Restoring gradients from fossil communities: a graph theory approach

Petr ČEJCHAN

Geological Institute, Academy of Sciences of the Czech Republic, Rozvojová 135, CZ – 165 02 Praha 6 – Lysolaje, Czechia

ABSTRACT. Gradients are often involved in patterns of species distribution in space and time. They are characterised by a specific pattern of species abundance, which in turn can be used to reconstruct fossil gradients from fossil biotic communities. Rather than the classical approach to gradient retrieval, based essentially on clustering or eigenvector methods, a graph theory was used as closer to the nature of the problem. Late Devonian marine benthic communities were successfully used to search for the latent gradient corresponding to the intensity of environmental stress in the vicinity of the Kellwasser extinction event. The reconstructed gradient was then in turn applied to elucidate the history of environmental stress during the Kellwasser crisis interval.

KEYWORDS: quantitative palaeoecology, gradients, methods, graph theory, travelling salesman problem.

Prerequisites

Continuous, directed environmental change, a time trend, or other similar processes often produce a typical pattern when reflected in fossil record. Such a directed smooth change is a **gradient**. What is its typical pattern? Usually, different species meet their optimal conditions at different places along the gradient; these optima are characterised by the highest abundance of the corresponding species. In both directions away from the species optimum, its abundance usually decreases monotonically (Fig. 1). How the species optima are spaced along the gradient depends on the process itself, as well as on other controls; however, the positions of optima usually do not coincide, but tend to be spaced in a spectrum of possibilities from almost random to almost regular. Examples include a depth gradient in a marine environment: some organisms are adapted to shallower conditions, others to deep water ones, each having its respective optimum on a defined point on the gradient (maximum abundance), and its tolerance limits (towards them the abundance is decreasing to zero). However, this is a very simple example.

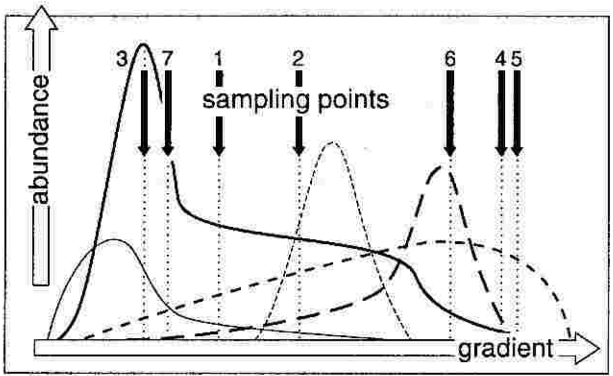

Figure 1. Typical distribution of species abundance along the gradient. Abundance curves are usually continuous, smooth, unimodal. The gradient is sampled more or less randomly at sampling points (1 to 7) of unknown position on the gradient. The sequence of samples 3-7-1-2-6-4-5 is the natural ordering, corresponding to the order of sampling points along the gradient. The aim of the study is to find the natural ordering using species abundances only.

Aims

In a fossil record, such gradients are commonly latent and are recorded only indirectly via biotic communities. This study deals with such a situation. The record may be (and commonly is) incomplete, or biased, but that is another story. Assume for now that the fossil communities preserve a usable record of the original gradient, and that the only thing we possess are samples of randomly chosen communities from different places on the assumed gradient. The crucial point is that we do not know anything about the nature of the gradient itself (i.e., the controlling process is compound or unknown) , nor do we have access to its direct record. Now we would like to arrange our samples in a series that corresponds to a natural section along that gradient`(Fig. 1).

Classical approach

Classically, this task was addressed via different varieties of cluster analysis, or via a primitive eigenvector procedure, the principal components analysis. Both branches evolved into more advanced techniques (e.g. fuzzy clustering, supervised clustering, etc., on the one side, and factor analysis and multidimensional scaling on the other). Yet all of these methods, although advanced, suffer from a common shortcoming: they make unrealistic assumptions about the data. While clustering methods essentially assume that distinguishable natural grouping exists in the data and believe that forcing the data into clusters will uncover the latent information, the eigenvector-based methods assume the linear response of the data to the latent factors. Both assumptions are unrealistic for our case. We chose another approach: from "realistic" model to suitable algorithm.

