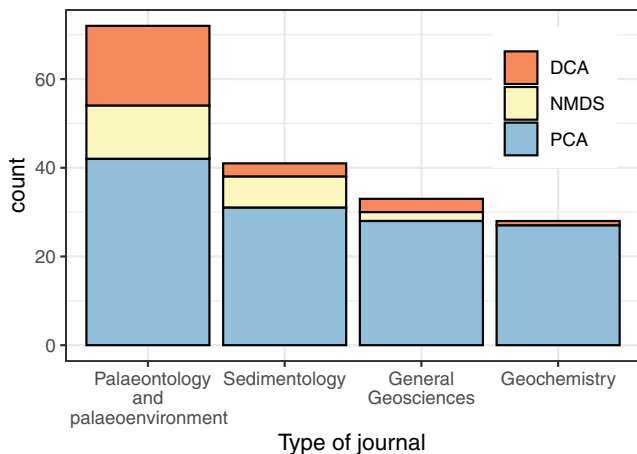
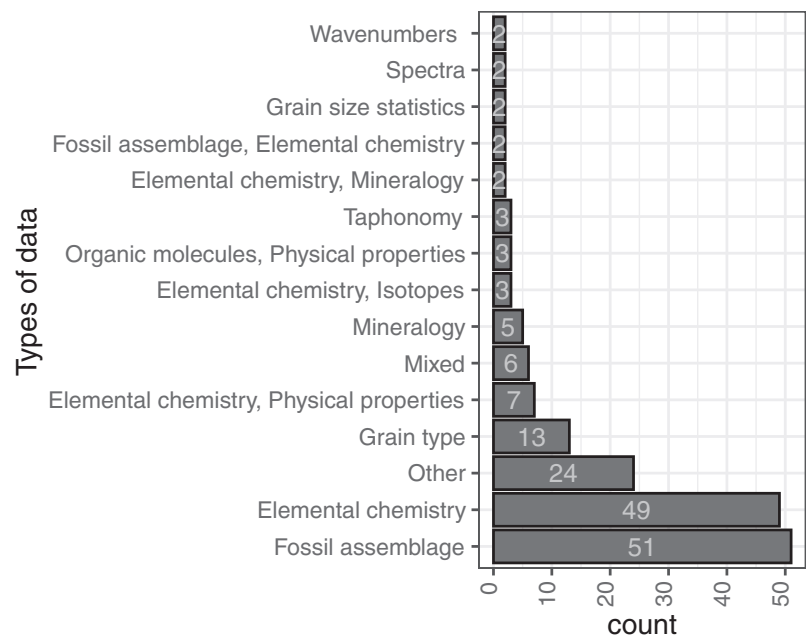


FIGURE 5 Frequency of ordination types across journal disciplines ($n = 174$)

of the references. In journals oriented towards palaeontology and palaeoenvironment NMDS and DCA were encountered more often (16% and 22% of references, respectively) whereas in sedimentological, geochemical and general geoscience journals, PCA exhibited the highest proportion (>90%, Figure 5). Interestingly, in palaeontology and palaeoenvironment journals, DCA was encountered more often than NMDS, whereas in all other groups this method was encountered in fewer than 1% of all references, and completely absent from some journals. The average (\pm standard deviation) number of articles per journal noting these ordination methods was 50 ± 31 for the sedimentological journals ($n = 6$), 270 ± 212 for palaeontology and palaeoenvironment ($n = 7$), 262 ± 124 for geochemistry ($n = 4$), and 140 ± 173 for general geosciences ($n = 6$). Use of multiple ordination methods, an idea promoted in other disciplines (van Son & Halvorsen,

2014), was rarely observed in this database and searches, even when specifically queried. The use of multiple methods was mainly for analysis of ecological assemblages and applied to the same sample set (Abdelhady & Fürsich, 2014; Tyler & Kowalewski, 2014; Zuschin et al., 2007), with the common pairing being PCA and NMDS. One exception to this was Lanci et al. (2001) which used different methods for different data sets, although the reasoning was not explained in the text. Of 128 uses of PCA, only five noted that they have tested for normality.

3.2.1 | Documentation of the analyses

Only 84% of all analyses provided complete information on the dimensions of the data set. Among those, 24 analyses used data sets where the number of variables was larger than the number of observations (Figure 6). Among the 20 analyses using NMDS, nine reported the similarity index used, with Bray–Curtis being the most common. Lack of information about software used for analysis was the most common case (38%, Figure 7), followed by PAST (14%) and R Software (10%). Among analyses which did report the software used, none of the 42 performed using major coding-based packages (i.e. R, Matlab, SAS or SPSS), made the code available.

3.2.2 | Data handling

The majority (53%) of surveyed studies did not provide the data used for analysis at all (Figure 8). The second most common approach (35%) was providing the data set directly in the article text or in the supplement, but in formats such as PDF or DOC, which are not readily imported into statistical

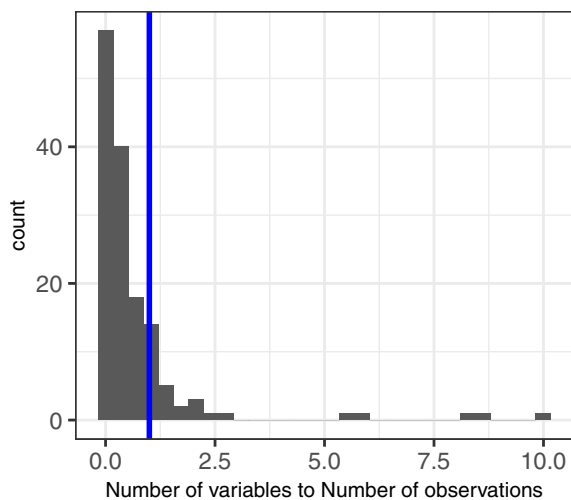


FIGURE 6 Distribution of the ratio of the number of variables to the number of observations in analysed data sets ($n = 146$). The blue line indicates the threshold value of 1

software. The ‘golden standard’ of placing a curated data set in an open access repository was followed only by six (3%) of the analyses (Figure 9). Throughout the entire time series, in each year, no data or data that are not readily processed (score 1) constituted more than a half of surveyed analyses. Only in the last two years (2019 and 2020) did the average score rise above 1. Except for sedimentological journals (49%, $n = 41$), in all other categories articles with no data (score = 0) formed the majority (Figure 10). Articles with data in repositories were noted only in journals in general geosciences and in palaeontology and palaeoenvironment, but no significant differences could be detected in the data handling between journal types ($p = 0.64$, $n = 174$, Fisher's exact test).

3.3 | Discussion

3.3.1 | Trends and patterns in the surveyed published ordination analyses

The use of ordination in the subdisciplines of Earth sciences investigated here is very diverse. Despite reviewing only three types of ordination types, a fair variety of input, application, methodology and objective was still noted.

There is a wide distribution of types of data. Most of the data types encountered are compositional, but non-compositional and mixed data types are not uncommon. It is expected that as more research will use a multiproxy approach, the fraction of the mixed data will further increase. This illustrates the need for diverse methodologies given the different nature of the type of data used and transparency about the method application, as different data types require different approaches.

Principal component analysis is not only the most commonly used method, it is also the one longest in use with the earliest reference from the late 1960s (Briggs, 1965; Read & Dean, 1968) whereas other methods were not encountered earlier than the late 1980s (Table S2, Bialik et al., 2021). While all three methods (or their precursors) were already discussed in the 1960s (Whittaker & Gauch, 1978), NMDS and DCA are more computationally demanding, which was a significant consideration at the time, leading to favouring PCA early on. This early adaptation probably had a significant role in cementing the popularity of PCA.

The higher numbers of papers referencing ordination methods over time could be the result of the increase in total number of publications, owing to increased access to publication and the addition of new journals (Bornmann & Mutz, 2015; Steen et al., 2013). However, individually examined examples in the subset of journals analysed do not seem to support this interpretation, as the increase is also observed at the single journal level, without a corresponding increase in publication rate. For example, the number of articles

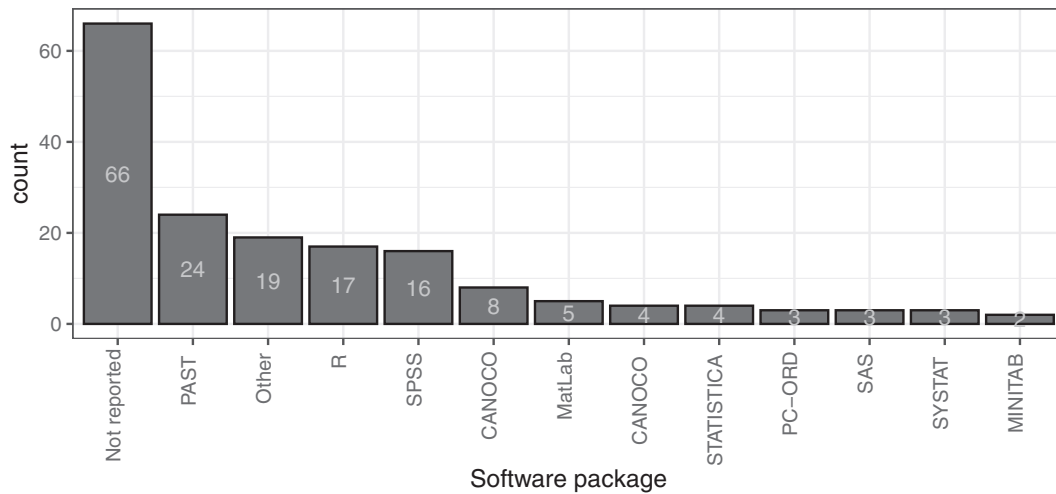


FIGURE 7 Distribution of software packages reported by authors in the survey ($n = 174$)

