

A statistical procedure for unsupervised classification of nutrient limitation bioassay experiments with natural phytoplankton communities

Tom Andersen¹, Tuomo M. Saloranta², and Timo Tamminen³

¹Dept. Biology, University of Oslo, P.O. box 1066, Blindern, 0316 Oslo, Norway

²Norwegian Institute for Water Research, Gaustadalléu 21, 0349 Oslo, Norway

³Finnish Environment Institute, Mechelininkatu 34a, P.O. box 140, FIN-00251, Helsinki, Finland

Abstract

We describe a novel method for statistical analysis of bioassay experiments with natural plankton communities. The procedure allows unsupervised classification of the type of nutrient limitation in factorial experiments with two limiting nutrients, based on objective selection between several generic limitation patterns with direct biological interpretation, using the Akaike Information Criterion (AIC), which balances the concerns of model fit and model robustness. As such, it avoids the interpretation of nuisance parameters related to time effects in classical factorial design analysis of experiments with repeated measurements over time. The proposed limitation patterns discriminate between exclusive and primary limitation, depending on whether the initially nonlimiting nutrient has an effect or not. They also discriminate between two types of colimitation, depending on whether both nutrients only have an effect in combination, or also separately. The latter response is interpreted as the result of different phytoplankton community components being simultaneously limited by different nutrients. The capabilities of the classification procedure is demonstrated on a comprehensive set of 163 bioassay experiments with N and P additions to natural phytoplankton communities from the coast of Finland (NE Baltic Sea).

Bioassay experiments with natural plankton communities have been a common approach for examining nutrient limitation of growth rates in aquatic systems (Howarth 1988; Elser et al. 1990, and references therein). Such assays are normally carried out as factorial experiments: subsamples of an initial water sample are spiked with all possible combinations of the relevant nutrients, and the responses are recorded after some incubation time. In the simplest, most clear-cut case, the limiting nutrient will produce a significant biomass increase compared with the control treatment (no nutrients added), both alone and in combination with others. Furthermore, there will be no effect caused by any of the other nutrients.

Factorial design experiments have become a standard approach for exploring yield relationships in industrial

processes depending on many independent parameters (Box and Draper 1987). A factorial design will be optimal and cost efficient in the sense that it identifies all additive and multiplicative effects with the least number of treatment combinations. As such, factorial designs should be ideal for investigating the effects of several potentially limiting nutrients on plankton communities. Unfortunately, nutrient limitation is a generically nonlinear process: limitation by essential resources is often found to be better described by a threshold model with abrupt switches at critical transition points (Droop 1974; Tilman 1982). This can lead to asymmetrical responses where one nutrient has an effect only in combination with another, but not alone. Such noncommutative relations will normally violate the standard assumption used in interpreting factorial designs.

Any method dependent on a certain incubation time to produce a response suffers from an inescapable trade-off between the concerns of keeping the incubation time short and making the signal-to-noise ratio high. Short incubation time is dictated by the expectation that the community will diverge from the initial one with time, and the desire to produce a result that can be extrapolated back to the initial condition with minimal bias. On the other hand, the divergence with time will also increase statistical confidence in differ-

Acknowledgments

This work was financed through the research contract "Nitrogen Discharge, Pelagic Nutrient Cycles, and Eutrophication of the Northern Baltic Coastal Environment – PELAG III" jointly funded by the Finnish Ministry of the Environment, SYKE, Academy of Finland, University of Helsinki, Maj and Tor Nessling Foundation, and 25 other parties of Finnish coastal industries and towns, in 1991–1995. Different aspects of the in-depth data analysis have been funded by the EU fifth Framework Project DANLIM (EVK3-CT-2001-00049) and the sixth framework project THRESHOLDS (GLOBAL/IP/02/0257).

ences between treatments, which is desirable. Normally, one cannot predict the time scale of the response in advance, especially since phytoplankton growth often exhibits significant lags before growth commences after a perturbation. In light of this, it makes sense to make repeated observations to ensure that the relevant time scale is covered. Time effects of repeated sampling at regular intervals can be represented by orthogonal polynomials in statistical models of factorial designs. Unfortunately, this inclusion inflates the number of parameters in the statistical model, and also introduces interaction terms between time and treatment effects, which are notoriously hard to interpret (e.g., Kivi et al. 1993; Lignell et al. 2003)

The rationale behind this work is that it is often not the actual quantitative response itself, but some derived nominal classification such as “P limited,” “N limited,” etc., which is of interest when analyzing a factorial nutrient limitation bioassay. We present a statistical model selection procedure, which implicitly classifies a factorial experiment response into a small number of biologically interpretable limitation classes without the need for arbitrary and heuristic decision rules.

