Coenocline reconstruction using graph theory and Bayesian probability data generator

Čejchan, Petr
2007

Coenocline reconstruction using graph theory and Bayesian probability data generator

Peter A. Cejchan, Institute of Geology, Academy of Sciences of the Czech Republic, Rozvojova 269, CZ 16502 Prague, <cej@gli.cas.cz>, fax +420-2-20922670

19th February 2007

Keywords: paleobiology, paleoecology, indirect gradient analysis, coenocline, continuum concept, gradient reconstruction, fossil assemblages, Bayesian inference, population density, abundance, species performance, Quadratic Assignment Problem.

'There are three kinds of lies: lies, damned lies, and statistics.'
Mark Twain, 1924

1 The problem

The goal is to design a novel way to reconstruct the 'species response function (SRFs) of unitary organisms on an unidimensional environmental / time gradient solely from abundance data on species communities ('indirect gradient analysis'[**?**]). The word 'indirect' means that the gradient itself is latent, is not accessible, or even is presumed. The study will focus on the topology of the problem, thus disregarding the quantitative measuring of 'sample scores'. Formally, this task can be termed 'seriation of fossil communities (assemblages)'.

Bayesian inference is used here to 'squeeze' the population density distributions out of observed species' abundances. A large number of generated hypotheses serve to derive probability density functions of sample ranks, which are the result of the analysis.

2 Seriation

We use the term "seriation" here simply for topological (preserving just order, not distances) ordination in one dimension, usually, but not necessarily environment, or time.

Seriation is founded upon the assumption of humped unimodal 'response functions'. The task might seem trivial at first glance: permute rows of an abundance matrix until the humped pattern appears in columns. Unfortunately, the number of permutations increases with $n!/2$, where n is the number of samples. Thus, having to seriate 20 samples by exhaustive search means to evaluate 1216451004088320000 different permutations.

3 Need for an improved method

It appears that there is up to now no method in use, which would take the nonlinearity of species response functions into account from the very beginning. Instead, methods developed for the linear model are used, their misapplication being 'cured' by subsequent 'detrending', in the best case. A need for an improved method is thus obvious.

4 Assumptions

- We will consider only unitary organisms which do not form clusters, spread the progeny using (actively or passively) moving propagules, and having defined, unchanging demographic functions.
- We will adopt the population density as a feasible measure of species' performance.
- We require that a defined proportion of the population gets into the sediment, where it fossilizes. Actually, we will hide several processes under the term "fossilization chance" (see below).
- We will suppose the combined effect of these processes to be 'unpredictable' enough to form a stochastic variate that is characteristic for a given environment.
- We will suppose that sampling is done by taking a constant (up to a measuring error) volume of the sediment, extracting the fossils out of it, and counting individuals belonging to different species.
- Due to the complexity of the problem, we will assume that unidimensional analysis is appropriate. Although often unrealistic oversimplification, environments governed by an univariate gradient may occur quite frequently (cf. typical zonal communities of the mountain vegetation, inter-tidal biota, vegetation on moisture gradient, stratified oceanic communities depending on bathymetry, etc...).
- We assume that the typical 'species response function' on an environmental / time gradient is a unimodal function.

5 Processes contributing to abundance

5.1 Population density

Species performance, as determined by the environmental factor(s) and inter-specific competition, maps onto the population density, which is the essential component of the abundance model. There are also other measures of performance, like biomass production (standing crop), and other. However the present study deals with unitary organisms and their counts only; thus we confine to density as a performance measure. The population density is then used as the dependent variable of the SRF.

5.2 Volume sampling

Ideally, we would like to take samples representing constant $area \times time$. We assume we are able to take samples, or 'snapshots', of the communities that represent constant *volume* of the sediment, within the sampling error $N(\mu, \sigma)$. Equation 5shows the relation of the sample volume to time, using sedimentation rate, and sample basal area. Equation 16 incorporates the sample volume into a probabilistic estimate of abundance, y.

Whenever abundance (and hence, density) is estimated by means of counting individuals (a discrete variable), we are to deal with a so-called "Poisson point process" [**?**] [**?**] [**?**] [**?**] [**?**], where the number of individuals in constant amount of sampled $area \times time$ has a Poisson distribution [**?**]. Unfortunately, this Poisson sampling process is a source of a substantial error in abundance estimate, unless the samples are very large.