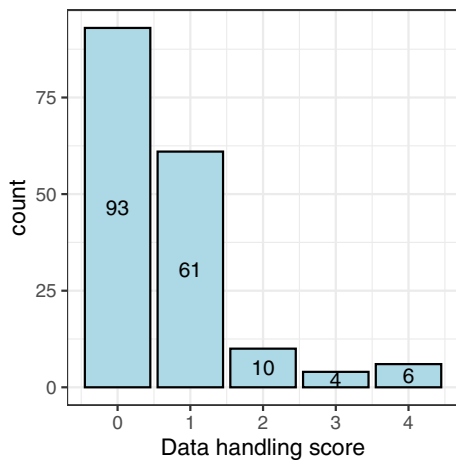


FIGURE 8 Data handling score distribution ($n = 174$). 0—no data; 1—data in article text or in a non-machine-readable format in the supplement; 2—machine-readable format in a supplement; 3—2, but no dead cells; 4—data curated in a repository

published per year by *Palaeogeography*, *Palaeoclimatology*, *Palaeoecology* increased 300% over the past 20 years (from 444 in 1999 to 1,334 in 2019), whereas the total number of papers found in this survey referencing PCA had increased by an order of magnitude (from three in 2000 to 34 in 2020, Table 2). The *Journal of Sedimentary Research* has seen a decline in publication volume over the comparable period of around 40% but still exhibited an increase in mentions for PCA from between 0 and 1 per year in the late 1990s and early 2000s to 0–5 in the late 2010s. These trends indicate an overall increase of awareness and use of ordination methods. Looking at the distribution of mentions per journal (Table S2, Bialik et al., 2021) in each field, and considering that the time span covered for all of these journals was rather similar (*ca* 38 years), even with the very large standard deviations it is clear there is still a lower inclination to use these methods in sedimentary research compared to other Earth science fields.

3.3.2 | Common mistakes and issues

Basic metadata

The most common problem encountered in this review was poor documentation of the workflow used by the authors. Out of the 174 analyses evaluated (Table S2, Bialik et al., 2021), 66 do not include information on the software used. The most commonly used statistical programs were PAST (Hammer et al., 2001), R (R Core Team, 2020), SPSS (IBM Corp., 2017) and CANOCO (ter Braak, 1989; ter Braak & Smilauer, 2012), in that order (Figure 6). Matlab (Matlab, 2020) and Statistica (TIBCO Software, 2018) were also reported along in-house, niche or non-statistical specific programs (e.g. ArcGIS). The preference for PAST and R Software, both freely available, as well as non-statistical software, such as Matlab (which many institutions acquire via bulk license), suggests that accessibility plays an important role in determining software usage. The majority of authors provided no information on data pre-treatment (if at all), how the analysis was carried out, or if the process was iterative or not, 16 did not even report the number of data points. This withholding of information made the evaluation of the validity of the ordination or its replicability impossible. Moreover, of the reviewed manuscripts that used code-based environments (such as Matlab or R), none included their code in the supplement or provided it through a repository. This later part will change in the future as journals are adopting new transparency standards. Journals like *Palaeoceanography* and *Palaeoclimatology* adhere to the FAIR (Findability, Accessibility, Interoperability and Reusability; sensu Wilkinson et al., 2016) principles with regard to data and require all data presented in the article to be available via a data repository. Other journals, like *PeerJ*, now also require the code to be included in the supplement or external repository. Another minor reporting problem encountered was a mismatch between the reported use of ordination and their actual use. For example, the authors might

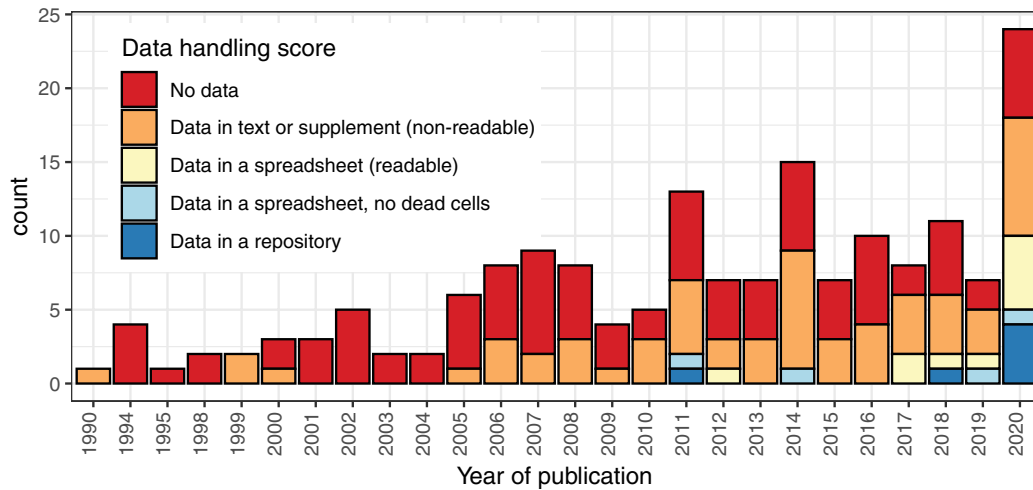


FIGURE 9 Distribution of data handling scores across types of journals ($n = 174$). For legend, see Figure 8

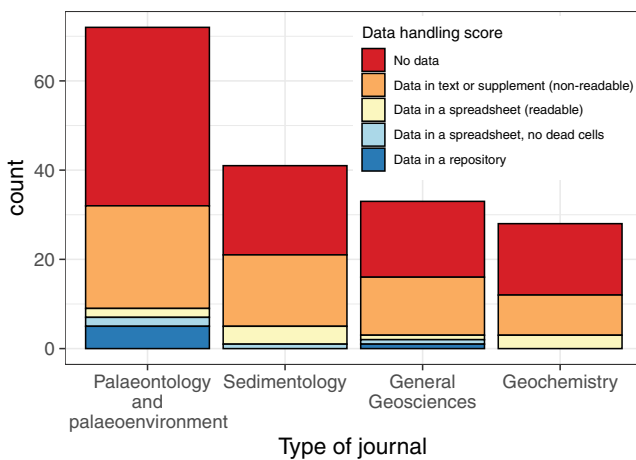


FIGURE 10 Frequency of data handling scores across journal disciplines ($n = 174$). For legend, see Figure 8

write in the methods that they used the ordination method to group the results, but actually used it to examine the relationships between variables.

In some cases, where the data set was provided, it was not annotated or clearly marked, for example variables have mysterious names, the supplement is a folder of loose spreadsheets with cryptic names, etc. The use of data archives with established metadata structures is suggested, for example PANGAEA. Although data publication in PANGAEA is relatively new and none of the studies in this survey used it, it has several advantages over repositories that are not curated such as Dryad: data curation assures that information is automatically exchanged with other databases, for example any biodiversity records are automatically transferred to Global Biodiversity Facility (GBIF) and the Ocean Biogeographic Information System (OBIS). This applies to many inventories, increasing the chances that the data set will be found and re-used. This is further facilitated by automatic registration of PANGAEA data sets in major

scholarly databases and search engines, such as ORCID and Google Search.

Only a handful of the data sets in surveyed analyses were fully compliant with the FAIR data principles (Wilkinson et al., 2016), that is it was possible to find them through a database search, identify the data structure and reuse them based on the metadata and the license. These positive examples included Watkinson and Hall (2019) and four others. The use of a repository addresses several issues. It offers an external quality control on the arrangement of the data, it makes it discoverable and available to the rest of the scientific community, and it does not lock it behind a paywall. The need for quality control is illustrated by the fact that in some studies the data set does not match what is stated about the analysis. For example, the analysis has been produced on a cleaned data set with dead cells, such as values below detection limit, removed, but only the raw, not cleaned data set is provided. Data sets provided in the PDF format directly in the article (Bialik et al., 2012; Jarochovska, 2012) or in the supplement are not strictly machine-readable, as exporting into an editable format typically introduces mistakes and requires extensive cleaning of formatting.

Dimensions

Another issue encountered with the misuse of ordination was the number of variables being larger than the number of data points. This was encountered in 25 analyses, in addition to another 26 in which either the number of data points or variables was unknown. Since there are only $n-1$ degrees of freedom for n variables, the total number of variables should not exceed $n-1$ (Legendre & Legendre, 2012). Analysis of the impact of the ratio between the number of variables to the number of samples found the optimal ratio should be at least two, with some going up to six (Cattell, 1978; Kline, 1979). Most studies found the ratio of three samples per variable is optimal (Björklund, 2019; Shaikat et al., 2016), but 41% of

TABLE 2 Examples of compositional and non-compositional data types

Compositional	Non-compositional
Concentrations (% ,‰, ppm, M, m, etc.)	Instrumental raw counts
Fraction of area	Absolute area units (e.g. m ²)
No. of counts when total sum is constant	Morphometric measurements (length, angle, diameter, No. of warts, etc.)
Relative abundances	

the studies analysed here that used PCA had lower ratios. It is not uncommon that, in multiproxy studies, multiple analyses are carried out on the samples and, at times, this leads to the analyses producing a greater number of variables than there are samples available. This is especially true for studies where the number of samples is limited or applications where the pre-treatment is especially long and complex, such as organic geochemistry. Analyses with more variables than samples could have been avoided in some studies. For example, in studies where the number of variables was increased by the authors including the measurements and ratios between said measurements as separate variables.