Theory—We will consider nutrient limitation assays where a set of sub-samples from the same source are individually spiked in all factorial combinations of nutrients, and the responses are followed by repeated sampling over time. For such a design with n possibly limiting nutrients, there will need to be 2^n different treatments. In this particular context, we consider an assay with two limiting nutrients, but the method can be generalized to more than two limiting factors. We will also, for the present, assume that the two limiting nutrients are phosphorus (P) and nitrogen (N), but these can of course be replaced by other essential resources.

The responses of a single-factor experiment with only $2^1 = 2$ treatments can be described by only two parameters: the mean response and the response difference among treatments. This can be represented by a design matrix with orthogonal columns:

$$M_1 = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \quad (1)$$

We can fit measured responses (y) to the two treatment levels by a model $y = M_1 b + \varepsilon$ with ε being zero-mean, constant variance noise. The fitted vector b will have two elements, of which the first estimates the mean of the 2 treatment levels, whereas the second estimates the difference between them. Thus, the first column of M_1 represents the mean, and the second represents the contrast between treatments.

For a 2-way design (2^2), we will need to consider the contrasts between treatment levels of one factor for both levels of the other factor, which effectively means that the second factor effects become embedded in the design matrix of the first factor. It has been known since a publication by F. Yates in 1937 (Cox and Reid 2000) that this embedding can be efficiently expressed by using the Kronecker tensor product operator (\otimes). In general, if U is an $n \times n$ matrix and V is an $m \times m$ matrix, then $U \otimes V$ is an $nm \times nm$ matrix with block elements.

That is, each $m \times m$ block element of $U \otimes V$ will be the corresponding element of U times the whole of V . The orthogonal contrast matrix for a 2^2 design will thus become

$$M_2 = M_1 \otimes M_1 = \begin{bmatrix} M_1 & -M_1 \\ M_1 & M_1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad (2)$$

The first column of M_2 will still be the mean of all responses while the subsequent ones will contain the contrasts among treatments. If we consider a 2-factor experiment with P and N addition, we can label the $2^2 = 4$ treatments as 0 + 0 (control, no addition), P + 0 (single P addition), 0 + N (single N addition), and P + N (combined P and N addition). If we label the rows of M_2 correspondingly (0 + 0, P + 0, 0 + N, P + N), the second column will represent the difference between the mean of the 0 + 0 and 0 + N treatments, and the P + 0 and P + N treatments. So this column will represent the effect of P, alone or in combination with N. Similarly, the third column will represent the effect of N, alone or in combination with P.

Repeated sampling at m equally spaced points in time can be represented by polynomials of order up $m - 1$. The Chebyshev polynomials of order $m - 1$, defining an orthogonal matrix of rank m , can be efficiently computed by a recursive relationship given in Abramowitz and Stegun (1974). For example, the matrix of Chebyshev polynomials up to order 3 for four equally spaced time intervals will be

$$T_3 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \frac{1}{3} & -1 & -3 \\ 1 & -\frac{1}{3} & -1 & 3 \\ 1 & -1 & 1 & -1 \end{bmatrix} \quad (3)$$

Lower order time effects can be constructed as subsets of the columns of T_3 . If we label the columns as in $T_3 = [t_0 \ t_1 \ t_2 \ t_3]$, the linear time effect matrix can be written as $T_2 = [t_0 \ t_1]$ whereas the quadratic time effect becomes $T_2 = [t_0 \ t_1 \ t_2]$. The relationship between the orthogonal time effect matrix columns and the corresponding interpolating polynomials are illustrated in Fig. 1. Time effects from repeated sampling

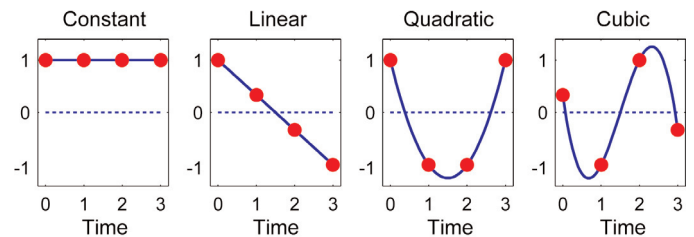


Fig. 1. Orthogonal (Chebyshev) polynomials of order 0 to 3 on four equidistant points (blue curves). Red dots represent the corresponding columns in the orthogonal time effect matrix given by Eq. 3.