5.3 Sedimentation rate

Sedimentation rate influences the abundances via mapping of time to sediment thickness, i.e., also the sample volume. Thus, it is critical not only to take care of the volume of all samples be (as much as possible) equal, but also the sedimentation rate to be known, together with its variance.

5.4 Fossilization

Several processes influence the chance to fossilize. Predation removes living organisms from the environment, and from the reproduction process, modifying the ecological SRF of the species in a manner similar to competition. If the individual avoids predation, its remains must quickly get buried by the sediment, in order to have a nonzero chance to fossilize. The organismal remains not buried quickly are subject to decay and wash-away, that does not let the remain to enter the sediment. Even if buried, erosion, or dissolution of the sediment removes the remains together with the sediment. After surviving all the previous processes, the organismal remains must cope with bioturbation, and dissolution, in order not to be removed from the fossil record.

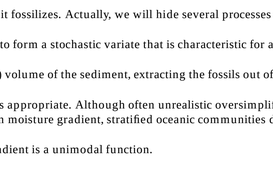
5.5 Contamination and reworking

Although perhaps far too conservative, these effects are modeled here as a constant 'contamination intensity' to allow for nonzero abundances to be sampled in arbitrarily large distance from the species' optimum. This 'contamination intensity' is modeled here as dependent on the species' mean performance; thus, for euryecous and common species it is high, whereas for stenoecous and rare species it is considered low. Although this model is rough, its ability to allow for substantial nonzero densities far from the optimum make it a suitable one for the purpose of testing the gradient's reconstructability under a wide spectrum of conditions.

Reworking of fossils from older strata may produce, if not recognized, a similar effect as contamination does. However, the resistance of fossils to reworking phenomena, not directly the species commonness / scarcity, plays a key role here. Even a prolific species may easily be destroyed during the process(es) of reworking; thus, a reworked scarce species may finally become more abundant than is a reworked common species.

5.6 Generation length

Generation length maps density and time onto abundance. The lower is the generation length of given species, the higher is the abundance compared with another species with the same population density, but longer length of one generation.



Volume sampled constant density process with varying sedimentation rate, fossilization, contamination, and varying length of generation. Parameter setting was μ_r = 0.001, σ_v= 0.0004, r_l= 0.008, r_h= 0.012, c_l= 0.45 c_h= 0.55, o_l= 0.01, o_h= 0.1, μ_g= 1.00 σ_g= 0.15.

6 Deterministic model

l.

Population densities are measured on the unit area. Density is here vaguely defined as a number of individuals encountered in a "time cut" on a unit area. Total number of individuals (of the species under consideration) that inhabited some area during a time interval, is

$$e = \frac{dta}{g} \tag{1}$$

where
e is the total number of individuals encountered in a time interval on an area,
d is a 'realized' density,
t is time interval within which the sampled volume of sediment was formed ,
a is the basal area of the sample, and
g is duration of one generation for the given species. We expect that g is constant for the given species.
The model should account also for a part of density caused by 'alien' individuals. We model the contamination as proportional to both contamination intensity, and mean population density of the proper species

$$d = d_p + o \cdot d_m \tag{2}$$

where
d_p is 'pure' population density,
d_m is mean population density of given species across all samples, and,
o is the contamination intensity.
Then,

$$e = \frac{dta}{g} = \frac{(d_p + od_m)ta}{g}$$

Usually, we cannot measure time directly, instead, we can estimate it from the volume of the sediment that has deposited under varying sedimentation rate. Thus,

$$v = ah \tag{3}$$

$$r = \frac{h}{t} \tag{4}$$

$$t = \frac{h}{r} = \frac{v}{ar} \tag{5}$$

$$e = \frac{(d_p + od_m)ta}{g} = \frac{(d_p + od_m)va}{arg} = \frac{(d_p + od_m)v}{rg} \tag{6}$$

where
v is the volume of the sediment,
h is height (thickness) of the sedimentary column,
r is the sedimentation rate,
t is time under which the sampled volume of sediment was formed.
Unfortunately, we usually **cannot** arrange sampling so that the sample represents a constant time interval.

Fossilization chance c is a proportion of individuals which actually fossilizes. Includes predation, gaps in sedimentation, and removal from the sediment due to any conceivable process. Observed abundance y is then:

$$y = ec = \frac{(d_p + od_m)vc}{rg} \tag{7}$$

However the above variables are not deterministic, rather stochastic. With unchanged connotation we will now proceed to the probabilistic (stochastic) model.