The Travelling Salesman: our approach

We started from the assumptions about the biotic responses to the continuous change along the gradient, stated above. According to the model, the gradient is crowded by continuous, smooth, unimodal "abundance curves" of all the species involved. The curves have their maxima at different points along the gradient. The sampling points are spaced more or less randomly on the gradient; quantitative samples of communities originate here yielding species abundances.

Define the distance between two samples as assemblage dissimilarities, calculated somehow from differences in abundances of individual species between the samples (voluminous literature exists on this topic). Do not confuse these with spatial or temporal distances (they are not used here)! Then try to examine the distances between contiguous points along the gradient: when the gradient becomes more and more densely sampled, the abundance differences, and hence the distances between neighbouring samples, become naturally shorter and shorter. Moreover, the distance from each sample to the nearer (more similar) of its neighbours also tends to be absolutely minimal, i.e. the smallest of the distances from that point to any other sample (Fig. 2). The sum of terms which tend to be the minimum possible also tends to be the minimum possible. Hence, the sum of distances between contiguous samples along the gradient (the "length" of the natural ordering of samples) should also tend to be the minimum of all those of other possible permutations (artificial orderings). This statement, although unproved, is the basis of the present study. We are searching for such a permutation of sample series that has the minimum possible length (sum of distances between contiguous samples).

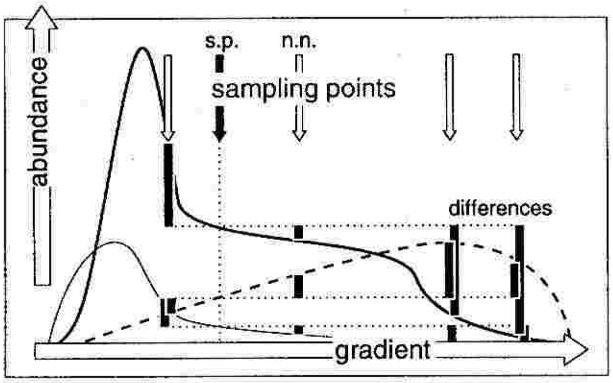

Figure 2. Differences in abundances along the gradient. Inter-sample distances are calculated from these differences. If the gradient is sufficiently densely sampled, then the distance from any selected sampling point (s.p., black arrow) to one of its neighbors (n.n., nearer of the neighbors) shows the tendency to be the smallest of all distances from that point to any other sampling point.

This problem is one of the well-known, classical problems of the graph theory, known as the **travelling salesman problem** (TSP). In terms of graph theory, our problem is represented by a complete graph, with nodes at samples and with edges represented by inter-sample distances. The opened variant of the TSP searches for the shortest path through all vertices (shortest Hamiltonian path). Although the problem has been well-known for a long time, its solution is not straightforward. It belongs to the class of so-called NP – complete tasks, which, with increasing number of vertices, become unavoidably very quickly limited by computing time. When the direction of the path does not matter (i.e., when we don't care whether we go through the ordered vertices in one or the opposite direction), the number of all possible different paths through n vertices is $n!/2$ (half of n factorial, where n is the number of vertices). If the number of vertices is low, say up to 10, the problem can be treated by brute force, i.e., by a complete search (the number of possibilities is 1 814 400 for 10 vertices). As the number of vertices grows, more sophisticated search algorithms have to be used, as the number of possibilities increases explosively (1.22×10^{18} for 20 samples). Very soon, the problem becomes intractable with searching methods: for 30 vertices the number of possibilities is 1.33×10^{32}; that means that if one checks one million paths per second for its length (a reasonable number), 4.2×10^{18} years would be needed to complete the search (this time exceeds about two hundred million times the total history of the Universe). Even sophisticated search methods, e.g., branch-and-bound one, often fail when dealing with such a large task. It is clear that other methods have to be applied when dealing with a comparable or higher number of vertices. Unfortunately, our premise is to have as many sampling points (= vertices) as possible, and that means very often >30 in practice.

As the NP – completeness of the problem excludes the possibility of a real-time analytical solution for large number of vertices, heuristics are often used to provide some "good" solution to the problem. Heuristic is an algorithm that is based on a "guess" of how to solve the problem, as opposed to an exact, analytical approach. There is no guarantee that a heuristic provides the best solution (i.e., the global optimum), nor is there a statement about how close the provided solution is to the best one. However, in practice the heuristics often provide acceptable solutions to otherwise intractable large-scale problems.

