Vol. 501: 291–296, 2014
doi: 10.3354/meps10705

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published March 31

NOTE

Testing for simple structure in a spatial time series with an application to the distribution of *Alexandrium* resting cysts in the Gulf of Maine

Andrew R. Solow[1,*], Andrew R. Beet[1], Bruce A. Keafer[1], Donald M. Anderson[1]

[1]Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

ABSTRACT: We describe a test of the goodness of fit of a simple model for the mean of a spatial time series. Under this model, the mean consists of a temporally varying scaling of a fixed spatial map. The test is applied to the distribution of cysts of the harmful algal bloom organism *Alexandrium fundyense* in the top 1 cm of sediment in the Gulf of Maine over the period 2004 to 2012. The analysis, which was motivated by a practical problem encountered when designing an annual cyst survey, supports this simple model. The test has broad applicability in marine ecology and related fields.

KEY WORDS: *Alexandrium fundyense* · Cyst distribution · Goodness of fit test · Sampling design · Space-time interaction

INTRODUCTION

Data having the form of measurements over time at a fixed collection of spatial locations are common in marine ecology and related fields. Efficient designs for collecting such data and appropriate methods for their subsequent analysis depend on the nature of the underlying variability among the data. In the present paper, we describe a goodness of fit test of a particularly simple form of space-time variability and apply it to the distribution of the harmful algal bloom (HAB) organism *Alexandrium fundyense* — commonly referred to as red tide — in the Gulf of Maine over the period from 2004 to 2012. As described below, this analysis was motivated by a practical problem in sample design. Although the focus here is on this specific problem, the statistical approach has the potential for a broader applicability.

In the Gulf of Maine, blooms of the toxic dinoflagellate *Alexandrium fundyense* can cause paralytic shellfish poisoning, shellfish-harvest closures, and mortality in fish and other animals, all with substantial economic costs (Hoagland & Scatasta 2006). There has been considerable progress in recent years in predicting seasonal *Alexandrium* blooms in the Gulf of Maine using coupled biological-physical models (e.g. He et al. 2008, McGillicuddy et al. 2011). An important input to these predictive models is the autumn distribution in the sediment of resting *Alexandrium* cysts that provide an inoculum for the subsequent spring bloom. To predict the spring *Alexandrium* bloom, the cyst distribution has been estimated each year from cyst counts in sediment samples collected during the previous autumn at a set of stations covering the region from southern Maine to the Bay of Fundy (Fig. 1). These annual cyst surveys are costly, and there is interest in reducing the number of sampling stations and using the reduced set to estimate the density at unsampled stations. Designing a cyst survey requires an under-

*Corresponding author: asolow@whoi.edu

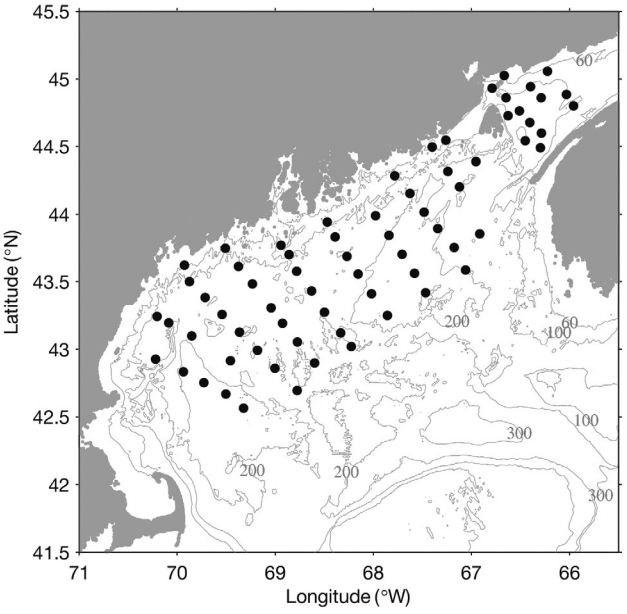

Fig. 1. Locations of cyst-sampling stations

standing of the spatial and temporal variability in cyst density. The present paper reports some statistical work aimed at evaluating a simple model of the cyst distribution in the Gulf of Maine that allows an economical survey design.