7 Probabilistic model

Density, d, is a Poisson variate (output of a Poisson process of "colonization", with intensity d_p, of the area under consideration). The sample volume, v, although tried to be kept constant, is subject to the sampling error, which has (approximately) a Normal distribution (this is not clean, as Gaussian distribution was derived for location, not scale variables, cf. Sivia 1996, p. 112). Sedimentation rate, r, is unknown, but supposed to lie within the limits r_l, r_h, within one sample. Fossilization chance, c, is unknown as well, but is supposed to lie within the limits c_l, c_h, within one sample. Contamination intensity, o, is unknown as well, but is supposed to lie within the limits o_l, o_h, for all samples and species. We model r,and c by a uniform distribution as well. The length of generation for a given species is supposed to be known with certain uncertainty, which has (approximately) a Normal distribution (see above). Thus we put:

$$d = Po\,(d_p) + Po\,(d_m \cdot U(o_l,\, o_h)) \tag{8}$$

$$e = Po\left(\frac{(d_p + d_m \cdot U(o_l,\, o_h)) \cdot v}{rg}\right) \tag{9}$$

$$y = Bi\,(e, c) \tag{10}$$

$$v = N(\mu_v,\, \sigma_v) \tag{11}$$

$$r = U(r_l,\, r_h) \tag{12}$$

within sample,

$$c = U(c_l,\, c_h) \tag{13}$$

within sample,

$$o = U(o_l,\, o_h) \tag{14}$$

$$g = N(\mu_g,\, \sigma_g) \tag{15}$$

within as well as among samples.

From 10 we have:

$$y = Bi\left(Po\left(\frac{(d_p + d_m \cdot U(o_l,\, o_h)) \cdot N(\mu_v,\, \sigma_v)}{N(\mu_g,\, \sigma_g) \cdot U(r_l,\, r_h)}\right), U(c_l,\, c_h)\right) \tag{16}$$

where
$Po(\lambda)$ is a Poisson variate with parameter λ ,
$Bi(n,\, m)$ is a Poisson variate with parameters n, m,
$N(\mu,\, \sigma)$ is a Normal (Gaussian) variate with mean μ, and standard deviation σ , and
$U(l,\, h)$ is a uniform variate on the interval $(l,\, h)$.

Estimating density

Abundance is a result of a compound process, with density being only one of its parameters. It should be stressed here again that abundance and density is not the same, and the two terms should not be mixed! *Only density has an ecological meaning*, however, for the paleobiologist, it is accessible only via abundance.

We shall use abundance pdf as the likelihood function for estimation of density posterior pdf from Eq. **??**. Thus, we need to derive the abundance pdf first. We will use Eq. 16; hence, using product rule (Eq. **??**):

$$prob(y|c, e, I) = \frac{e!}{y!\,(e-y)!} c^y\,(1-c)^{e-y} \tag{17}$$

$$prob(y, c|e, I) = prob(y|c, e, I) \times prob(c|I) = \frac{e!}{y!\,(e-y)!} c^y\,(1-c)^{e-y} \times \frac{1}{c_h - c_l} \tag{18}$$

$$prob(y, c, e|o, v, r, I) = prob(y, c|e, I) \times \frac{\left(\frac{d_p + d_m \cdot v}{rg}\right)^e \exp\left(-\frac{d_p + d_m \cdot v}{rg}\right)}{e!} \tag{19}$$

$$prob(y, c, e, o|v, r, g, I) = prob(y, c, e|o, v, r, g, I) \times \frac{1}{o_h - o_l} \tag{20}$$

$$prob(y, c, e, o, v|r, g, I) = prob(, c, e, o|v, r, g, I) \times \frac{1}{\sigma_v\sqrt{2\pi}} \exp\left(-\frac{(v-\mu_v)^2}{2\sigma_v^2}\right) \tag{21}$$

$$prob(y, c, e, o, v, r|g, I) = prob(y, c, e, o, v|r, g, I) \times \frac{1}{r_h - r_l} \tag{22}$$

$$prob(y, c, e, o, v, r, g|I) = prob(y, c, e, o, v|r, g, I) \times \frac{1}{\sigma_g\sqrt{2\pi}} \exp\left(-\frac{(g-\mu_g)^2}{2\sigma_g^2}\right) \tag{23}$$