Case study

We used a TSP heuristic developed by us to carry out a case study on Upper Devonian marine benthic communities subject to a coming environmental crisis accompanied by mass extinctions (Čejchan and Hladil, in print). A series of inhabited fossil seafloors of a Late Devonian carbonate ramp environment is derived from the Mokrá, Western Quarry section, ca. 15 km E of Brno, Czechia (Fig. 3.). The limestone sequence includes the Frasnian – Famennian (F – F) Kellwasser events (Dvořák et al., 1987, Hladil et al., 1989) . The exact position of the F-F boundary is not known here, and the Kellwasser events themselves are correlated only tentatively.

Virtual 4 m^2 quadrats of the palaeo-seafloor were assembled at distinct levels of this section, as a perfect exhumation of such an area of the seafloor is rare. Twenty-eight reconstructed virtual quadrats were scrutinised for abundance of each of the eighty five species (mainly corals, stromatoporoids, brachiopods, gastropods, sponges, and algae) involved. Abundance was measured as the number of individuals for unitary organisms or as the number of colonies for modular ones.

The TSP algorithm was applied to reconstruct the probable sequence corresponding to the latent gradient. The consecutive assemblages in the resulting sequence, which represents the underlying gradient, were inspected in order to decide whether the sequence corresponds with the presumed extinction gradient. As the reconstructed sequence displayed poorly structured and ill diversified assemblages near one of its ends, and thriving assemblages near the other end, we concluded that a gradient of overall stress was well represented by the reconstructed sequence. Finally, we used the position of each sample on the reconstructed gradient as a measure of the intensity of stress, hence the environmental crisis progression / regression at the respective time (Fig. 3).

Criticism

Several assumptions have been made, but not proved:

(1) species abundance distributions along the gradient are unimodal,

(2) maxima of abundance are not strongly clustered,

(3) real abundance values are well represented by sampled data,

(4) squared Euclidean distance is a suitable measure for assessing dissimilarities of biotic communities,

(5) the gradient is sufficiently densely sampled, and

(6) sum of inter-sample distances along the gradient is, or approaches, the minimum possible.

Assumption (1) is usually accepted. Violations are possible, e.g., when strong competition occurs and the stenoecous stronger competitor outcompetes the euryecous weaker competitor species from a part of its span over the gradient. We do not know how serious assumption (2) is, i.e., how much does the clustering of maxima affect assumption (6). Modelling should give answer to this, but it is a matter for future research. Assumption (3) can be tested by resampling or using a binomial distribution. Neither was done during this study. We have no answer to (4), perhaps modelling will say more. We have no answer to (5) as well, no idea how to address this. Intuitively, we feel that (6) holds, but we did not prove it. Perhaps proof is possible, or we should use modelling to give us some insight.

