Using dynamic factor analysis to show how sampling resolution and data gaps affect the recognition of patterns in limnological time series

Rosana Aguilera,1*† David M. Livingstone,2 Rafael Marcé,1 Eleanor Jennings,3 Jaume Piera,4 and Rita Adrian5,6
1 Catalan Institute for Water Research (ICRA), Girona, Spain
2 Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
3 Department of Applied Sciences, Dundalk Institute of Technology, Dundalk, Ireland
4 Physical and Technological Oceanography Department, Institute of Marine Sciences (ICM-CSIC), Barcelona, Spain
5 Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
6 Free University Berlin, Department of Biology, Chemistry and Pharmacy, Berlin, Germany
† Current address: Marine Science Institute, University of California Santa Barbara, Santa Barbara, CA, USA
* Corresponding author: raguilera@icra.cat

Received 11 November 2015; accepted 23 February 2016; published 17 May 2016

Abstract

Currently, many water quality variables can be automatically monitored at subdaily frequencies, but most environmental datasets still rely heavily on comparatively low-frequency (e.g., monthly) sampling campaigns. Taking advantage of the long-term ecological data available from polymictic Lake Müggelsee (Germany) at weekly temporal scales, we tested whether dynamic factor analysis (DFA), a dimension-reduction technique especially designed for time series with data gaps, was able to reproduce the same underlying patterns if the resolution of the observations was artificially reduced from weekly to once every 2 or 4 weeks (2-weekly or 4-weekly, respectively) and whether the same results were obtained for different variables (water temperature, water chemistry, and plankton). As expected, the reduction in data resolution and the addition of gaps increased the amount of variance in our case studies; however, water temperature patterns, which generally vary only slowly if considered at broader temporal scales, were represented well in the 4-weekly data model. By contrast, total algal biomass and nutrient dynamics, which usually fluctuate more rapidly, were not. We further assessed the representativeness of spot samples taken by selection at 4-weekly and seasonal intervals compared to samples obtained by averaging the weekly data over these 2 intervals. In this case, the resulting observation variance was more pronounced in DFA models of the samples taken by selection than of those obtained by averaging. The introduction of additional artificial gaps in one univariate and one multivariate model decreased the amount of information that could be explained by the pattern(s) extracted by DFA. Overall, we found that a too low sampling frequency could be a limiting factor when studying interactions and responses to changing conditions in aquatic systems, particularly when biological variables are involved.

Key words: dynamic factor analysis, Müggelsee, sampling resolution, time series

Introduction

The Nyquist-Shannon sampling theorem implies that a continuous signal can be properly sampled only if it contains no power at frequencies above the Nyquist frequency, defined as half the sampling frequency (Zayed 1993). In other words, the sampling frequency should be at least twice the highest frequency contained in the signal. An inadequate sampling rate results in the phenomenon known as aliasing, in which power at frequencies higher...
than the Nyquist frequency is interpreted as appearing at frequencies lower than the Nyquist frequency, which results not only in a loss of information about the system but also distorts the information obtained. In limnology, these signals might represent processes and interactions between variables within a given ecosystem or region. Processes in lakes are complex, involving overlapping temporal and trophic-level scales. Thus, when measuring ecological variables in lakes, the sampling frequency and the lengths of the time series obtained are critical for capturing and understanding patterns of temporal evolution in these variables. These patterns include (but are not limited to) long-term trends and seasonal variability, which can usually be identified using lower frequency data (e.g., monthly) as well as high-frequency fluctuations, which typically require finer data resolution.

In the past few decades, studies relying mainly on weekly to once every 4 week (4-weekly) spot samples (Halliday et al. 2012) have revealed the complex nature of ecosystems response to environmental change and the time scales involved (review in Adrian et al. 2012). Such studies have encouraged the measurement of long-term, high-frequency datasets (Parr et al. 2003). The recent development and deployment of high-frequency sensor monitoring equipment and the increasing duration of observational time series available from lakes have in turn allowed more detailed study of lake responses and processes (Jennings et al. 2012, Solomon et al. 2013, Rose et al. 2014). These technological developments also guarantee that the sampling frequency will normally exceed the Nyquist frequency, avoiding any numerical artifacts associated with aliasing.