From marginalization we have:

$$prob\,(y, c, e, o, v, r \mid I) = \int_0^{+\infty} prob(y, c, e, o, v, r, g|I) dg$$

$$prob\,(y, c, e, o, v \mid I) = \int_0^{+\infty}\int_0^{+\infty} prob(y, c, e, o, v, r|I) dg\, dr$$

$$\dots$$

$$e = \frac{\lambda^y \exp(\lambda)}{y!}$$

$$\lambda = d_p + \frac{d_m}{o_h - o_l} \cdot \frac{v}{rg}$$

$$\dots$$

so that finally we have

$$prob\,(y \mid I) = \int_0^{+\infty}\int_{r_l}^{r_h}\int_0^{+\infty}\int_{o_l}^{o_h}\int_0^{d_{max}}\int_0^1 prob\,(y, c, d, o, v, r, g|I)\, dg\, dr\, dv\, do\, dd\, dc \times \frac{1}{y} \tag{25}$$

From the Bayes' theorem, as given in Eq. **??**:

$$prob\,(hypothesis \mid data, I) \propto prob\,(data \mid hypothesis, I) \times prob\,(hypothesis \mid I)$$

Using Eq. 24 as the likelihood, and Jeffreys' prior (Eq. **??**) for abundance, we get the posterior pdf for d_p:

$$prob\,(d_p \mid y, I) \propto \int_0^{+\infty}\int_{r_l}^{r_h}\int_0^{+\infty}\int_{o_l}^{o_h}\int_0^{d_{max}}\int_0^1 prob\,(y, c, d, o, v, r, g|I)\, dg\, dr\, dv\, do\, dd\, dc \times \frac{1}{y} \tag{25}$$

Reconstruction

We have designed the measure that quantifies a 'compactness' of a pattern created by a series of SRFs, and we will try to show that it approximates, or quantifies, unimodality quite well. This is because the unimodal pattern is 'compact' in the sense that high values tend to be clustered together, whilst low values surround them.

Formalizing the task we get an objective function:

$$c = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{n} y_{\pi(i)k} \cdot y_{\pi(j)k} \cdot a_{ij} \tag{26}$$

thus

$$c = \sum_{i=1}^{n}\sum_{j=1}^{m} \delta_{ij} \cdot \varphi_{\pi(i)\pi(j)} \tag{27}$$

where

$$\varphi_{\pi(i)\pi(j)} = \sum_{k=1}^{n} y_{\pi(i)k} \cdot y_{\pi(j)k} \tag{28}$$

thus

$$\varphi_{ij} = \sum_{k=1}^{n} y_{ik} \cdot y_{jk} \tag{29}$$

However, Eq. 27 is an instance of a notoriously known graph-theoretical problem, the Quadratic Assignment Problem (QAP), where c is the cost function, $|\delta|$ is the distance matrix, $|\varphi|$ is the flow matrix, $\pi\,(i)$ is the position of the vertex i in the permutation π, and y_{ik} is the performance of species k in sample i.

There exists a vast literature concerning the QAP.

Efficiency

To assess the efficiency of the QAP method of seriation, a series of comparisons with several traditional methods have been run.

The QAP method appeared superior to the traditional ones in a well-informed case: densities were a priori known, and the gradient was sampled using regular spacing of samples.

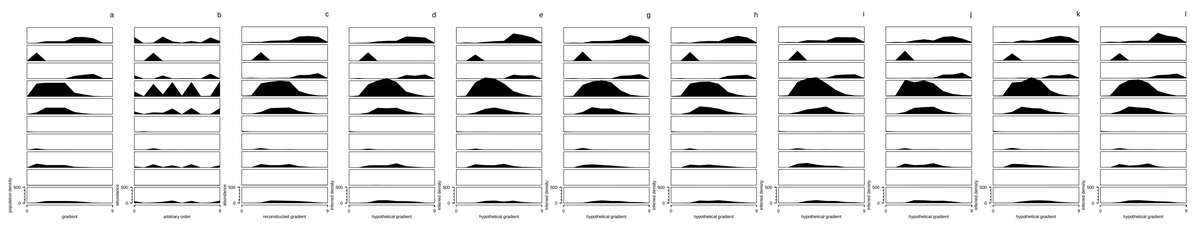
The testing array consisted of 10 samples, spaced *regularly* on the gradient, composed of up to 200 species.

8 Results

8.1 Artificial example

To test the whole procedure, an artificial example has been generated. Gaussian model with randomly spaced samples has been used. Thirty independent estimates were generated for density of every species at each site. From this array, 62 matrices of inferred Bayesian densities (cases 000 to 061) were generated.

The first 62 cases produced a handful of hypotheses, which are presented below.