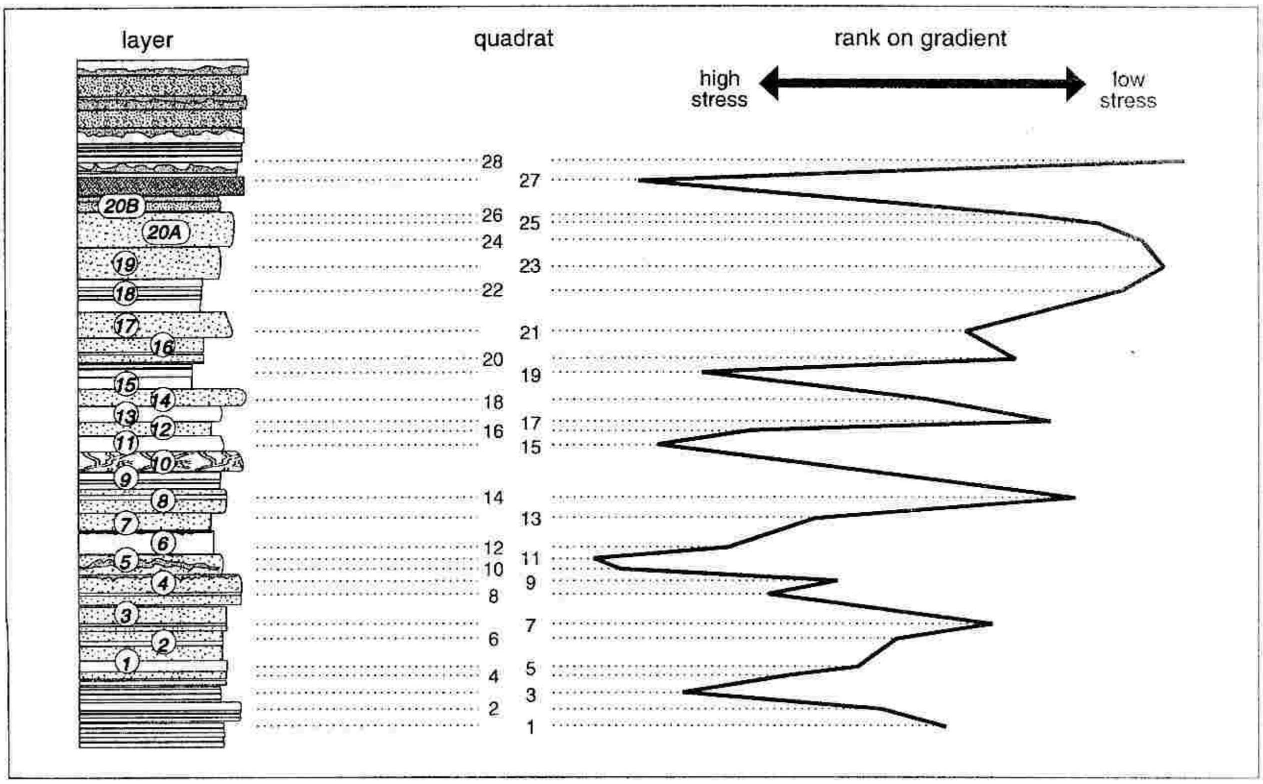

Figure 3. Mokrá, Western Quarry section. Late Devonian marine limestone sequence, vicinity of the Frasnian – Famennian boundary. Ranks of quadrats on presumed extinction gradient, derived from the TSP-reconstructed order of quadrats. Squared Euclidean distance based on abundances was used as a distance measure. Abundances were measured as numbers of individuals, or colonies. The curve of ranks is supposed to show the intensity of environmental crisis. Numbers in circles (italics) are numbered layers of the section, numbers outside the column are quadrat numbers.

Conclusion

The TSP algorithm showed a good capability to recover a meaningful gradient involved in the species abundance data collected from a medium-sized pool of samples, derived from fossil benthic communities. It is the first attempt to use this method in palaeobiology known to us. It seems to be a promising addition to the traditional instruments of ordination.

Acknowledgements

The study was carried out under the support of the Grant Agency of the Czech Republic (GAČR), Project 205 / 93 / 0723, and the Geological Institute of the Academy of Sciences of the Czech Republic Project 5213. It was a part of the UNESCO – IGCP Project 335. Thanks are due to Eric A. Oches (University of Minnesota, U.S.A.), Doug Erwin (Smithsonian Institution, U.S.A.), and James Ebert (University of Oneonta, N.Y., U.S.A.) for a critical review and language correction of the text.

References

ČEJCHAN P. and HLADIL J. (in print). Searching for extinction / recovery gradients: the Frasnian – Famennian interval, Mokrá Section, Moravia, Central Europe. *In* M.B. Hart (ed.): *Biotic Recovery from Mass Extinction Events,* 135–161. *Geological Society Special Publication,* 102. London.

DVOŘÁK J., FRIÁKOVÁ O., HLADIL J., KALVODA J. and KUKAL Z. 1987. Geology of the Palaeozoic rocks in the vicinity of the Mokrá Cement Factory quarries (Moravian Karst). *Sborník geologických věd, Geologie,* 42, 41–88.

HLADIL J., KALVODA J., FRIÁKOVÁ O., GALLE A. and KREJČÍ Z. 1989. Fauna from the limestones at the Frasnian / Famennian boundary at Mokrá (Devonian, Moravia, Czechoslovakia). *Sborník geologických věd, Paleontologie,* 30, 61–84.