Despite such advances, most limnological datasets are still prone to uneven sampling intervals and missing observations. These sampling campaigns are, in most cases, subject to logistical and financial limitations and often do not take into account the ideal frequency requirements necessary to obtain adequate sample representativeness of the system being studied. These commonly encountered problems can represent a limitation in the study of the underlying patterns in ecosystem responses to climatic and anthropogenic forcing.

Irregularities in sampling frequency and the presence of data gaps are often the first challenges encountered when selecting an appropriate method for analyzing time series. Consequently, the choice of analytical methodology becomes crucial to taking advantage of all available information. Of the various time-series analysis techniques that can be used to study responses and patterns, few are able to cope with these challenges. For instance, Schoellhamer (2001) applied a modified singular-spectrum analysis algorithm to obtain spectral estimates from records with a large fraction of missing data. The weighted wavelet Z-transform (WWZ) method can handle unevenly sampled data, but large gaps in the data would complicate the identification of temporal patterns (Foster 1996). Another example is Seasonal Kendall trend analysis, which can cope with missing values (Hirsch et al. 1982). Nevertheless, commonly used methods such as spectral analysis, wavelet analysis, and Box-Jenkins modeling usually require time series to be stationary and evenly sampled (Zuur et al. 2003a). In addition, these methods are not particularly suitable for unraveling patterns and interactions between variables over time.

Dynamic factor analysis (DFA) is a dimension-reduction method designed for time series (Zuur et al. 2003a) used mainly to estimate underlying common patterns in a set of time series. In this method, observations are related to the hidden patterns or states by the state-space model. This framework provides a means for modeling nonlinear effects and structural change in time series and is able to cope with missing observations (Harvey 1989). DFA has been used widely in econometric and psychological studies and has recently been applied to fisheries datasets (Zuur et al. 2003b, Zuur and Pierce 2004, Vilizzi 2012), groundwater-quality data (Muñoz-Carpena et al. 2005, Kuo et al. 2013), and river water-quality data (Aguilera et al. 2015).

To address the issue of the temporal scale of sampling and gaps in limnological time series, we took advantage of an ongoing long-term research program at Müggelsee, a shallow, polymeric lake in Berlin, Germany, in which data on physicochemical variables and plankton are collected weekly. We chose a shallow, polymeric lake because this type of lake is known to be more strongly influenced by short-term ambient weather conditions than are dimictic lakes (Gerten and Adrian 2000). The ultimate goal of this study was to assess the potential usefulness of DFA in analyzing patterns in limnological time series with differing temporal sampling resolutions and data gaps. To achieve this, we specified 4 objectives. (1) We tested whether DFA was able to reproduce the same underlying patterns in the data when the sampling resolution, in univariate analyses, was reduced from weekly to 4-weekly. (2) We compared and contrasted analyses of physicochemical and biological variables because these 2 types of variables may differ with regard to their frequency spectra. (3) We aimed to determine whether the complex dynamics of a set of time series with well-known interactions could be described using DFA and differing sampling frequencies. (4) Finally, we assessed the ability of DFA to cope with missing observations in 2 ways: (a) by introducing artificial gaps into the time series and investigating the effect of this on the patterns extracted; and (b) by resampling the available weekly data to obtain lower-frequency (i.e., seasonal and 4-weekly) datasets.
Methods

Study site and time-series data

Müggelsee is a shallow, polymictic lake in Berlin (Germany) with a surface area of 7.3 km², a mean depth of 4.9 m, and a water retention time of ~6–8 weeks. Data collection in the lake has been undertaken weekly (from spring to autumn) or once every 2 weeks (2-weekly; during winter) since 1979 (Köhler et al. 2005). A comprehensive description of the characteristics of Müggelsee and its surroundings can be found in Driescher et al. (1993). Physicochemical and biological variables from the Müggelsee dataset were selected to carry out DFA. Variables included water temperature, dissolved oxygen (DO), and pH (all measured at 1 m depth), nutrient concentrations (ammonium), and the abundances of phytoplankton (both taken from the upper 4 m of the water column) and zooplankton (taken from the entire water column). A detailed description of sampling and sample processing is given in Driescher et al. (1993) and in Gerten and Adrian (2000).