can be represented by applying the Kronecker tensor product operator again, such that the treatment contrasts become embedded in every time effect. The combination of the orthogonal time effect matrix T and the orthogonal contrast matrix M_2 yields

$$X = T \otimes M_2 \quad (4)$$

For the combination of four repeated measurements on a 2^2 factorial design, the design matrix X will be 16 rows by 16 columns. If measurements are organized so that an element y_i represents a column vector with the 4 measurements from a single 2^2 design on day i :

$$y_i = [y_{0+0,i} \ y_{P+0,i} \ y_{0+N,i} \ y_{P+N,i}]^T \quad (5)$$

then four repeated measurements on the same design can be written as

$$y = [y_0 \ y_1 \ y_2 \ y_3]^T \quad (6)$$

This finally leads to the formulation of a linear model for factorial experiments with repeated measurements:

$$y = Xb + \varepsilon, \quad (7)$$

where b is an unknown parameter vector that must be estimated, and ε represents a vector of independent, normally distributed errors with constant variance (σ). Experience has shown that many standard response parameters like Chlorophyll a and carbon fixation will appear to be log-normally distributed (Tamminen and Andersen, in press); in other words, this condition is more likely to be fulfilled when the response data (y) are log-transformed.

As the matrix X is constructed from orthogonal elements, it will always be full rank, which implies that Eq. 7 will have just as many equations as unknowns. This will normally not give a consistent set of equations unless the measurement noise (ε) is negligibly small, and the data fit the model perfectly. In practice, the parameters of b cannot be reliably estimated unless there are additional degrees of freedom, for example by performing 2 blocks of identical experiments A and B, so that

$$\begin{bmatrix} y_A \\ y_B \end{bmatrix} = \begin{bmatrix} X \\ X \end{bmatrix} b + \varepsilon \quad (8)$$

Standard procedures for analyzing a data set according to Model 8 would involve least squares fitting of the unknown parameter vector b , post-hoc screening of statistically significant elements of b , and interpreting the pattern of non-zero b elements in terms of treatment effects and interactions. In the present case, this will be complicated by the mixing of time and treatment effects implied by Eq. 4, giving in Model 8 up to 16 non-zero parameters, that is, $2^{16} = 65536$ possible patterns, whereas the treatment effect Model 2 alone has only 4 parameters, or 16 possible patterns. Of these 16, we can identify 7 different, biologically meaningful patterns that can be

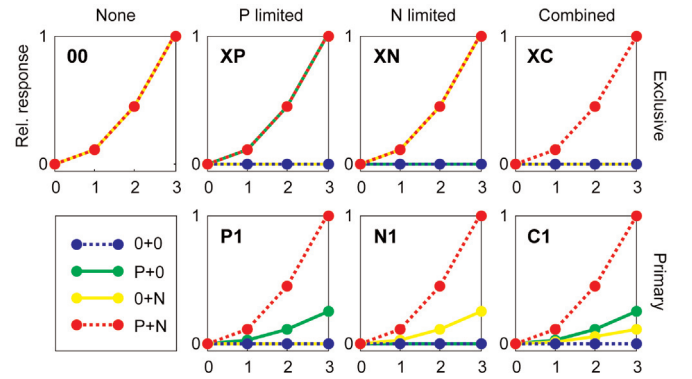


Fig. 2. The seven limitation patterns described by Eq. 9 shown for the case of quadratic time effects. Different treatments are symbolized by colors (blue: control, green: P alone, yellow: N alone, red: N and P combined).

expressed as linear combinations of the columns of the original design matrix $M_2 = [m_1 \ m_2 \ m_3 \ m_4]$

$$\begin{aligned} 0+0 = P+0 = 0+N = P+N &\Rightarrow M_{00} = [m_1] \\ (0+0 = 0+N) \neq (P+0 = P+N) &\Rightarrow M_{XP} = [m_1 \ m_2] \\ (0+0 = 0+N) \neq P+0 \neq P+N &\Rightarrow M_{P1} = [m_1 \ m_2 \ (m_3 + m_4)] \\ (0+0 = P+0) \neq (0+N = P+N) &\Rightarrow M_{XN} = [m_1 \ m_3] \\ (0+0 = P+0) \neq 0+N \neq P+N &\Rightarrow M_{N1} = [m_1 \ m_3 \ (m_2 + m_4)] \\ (0+0 = P+0 = 0+N) \neq P+N &\Rightarrow M_{XC} = [m_1 \ (m_2 + m_3 + m_4)] \\ 0+0 \neq P+0 \neq 0+N \neq P+N &\Rightarrow M_{C1} = [m_1 \ m_2 \ m_3 \ m_4] \end{aligned} \quad (9)$$

In the simplest case ('00'), there is no effect of any treatment so only the mean term (m_1) of the full model is retained. The time effects can be nonlinear but the lack of significant treatment contrasts dictate that all four treatments will behave identically. The remaining six cases can be arranged in three groups corresponding to P limitation (XP and P1), N limitation (XN and N1), and combined P and N limitation (XC and C1). Each of these groups comprises two cases that we will call exclusive limitation (XP, XN, and XC) and primary limitation (P1, N1, and C1). The exclusive cases are such that one nutrient produces an identical response alone and in combination with the other, while the response to the other nutrient alone is identical to the control. For the XC case, this means that only the combined addition produces a response whereas both single additions are indistinguishable from the control. The primary cases are such that the other nutrient alone produces no response, whereas the response to the primary limiting nutrient alone is different from the full addition. The primary combined case (C1) is such that both the single additions are different from both the control and the full addition.