Artificial example a, simulated coenocline, densities along the gradient; b, abundances sampled at sampling sites of (a); c, coenocline and the gradient reconstructed from sampled abundances using the QAP algorithm; d - l, several hypotheses that occurred first when running probabilistic QAP seriation on inferred densities; note that the coenocline pattern generally holds while there are minor changes in sample ordering (the seriation of the samples) between hypotheses.

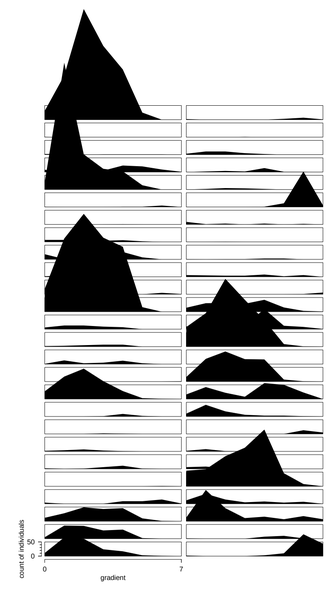
8.2 Case study

Data for this case study are derived from the fauna of land snails of the Holocene locality of Za krizem Cave near Svaty Jan pod Skalou, Bohemia. Age of the fauna is Holocene. The locality and the malacofauna was described by Lozek & Horacek 1993 [**?**], where also the quantitative data on fossils' abundances come from.

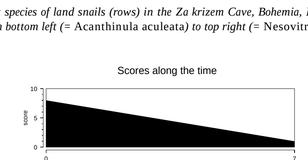
LAYER	9	8	7	6	5	4	3	2	1	
SPECIES / SAMPLE	-	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	
Acanthinula aculeata	0	0	1	3	17	24	61	66	11	
Aegopinella pura	0	0	0	2	32	29	45	46	5	
Cochlodina laminata	0	0	1	10	46	43	49	28	12	
Discus ruderatus	0	1	15	9	9	0	0	0	4	
Ena montana	0	0	1	1	0	0	0	0	0	
Ena obscura	0	0	0	1	11	6	1	0	2	
Helicodonta obvoluta	0	0	0	0	1	3	7	4	2	
Isognomostoma isognomostoma	0	0	0	0	1	2	0	0	0	
Macrogastra plicatula	0	0	0	0	2	9	1	0	0	
Monachoides incarnata	0	0	0	1	3	28	63	107	27	
Oxychilus cf. depressus	0	0	1	0	0	0	0	79	0	
Sphyradium doliolum	0	0	0	0	3	12	5	13	1	
Trichia unidentata bohemica	0	0	0	0	0	7	7	3	2	
Vertigo pusilla	0	0	0	0	0	7	9	13	6	
Alinda biplicata	0	0	0	0	16	239	261	258	80	
Arianta arbustorum	0	1	5	0	1	0	0	0	1	
Cepaea hortensis	0	0	0	0	0	1	3	4	2	
Discus rotundatus	0	0	0	0	7	26	21	3	1	
Limax cf. cinereoniger	0	0	0	0	2	6	4	2	7	
Oxychilus glaber	0	0	0	0	0	1	0	0	0	
Semilimax kotulae	0	0	0	5	1	1	0	0	1	
Aegopinella minor	0	0	0	1	15	64	73	124	26	
Bradybaena fruticum	0	0	1	9	20	23	5	1	8	
Helix pomatia	0	0	1	1	1	1	1	0	2	
Clausilia pumilla	0	0	0	0	0	0	0	0	1	
Cruaxiaria frumentum	0	0	0	0	25	177	260	156	31	
Helicopsis striata	4	44	77	10	3	0	0	0	2	
Chondrula tridens	0	0	1	10	7	0	0	0	1	
Pupilla sterri	0	7	18	7	16	11	46	110	13	
Pupilla triplicata	0	0	7	4	8	5	15	35	12	
Chondrina avenacea	0	0	8	46	200	136	106	60	55	
Pyramidula rupestris	0	0	0	0	0	6	2	8	6	
Truncatellina claustralis	0	0	0	0	0	0	1	0	2	
Pupilla muscorum	0	5	13	0	0	0	0	0	0	
Truncatellina cylindrica	0	0	1	3	3	6	18	41	10	
Vallonia costata	0	0	23	50	56	8	22	42	16	
Vallonia pulchella	0	0	1	7	78	79	106	80	16	
Vertigo pygmaea	0	0	0	0	0	0	1	0	0	
Bulgarica nitidosa	0	0	1	9	93	165	239	118	69	
Cochlicopa lubricella	0	1	8	12	71	13	12	7	10	
Euomphalia strigella	0	1	3	15	42	28	31	30	14	
Cochlicopa lubrica	0	6	1	0	0	0	0	0	0	
Euconulus fulvus	0	0	6	2	8	4	4	5	6	
Nesovitrea hammonis	0	0	0	0	3	1	0	0	1	
Oxychilus cellarius	0	0	0	0	0	0	1	0	0	
Punctum pygmaeum	0	0	2	0	3	0	3	1	0	
Trichia sericea	0	6	125	13	1	0	0	0	9	
Vitrea contracta	0	0	0	0	2	4	5	2	0	
Clausilia dubia	0	0	1	1	14	2	4	2	0	
Helicopsis lapicida	0	0	0	0	0	2	5	11	3	
Laciniaria plicata	0	0	0	0	0	0	1	0	0	
Vertigo alpestris	0	1	7	3	0	0	0	0	2	
Carychium tridentatum tridentatum	0	0	0	0	0	2	8	12	4	
Nesovitrea petronella	0	0	4	8	6	0	0	0	0	
TOTAL	**4**	**75**	**346**	**325**	**1333**	**1304**	**1797**	**1693**	**496**	**SR**
CHRONOLOGY	**V**	**LV**	**PB**	**B**	**A**	**EA**	**EA**	**SB-SA**		