MODEL AND METHODS

Let N_{jt} be the observed cyst count in a unit sample volume at Stn j ($j = 1, 2, \ldots, J$) in Year t ($t = 1, 2, \ldots, T$), and let $Y_{jt} = \log N_{jt}$. In the application below, we will use the natural logarithm. Interest here centers on the validity of the space-time separable model:

$$Y_{jt} = \beta_j + \gamma_t + \varepsilon_{jt} \quad (1)$$

where $\beta_1, \beta_2, \ldots, \beta_J$ represent a discrete spatial map that does not depend on t, $\gamma_1, \gamma_2, \ldots, \gamma_T$ represents a series of annual factors common to all stations, and ε_{jt} is a normal error with mean 0 and variance σ^2 that does not depend on j or t. This error includes both measurement error and natural variability. For identifiability, we impose the constraint $\sum_{j=1}^{J} \beta_j = 0$. Under this model, the mean spatial map of cyst density in each year is an annually varying multiplicative scaling of a fixed spatial map. Briefly, this model would be appropriate if the total supply of cysts varied from year to year but the process by which they are distributed spatially is, on average, the same in each year. As discussed below, this model is of interest because it allows the prediction of the cyst density map in the current year $T + 1$ using an estimate of $\beta_1, \beta_2, \ldots, \beta_J$ based on historical data and as few as a single sample from the current year to estimate γ_{T+1}.

Although it is written slightly differently, the model in Eq. (1) corresponds to the classical 2-way analysis of variance (ANOVA) with no interaction and no replication (i.e. 1 sample at each station in each year). Assessing the validity of this model can be formulated as testing it against the non-separable model:

$$Y_{jt} = \beta_j + \gamma_t + \delta_{jt} + \varepsilon_{jt} \quad (2)$$

where δ_{jt} is, in the terminology of ANOVA, an interaction term at Stn j in Year t with, for identifiability, $\sum_{j=1}^{J} \delta_{jt} = 0$ for all t and $\sum_{t=1}^{T} \delta_{jt} = 0$ for all j. Formally, the problem is to test the null hypothesis H_0 that $\delta_{jt} = 0$ for all j and t against the general alternative hypothesis H_1 that H_o is false.

It is well known that the standard test for interaction in a 2-way ANOVA is not applicable when there is no replication, and specialized methods are needed (Alin & Kurt 2006). Briefly, these methods fall into 2 broad categories. Parametric methods are based on a specific model of the interaction terms. The best known of these is Tukey's method in which δ_{jt} is assumed to be proportional to the product $\beta_j \gamma_t$ (Tukey 1949). In the absence of a parametric model for interaction, nonparametric methods can be used. These methods are based on identifying structure in the residuals from fitting the model in Eq. (1) that indicate the presence of interaction. Here, we will use the so-called locally best invariant test proposed by Boik (1990, 1993).

For the model in Eq. (1), the point estimate of β_j is:

$$\hat{\beta}_j = \sum_{t=1}^{T} Y_{jt}/T - \sum_{j=1}^{J}\sum_{t=1}^{T} Y_{jt}/JT \quad (3)$$

the point estimate of γ_t is:

$$\hat{\gamma}_t = \sum_{j=1}^{J} Y_{jt}/J \quad (4)$$

and the estimate of σ^2 is:

$$\hat{\sigma}^2 = \sum_{j=1}^{J}\sum_{t=1}^{T} (Y_{jt} - (\hat{\beta}_j + \hat{\gamma}_t))^2 \Big/ (JT - J - T + 1) \quad (5)$$

Define the J-by-T matrix $R = [R_{jt}]$ of residuals, where the following relation:

$$R_{jt} = Y_{jt} - (\hat{\beta}_j + \hat{\gamma}_t) \quad (6)$$

is the residual for Stn j in Year t. Let $\Lambda_1 > \Lambda_2 > \ldots > \Lambda_p$ be the ordered non-zero eigenvalues of $R'R$, where $'$ denotes matrix transpose and $p = \min(J-1, T-1)$, and let $\lambda_k = \Lambda_k \Big/ \sum_{k=1}^{p} \Lambda_k$ be the kth standardized eigen-

value. The test statistic proposed by Boik (1990, 1993) is:

$$S = (1 + p\sum_{k=1}^{p}(\lambda_k - p^{-1})^2)^{-1} \quad (7)$$

with the null hypothesis rejected for small values of S. Briefly, $R'R$ contains unscaled residual covariances and, under the null hypothesis, should be proportional to the p-by-p identity matrix whose standardized eigenvalues are all equal to p^{-1}. The distribution of S under the null hypothesis is complicated, and provided $q = \max(J-1, T-1)$ is large, it is more convenient to use as a test statistic:

$$S_1 = (S^{-1} - 1)/(p-1) \quad (8)$$

which under the null hypothesis has an approximately chi-squared distribution with $(p+2)(p-1)/2$ degrees of freedom (Sugiura 1972). The null hypothesis is rejected for large values of S_1. The power of this test was discussed by Boik (1993), who showed that it compares favorably to other nonparametric tests.

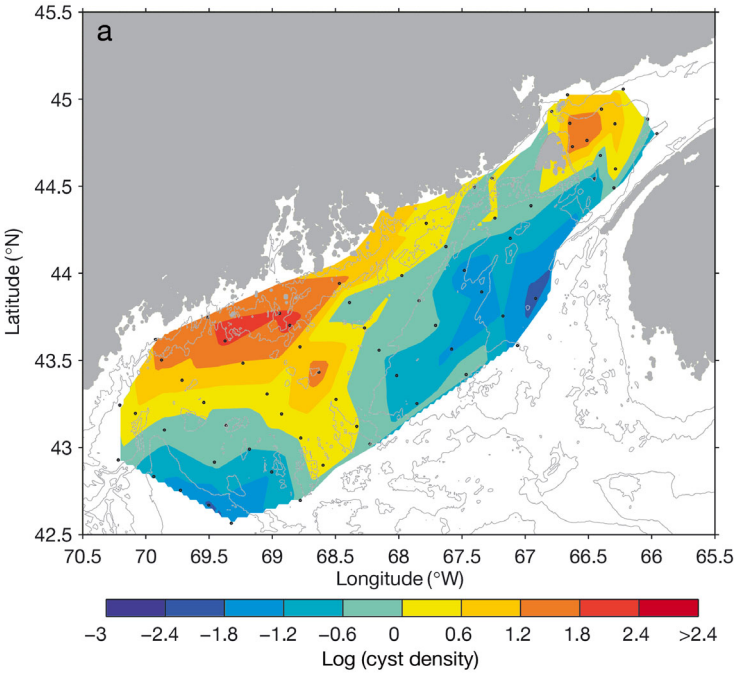

Fig. 2. *Alexandrium fundyense*. Estimated mean spatial map

Application

We applied the test described in the previous section to data from annual autumn cyst surveys in the Gulf of Maine for the $T = 9$ yr period from 2004 to 2012. The data in each year consist of cyst counts in samples of the top 1 cm of sediment at each of the $J = 68$ stations shown in Fig. 1. Details of the sampling are described by Anderson et al. (2005). We fit the separable model (Eq. 1) to these data as outlined above. To accommodate occasional zeros, we added 1 to each cyst count. The estimated mean spatial map $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_J$ is shown in Fig. 2, and the time series of estimated annual effects $\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_T$ is shown in Fig. 3. The estimate $\hat{\sigma}$ of the standard deviation of the error is 0.76. The value of the test statistic S_1 is 0.04 with observed significance level (or p-value) of essentially 1, so that the separable model cannot be rejected. The fitted model explains 70% of the variance in log cyst density ($R^2 = 0.70$).

In Fig. 4, the estimated maps of mean cyst density are shown for each year along with the observed maps. The fitted values are given by the following relation:

$$\hat{N}_{jt} = \exp(\hat{\beta}_j + \hat{\gamma}_t + \hat{\sigma}^2/2) \quad (9)$$

with the extra term $\hat{\sigma}^2/2$ arising from the expression for the mean of a lognormal random variable. The fitted and observed maps are in good agreement. The test for interaction is known to be sensitive to correlation among the errors ε_{jt}. To check for correlation, we formed a sample variogram of the residuals. The sample variogram, which is a standard measure of spatial correlation (Cressie 1993), is shown in Fig. 5 and shows no evidence of spatial correlation. It is, of course, possible that spatial correlation in cyst density exists at smaller spatial scales, but this is irrelevant to this application. We conclude that the separable model is a good one for these data.