The time spans covered by the time series ranged from ~19 yr (pH) to 33 yr (water temperature; Table 1). The finest resolution was weekly. The reduction in resolution from weekly to 2-weekly and 4-weekly was accomplished by simple decimation (i.e., by selecting every second or every fourth value from the original weekly dataset).

Dynamic factor analysis (DFA)

Dynamic factor analysis is based on structural time-series models viewed from within the state-space model framework (Harvey 1989). In DFA, all common patterns in a set of time series are estimated simultaneously using the maximum likelihood method. These patterns are modeled as random walks and thus are allowed to be stochastic, meaning that their shape can change over time and is not necessarily restricted to a straight-line trend or a cosine-function cyclic or seasonal component (Zuur et al. 2007). The state-space formulation, and thus DFA, also allows a univariate time series to be decomposed into a pattern and an error term.

DFA was performed using the Multivariate Autoregressive State-Space (MARSS) R-package 3.4 (Holmes et al. 2013), run in R 2.15.1 (R Core Team 2012). Each time series was standardized by subtracting its mean and dividing by its standard deviation to facilitate the interpretation and comparison of resulting patterns and factor loadings. The basis of the DFA formulation in MARSS includes a process model (x) and an observation model (y):

$$x_t = x_{t-1} + w_t \quad w_t \sim MVN(0,Q)$$
$$y_t = Zx_t + v_t \quad v_t \sim MVN(0,R).$$

(1)

The observations ($y_t$) are modeled as a linear combination of hidden patterns ($x_t$) and factor loadings ($Z$; Holmes et al. 2012). The terms $w_t$ and $v_t$ describe the errors associated with the hidden patterns and the observations, respectively, normally distributed with expectation 0 and with covariance matrices $Q$ and $R$. The $Q$ matrix is set to an identity matrix, and the $R$ matrix is usually taken to be diagonal. The diagonal values in the $R$ matrix contain the variances of the time series involved, and the off-diagonal values contain the covariances between the time series. The MARSS algorithm uses the Kalman filter and smoother to compute the best estimate for the hidden patterns, given a set of assumed values for the unknown parameters $Z$, $Q$, and $R$. The output of the Kalman filter is then used to compute the likelihood of the data given the better estimates for the set of parameters. MARSS estimates the maximum likelihood using the expectation–maximization (EM) method (Holmes 2013). The number of iterations necessary is determined by the log-likelihood of reaching convergence (slope tolerance for log-log test = 0.1; absolute tolerance test = 0.001).

A considerable advantage of the state-space approach is the ease of addressing missing observations (Durbin and Koopman 2012). Holmes (2013) performed a series of modifications in the EM update equations for missing values in MARSS models, based on the work of Shumway and Stoffer (1982, 2010) and Zuur et al. (2003a), and presented a thorough explanation of the methodology. The

Table 1. Time period and number of observations for the Müggelsee variables included in the study. (Note that for pH the number of observations was substantially lower than for the other variables).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of observations</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weekly 2-weekly 4-weekly</td>
<td>Start End</td>
</tr>
<tr>
<td>Dissolved oxygen</td>
<td>1722 861 431</td>
<td>22 Jan 1979 16 Jan 2012</td>
</tr>
<tr>
<td>Ammonium</td>
<td>1721 861 431</td>
<td>01 Jan 1979 19 Dec 2011</td>
</tr>
<tr>
<td>Total algal biomass</td>
<td>1721 861 431</td>
<td>01 Jan 1979 19 Dec 2011</td>
</tr>
<tr>
<td>Copepod species case study</td>
<td>1614 807 404</td>
<td>21 Jan 1980 20 Dec 2010</td>
</tr>
</tbody>
</table>
main disadvantage of DFA is its intensive computations, translating into long computational times ranging from hours to days and weeks, depending on the complexity of the model. In our study, the High-Performance Computing Cluster Undarius allowed the simultaneous computation of multiple independent analyses. Nevertheless, the computation time for the multivariate analyses (based on the copepod species and cryptophytes case study described below) ranged from 1 to 4 weeks.