The contrasts between the different patterns are illustrated graphically in Fig. 2. We see that the difference between the XP and P1 patterns is that in the first case the P alone treat-

Table 1. Number of unknown parameters in different limitation patterns (columns, notation as in Eq. 9) for different time order effects (rows)

Nr param.	00	XP	P1	XN	N1	XC	C1
Linear	2	4	6	4	6	4	8
Quadratic	3	6	9	6	9	6	12
Cubic	4	8	12	8	12	8	16

ment is bound to the NP treatment, whereas in the second case, it is allowed to move freely with respect to the NP treatment. The common feature of the XP and P1 treatments is that the N alone treatment is bound to the control, i.e., there cannot be any effect of N alone, only in combination with P. Similar arguments separate the XN and N1 patterns. The XC pattern represents a completely balanced situation where a combined treatment with both nutrients is needed to provoke a response different from the control. The C1 pattern represents the full model where none of the responses are bound to any of the others. This pattern is expected to occur if the plankton community is composed of two populations limited by different nutrients.

Combining seven generic effect contrast matrices with time effects up to third order (linear, quadratic, or cubic: T_1 , T_2 , and T_3) gives a total of 21 candidate design matrices. As illustrated in Table 1, the number of unknown parameters increases from 2 in the simplest (no treatment effect, linear time effect) to 16 in the most complex case (all treatments different, cubic time effect). If we could choose by some decision rule which of these 21 models is the “best” for a given data set, we would also have an implicit classification of the nutrient limitation response by considering just the chosen contrast matrix and disregarding the time effect.

The problem of model selection has a long tradition in statistical theory. It is evident that a model with many parameters would typically give a better fit to a given data set than one with fewer. On the other hand, many-parameter models would also be more sensitive to spurious errors in the input data, so that it is normally considered desirable to keep the number of parameters as low as possible (a variant of the principle of parsimony). Several indicators have been proposed for representing this trade-off between goodness-of-fit and model complexity (Burnham and Anderson 2002, see also Hilborn and Mangel 1997 for a less technical account). We will concentrate on one indicator that seems to have a firmer theoretical foundation than many others: the Akaike Information Criterion (AIC).

If we have a linear model with n measurements, p parameters, and with additive normally distributed noise with constant standard deviation, then the maximum likelihood and least squares estimates of the unknown parameter vector b and standard deviation will be identical and the AIC (with the

Table 2. Akaike information criterion (AIC) for different candidate models confronted with synthetic data corresponding to exclusive P limitation with quadratic time effects [XP(2)], listed with different limitation patterns in columns (notation as in Eq. 9) and time order effects in rows

AIC	00	XP	P1	XN	N1	XC	C1
Linear	2.10	-0.29	-0.10	2.26	1.83	1.89	0.17
Quadratic	2.13	-31.5*	-31.3	2.40	2.05	2.00	-30.9
Cubic	2.22	-31.3	-30.8	2.62	2.51	2.22	-30.0

*Candidate model with lowest AIC

small sample bias adjustment of Bedrick and Tsai, 1994) can be written as

$$\hat{b} = (X^T X)^{-1} X^T y$$

$$\hat{\sigma} = y^T \left(I - X(X^T X)^{-1} X^T \right) y \quad (10)$$

$$AIC = \log(\hat{\sigma}) + \frac{n+p}{n-p-2}$$

As a practical illustration of the application of the AIC-based model selection rule, we can look at a synthetic data set corresponding to exclusive P limitation with quadratic time effect XP(2). If we present this data set to all 21 candidate models, we get the values shown in Table 2 for the AIC criterion. We see that a decision rule prescribing to choose the model with the lowest AIC would have selected the correct model in this case. We also see that models like P1(2) and C1(2) give practically the same goodness-of-fit, but are discriminated against because they contain more parameters than XP(2).

The reason why models not involving P limitation rank so poorly is illustrated in Fig. 3. In the XP model P + 0 and N + P treatments are indistinguishable, whereas the 0 + N treatment is identical to the control. In the case of the XC model, all 0 + 0, P + 0, and 0 + N treatments are bound together, making it impossible at the same time to let the P + 0 treatment follow the N + P while the 0 + N treatment follows the control. The result of this compromise leads to bad fit for all the 0 + 0, P + 0, and 0 + N treatments, a high error variance, and, correspondingly, a high AIC. In the extreme case of fitting a pure N limitation model (XN) to a pure P limitation data set (XP), all fitted values for treatments end up superimposed because the single nutrient treatments are bound completely opposite to control and full treatment.