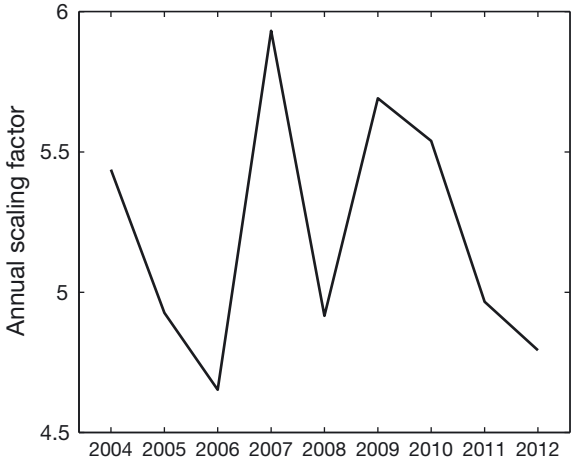

Fig. 3. Estimated annual scaling factors

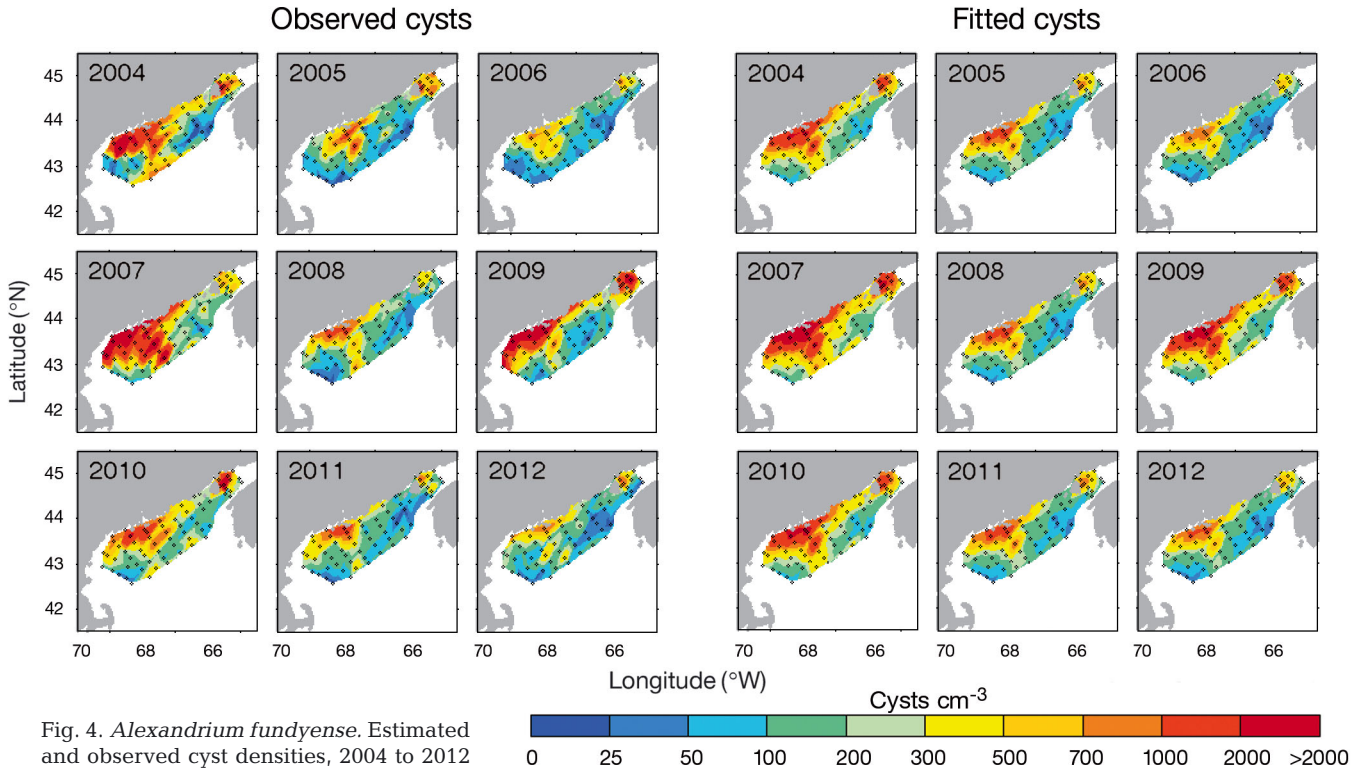

Fig. 4. *Alexandrium fundyense*. Estimated and observed cyst densities, 2004 to 2012