**Effect of sampling resolution**

Univariate time-series analysis

Univariate time-series analyses were performed so that a pattern plus an error term would be extracted from each variable along a trophic hierarchy (water temperature, nutrients, DO, pH, phytoplankton, zooplankton) and at different data sampling resolutions (weekly, 2-weekly, and 4-weekly). Thus, not only the effect of differences in the temporal resolution of the time-series data was assessed, but also the behavior of different system levels across the trophic hierarchy, which is affected by biotic interactions and thus typically incorporates processes of both high-frequency and low-frequency variability.

Multivariate time-series analysis

A well-established and well-studied relationship between variables measured in Müggelsee was selected for the application of DFA to a set of relevant time series. The case study dealt with the abrupt increase in the abundance of a copepod species, *Cyclops kolensis*, when the abundance of a competing, larger species, *Cyclops vicinus*, fell below a critical threshold triggered by, for example, a change in the abundance of cryptophytes, a major food source for the juvenile stages of both *Cyclops* species (Scharfenberger et al. 2013).

A set of models was evaluated for the copepod species case study to test different numbers of common patterns (*m*) ranging from 1 to *n* − 1 (where *n* is the number of time-series involved in a particular analysis, in this case 3) as well as different structures for the covariance matrix R. Because the number of model parameters increases drastically when using an unconstrained R matrix, we assumed the time series involved in our multivariate analyses had either an equal (diagonal and equal) or different (diagonal and unequal) observation variance but no covariance between time series. The best model was then selected based on the Akaike information criterion (AIC; Akaike 1974) as well as on model inspection of plotted patterns versus observed values and a visualization of the magnitudes and signs of factor loadings.

Effect of gaps on pattern extraction in time-series analysis

Gaps were introduced in one representative univariate time series (total algal biomass) and in the multivariate copepod species case described earlier (Scharfenberger et al. 2013). Gaps were systematically introduced by deleting 3 months of available data per year. The deleted data consisted of a large gap comprising 2 consecutive spring months (Apr and May) and a shorter gap in winter (Dec). Three different analyses with different sampling resolutions (weekly, 2-weekly, and 4-weekly) were performed to determine changes to the observed patterns (previous models without artificial gaps) when data gaps are larger or evenly distributed across time (i.e., systematic missing observations at specific times of the year). In the copepod species case study, gaps were introduced only in the driving variable (i.e., cryptophyte availability).

Effect of sampling representativeness

Using the same univariate time series (total algal biomass) and multivariate copepod species examples mentioned earlier, we compared the results of DFA models for 4-weekly and seasonal observations taken at those specific intervals (i.e., discrete, equidistant data points measured every 4 weeks and every season, respectively) with the results of models that averaged the weekly data over 4-weekly and seasonal intervals. This simple exercise aimed to assess how representative a sample could be in the case of discrete, low-frequency sampling compared to the case in which all available information (i.e., weekly data) was used by simply averaging the observations over the 4-weekly or seasonal intervals.

DFA model evaluation

Comparing weekly, 2-weekly, and 4-weekly data-resolution models might not be straightforward because these models seem to extract different temporal patterns, including long-term trends and seasonal and other cyclic patterns, although the patterns extracted might vary depending on the type of variable. The resulting models for differing data resolutions can first be assessed visually; because DFA is analogous to linear regression, the same tools can be applied for model validation (Zuur et al. 2007). The first tool plots the resulting best fits to the observed values to identify the extent to which the model can capture the patterns in the original time series.

The MARSS R-package assesses different model fits by means of AIC. AIC was used in the copepod species case study, where the best model fit with *m* patterns had to be selected from among a set of different models. In addition, factor loadings and values in the covariance matrix R (i.e., the observation error matrix containing
variance values) were used to assess the strength of a model in extracting underlying patterns in our univariate and multivariate time-series analyses.

Results

Effect of reducing sampling resolution in univariate time-series

The reduction in resolution (from weekly to 2-weekly to 4-weekly) did not significantly change the DFA results for near-surface water temperature (Fig. 1, 4-weekly data), which exhibited relatively little high-frequency variability (at least at weekly sampling resolution).

For the ammonium concentration data, the analysis at a 4-weekly data resolution revealed a slight decrease over time since the beginning of the sampling program (Fig. 2a), but neither interannual nor seasonal variability could be captured (Fig. 2b). The amount of unexplained information was fairly large in the 4-weekly resolution model, which was also evident in the fit (Fig. 2b). For DO, the 4-weekly resolution data analysis resulted in a constant zero pattern across the 3 decades of data (Fig. 3), also the case for pH (results not shown). The amount of unexplained information, as reflected in the covariance matrices, was thus significantly larger for the 4-weekly data models.