Principal component analysis

In this review, most of the issues found were with the use and application of PCA. As stated above, normality, or at least symmetry of the distribution, is not needed for NMDS or DCA, but it is important when using PCA. Yet very few studies ($n = 5$) clearly reported having tested for normality of individual variables (Allafta & Opp, 2020; Klubi et al., 2018) or multivariate normality of the data set (Abdulla et al., 2013). The issue is further complicated when closed sum (compositional) data are involved. Here it was observed that most sedimentological and palaeoenvironmental studies sampled had compositional or a mixture of compositional and non-compositional data. These data sets have variables that are not independent of each other, as they are part of a sum of a constant (Aitchison, 1982). Examples of this (Table 2) would be percent, which describe a very large fraction of all sedimentological, geochemical and palaeoecological data sets. True central moments of compositional data are not straightforward Euclidean geometrical products, such as mean and standard deviations (Filzmoser et al., 2009; Tolosana-Delgado, 2012). These issues can be addressed using transformations (Auer et al., 2019; Caron et al., 2020; Dunkley Jones et al., 2008), but these were rarely implemented. Furthermore, in some data sets the distribution of variables could not be made multinormal even if these transformations would be applied, such as mixtures of closed and open data sets. A non-parametric approach would have been preferable in those cases, rather than those making assumptions about the distribution of variables (here referred to as quasi-parametric). However, parametric or quasi-parametric methods are those which are usually implemented. It should

be noted that not all geochemical or sedimentological data sets are necessarily closed. Raw X-ray fluorescence measurements, for example, reported in counts per second, are not part of a sum of a constant and could be considered as not compositional.

Non-metric multidimensional scaling

With NMDS, the most common issue was the absence of reporting on the distance matrix and the methods used to generate it. This is of particular importance with respect to the data type analysed as different distance indexes will define the stress value. However, some distance methods, such as Gower, will be more informative with mixed data sets (van de Velden et al., 2019). Another issue was the use of the NMDS axes as indices. This is not wrong per se, information about the separation between the data points is present in the NMDS output. Unlike PCA, where the axes represent scaled linear combinations of the variables, NMDS axes in default implementations are more qualitative and not measured on a ratio scale. However, some implementations allow scaling of NMDS axes as 'half-change', for example the function `postMDS` in the `vegan` package (Oksanen et al., 2019) centres and scales the axes so that one unit means halving the similarity (Jarochowska et al., 2017). To substantiate interpretation of axes, their scores can be tested for correlation with variables (Tyler & Kowalewski, 2014). Moreover, the choice of an implementation can significantly impact the replicability of the axial direction. The `metaMDS` implementation in the `vegan` package for R Software, testing multiple starting configurations, which can be made reproducible by initiating the random number generator by a fixed seed (Oksanen et al., 2019) can mitigate these issues, but information about these steps is often not specified in methods.

Hypothesis testing

In many cases, the output of the ordination is not substantiated or tested, notably when relationships between variables are concerned, but also when clustering. While some studies did combine ordination with other Euclidean and non-Euclidean clustering methods, and then performed a statistical test to validate that the resulting groups are significantly different (Gosling et al., 2019; More et al., 2018), others did not. Furthermore, some circular reasoning was encountered with a priori groups preassigned to the data

and then treating the ordination clusters as validation. This last is fundamentally erroneous as ordination is a data exploration tool set and not a tool to validate dissimilarity. An intermediate state is present with a priori groups where the researchers seek to understand the relationship of the groups they predetermined to the n -dimensional variable space. For the research question in this situation, it does not matter if the groups are dissimilar, just how they relate to each other with respect to the variables. But the a priori assignment can result in groups having no relationship to the clustering observed in the ordination—which undermines this application.

Graphical presentation

A persistent issue in many publications reviewed was the presentation of the ordination. One common issue was not presenting the ordination outputs, but rather only a partial presentation of the resulting outcome. These cases often do not show any graphical presentation of the ordination output and auxiliary information such as the scree plot in PCA is not shown. This was very common in papers where ordination was used to generate indices, in these it was common for elements of the ordination not to be shown and only the index value to be presented. In other cases, only the loading was reported in a table or variables were shown in the ordination space without the data points. This sort of presentation limits the possibility to evaluate how the variables interact with the data set. The opposite issue was also encountered, where ordination presented with just data points but without the variables/biplot (for PCA and DCA). Presentation of ancillary information, such as the scree plot for PCA or Shepard plot for NMDS was rarely encountered in the main text or supplement. More often, the authors would only report the total variance (for PCA) for an axis, whereas these plots can inform the reader of the level of importance of each component, including the ones not shown.

Another issue observed with graphical presentation, encountered less but still present, drawing of ‘blobs’ around groups arbitrarily, according to authors’ own preferences. This does not offer a reproducible and objective evaluation of separation. A better option would have been to draw convex hulls (Arreguín-Rodríguez & Alegret, 2016; Tomašových, 2004) if the groups are known. When they are not known, clustering/machine learning without supervision for unknown groups (Bertolini et al., 2020; Höltnke et al., 2016) would be a preferred way of identifying the groups.

We also encourage considering common colour blindness in data presentation. An example of how to maintain a consistent colour scheme legible to most colour-blind readers using RColorBrewer (Neuwirth, 2014) is offered in the proposed workflow illustrated in S5 (Bialik et al., 2021).

4 | PROPOSED WORKFLOW

4.1 | Performance of ordination methods in a sedimentological—case study

To illustrate a workflow fitting a typical sedimentological analysis, a data set from Bialik et al. (2018) was used, available as S4 in Bialik et al. (2021). This data set included geochemical and sedimentological information from a section carried out on an Albian carbonate sequence. A summary of analyses is presented below, but codes and full results of each step are provided in Bialik et al. (2021). The initial data set consisted of 90 observations and 17 variables. Of these, eight were compositional (concentrations given in ppm or percent, both of elements and mineralogies), another three were ratios of compositional variables, five were independent ratios considered non-compositional (isotopic data) and two were classifications (lithology and texture). Additionally, the isotopic data also included error values. This data set was selected for its complexity to illustrate a sort of worst-case scenario. The test set is composed of a mixture of open and closed sets and the variables are not normally distributed. Types of variables and approaches to handling them are discussed in Section 4.2. The analysis was carried out in R Software (R Core Team, 2020), using routines which are also available in other popular softwares such as Matlab or PAST.

Variables and observations with empty cells were identified using the package *phreatmap* (Kolde, 2019). As default implementations of ordination analyses cannot handle empty cells, the decision was made to exclude two variables with the highest number of empty cells: $\delta^{25}\text{Mg}$ (‰ DSM3) and inorganic carbon (IC, %wt). This left a 61×12 matrix, which was used for further analysis. Two variables were defined on a categorical scale (lithological description and packing according to Embry & Klovan, 1971). They were excluded from ordination and used as descriptors of samples.

4.1.1 | Principal component analysis

To fulfill the assumptions of PCA, tests for multivariate normality were conducted using the *MVN* package (Korkmaz et al., 2014), including Mardia's, Henze-Zirkler and Royston's tests. The normality of individual variables was additionally tested using the Shapiro–Wilk test. Variables with extreme positive skewness were transformed as follows: Calcite (%) and Ni/Co using cubic root transformation; Sr (ppm), Zr (ppm) and Mn (ppm)—using square root transformation. Transformation was chosen iteratively by measuring the resulting skewness, as skewness, or deviance from symmetry, is the parameter of the distribution that is the most confounding for PCA (Legendre & Legendre, 2012). The PCA was

performed using the *princomp* function of R Software and the correlation matrix, because the variables in the data set were mixed (i.e. measured in different units). For comparison, output from analysis carried out on untransformed variables has been added in the supporting information. It is not possible to give a rule of thumb on how the ordination will change prior and post-transformation, but typically a lower proportion of the entire data set will be represented by PC1 and PC2, as all principal components are linear combinations of the variables.

4.1.2 | Non-metric multidimensional scaling

As NMDS does not require normally or symmetrically distributed variables, the original, un-transformed data set as input for the *metaMDS* function in the *vegan* package (Oksanen et al., 2019) was used. This implementation differs from the original NMDS algorithm in that it tries to find a stable solution using several random starts and standardises the scaling in the result. The random number generator was seeded at a fixed number to assure that the same results are obtained every time, otherwise each new run of the NMDS would result in a different ordination or flipping of the axes.

4.1.3 | Detrended correspondence analysis

Detrended correspondence analysis was performed on the data set after normalising transformation (see PCA above). Additionally, the variables were standardised to the same range [0, 1], but without standardising their variance. The *decorana* function of the *vegan* package (Oksanen et al., 2019) was used.

4.1.4 | Results of the case study

For the initial data set, multivariate normality, as well as normality of individual variables, was rejected at $\alpha = 0.05$. Transformations of the most skewed variables did not improve the multivariate distribution and the transformed data set still failed a test for multinormality, but univariate normality could not be rejected for the transformed Zr content (test statistic 0.9839, $p = 0.6044$ compared to 0.9048 and $p = 0.0002$ prior to transformation, $n = 61$). The transformations reduced the skewness of the univariate distributions, but the non-linear relationships between variables (table S5 in Bialik et al., 2021) is a warning that PCA may not be suitable for this data set.

A PCA ordination plot is shown in Figure 11A. The same plot with observations labelled by the Embry and Klovan (1971) classification, as well as the scree plot and

a visualisation of loadings, is available in S6 (Bialik et al., 2021). PC1 explained only 31.7% of the total variance, its highest loadings were $^{87}\text{Sr}/^{86}\text{Sr}$ Initial Value (0.36) and %calcite (loading 0.34) and lowest—Ce/Ce* (−0.42) and Zr content (−0.40). These pairs of variables defined therefore the largest proportion of variance in the data set, with limestone samples grouping at high values of PC1 and low values of PC1 corresponding to a mixture of dolomite, marly dolomite and dolomitic marl samples. PC2 explained 19.1% of the total variance and was most influenced by Sr content (loading 0.41) and MgCO_3 content in dolomite (loading −0.51).