Using the separable model

As noted, our interest in the separable model arose from its potential to reduce annual sampling effort. If the non-separable model in Eq. (2) holds, then information about the interaction term $\delta_{j,T+1}$ needed for bloom prediction in Year $T+1$ comes only through observing $N_{j,T+1}$. However, if the separable model in Eq. (1) holds, the situation is different. In this case, historical data can be used to estimate the fixed station effect β_j and a subset of stations can be sampled in Year $T+1$ to estimate the common annual effect γ_{T+1}.

For example, suppose that the original sample has been supplemented with observations $N_{k,T+1}$ with $k =$ 1, 2, …, K and $K < J$. No generality is lost by labeling the stations so that the first K correspond to the subset that are sampled in Year $T+1$. The data now have the form of an unbalanced 2-way ANOVA without replication. Although the parameter estimates no longer have a simple form, they can be found by regressing Y_{jt} against binary regressors representing station and year with the side condition that $\sum_{j=1}^{J}\beta_j = 0$ (Monahan 2008). Details are provided in the Appendix. The estimated mean cyst density at unsampled Stn j in Year $T+1$ is determined as follows:

$$\hat{N}_{j,T+1} = \exp(\hat{\beta}_j + \hat{\gamma}_{T+1} + \hat{\sigma}^2/2) \quad (10)$$

Note that the observations in Year $T+1$ are being used to update the estimates of both the spatial map of mean density and the annual effects in previous years. As outlined in the Appendix, it is also possible to construct a prediction interval for cyst density at an unsampled location.

As an illustration, Fig. 6 shows estimated cyst maps for 2012 based on 2 different subsets of $K = 26$ stations. The first subset (Fig. 6a) consists of stations along 4 widely spaced cross-shore transects, while the second (Fig. 6b) consists of 2 along-shore transects.

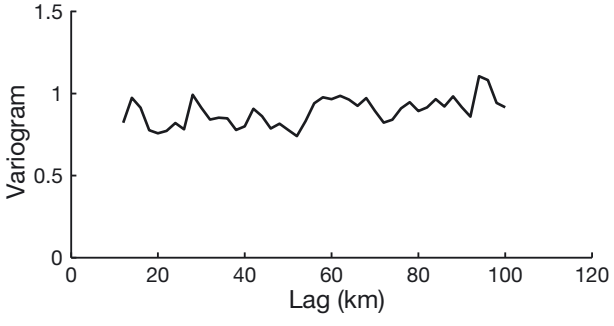

Fig. 5. Sample variogram of residuals from the fitted model

DISCUSSION

The main contribution of this study has been to describe a test for space-time separability in unreplicated spatial time series data and to apply it to a practical problem of survey design in marine ecology. As noted, although we have focused on a specific application, the test seems to have broad applicability in marine ecology and related fields. Potential examples include the analysis of spatial time series of abundance data to detect a shift in the range of a species and of chemical measurements to differentiate between local and global sources.

Turning to our specific application, the statistical properties of the estimate $\hat{\gamma}_{T+1}$ of the annual effect in Year $T+1$ depend only on the number K of stations sampled in Year $T+1$ and not on their location. Although this estimate improves as K increases, the returns are diminishing. Results not presented here indicate that the improvement is only marginal as K increases beyond 20. As ship-time represents a significant portion of the cost of cyst survey, these results favor sampling stations that are close together. However, there is no guarantee that the stability of the spatial map over the 9 yr period considered here will persist over a longer period. It would therefore be prudent to continue to sample in different parts of the overall region to allow the identification of future changes in the spatial structure of the cyst distribution.

Acknowledgements. Helpful comments from D. McGillicuddy and 3 anonymous reviewers are acknowledged with gratitude. This work was supported in part by NOAA Grants NA06NOS4780245 and NA09NOS4780193, NSF Grant OCE-0934653, and the State of Maine.