The analysis of total algal biomass in Müggelsee revealed a drastic change in the overall pattern of the time series after the resolution was reduced from weekly (Fig. 4a) to 2-weekly (Fig. 4b), and 4-weekly (Fig. 4c). Further, the variance value was higher for the 4-weekly resolution model (0.84) than for the 2-weekly (0.51) and weekly (0.25) models, implying that the observation variance in the 4-weekly resolution model was at least 3 times higher than in the weekly resolution model.

Fig. 1. Pattern (line) derived from dynamic factor analysis fitted to the standardized 4-weekly data points (dots) for near-surface water temperature in Müggelsee.

Fig. 2. Patterns (lines) derived from dynamic factor analyses fitted to the standardized (a) weekly and (b) 4-weekly data points (dots) for ammonium concentration in Müggelsee (measured at 1 m depth).

Fig. 3. Patterns (lines) derived from dynamic factor analyses fitted to the standardized (a) weekly and (b) 4-weekly data points (dots) for dissolved oxygen concentration in Müggelsee.
Reducing sampling resolution in multivariate time-series analysis

In the case study of the coexisting copepod species *C. vicinus* and *C. kolensis* and the cryptophyte biomass (Scharfenberger et al. 2013), model selection based on AIC implied that both weekly and 2-weekly data were best described by 2 common patterns and a diagonal and unequal covariance matrix. According to the factor loadings, *C. vicinus* and the cryptophyte biomass (Fig. 5b) seemed to be best described by Pattern 1, which decreased across time. By contrast, *C. kolensis* (Fig. 5b) was best characterized by Pattern 2, which tended to be the inverse of Pattern 1. The diagonal values in the R matrix indicated that the selected model satisfactorily explained the variability in the time series of the 2 copepod species (0.54 for *C. vicinus* and 0.12 for *C. kolensis*) but not the variability in the cryptophyte biomass time series, which had a high value of unexplained variance in the covariance matrix R (0.76).

The 4-weekly data yielded one common pattern; *C. vicinus* and the cryptophyte biomass followed a decreasing trend with time, and *C. kolensis* followed an increasing trend related to a negative factor loading (dotted line in Fig. 6a). The factor loadings of the 4-weekly resolution model (Fig. 6b) were about one order of magnitude lower than those in the weekly and 2-weekly resolution models. Overall, the variance reflected in the 4-weekly resolution model (average diagonal values in R matrix = 0.92) was about twice that in the weekly model, indicating that a significant amount of information was lost by using 4-weekly biomass data rather than data of higher resolution.

Unexplained information in time-series analysis with additional data gaps

In the univariate model of total algal biomass, the amount of unexplained information (diagonal values in the R
covariance matrix) in the weekly resolution model with artificial gaps increased by 40% (weekly data) compared to the previous weekly resolution model without such gaps (Table 2). The increase in unexplained information was lower in the case of the 2-weekly (27.5%) and 4-weekly (4.8%) resolution models. These sampling resolutions initially yielded a large amount of unexplained information in the previous models (0.51 and 0.84; Table 2), and therefore the difference in the corresponding artificial gap models was not as drastic as in the case of weekly data. Nevertheless, the explanatory power of the 2-weekly and 4-weekly data was much lower than that of the weekly data.

In the multivariate copepod species model, the covariance R value increased by 44.4% and 43.9% for the weekly and 2-weekly data resolution models, respectively (Table 2). The 4-weekly model increased the amount of unexplained information by only 1.1% because the initial variance was already high (from R = 0.84 to 0.88). The total length of gaps introduced in all 3 temporal resolution cases decreased the number of observations by an average of 24.6% (range 24.3–25.0%) compared to the corresponding models of the same case study without artificial gaps. Despite these differences, the extracted patterns and resulting best fits for models with and without artificial gaps did not differ substantially from one another; in other words, the introduction of gaps in the cryptophyte time series did not substantially alter the shape and behavior of the patterns extracted by DFA.