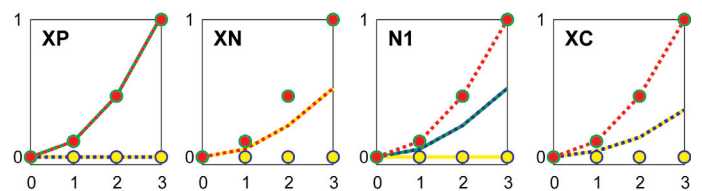


Fig. 3. Limitation patterns XP, XC, XN, and N1 fitted to data generated by an XP(2) model (same as in Table 2), with the same treatment color coding as in Fig. 2.

Methods and procedures

We will not give a rigid procedure for the performance of nutrient limitation bioassays with natural plankton communities. Rather, we will go through some critical points to consider before designing a nutrient limitation study and specify the minimum requirements for the proposed classification method to be applicable. Finally, we will describe in some detail the software tools that are provided with this contribution.

The following elements are required for performing a bioassay: (1) a water sample containing the intact planktonic community of interest. Arguments can be made for and against screening to remove large herbivores and other pre-treatments. (2) Incubation facilities for retaining suitable in situ temperature and light conditions for the duration of the experiment. Constant, artificial light with in situ light: dark cycle is preferable, but natural sunlight reduced with neutral density screening is also usable as long as all treatments are ensured equal conditions. Unless light is considered as a treatment factor, light intensity should be nonlimiting and noninhibiting to algal growth. (3) Nutrient solutions for spiking the sub-samples with fixed amounts according to the experimental design. Nutrient additions should be sufficient to potentially produce a response clearly distinguishable from the control, but low enough that chemical equilibria are not disturbed. For P and N treatments, we suggest the following a rule of thumb: P addition (as PO_4^{3-}) equivalent to the initial total P, with N addition (as NH_4^+) in Redfield proportions to this amount (16:1, by atoms). (4) Laboratory equipment for preparing and analyzing the variable(s) chosen to represent the community response. Parameters specific to phytoplankton, which also can be measured with high precision and sensitivity at reasonable ease, are to be preferred. Chlorophyll *a* is a natural choice, but other options are possible. The accumulated water volume needed to analyze the chosen response parameters sets a lower limit to the size of the experimental units: at least half of the initial volume should be left at the termination of the experiment.

The present classification software can only handle two-factor limitation experiments, although future releases may change this. The classification procedure depends on all treatment combinations being present in equal numbers in the experimental design, and that this structure is maintained throughout the experiment. All treatment combinations must be, at least, duplicated. Unequal replication of treatment combinations is possible but will require software modifications that are not treated here. As the software is presently unable to handle missing values, some safeguarding against accidentally lost response measurements is advised. Taking duplicate samples from each experimental unit can be a cost-efficient precaution against missing values. Due to potential problems of pseudo-replication, such bottle-level duplicates should not be used directly in the analysis. But they can be used for investigating error variance components, general quality assurance,

and for resampling-based robustness checks (Tamminen and Andersen, in press). The classification procedure requires equal sampling intervals, at least to the extent that variability between sampling times must be less than 5% of the sampling interval (i.e., a variability of ± 1 h is acceptable for daily samples). Allowing for unequal sampling times is possible, but will require major software modifications. Daily sampling is usually suitable for natural phytoplankton communities, but other intervals can be considered under, e.g., particularly low or high temperatures.

Details on software implementation of the theory presented here are given in Web appendix 1, with two Multimedia Appendices containing (1) a modifiable Matlab source code and (2) a ready-to-run application for one particular design implemented as an Excel add-in generated with the Matlab Excel Builder.

Assessment

We use a subset of the data reported in Tamminen and Andersen (in press) for assessing the potentials of the method described here. Briefly, the data set consists of 163 nutrient limitation bioassay experiments performed over a 3-year period at 6 locations along the coast of Finland (NE Baltic Sea). In each experiment, a natural water sample was divided among 8 experimental units, spiked with P (PO_4^{3-}) and N (NH_4^+) in a duplicated 2^2 factorial design, and sampled daily under incubation at in situ temperature and photo period. We will only consider Chlorophyll *a* as response parameter, which was sampled in duplicate from each unit for four successive days, including the initial value. Chlorophyll *a* was measured spectrofluorometrically on ethanol extracts of material collected on glass fiber filters (GF/F), using pure Chlorophyll *a* (Sigma) as standard. See Tamminen and Andersen (in press) for further details on experimental details and quality assurance procedures.