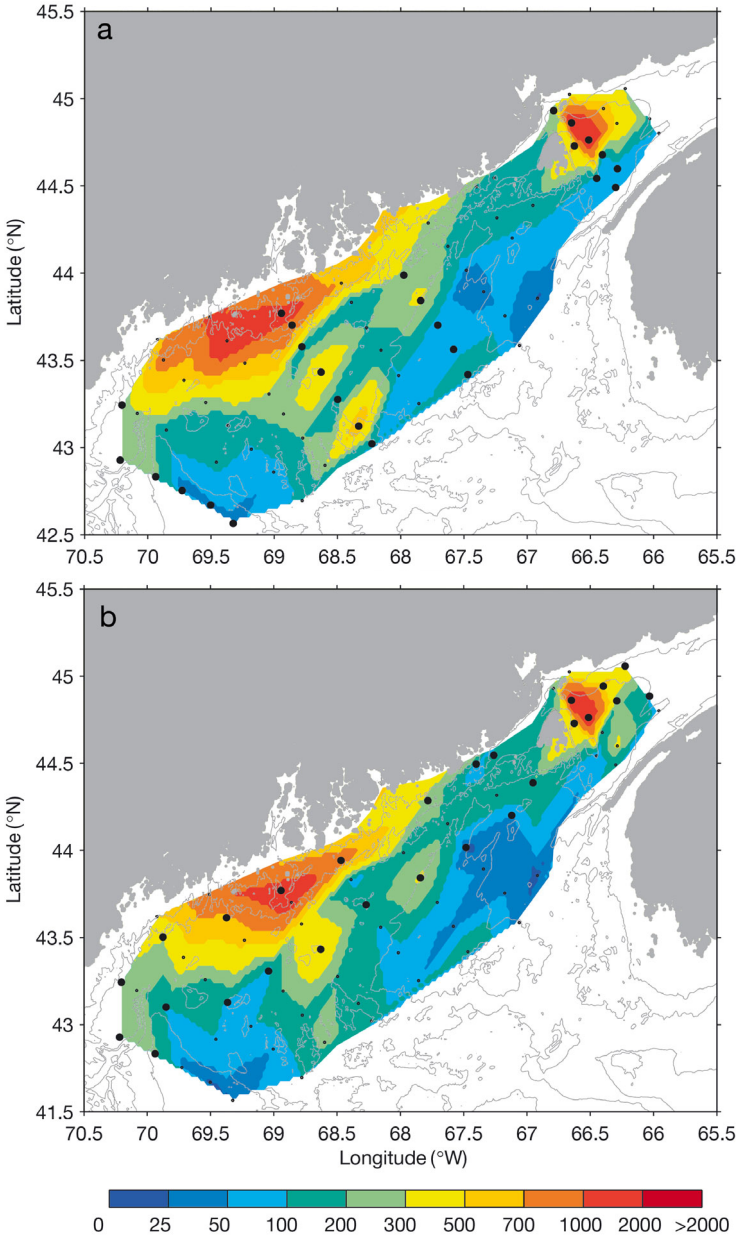

Fig. 6. *Alexandrium fundyense.* Estimated 2012 cyst density based on 2 different subsets of 26 sampling stations indicated by large dots

LITERATURE CITED

Alin A, Kurt S (2006) Testing non-additivity (interaction) in two-way ANOVA tables with no replication. Stat Methods Med Res 15:63–85

Anderson DM, Stock CA, Keafer BA, Nelson AB and others (2005) *Alexandrium fundyense* cyst dynamics in the Gulf of Maine. Deep-Sea Res II 52:2522–2542

Boik RJ (1990) Inference on covariance matrices under rank restrictions. J Multivariate Anal 33:230–246

Boik RJ (1993) A comparison of three invariant tests of additivity in two-way classifications with no replications. Comput Stat Data Anal 15:411–424

Cressie NAC (1993) Statistics for spatial data. Wiley, New York, NY

He R, McGillicuddy DJ, Keafer BA, Anderson DM (2008) Historic 2005 toxic bloom of *Alexandrium fundyense* in the western Gulf of Maine. 2. Coupled biophysical numerical modeling. J Geophys Res 113:C07040, doi: 10.1029/2007JC004602