The NMDS yielded an ordination with a stress value of 0.154 (Figure 11B). Squared correlation between fitted values and ordination distances exhibit $R^2 = 0.905$, indicating a good representation of the distances between samples. Typically, only observations are plotted in NMDS and DCA as these methods were designed to ordinate community matrices, but it is possible to obtain variable scores (table S5 in Bialik et al., 2021). The NMDS axis 1 corresponds to a gradient between limestone samples (high values of $\delta^{13}\text{C}$) and marly dolomite samples (high $\delta^{18}\text{O}$ values), with dolomite and marly limestone occupying intermediate positions along the gradient. High values of NMDS axis 2 corresponded to high Ni/Co and V/Cr content, represented by some dolomite samples. Low values of the NMDS axis 2 corresponded to high Mn content.

The DCA (Figure 11C) resolved a similar gradient as NMDS, with limestone and marly dolomite defining the axis of the largest variance. Note that the gradient is flipped between Figure 11B and C, as the sign of the ordination axis has no meaning. In this analysis, DCA and NMDS cross-validate each other. It is an illustration of a case where PCA would not perform as well as NMDS or DCA, as initially identified by the non-linear relationships between variables and by the high skewness of individual variables.

4.2 | Consideration in compiling a multivariate data set—Types of variables and how they are coded

Most introductory textbooks on multivariate data analysis implicitly assume variables are continuous, defined on a ratio scale and have no missing records or bounds. The best-known parametric methods are also designed for such variables. But in geosciences, deviations from these assumptions are plentiful.

Variables can be defined on four scales, sometimes called levels of measurement: nominal, ordinal, interval and ratio variables. Nominal variables are categories which have no particular order. They are common in sedimentology and include for example rock types which do not represent any particular gradient (e.g. sedimentary, igneous and metamorphic

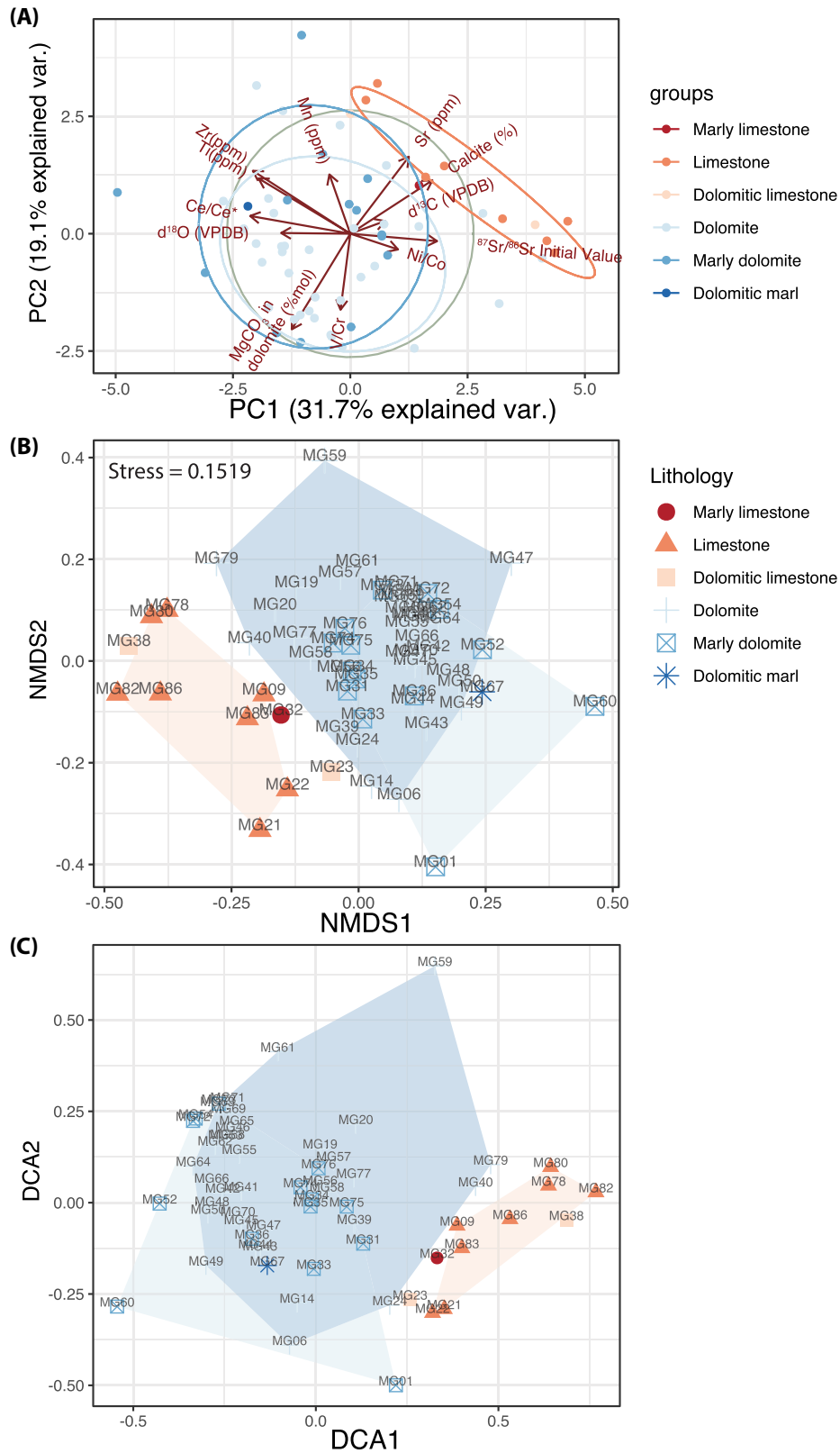


FIGURE 11 Comparison of ordination results applied to the case study data set from Bialik et al. (2018), with observations grouped by lithology (61 observations, 12 variables). (A) Principal component analysis with ellipses marking 68% confidence interval based on a multivariate t distribution. (B) Non-Metric Multidimensional Scaling with convex hulls. (C) Detrended correspondence analysis with convex hulls

rocks). That is not to say there are no categories of rocks that have a natural order such as carbonate mudstone, wackestone and packstone according to the proportion of skeletal

components. Such variables are defined on an ordinal scale, but it is not possible to measure the distance between them, for example it is not possible to say that packstone always

has twice as many components as wackestone. A mean or median calculated from a variable assigning samples to various categories in Dunham (1962) classification would be meaningless.

Interval variables are ones whose values are measured along a scale, that is values are equidistant. Their values can be added and subtracted and means and medians calculated from them are meaningful. But they do not have a true zero. The best-known example is temperature measured in °C, but in sedimentology perhaps the most common case are isotope ratios such as $\delta^{13}\text{C}$. If the variable can take negative values, it does not have a 'true' zero. This is important, for example for PCA, which cannot handle negative values, because a variable defined on an interval scale can be standardised (here by moving, i.e. translating, the distribution to positive values without changing its shape) to meet PCA requirements. Finally, values which can be added, subtracted and have a 'true' zero are called ratio variables and include concentrations of elements or grain sizes.

Recognising the scale at which a variable is measured is important to 'code' it properly, that is indicate the order and distance of categories in categorical variables. Coding refers here to assigning numerical values to categories so that they can be processed by an ordination method. For example, there are at least three different ways of coding grain size recorded in the categories on the Wentworth (1922) scale: using phi (\log_2 of the diameter) would make them almost equidistant (nearly interval), whereas assigning its middle value in metric units to each category would produce a variable measured on an ordinal scale. This would produce different distances between samples in NMDS, depending also on the similarity index used. How the weights are assigned to variables may determine the conclusions of a study (Peng, 2015).

Some software packages allow the type of variable to be defined, which determines how it is processed by the ordination method. For example, categorical variables in R Software are stored as factors. These factors can be ordered, which corresponds to ordinal variables, or not ordered, which corresponds to nominal variables. A common issue is that the type of variable is not recognised correctly, for example because of a typographical error, and it can affect how the variable is handled by ordination. It is recommended that types of variables are defined explicitly, especially when types are mixed within the data set (see workflow in Section 4.1). This allows problems caused by visual spreadsheets to be avoided, which otherwise may introduce errors by attempting to identify the types of variables automatically (Ziemann et al., 2016). This automatic recognition is also a great hindrance to reproducibility, since the same variable may be recognised as a different type depending on the version and language settings of a particular computer.

A related problem is correct coding of zeros and missing values. The ordination analyses described here, in their basic

implementations, cannot handle missing values. There are tools allowing imputation of missing values. Some of them are specifically designed to assist the ordination algorithm (Filzmoser et al., 2018; Stacklies et al., 2007; Zhu et al., 2019), some may be specifically designed to impute missing data of a given format. For example, software packages for acquiring a diffraction signal will typically include imputation of single missing data points, and such specialised algorithms are likely to perform better than software for general use with any type of data. Compositional data in particular, due to their properties, lend themselves to missing value imputation (Hron et al., 2010; Palarea-Albaladejo & Martín-Fernández, 2015). A very common error is to code missing values as zeros. Zeros are not 'visible' in descriptive statistics such as the mean and standard deviation and therefore it often goes undetected that they do contribute to the distribution of a variable. A none-too-rare example would be geochemical analysis where concentrations are below the detection limit. Should these be coded as zeros, their distribution will artificially deviate from the normal distribution, which they might otherwise follow. If the software allows this, missing cells and zeros should be clearly distinguished. Some environments, such as R Software, can handle not applicable/available (N/A) statements.