All 21 candidate models get selected at least once when we run the proposed classification procedure on the whole data set. By selecting one experiment as representative for each candidate model, we can compile a 7 by 3 matrix of graphs illustrating each of the 21 possibilities. The overall impression from inspecting Fig. 4 is that the unsupervised classification procedure is largely in accordance with the subjective impression one gets from looking at the different responses. The most controversial cases are perhaps the cases classified as primary combined C1(1–3). A human operator would perhaps rather have classified the C1(3) case (Kokkola district, week 32–92) as XC(3) and the C1(2) case (Kymi district, week 33–93) as XN(2). Such ambiguities seem to be unavoidable with any automated classification procedure, but we find the number of such cases to be small, and their presence more than weighted up by the inherent objectivity of unsupervised classification. Ambiguities can even be quantified by doing classifications on repeated resamplings of the original data to get probabilities of different limitation classes, as illustrated

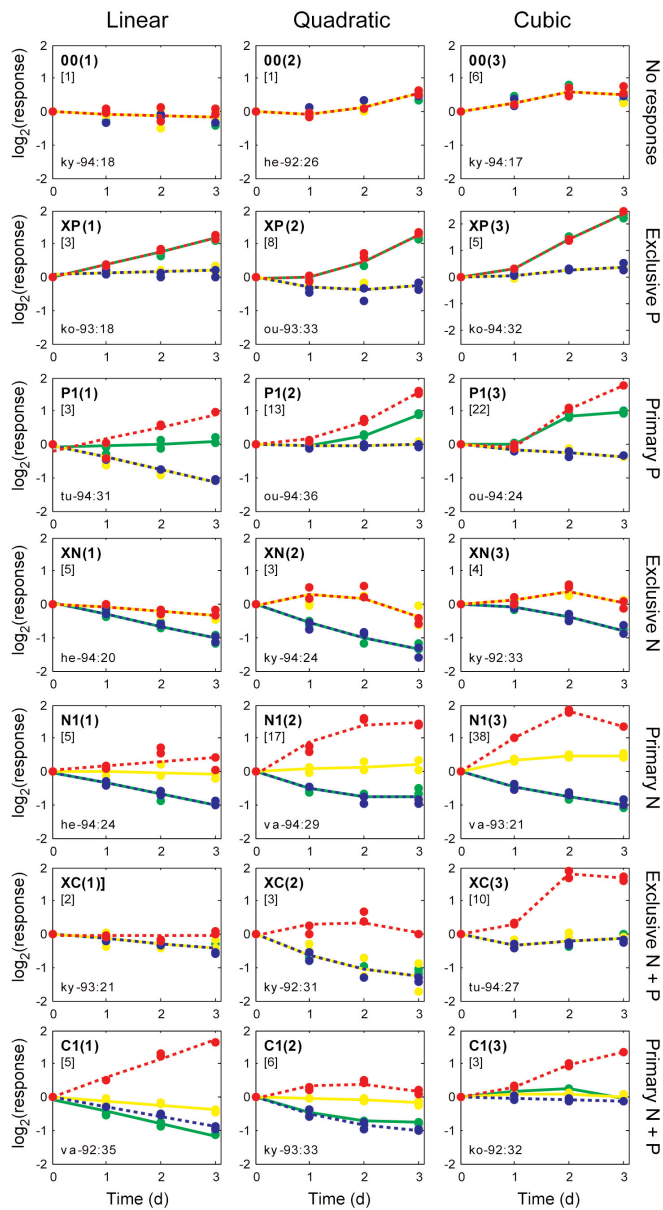


Fig. 4. Examples of different nutrient addition responses in plankton communities along the coast of Finland, selected from 161 nutrient limitation bioassays using Chlorophyll a as response parameter (see Tamminen and Andersen, in press, for details). Limitation patterns in rows and time effect orders in columns, using the same treatment color coding as in Fig. 2. Responses are shown scaled to the initial condition and log-transformed. The code in the upper left corner represents limitation pattern (as in Eq. 9) and time effect order, with the total number of cases with this response in parentheses below. The identification code in the lower left corner refers to location, year, and week of the experiment (Ou: Oulu, Ko: Kokkola, Va: Vaasa, Tu: Turku, He: Helsinki, Ky: Kymi, see Tamminen and Andersen, in press, for details).

in the Web appendix 1, and more fully developed in Tamminen and Andersen (in press).

The proposed model selection classification procedure only considers whether different treatments behave identically or not. It does not take into account the magnitudes and signs of

differences between treatments, which some would consider a weakness, for example that N should be more limiting than P in the C1(2) case in Fig. 4 (Kymi district, week 33–93) because there is a stronger response to N than to P, although both differ from the control. We think this is just a matter of interpretation: if we assume the plankton community in this particular case to be composed of 2 functional groups, of which one is N-limited and fast-growing and the other P-limited with lower growth rate or initial biomass, then the observed pattern makes perfect sense. But it would be impossible to say whether one of the groups was more limited than the other, without knowledge of their respective growth and loss rates in the initial community. As such, we think it defensible to say that the community as a whole was limited by both nutrients, which is in line with the automated classification.