Hoagland P, Scatasta S (2006) The economic effects of harmful algal blooms. In: Granéli E, Turner JT (eds) Ecology of harmful algae. Springer-Verlag, Berlin, p 391–401

McGillicuddy DJ, Townsend DW, He R, Keafer BA and others (2011) Suppression of the 2010 *Alexandrium fundyense* bloom by changes in physical, biological, and

chemical properties of the Gulf of Maine. Limnol Oceanogr 56:2411–2426

Monahan JF (2008) A primer on linear models. Chapman & Hall/CRC Press, London/Boca Raton, FL

Sugiura N (1972) Locally best invariant test for sphericity and the limiting distributions. Ann Math Stat 43:1312–1316

Tukey JW (1949) One degree of freedom for additivity. Biometrics 5:232–242

Appendix 1

This appendix outlines the fitting of the separable model and the elements of inference under it when not all stations are necessarily sampled in all years. The results presented here are standard and can be found in Monahan (2008) and other texts on linear models.

The model:

$$Y_{jt} = \beta_j + \gamma_t + \varepsilon_{jt} \quad \text{(A.1)}$$

with the side condition $\sum_{j=1}^{J} \beta_j = 0$ can be written in matrix form as follows:

$$Y = X\delta + \varepsilon \quad \text{(A.2)}$$

where Y is a vector of length n containing the observed log cyst densities, and ε is the corresponding vector of errors. In the balanced case, $n = JT$. For convenience, suppose that Y is formed by stacking the time series of observations at the J stations (so that, in the balanced case, the first T elements of Y are $Y_{11}, Y_{12}, \ldots, Y_{1T}$ and so on) and let $\delta = (\beta_1 \beta_2 \ldots \beta_{J-1} \gamma_1 \gamma_2 \ldots \gamma_T)'$, where $'$ denotes transpose. The row of the n-by-$(J - 1 + T)$ matrix X corresponding to Y_{jt} with $j \neq J$ consisting of zeros except for 1's in columns j and $J - 1 + t$. To satisfy the side condition, when $j = J$, the first $J - 1$ elements of this row are all equal to −1 with the remaining elements all 0 except again a 1 in column $J - 1 + t$.

The point estimates of δ and the error variance σ^2 are as given in Eq. (A.3) and (A.4):

$$\hat{\delta} = (X'X)^{-1} X'Y \quad \text{(A.3)}$$

$$\hat{\sigma}^2 = (Y - X\hat{\delta})'(Y - X\hat{\delta}) \big/ (n - (J - 1 + T)) \quad \text{(A.4)}$$

The variance matrix of $\hat{\delta}$ is the following:

$$\mathrm{Var}(\hat{\delta}) = \sigma^2 (X'X)^{-1} \quad \text{(A.5)}$$

from which confidence intervals for the elements of δ can be constructed.

Finally, the predicted log cyst density at unsampled station j_o in year t_o is:

$$\hat{Y}_o = x_o' \hat{\delta} \quad \text{(A.6)}$$

where, for $j_o \neq J$, x_o is a vector of length $J - 1 + T$ with all elements 0 except 1 in rows j_o and $J - 1 + T_o$. For $j_o = J$, the first $J - 1$ elements of x_o are −1, and the rest are 0 except 1 in row $J - 1 + T_o$. The variance of the prediction error $Y_o - \hat{Y}_o$ is as follows:

$$\mathrm{Var}(Y_o - \hat{Y}_o) = \sigma^2 (1 + x_o' (X'X)^{-1} x_o) \quad \text{(A.7)}$$

Let $c_o = (1 + x_o'(X'X)^{-1} x_o)$. An approximate $1 - \alpha$ prediction interval for Y_o is $\hat{Y}_o \pm t_{n-(J-1+T)}(\alpha/2)\hat{\sigma} c_o^{1/2}$, where $t_{n-(J-1+T)}(\alpha/2)$ is the upper $(\alpha/2)$-quantile of the t distribution with $n - (J - 1 + T)$ degrees of freedom. The endpoints of a $1 - \alpha$ prediction interval for $N_o = \exp(Y_o)$ are found by exponentiating the corresponding endpoints for Y_o.

Editorial responsibility: Katherine Richardson, Copenhagen, Denmark

Submitted: June 7, 2013; Accepted: December 17, 2013
Proofs received from author(s): February 20, 2014