Values below the detection limit are a common example of censored data, where a part of the distribution is not known. The sole coding of them (often not numeric e.g. '<0.01', 'LOD') is a common cause of errors in how the variable type is recognised. More importantly, it may be decisive for the outcome of the analysis whether these values are coded as a very small value, zeros or as absent. If very low values of a particular variable are an important characteristic of a set of samples, this information would be lost by replacing these values with empty cells. Furthermore, it would violate the assumptions of algorithms imputing missing values, as they assume that empty cells are randomly distributed. As values outside of detection limits are most common in compositional data, such cases are best treated with dedicated packages which allow them to be handled with less information loss (Filzmoser et al., 2018; Templ et al., 2016 and others cited below). There is no one set of operations to deal with missing cells, as the analysis environment and the nature of the data set may dictate different solutions. If possible, eliminating the rows with the missing cells will be the preferred solution.

4.3 | Workflow recommendation

A researcher interested in using ordination can use the following workflow (Figure 12). The first stage of analysis of data should be identifying what kind of variables it contains (compositional, independent or mixed; nominal, ordinal,

interval and ratio variables), as that will dictate many of the following steps. Selection of variables might be necessary as ideally the number of samples should be three times the number of variables (Shaukat et al., 2016), and under no circumstances equal or larger than the number of samples (Legendre & Legendre, 2012 and Section 3.3.2.2 here). Data sets which do not fulfil this criterion are sometimes referred to as 'wide' and dedicated variations of PCA are available for them to allow exploratory analysis without a priori trimming the variables (Croux et al., 2013; Todorov & Filzmoser, 2013). If the data set contains nominal variables (e.g. facies names, most common minerals), they can either be ranked into an ordinal element or broken down to individual columns and turned into Boolean (presence-absence) variables. If the former is employed, this will impact the data structure down-the-line and must be accounted for. Following that, the variables should be examined individually to identify the shape of distributions (either graphically or by calculating the skewness) and if any transformation might be needed. The variables can be plotted as histograms, but in most Earth science data sets, a spatial/temporal component exists. Plotting the variables along these axes can give a first estimation of type of variability and inform later interpretation of clusters. If the variables appear to have normal distributions, they should be tested for multinormality. If they are clearly not normally distributed, examine the shape of the distribution for pre-treatment in the following stages.

Next, generation of a correlation matrix is recommended. This can be done in any statistical software. Given the non-linear nature of some sedimentological and geochemical data sets, and because many of them are closed sets, a use of a ranked series correlation coefficient such as Spearman's or Kendall's coefficients rather than Pearson's for initial reconnaissance (Tolosana-Delgado, 2012) is recommended. Any pair of variables with a high correlation coefficient should be plotted to evaluate the correlation, p -values are not a reliable indicator with very large (1,000s or more data points) databases.

With an initial expectation of the outcome established, it is then possible to select the ordination method and pre-treat of the data. Principal component analysis relies on detecting linear relationships between variables, since principal components are linear combinations of variables (Minchin, 1987). In contrast, NMDS and DCA perform better when relationships between variables are not linear (such as redox-sensitive trace metals, which typically show logistic responses). As this is not always known a priori, it is recommended that the results of two or more ordination methods are compared (Patzkowsky & Holland, 2012, see also Abdelhady & Fürsich, 2014; Tyler & Kowalewski, 2014). If the variables are measured on a ratio, interval or ordinal scale, are non-compositional and symmetrically (e.g. normally) distributed,

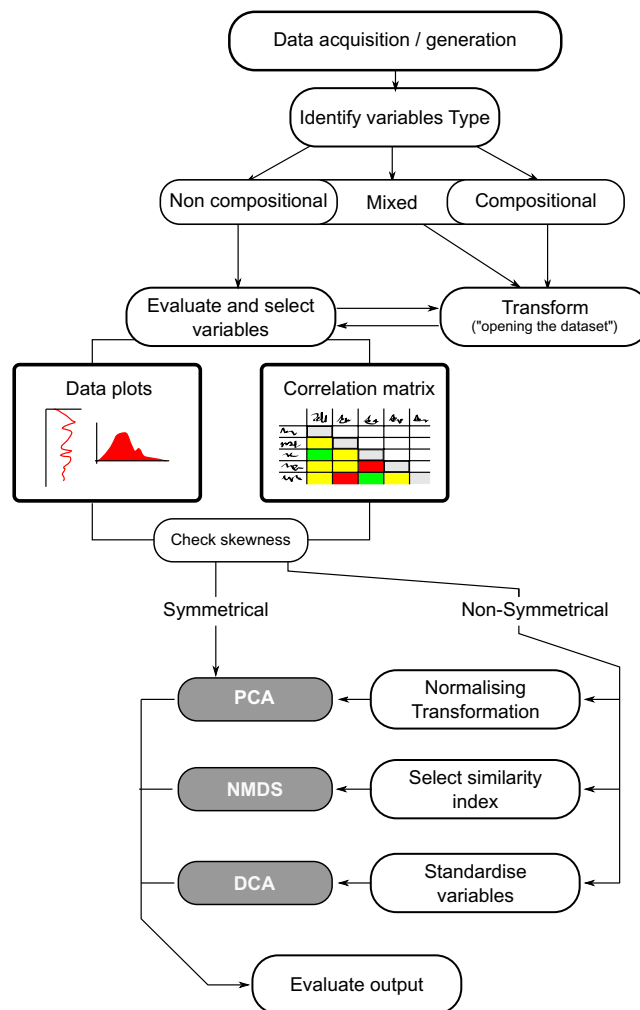


FIGURE 12 Schematic workflow for the use of ordination in sedimentological, sedimentary geochemistry or palaeoenvironmental research

they can be used for PCA as is. Nevertheless, other ordination methods should also be used in addition. If not all variables are symmetrically distributed, further processing is needed. Compositional or mixed variables should be transformed, for example, using a root arcsin or an isometric log-ratio transformation (Filzmoser et al., 2009). After transformation, check the outcome and distribution shape of the data by plotting the histograms or testing for normality. If the result is suboptimal, attempt to use a different type of transformation. With the data optimised, PCA could be performed. After performing PCA, examine the scree plot to see how much of the explained variance is accounted for in each component to select the ones for evaluation.

If the distribution of variables cannot be adjusted sufficiently, or the eigenvalues are very low, a method that does not require symmetrically distributed variables, such as NMDS or DCA, should be employed. Simulations by Minchin (1987) and Patzkowsky and Holland (2012) provided an empirical evaluation of the utility of different ordination

methods for different types of data sets. Patzkowsky and Holland (2012), for example, found that NMDS performed better than DCA when two gradients (NMDS or DCA axes 1 and 2, respectively) represented similar amounts of variation, whereas DCA performed better when one dominant gradient was present. Empirical evaluations can be found in Bush and Brame (2010) and Tyler and Kowalewski (2014), but clearly more are needed, especially for sedimentological data sets. When generating the distance matrix for NMDS, see Section 1.2 for a discussion of the choice of similarity index. As a rule of thumb, if the data type is mixed, use Gower (van de Velden et al., 2019), if the data are Boolean, use Jaccard, but consider how zeros and missing values are coded, as they may influence the results (see Section 4.2 here).

With the ordination performed, its results should be evaluated against the initial expectations from the survey of the data. If they are different, first check if the output makes sense with the data and if the initial expectations were wrong. Also consider if ordination output makes sense from a geological perspective. In either case, it might be prudent to redo the analysis using a different transformation or with some of the variables excluded to evaluate if the results replicate. If the desired output is clustering, perform the appropriate statistical test to make sure the difference between the resulting groups is significant (see Section 2.4).

With the ordination generated, it is important to report the detail of the workflow in the publication. It is strongly recommended to include the graphical representation of the ordination, loading tables (for PCA and DCA), scree plot (PCA), etc. in the supplement if not the main text of the manuscript. As advocates for open science, the call is made here upon authors, when possible, to include the original data, preferably in a data repository.

5 | CONCLUDING REMARKS

Exploratory statistics and ordination in particular are of growing interest within the Earth science community. These methods offer an opportunity to analyse large multivariate data sets, particularly with increased digitization of archival data and increasing data set sizes in modern studies. The proliferation of freeware software and databases in Earth sciences and rising interdisciplinarity in research creates an environment in which Earth scientists can benefit from these tools. Students in these fields, in turn, could benefit from increased training in multivariate statistics. Here, a review is presented of what ordination is, how it could be used and what are some potential pitfalls in its application.

This work presents a survey of a large swath of studies using ordination in a sedimentology related context. A

diverse range of uses and applications was found. Although most of the surveyed analyses used PCA, NMDS and DCA are probably more suitable for most geological data sets. The observations here indicate many small mistakes that could be avoided. Data reporting and access could benefit from new tools and policies which would enhance reproducibility. Based on the finding from this review and survey a workflow (Figure 11) is proposed for researchers new to ordination that are interested in unlocking the potential in their sedimentological, sedimentary geochemistry or palaeoenvironmental data.

For those who seek to deepen their understanding of the topics covered here, there are many excellent reference books (Filzmoser et al., 2018; Greenacre, 2018; Hammer & Harper, 2007; Patzkowsky & Holland, 2012; Reimann et al., 2008; Rencher, 2003; Zar, 2010). All of these are useful introductory books written with geoscientists in mind as well as discussing the use of accessible statistical software, notably PAST (Hammer et al., 2001) and R Software (R Core Team, 2020).

ACKNOWLEDGEMENTS

This work received no dedicated funding. OMB is supported by Marie Skłodowska Curie Fellowship No. 101003394 (RhodoMalta). MG is supported by DFG Priority programme (project DynaTrait, phase 2, 1704). EJ received funding from Deutsche Forschungsgemeinschaft (project Ja 2718/3-1). The first two authors have made equal contribution to this study. The authors would like to thank Rute Coimbra and Daniel Killam for their detailed review and helpful comments, as well as associate editor Benjamin Walter for his handling of this manuscript.