The C1(1) example in the lower left corner of Fig. 4 (Vaasa district, week 35–92), is perhaps what most people would consider most controversial, since the automated classification cannot take into account that the P alone treatment actually produces less chlorophyll response than the control. Some would consider such a response erratic or inconsistent, and just dismiss the experiment. As pointed out by Thingstad et al. (2005) this response is perfectly logical if bacteria are P-limited but have access to an organic N source, which is unavailable to the phytoplankton which is N-limited. In this respect, it is notable that Tamminen and Andersen (in press) observe several instances this below-control response for P but none for N.

Discussion

Several approaches have been proposed for systematic classification of nutrient limitation bioassays. Morris and Lewis (1988) used a 2^2 factorial N and P design with 3 replicated treatments and single endpoint measurement after 4–7 d of incubation. Their design was analyzed by one-way analysis of variance (ANOVA) and post-hoc tests for significant treatment differences. From this, they could differentiate 4 limitation classes in addition to a no-response class (no limitation), corresponding to a nonsignificant F-test in the ANOVA. Morris and Lewis (1988) distinguished between two types of combined N and P responses: concomitant and reciprocal N and P limitation, which correspond to what we have termed exclusive (XC) and primary (C1) combined responses.

Fisher and Gustafson (2002), in their impressive > 10-year study of nutrient limitation in Chesapeake Bay, most of the time used a 2^2 factorial N and P design with a single end-point measurement after 2–5 d of incubation. As their design had no replication except for the control, which was duplicated, they could not apply any classification scheme based on standard ANOVAs or other linear models. Instead they used a predefined decision tree scheme based on exceeding a fixed, season-specific percentage of the control treatment. Fisher and Gustafson (2002) differentiated between 2 no-response classes (inconsistent and no limitation) and 5 limitation classes. They

distinguished between exclusive and primary limitation responses depending on whether the nonlimiting nutrient has any effect or not.

Our nutrient addition response classification combines two features of Morris and Lewis (1988) and Fisher and Gustafson (2002) in that we distinguish between exclusive and primary effects for both combined and single nutrients. The former is important because it separates two qualitatively contrasting situations when either the whole phytoplankton community is close to balanced limitation, or when different community components are limited by different nutrients. This distinction corresponds to what Arrigo (2005) terms multi-nutrient and community colimitation in a recent review. The latter distinguishes between the classical, textbook style of exclusive limitation where the nonlimiting nutrient has no effect whatsoever, and the situation commonly encountered in practice where addition of a single nutrient soon forces the community into limitation by the initially nonlimiting one. It can be argued that such secondary limitation of the initially nonlimiting nutrient is an artifact imposed by containing and perturbing the community. Tamminen and Andersen (in press) found a decreasing fraction of cases of exclusive limitation with incubation time, suggesting that exclusive and primary limitation can both be considered indicators of single-nutrient limitation in the initial community, even though they differ in the extent to which the non-limiting nutrient is truly in excess.

The proposed method is completely self-contained in the sense that it bases decisions only on information contained in the data themselves (i.e., the variation between replicated treatments on the same day). Thus it avoids the introduction of external, ad hoc-based decision rules as in Fisher and Gustafson (2002), as well as the potential pit-falls of multiple comparisons, as used by Morris and Lewis (1988): there are many ways of doing multiple comparisons, and most of them are too optimistic and often mutually inconsistent (e.g., Cook and Farewell, 1996).

Compared to single endpoint approaches, which can be analyzed by standard ANOVAs, the method proposed here is designed to handle repeated measurements. This is important because of the inescapable conflict between keeping the incubation time short and getting a high signal-to-noise ratio, and because we cannot know in advance how long a plankton community will need to produce a detectable response. Cascading responses on higher trophic levels can efficiently consume an initial increase in primary producer biomass, indicating no limitation in a single-endpoint design while showing a clear limitation effect when measurements are repeated over time.

The presentation here has been focused on 2-factor experiments, mainly because this is a design which we have been using extensively in practice (Tamminen and Andersen, in press). Still, as mentioned above, there are no limitations to extending the model selection procedure to more than 2 limiting factors. If we consider an experiment with 3 potentially

limiting nutrients (for example N, P, and Si or N, P, and Fe) there will be $2^3 = 8$ treatment combinations in a full factorial design, and at least twice that amount for necessary treatment replication. The treatment effect matrix corresponding to Eq. 2 will now be an 8 by 8 matrix ($M = M_1 \otimes M_2 = M_1 \otimes M_1 \otimes M_1$), and a number of possible non-zero parameter patterns a baffling $2^8 = 256$ (compared to $2^4 = 16$ for the 2-factor model). Fortunately, it turns out that the biologically relevant ones are much fewer and that many of them are redundant (functionally identical), resulting in altogether 15 candidate models (details available upon request). In other words, doubling the number of treatments from 4 to 8 will only increase the number of biologically relevant candidate models from 7 to 15.