*Abundances of different species of land snails in the **Za krizem Cave** Section, Bohemia, Holocene. Samples were taken from numbered layers, 9 being the oldest, 1 the youngest; 0 is recent. Chronology: V = Vistulian pleniglacial, LV = late Vistulian glacial, PB = Preboreal, B= Boreal, A = Atlantic, EA = Epiatlantic, SB = Subboreal, SA = Subatlantic, SR = Subrecent. Data from Lozek & Horacek 1993:114, tab. 1[?].*

An excellent unimodal pattern of SRFs yielded by the seriation performed on abundance data promised a real gradient to be included. The driving gradient was *time*, i.e., the faunal succession, not ecology itself (compare figs 9.24 and 9.28).



Coenocline pattern. SRFs obtained from the QAP seriation on abundances of different species of land snails (rows) in the Za krizem Cave, Bohemia, Holocene. Very nice unimodal pattern was revealed. Reconstructed sample order along the gradient was 8 7 6 5 4 3 2 1. Species are ordered according to tab 9.7, column-wise from bottom left (= Acanthinula aculeata*) to top right (*= Nesovitrea petronella*).*



Sample scores vs. time. Scores obtained from the QAP seriation on abundances of different species of land snails in the Za krizem Cave, Bohemia, Holocene. Time runs from left to right. PV = Vistulian glacial, LV = late Vistulian, PB = Preboreal, B = Boreal, A = Atlantic, EA = Epiatlantic, SB = Subboreal, SA = Subatlantic, SR = Subrecent. Note that scores change monotonically with time. The gradient could not be interpreted as ecological anymore, though there is strong evidence for presence of some kind of an unidimensional gradient (see fig. 9.27). One possible interpretation is that faunal migrations overrided the ecological pattern: new assemblages in a similar context are composed of a new fauna.

9 Main results

A new probabilistic algorithm for indirect gradient analysis was presented. The task of indirect gradient analysis was formalized as an instance of a long known combinatorial problem, the Quadratic Assignment Problem (QAP). It was shown empirically that seriation using the QAP solver algorithm outperforms that of DECORANA (detrended correspondence analysis, perhaps the most popular ordination algorithm so far) significantly.

It was shown that several stochastic processes map population densities onto species abundances. To make an inference about densities we used probabilistic (Bayesian) inference.

We derived the Bayesian estimate of population density. **It is for the first time** (as far as the author knows) in the paleoecological ordination when such parameters as variance of sampling volume, sedimentation rate and the uncertainty about it, fossilization chance, contamination risk, length of a generation, and associated uncertainties **are incorporated directly into the model** to be used with ordination.

We used a Monte Carlo generator based on equation **??** to carry out a stochastic seriation, i.e., the seriation based on Bayesian estimates of densities. The QAP seriation served as a 'multiple hypotheses generator' for a stochastic solution. Although believed to be the proper way to present the results, the author never saw it applied in recent (paleo)ecological literature.

Presented at The Royal Society UK-V4 Frontiers of Science Symposium at Krakow, Pl, EU, 21-23. 02. 2007