DATA AVAILABILITY STATEMENT

All the data sets, codes and RMarkdown documents are available as Bialik et al. (2021). Link for review: <https://figshare.com/s/de2467546265bd867642>

ORCID

Or M. Bialik  <https://orcid.org/0000-0003-1915-7297>

Emilia Jarochovska  <https://orcid.org/0000-0001-8937-9405>

REFERENCES

- Abdelhady, A.A. & Fürsich, F.T. (2014) Macroinvertebrate palaeo-communities from the Jurassic succession of Gebel Maghara (Sinai, Egypt). *Journal of African Earth Sciences*, 97, 173–193. <https://doi.org/10.1016/j.jafrearsci.2014.04.019>
- Abdulla, H.A.N., Minor, E.C., Dias, R.F. & Hatcher, P.G. (2013) Transformations of the chemical compositions of high molecular weight DOM along a salinity transect: using two dimensional correlation spectroscopy and principal component analysis approaches. *Geochimica et Cosmochimica Acta*, 118, 231–246. <https://doi.org/10.1016/j.gca.2013.03.036>

- Aitchison, J. (1982) The statistical analysis of compositional data. *Journal of the Royal Statistical Society Series B (Methodological)*, 44(2), 139–177.
- Allafta, H. & Opp, C. (2020) Spatio-temporal variability and pollution sources identification of the surface sediments of Shatt Al-Arab River, Southern Iraq. *Scientific Reports*, 10(1), 6979. <https://doi.org/10.1038/s41598-020-63893-w>
- Alotaibi, K., Rayward-Smith, V.J. & de la Iglesia, B. (2011) Non-metric multidimensional scaling for privacy-preserving data clustering. In: Yin, H., Wang, W. & Rayward-Smith, V. (Eds.) *Intelligent data engineering and automated learning*, vol. 6936. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 287–298. https://doi.org/10.1007/978-3-642-23878-9_35
- Amrhein, V., Trafimow, D. & Greenland, S. (2019) Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. *The American Statistician*, 73(sup1), 262–270. <https://doi.org/10.1080/00031305.2018.1543137>
- Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1), 32–46. <https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x>
- Arreguín-Rodríguez, G.J. & Alegret, L. (2016) Deep-sea benthic foraminiferal turnover across early Eocene hyperthermal events at Northeast Atlantic DSDP Site 550. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 451, 62–72. <https://doi.org/10.1016/j.palaeo.2016.03.010>
- Auer, G., De Vleeschouwer, D., Smith, R.A., Bogus, K., Groeneveld, J., Grunert, P., Castañeda, I.S., Petrick, B., Christensen, B., Fulthorpe, C., Gallagher, S.J. & Henderiks, J. (2019) Timing and pacing of Indonesian throughflow restriction and its connection to Late Pliocene climate shifts. *Paleoceanography and Paleoclimatology*, 34(4), 635–657. <https://doi.org/10.1029/2018PA003512>
- Austin, M.P. (1968) An ordination study of a chalk grassland community. *The Journal of Ecology*, 56(3), 739. <https://doi.org/10.2307/2258104>
- Bertolini, G., Marques, J.C., Hartley, A.J., Da-Rosa, A.A.S., Scherer, C.M.S., Basei, M.A.S. & Frantz, J.C. (2020) Controls on Early Cretaceous desert sediment provenance in south-west Gondwana, Botucatu Formation (Brazil and Uruguay). *Sedimentology*, 67(5), 2672–2690. doi: <https://doi.org/10.1111/sed.12715>
- Bialik, O.M., Jarochovska, E. & Grossowicz, M. (2021) Supporting data for: on the application and utility of ordination methods in sedimentology, paleontology and geochemistry. Figshare. <https://doi.org/10.6084/m9.figshare.c.5250993>
- Bialik, O.M., Korngreen, D. & Benjamini, C. (2012) Lithofacies and cyclicity of Mohilla evaporite basins on the rifted margin of the Levant in the Late Triassic, Makhtesh Ramon, southern Israel. *Sedimentology*, 59(7), 2097–2124. <https://doi.org/10.1111/j.1365-3091.2012.01336.x>
- Bialik, O.M., Wang, X., Zhao, S., Waldmann, N.D., Frank, R. & Li, W. (2018) Mg isotope response to dolomitization in hinterland-attached carbonate platforms: outlook of $\delta^{26}\text{Mg}$ as a tracer of basin restriction and seawater Mg/Ca ratio. *Geochimica et Cosmochimica Acta*, 235, 189–207. <https://doi.org/10.1016/j.gca.2018.05.024>
- Birks, H.J.B. (1985) Recent and possible future mathematical developments in quantitative palaeoecology. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 50(1), 107–147. [https://doi.org/10.1016/S0031-0182\(85\)80008-0](https://doi.org/10.1016/S0031-0182(85)80008-0)
- Bitušik, P., Trnková, K., Chamutiová, T., Sochuliaková, L., Stoklasa, J., Pipík, R., Szarłowicz, K., Szacilowski, G., Thomková, K., Šporka, F., Starek, D., Šurka, J., Milovský, R. & Hamerlík, L. (2018) Tracking human impact in a mining landscape using lake sediments: a multi-proxy palaeolimnological study. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 504, 23–33. <https://doi.org/10.1016/j.palaeo.2018.04.021>
- Björklund, M. (2019) Be careful with your principal components. *Evolution*, 73(10), 2151–2158. <https://doi.org/10.1111/evo.13835>
- Bloom, S. (1981) Similarity indices in community studies: Potential pitfalls. *Mar Ecol Prog Ser*, 5, 125–128.
- Bornmann, L. & Mutz, R. (2015) Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66, 2215–2222. <https://doi.org/10.1002/asi.23329>
- Bray, J.R. & Curtis, J.T. (1957) An ordination of the upland forest communities of Southern Wisconsin. *Ecological Monographs*, 27(4), 325–349. <https://doi.org/10.2307/1942268>
- Briggs, L.I. (1965) Heavy mineral correlations and provenances. *Journal of Sedimentary Research*, 35(4), 939–955. <https://doi.org/10.1306/74D713B6-2B21-11D7-8648000102C1865D>
- Brown, C.E. (1998) *Applied multivariate statistics in geohydrology and related sciences*. Berlin, Heidelberg: Springer, Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-80328-4>
- Bush, A.M. & Brame, R.I. (2010) Multiple paleoecological controls on the composition of marine fossil assemblages from the Frasnian (Late Devonian) of Virginia, with a comparison of ordination methods. *Paleobiology*, 36(4), 573–591. <https://doi.org/10.1666/07022.1>
- Buttigieg, P.L. & Ramette, A. (2014) A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiology Ecology*, 90(3), 543–550. <https://doi.org/10.1111/1574-6941.12437>
- Caron, M., Montero-Serrano, J.-C., St-Onge, G. & Rochon, A. (2020) Quantifying provenance and transport pathways of holocene sediments from the Northwestern Greenland Margin. *Paleoceanography and Paleoclimatology*, 35(5). <https://doi.org/10.1029/2019PA003809>
- Cattell, R.B. (1978) *The scientific use of factor analysis in behavioral and life sciences*. Boston, MA: Springer. <https://doi.org/10.1007/978-1-4684-2262-7>
- Chao, A., Chazdon, R.L., Colwell, R.K. & Shen, T.-J. (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics*, 62(2), 361–371. <https://doi.org/10.1111/j.1541-0420.2005.00489.x>
- Chatfield, C. (1995) *Problem solving: a statistician's guide*. London, UK: Chapman and Hall.
- Clarke, K.R. (1993) Non-parametric multivariate analyses of changes in community structure. *Austral Ecology*, 18(1), 117–143. <https://doi.org/10.1111/j.1442-9993.1993.tb00438.x>
- Collingridge, D.S. (2013) A Primer on quantized data analysis and permutation testing. *Journal of Mixed Methods Research*, 7(1), 81–97. <https://doi.org/10.1177/1558689812454457>
- Cooper, M.M. (2018) The replication crisis and chemistry education research. *Journal of Chemical Education*, 95(1), 1–2. <https://doi.org/10.1021/acs.jchemed.7b00907>
- Croux, C., Filzmoser, P. & Fritz, H. (2013) Robust sparse principal component analysis. *Technometrics*, 55(2), 202–214. <https://doi.org/10.1080/00401706.2012.727746>
- Devezer, B., Nardin, L.G., Baumgaertner, B. & Buzbas, E.O. (2019) Scientific discovery in a model-centric framework: reproducibility, innovation, and epistemic diversity. *PLoS One*, 14(5), e0216125. <https://doi.org/10.1371/journal.pone.0216125>
- Dexter, E., Rollwagen-Bollens, G. & Bollens, S.M. (2018) The trouble with stress: a flexible method for the evaluation of nonmetric