<table>
<thead>
<tr>
<th>DFA Model</th>
<th>Weekly</th>
<th>2-weekly</th>
<th>4-weekly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total algal biomass (no gaps)</td>
<td>1</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td>Total algal biomass (gaps)</td>
<td></td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>Copepod abundances (no gaps)</td>
<td>2</td>
<td>0.54</td>
<td>2</td>
</tr>
<tr>
<td>Copepod abundances (gaps)</td>
<td></td>
<td>0.78</td>
<td></td>
</tr>
</tbody>
</table>

$m =$ number of patterns; $R =$ amount of unexplained information (variance value in covariance matrix). Copepod abundance data from Scharfenberger et al. (2013).

Unexplained information and sampling representativeness

For both univariate and multivariate data, DFA models based on data obtained by averaging the original time-series over either 4-weekly or seasonal intervals outperformed DFA models based on data points simply selected from the original time series at the same 2 intervals. The values of variance in the R matrices for the data selection models are considerably larger than those for the data-averaging models (Table 3). These differences in variance are notably larger for the seasonal data than for the 4-weekly data. Overall, our results confirm that, as expected, averaging the available observations in a time series to obtain lower-frequency measurements is preferable to selecting individual values from the time series at specific target intervals.

Discussion

The responses of lakes to environmental change are complex, occur on a variety of temporal scales, and can be associated with changes that occur within critical short-term time windows within a season (Adrian et al. 2012). Many past and ongoing long-term research programs provide high-quality data at short temporal
Effects of sampling resolution and data gaps of limnological data on pattern recognition

Inland Waters (2016) 6, pp.284-294

Table 3. Amount of unexplained information in univariate and multivariate DFA models with differing resampling strategies (i.e., data selection as opposed to data averaging).

<table>
<thead>
<tr>
<th>DFA Model</th>
<th>4-weekly</th>
<th>Seasonal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(m)</td>
<td>(R)</td>
</tr>
<tr>
<td>Total algal biomass (data selection)</td>
<td>1</td>
<td>0.84</td>
</tr>
<tr>
<td>Total algal biomass (data averaging)</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>Copepod abundances (data selection)</td>
<td>1</td>
<td>0.92</td>
</tr>
<tr>
<td>Copepod abundances (data averaging)</td>
<td>0.88</td>
<td>0.81</td>
</tr>
</tbody>
</table>

\(m\) = number of patterns; \(R\) = amount of unexplained information (variance value in covariance matrix). Copepod abundance data from Scharfenberger et al. (2013).

Sampling resolution reduction in physicochemical and biological variables

Predictably, decreasing the data resolution from weekly to 4-weekly lowered the capability of the DFA models to explain the underlying short-term patterns; however, the decrease in the amount of unexplained information varied. In the case of an abiotic variable such as near-surface water temperature, which is usually dominated by low-frequency variability, 4-weekly data were sufficient to capture the main long-term pattern and seasonality. This was not the case for the pH and DO models, which presumably show more high-frequency variability than does water temperature when considered at broader temporal scales. The information contained in the time series of ammonium concentration and total algal biomass, however, was best explained by the weekly data model. The aliasing effect caused by undersampling is much stronger in processes with a significant power of high frequency components. Algal growth (Reynolds 1984) and nutrient dynamics operate on temporal scales of days or less (Lampert and Sommer 2007), indicating the need, in these cases, for a higher sampling frequency to avoid or reduce the effect of aliasing.

Overall, in the univariate time-series analyses of the chemical and biological variables, we observed that 4-weekly resolution might not suffice to describe the variability that occurs within short time windows. Addressing temporal scale is especially relevant in polymictic lakes, which, unlike dimictic and monomictic lakes, are subject to frequent changes in thermal regime during summer. In Müggelsee, for instance, although 79% of all stratification events last <1 day, the frequency and period of stratification are predicted to increase (Wilhelm and Adrian 2008). This change in stratification has potential consequences for plankton development in the lake; stratification periods exceeding 3 weeks, for instance, have been shown to favor the proliferation of cyanobacteria (Wagner and Adrian 2009), and severe climate-related changes in plankton diversity have occurred on temporal scales of 2–3 weeks (Wagner and Adrian 2011).