The proposed method uses inference based on model selection rather than the hypothesis testing approach of classical statistics. As such, it follows a general trend in modern statistics (Burnham and Anderson 2002). The strength of this approach in the present context is that it efficiently separates nuisance parameters due to time effects from the treatment effects, and that the treatment effects can be directly classified as biologically interpretable responses. The classification is objective in the sense that it blindly follows the model selection criterion, even if the AIC of competing models may be arbitrarily close. Burnham and Anderson (2002) recommends as a rule of thumb that model selection should be considered unequivocal when there is an AIC difference of at least two between the best and next-best model. Our experience is that sometimes this criterion is not fulfilled, and that the resulting ambiguity is also evident by inspection of the model fits. While we hardly can argue against using common sense and visual inspection of the model fits in general, this easily becomes unwieldy in data sets of > 100 experiments and 21 candidate models, as in Tamminen and Andersen (in press).

We have found it illustrative to assess the ambiguity in the resulting classification by resampling methods, either by random selection from measurement level replicates or by split-plot type sampling with replacement from replicated treatments. The former, which is shown in Multimedia Appendix 2 and also employed by Tamminen and Andersen (in press), would imply an element of pseudo-replication, but practical experiences show similar results for the two approaches when treatment-level noise and measurement noise are of similar magnitude. The results of a resampling analysis can be seen as a voting system, where an unambiguous experiment would be recognized by having the votes concentrated on one or a few closely related models (say, XP[2] and XP[3]), while an ambiguous experiment would have votes distributed between contradicting models. Tamminen and Andersen (in press) show how these voting probabilities can be aggregated to express the uncertainty in inferences on N versus P limitation over seasonal and spatial scales.

References

- Abramowitz, M., and I. A. Stegun. 1974. Handbook of mathematical functions. Dover.
- Arrigo, K. R. 2005. Marine microorganisms and global nutrient cycles. *Nature* 437:349-355.
- Bedrick, E. J., and C. L. Tsai. 1994. Model selection for multivariate regression in small samples. *Biometrics* 50:226-231.
- Box, G. E. P., and N. R. Draper. 1987. Empirical model-building and response surfaces. Wiley.
- Burnham, K. P., and D. R. Anderson. 2002. Model selection and multi-model inference: a practical information-theoretic approach. Springer.
- Cox, D. R., and N. Reid. 2000. The theory of the design of experiments. CRC Press.
- Cook, R. J., and V. T. Farewell. 1996. Multiplicity considerations in the design and analysis of clinical trials. *J. Royal Stat. Soc. A* 159:93-110.
- Droop, M. R. 1974. The nutrient status of algal cells in continuous culture. *J. Mar. Biol. Assoc. U.K.* 54:825-855.
- Elser, J. J., E. R. Marzolf, and C. R. Goldman. 1990. Phosphorus and nitrogen limitation of phytoplankton growth in the fresh-waters of North America—a review and critique of experimental enrichments. *Can. J. Fish. Aquatic Sci.* 47(7): 1468-1477.
- Fisher, T. R., and A. F. Gustafson. 2002. Nutrient-addition bioassays in Chesapeake Bay to assess resources limiting algal growth. http://www.dnr.state.md.us/bay/monitoring/limit/2002_level1_report.pdf
- Hilborn, R., and M. Mangel. 1997. The ecological detective: Confronting models with data. Princeton Univ. Press.
- Howarth, R. W. 1988. Nutrient limitation of net primary production in marine ecosystems. *Ann. Rev. Ecol.* 19:89-110.
- Kivi, K., et al. 1993. Nutrient limitation and grazing control of the Baltic plankton community during annual succession. *Limnol. Oceanogr.* 38(5):893-905.
- Lignell, R., J. Seppala, P. Kuuppo, T. Tamminen, T. Andersen, and I. Gismervik. 2003. Beyond bulk properties: Responses of coastal summer plankton communities to nutrient enrichment in the northern Baltic Sea. *Limnol. Oceanogr.* 48(1):189-209.
- Morris, D. P., and W. R. Lewis. 1988. Phytoplankton nutrient limitation in Colorado mountain lakes. *Freshwater Biol.* 20:315-327.
- Tilman, D. 1982. Resource competition and community structure. Princeton Univ. Press.
- Thingstad, T. F., et al. 2005. Nature of phosphorus Limitation in the ultraoligotrophic Eastern Mediterranean. *Science* 309: 1068-1071.

Submitted 6 November 2005

Revised 19 June 2006

Accepted 22 August 2006