- multidimensional scaling. *Limnology and Oceanography: Methods*, 16(7), 434–443. <https://doi.org/10.1002/lom3.10257>
- Duneman, G.H. (1989) Principal component analysis. *Journal of Classification*. Sage Publications – Quantitative Applications in the Social-Sciences, 69(1).
- Dunham, R.J. (1962) Classification of carbonate rocks according to depositional textures. In: Ham, W.E. (Ed.) *Classification of carbonate rocks – a symposium. AAPG Memoir* 1, pp. 108–121.
- Dunkley Jones, T., Bown, P.R., Pearson, P.N., Wade, B.S., Coxall, H.K. & Lear, C.H. (2008) Major shifts in calcareous phytoplankton assemblages through the Eocene-Oligocene transition of Tanzania and their implications for low-latitude primary production. *Paleoceanography*, 23(4). <https://doi.org/10.1029/2008PA001640>
- Egozcue, J.J. (2009) Reply to “On the Harker Variation Diagrams; ...” by J.A. Cortés. *Mathematical Geosciences*, 41(7), 829–834. <https://doi.org/10.1007/s11004-009-9238-0>
- Embry, A.F. & Klovan, J.E. (1971) A Late Devonian reef tract on northeastern Banks Island, NWT. *Bulletin of Canadian Petroleum Geology*, 19, 730–781.
- Field, J.G., Clarke, K.R. & Warwick, R.M. (1982) A practical strategy for analysing multispecies distribution patterns. *Marine Ecology Progress Series*, 8(1), 37–52.
- Filzmoser, P., Hron, K. & Reimann, C. (2009) Univariate statistical analysis of environmental (compositional) data: problems and possibilities. *Science of the Total Environment*, 407(23), 6100–6108. <https://doi.org/10.1016/j.scitotenv.2009.08.008>
- Filzmoser, P., Hron, K. & Templ, M. (2018) *Applied compositional data analysis*. Cham: Springer International Publishing (Springer Series in Statistics). <https://doi.org/10.1007/978-3-319-96422-5>
- Finch, H. (2005) Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated. *Methodology*, 1(1), 27–38. <https://doi.org/10.1027/1614-1881.1.1.27>
- Gauch, H.G. (1982) *Multivariate analysis in community ecology*. Cambridge, UK: Cambridge University Press.
- Gauch, H.G. & Whittaker, R.H. (1972) Comparison of ordination techniques. *Ecology*, 53(5), 868–875. <https://doi.org/10.2307/1934302>
- Gosling, W.D., Cornelissen, H.L. & McMichael, C.N.H. (2019) Reconstructing past fire temperatures from ancient charcoal material. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 520, 128–137. <https://doi.org/10.1016/j.palaeo.2019.01.029>
- Gower, J.C. (1987) Introduction to ordination techniques. In: Legendre, P.L.L. (Ed.) *Developments in numerical ecology*. Berlin, Heidelberg: NATO ASI Series (Series G: Ecological Sciences) 14. Springer, pp. 3–64. doi: https://doi.org/10.1007/978-3-642-70880-0_1
- Greenacre, M. (2018) *Compositional data analysis in practice*. London, UK: Chapman and Hall, CRC Interdisciplinary Statistics.
- Griffiths, J.C. (1947) Sedimentary petrography and the oil industry. *Journal of Sedimentary Research*, 17(1), 29–34. <https://doi.org/10.1306/D4269297-2B26-11D7-8648000102C1865D>
- Gupta, A.K. & Thomas, E. (1999) Latest Miocene-Pleistocene productivity and deep-sea ventilation in the northwestern Indian Ocean (Deep Sea Drilling Project Site 219). *Paleoceanography*, 14(1), 62–73. <https://doi.org/10.1029/1998PA900006>
- Hammer, Ø. & Harper, D.A.T. (2007) *Paleontological data analysis*. Hoboken, NJ: Blackwell Publishing Ltd. <https://doi.org/10.1002/9780470750711>
- Hammer, Ø., Harper, D.A.T. & Ryan, P.D. (2001) PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*, 4(1). http://palaeo-electronica.org/2001_1/past/issue1_01.htm
- Harzing, A.W. (2019) Publish or perish. <https://harzing.com/resources/publish-or-perish>
- Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer. <https://doi.org/10.1007/978-0-387-84858-7>. Series in Statistics (Springer Series in Statistics).
- Hill, M.O. (1979) Decorana—A Fortran program for detrended correspondence analysis an reciprocal averaging. Section of Ecology and Systematics, Cornell University.
- Hohmann, N. & Jarochovska, E. (2019) Enforced symmetry: the necessity of symmetric waxing and waning. *PeerJ*, 7, e8011.
- Höltke, O., Salvador, R.B. & Rasser, M.W. (2016) Paleobiogeography of Early/Middle Miocene terrestrial gastropods in Central Europe: an approach using similarity indices. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 461, 224–236. <https://doi.org/10.1016/j.palaeo.2016.08.027>
- Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441. <https://doi.org/10.1037/h0071325>
- Hron, K., Templ, M. & Filzmoser, P. (2010) Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, 54, 3095–3107. <https://doi.org/10.1016/j.csda.2009.11.023>
- Ibanez, F. (1971) Effet des transformations des données dans l’analyse factorielle en écologie planctonique. *Cah Oceanogr*, 23, 545–561.
- IBM Corp. (2017) *IBM SPSS statistics for windows*. Armonk, NY: IBM Corp. <https://hadoop.apache.org>
- Jarochovska, E. (2012) High-resolution microtaphofacies analysis of a carbonate tidal channel and tidally influenced lagoon, pigeon creek, San Salvador Island, Bahamas. *Palaios*, 27(3), 151–170. <https://doi.org/10.2110/palo.2011.p11-063r>
- Jarochovska, E., Viira, V., Einasto, R., Nawrot, R., Bremer, O., Männik, P. & Munnecke, A. (2017) Conodonts in Silurian hypersaline environments: specialized and unexpectedly diverse. *Geology*, 45(1), 3–6. <https://doi.org/10.1130/G38492.1>
- Jost, L., Chao, A. & Chazdon, R.L. (2011) Compositional similarity and β (beta) diversity. In: Magurran, A.E. & McGill, B.J. (Eds.) *Biological diversity: frontiers in measurement and assessment*. Oxford University Press, pp. 66–84.
- Johnston, J.W. (1976) Similarity indices I: What do they measure. No. BNWL-2152 (ADD. 1). Battelle Pacific Northwest Labs.
- Kline, P. (1979) *Psychometrics and psychology*. London: Academic Press.
- Klubi, E., Abril, J.M., Nyarko, E. & Delgado, A. (2018) Impact of gold-mining activity on trace elements enrichment in the West African estuaries: the case of Pra and Ankobra rivers with the Volta estuary (Ghana) as the reference. *Journal of Geochemical Exploration*, 190, 229–244. <https://doi.org/10.1016/j.gexplo.2018.03.014>
- Kolde, R. (2019) pheatmap: pretty heatmaps. R package version 1.0.12. 2019.
- Korkmaz, S., Goksuluk, D. & Zararsiz, G. (2014) MVN: an R package for assessing multivariate normality. *The R Journal*, 6(2), 151.
- Koutsodendris, A., Nakajima, K., Kaboth-Bahr, S., Berke, M.A., Franzese, A.M., Hall, I.R., Hemming, S.R., Just, J., LeVay, L.J., Pross, J. & Robinson, R. (2020) A Plio-Pleistocene (c. 0–4 Ma) cyclostratigraphy for IODP Site U1478 (Mozambique Channel, SW Indian Ocean): exploring an offshore record of

- paleoclimate and ecosystem variability in SE Africa. *Newsletters on Stratigraphy*, 54(2), 159–181. <https://doi.org/10.1127/nos/2020/0608>
- Kowalewski, M. & Novack-Gottshall, P. (2010) Resampling methods in paleontology. *The Paleontological Society Papers*, 16, 19–54. <https://doi.org/10.1017/S1089332600001807>
- Kruskal, J.B. (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2), 115–129.
- Lanci, L., Hirt, A.M., Lotter, A.F. & Sturm, M. (2001) A record of Holocene climate in the mineral magnetic record of Alpine lakes: Sägistalsee and Hinterburgsee. *Earth and Planetary Science Letters*, 188(1–2), 29–44. [https://doi.org/10.1016/S0012-821X\(01\)00301-6](https://doi.org/10.1016/S0012-821X(01)00301-6)
- Legendre, P. & Gallagher, E.D. (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129(2), 271–280. <https://doi.org/10.1007/s004420100716>
- Legendre, P. & Legendre, L. (2012) Ordination in reduced space. *Developments in Environmental Modelling*, 24, 425–520. <https://doi.org/10.1016/B978-0-444-53868-0.50009-5>
- Matlab. (2020) MATLAB and statistics toolbox. MathWorks. <https://www.mathworks.com/products/matlab.html>
- Minchin, P.R. (1987) An evaluation of the relative robustness of techniques for ecological ordination. In: Prentice, I.C. & van der Maarel, E. (Eds.) *Theory and models in vegetation science*. Advances in vegetation science. Dordrecht: Springer, pp. 89–107. doi: https://doi.org/10.1007/978-94-009-4061-1_9
- More, K.D., Orsi, W.D., Galy, V., Giosan, L., He, L., Grice, K. & Coolen, M.J.L. (2018) A 43 kyr record of protist communities and their response to oxygen minimum zone variability in the Northeastern Arabian Sea. *Earth and Planetary Science Letters*, 496, 248–256. <https://doi.org/10.1016/j.epsl.2018.05.045>
- Mulaik, S.A. (1985) Exploratory statistics and empiricism. *Philosophy of Science*, 52(3), 410–430. <https://doi.org/10.1086/289258>
- Neuwirth, E. (2014) RColorBrewer: ColorBrewer Palettes. R package. <https://CRAN.R-project.org/package=RColorBrewer>
- Nissen, S.B. et al. (2016) Research: publication bias and the canonization of false facts. *eLife*, 5, e21451. <https://doi.org/10.7554/eLife.21451.001>
- Nüst, D. & Pebesma, E. (2020) Practical reproducibility in geography and geosciences. *Annals of the American Association of Geographers*, 111(5), 1–11. <https://doi.org/10.1080/24694452.2020.1806028>
- Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D. et al. (2019) vegan: community ecology package. <https://cran.r-project.org/web/packages/vegan/index.html>
- Olson, C.L. (1974) Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 69(348), 894–908. <https://doi.org/10.1080/01621459.1974.10480224>
- Palarea-Albaladejo, J. & Martín-Fernández, J.A. (2015) zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143, 85–96. <https://doi.org/10.1016/j.chemo.2015.02.019>
- Patzkowsky, M.E. & Holland, S.M. (2012) *Stratigraphic Paleobiology: understanding the distribution of fossil taxa in time and space*. Chicago, IL: University of Chicago Press.
- Peng, R. (2015) The reproducibility crisis in science: a statistical counterattack. *Significance*, 12, 30–32. <https://doi.org/10.1111/j.1740-9713.2015.00827.x>
- Pinkham, C.F. & Pearson, J.G. (1976) Applications of a new coefficient of similarity to pollution surveys. *Journal (Water Pollution Control Federation)*, 717–723.
- Piper, D.Z. & Bau, M. (2013) Normalized rare earth elements in water, sediments, and wine: identifying sources and environmental redox conditions. *American Journal of Analytical Chemistry*, 04(10), 69–83. <https://doi.org/10.4236/ajac.2013.410A1009>
- Podani, J. (1999) Extending Gower's general coefficient of similarity to ordinal characters. *Taxon*, 48, 331–340. <https://doi.org/10.2307/1224438>
- Popper, K. (1934) *Logik der Forschung*. Berlin, Germany: Akademie Verlag.
- Puri, M.L. & Sen, P.K. (1971) *Nonparametric methods in multivariate analysis*. Hoboken, NJ: John Wiley and Sons.
- R Core Team. (2020) R: a language and environment for statistical computing. <http://www.r-project.org/>
- Read, W.A. & Dean, J.M. (1968) A quantitative study of a sequence of coal-bearing cycles in the Namurian of Central Scotland, 2. *Sedimentology*, 10(2), 121–136. <https://doi.org/10.1111/j.1365-3091.1968.tb01104.x>
- Reimann, C. et al. (2008) *Statistical data analysis explained*. Chichester, UK: John Wiley and Sons, Ltd. <https://doi.org/10.1002/9780470987605>
- Rencher, A.C. (2003) *Methods of multivariate analysis*. Hoboken, NJ: Wiley.
- Reyment, R.A. & Savazzi, E. (1999) *Aspects of multivariate statistical analysis in geology*. Amsterdam, Netherlands: Elsevier. <https://doi.org/10.1016/B978-0-444-82568-1.X5012-5>
- Romesburg, C. (2004) *Cluster analysis for researchers*. Morrisville, NC: Lulu Press.
- Sharma, S., Bawa, S. & Lomash, H. (2016) Proliferation of social computing: cultural computing paradigm. *International Journal of Computer Applications*, 137(9), 27–30. <https://doi.org/10.5120/ijca2016908922>
- Shaukat, S.S., Rao, T.A. & Khan, M.A. (2016) Impact of sample size on principal component analysis ordination of an environmental data set: effects on eigenstructure. *Ekológia (Bratislava)*, 35(2), 173–190. <https://doi.org/10.1515/eko-2016-0014>
- Shi, G.R. (1993) Multivariate data analysis in palaeoecology and palaeobiogeography—a review. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 105(3–4), 199–234. [https://doi.org/10.1016/0031-0182\(93\)90084-V](https://doi.org/10.1016/0031-0182(93)90084-V)
- Stacklies, W., Redestig, H., Scholz, M., Walther, D. & Selbig, J. (2007) pcaMethods a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23, 1164–1167. <https://doi.org/10.1093/bioinformatics/btm069>
- Sneath, P.H. & Sokal, R.R. (1973) Numerical taxonomy. The principles and practice of numerical, classification.
- Steen, R.G., Casadevall, A. & Fang, F.C. (2013) Why has the number of scientific retractions increased? *PLoS One*, 8, e68397. <https://doi.org/10.1371/journal.pone.0068397>
- Syms, C. (2008) 'Ordination'. In: Jørgensen, S.E. & Fath, B.D. (Eds.) *Encyclopedia of ecology*. Amsterdam, Netherland: Elsevier, pp. 2572–2581. doi: <https://doi.org/10.1016/B978-008045405-4.00524-3>
- Taguchi, Y.H. & Oono, Y. (2005) Relational patterns of gene expression via non-metric multidimensional scaling analysis. *Bioinformatics*, 21(6), 730–740. <https://doi.org/10.1093/bioinformatics/bti067>
- Templ, M., Hron, K., Filzmoser, P. & Gardlo, A. (2016) Imputation of rounded zeros for high-dimensional compositional data.

- Chemometrics and Intelligent Laboratory Systems*, 155, 183–190. <https://doi.org/10.1016/j.chemolab.2016.04.011>
- ter Braak, C.J.F. (1986) Canonical correspondence analysis: a new Eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5), 1167–1179. <https://doi.org/10.2307/1938672>
- ter Braak, C.J.F. (1989) CANOCO — an extension of DECORANA to analyze species-environment relationships. *Hydrobiologia*, 184(3), 169–170. <https://doi.org/10.1007/BF02392953>
- ter Braak, C.J.F. & Prentice, I.C. (1988) A theory of gradient analysis. *Advances in Ecological Research*, 18, 271–317. [https://doi.org/10.1016/S0065-2504\(08\)60183-X](https://doi.org/10.1016/S0065-2504(08)60183-X)
- ter Braak, C.J.F. & Smilauer, P. (2012) Canoco reference manual and user's guide: software for ordination, version 5.0. Microcomputer Power.
- TIBCO Software (2018) Statistica® 13.3.0. <https://www.tibco.com/>
- Todorov, V. & Filzmoser, P. (2013) Comparing classical and robust sparse PCA. In: Kruse, R. et al. (Eds.) *Synergies of soft computing and statistics for intelligent data analysis. Advances in Intelligent Systems and Computing*, vol. 190, pp. 283–291. London, UK: Springer. doi: https://doi.org/10.1007/978-3-642-33042-1_31
- Tolosana-Delgado, R. (2012) Uses and misuses of compositional data in sedimentology. *Sedimentary Geology*, 280, 60–79. <https://doi.org/10.1016/j.sedgeo.2012.05.005>
- Tomašových, A. (2004) Microfacies and depositional environment of an Upper Triassic intra-platform carbonate basin: the Fatric Unit of the West Carpathians (Slovakia). *Facies*, 50(1), 77–105. <https://doi.org/10.1007/s10347-004-0004-y>
- Tyler, C.L. & Kowalewski, M. (2014) Utility of marine benthic associations as a multivariate proxy of paleobathymetry: a direct test from recent coastal ecosystems of North Carolina. *PLoS One*, 9(4), e95711. doi:<https://doi.org/10.1371/journal.pone.0095711>
- van de Velden, M., Iodice D'Enza, A. & Markos, A. (2019) Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(3), e1456. <https://doi.org/10.1002/wics.1456>
- Van der Weijden, C.H. (2002) Pitfalls of normalization of marine geochemical data using a common divisor. *Marine Geology*, 184(3–4), 167–187. [https://doi.org/10.1016/S0025-3227\(01\)00297-3](https://doi.org/10.1016/S0025-3227(01)00297-3)
- van Son, T.C. & Halvorsen, R. (2014) Multiple parallel ordinations: the Importance of choice of ordination method and weighting of species abundance data. *Sommerfeltia*, 37(1), 1–37. <https://doi.org/10.2478/som-2014-0001>
- Vu, V.Q. (2011) ggbiplot: a ggplot2 based biplot. R package version 0.55. Available at: <http://github.com/vqv/ggbiplot>
- Wackernagel, H. (2003) *Multivariate geostatistics*. Berlin, Heidelberg: Springer, Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-05294-5>
- Warne, R.T., Lazo, M., Ramos, T. & Ritter, N. (2012) Statistical methods used in gifted education journals, 2006–2010. *Gifted Child Quarterly*, 56(3), 134–149. <https://doi.org/10.1177/0016986212444122>
- Wartenberg, D., Ferson, S. & Rohlf, F.J. (1987) Putting things in order: a critique of detrended correspondence analysis. *The American Naturalist*, 129(3), 434–448. <https://doi.org/10.1086/284647>
- Watkinson, I.M. & Hall, R. (2019) Impact of communal irrigation on the 2018 Palu earthquake-triggered landslides. *Nature Geoscience*, 12, 940–945. <https://doi.org/10.1038/s41561-019-0448-x>
- Wentworth, C.K. (1922) A scale of grade and class terms for clastic sediments. *The Journal of Geology*, 30, 377–392. <https://doi.org/10.1086/622910>
- Whittaker, R.H. & Gauch, H.G. (1978) Evaluation of ordination techniques. In: Whittaker, R.H. (Ed.) *Ordination of plant communities*. Dordrecht: Springer, pp. 277–336. https://doi.org/10.1007/978-94-009-7989-5_10
- Wickham, H. (2016) *ggplot2: elegant graphics for data analysis*. New York, NY: Springer. <https://doi.org/10.1007/978-3-319-24277-4>
- Wildi, O. (2018) Evaluating the predictive power of ordination methods in ecological context. *Mathematics*, 6(12), 295. <https://doi.org/10.3390/math6120295>
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1). <https://doi.org/10.1038/sdata.2016.18160018>
- Zar, J.H. (2010) *Biostatistical analysis*. London, UK: Pearson.
- Ziemann, M., Eren, Y. & El-Osta, A. (2016) Gene name errors are widespread in the scientific literature. *Genome Biology*, 17, 177. <https://doi.org/10.1186/s13059-016-1044-7>
- Zhu, Z., Wang, T. & Samworth, R.J. (2019) High-dimensional principal component analysis with heterogeneous missingness. *arXiv* 1906.12125
- Zuschin, M., Harzhauser, M. & Mandic, O. (2007) The stratigraphic and sedimentologic framework of fine-scale faunal replacements in the middle Miocene of the Vienna Basin (Austria). *Palaios*, 22(3), 285–295. <https://doi.org/10.2110/palo.2005.p05-023r>

How to cite this article: Bialik OM, Jarochovska E, Grossowicz M. Ordination analysis in sedimentology, geochemistry and palaeoenvironment—Background, current trends and recommendations. *Depositional Rec.* 2021;7:541–563. <https://doi.org/10.1002/dep2.161>