Sampling resolution reduction in multivariate analysis

In the multivariate copepod species case, the best DFA models usually yielded fewer common patterns for the 4-weekly data resolution, indicating a reduced capability of the models to capture interactions apparent in data sampled at higher resolution (i.e., weekly and 2-weekly). The 4-weekly data were thus unable to capture the...
that the decline in copepod abundances seemed to coincide with cryptophyte biomass. Note that the overall long-term pattern in copepod abundances seemed to coincide with the results of Scharfenberger et al. (2013), who stated that the decline in C. vicinus followed the driver threshold scenario of regime shift theory (Andersen et al. 2009), caused by an abrupt change in a driver (in this case the decline in cryptophyte prey availability), whereas concomitantly C. kolensis benefited from this situation and increased in abundance in the decades following the shift event. The failure to capture the dynamics in zooplankton development and interactions critically depends on species-specific life history traits. Fast-growing species such as cladocerans and rotifers exhibit life-history durations of only a couple of days compared to >4 weeks for copepods (Adrian et al. 2006). Thus, the population dynamics of cladocerans and rotifers are more easily missed if sampling is at low temporal resolution. Nevertheless, for copepods with long and complex life histories, entire generations can be missed if sampling intervals are in the range of a month (Adrian et al. 2006). Despite the relative dearth of information obtained by analyzing 4-weekly observations, DFA could still extract long-term, general patterns and, in some cases, intraannual and seasonal patterns in our data (e.g., Fig. 4c and 6a).

Dealing with data gaps and sampling representativeness

The introduction of additional artificial gaps into the data used for the univariate total algal biomass models and the multivariate copepod species case study models decreased the amount of information that could be explained by the pattern(s) extracted by DFA (Table 2). The increased observation variance, however, varied according to the temporal resolution of the data and was more pronounced in the weekly data models in both the univariate and multivariate cases. In real terms, the introduction of gaps for 2 consecutive spring months would be especially critical in real sampling programs for most biological variables such as algal or cladoceran biomass because spring is a key time of year when rapid, short-term changes in external driving variables such as air temperature and day length usually take place. Nevertheless, even in winter (a season much neglected because of the notion that not much happens), variability in the phytoplankton can be high at short temporal scales (Özkundakci et al. 2016). All of this was reflected in the loss of information found in the weekly models with data gaps.

Intuitively, averaging the highest-frequency observations to rescale the data to a lower frequency retains a larger portion of the available information, whereas the practice of selecting observations at specific target intervals excludes much of the available information. We acknowledge that the weekly sampling frequency in this study might be below the Nyquist frequency for many biotic and abiotic processes in limnology. Despite this limitation, however, we strongly recommend that the sampling frequency be tailored to include the potentially distorting effects of aliasing when designing monitoring schemes. Also noteworthy is that, although high-frequency (e.g., daily, hourly) measurements of biotic variables would be preferred to guarantee sampling representativeness, most measurements of biotic variables rely on spot-sample campaigns, which are undertaken at lower frequencies.

Several studies have suggested that long-term routine monitoring, even at low sampling frequencies, is important with respect to slow ecological processes, rare or episodic phenomena, highly variable processes, and subtle or complex phenomena. For instance, monthly averaged stream nitrate data in the tropical forest of the Luquillo Mountains (Puerto Rico) allowed researchers to quantify the magnitude of the impact of Hurricane Hugo on water chemistry and to assess the speed at which the system approached background conditions after the hurricane hit the island in 1989 (Schaefer et al. 2000, Dodds et al. 2012). In an analysis of watershed characteristics, Halliday et al. (2012) found that 4-weekly data allowed the analysis of long-term trends and provided new insights into the changing amplitude and phase of the seasonality of water quality variables. Furthermore, they observed that low-frequency (e.g., monthly) data provided background information on their study basin and on the processes occurring necessary to interpret the complexity of the high-frequency data.

Overall, low-frequency data allow the detection of long-term trends and seasonality in those cases in which the power in the low-frequency components of the processes is much higher than that in the high-frequency components, specifically in cases in which the aliasing effect does not significantly affect characterization of the large-scale processes. High-frequency data are key in the identification of extremes, short-term trends, and subdaily variability in limnological variables. High-frequency data are also needed when the power in the high-frequency components is comparable to that in the low frequency components; in this case the aliasing effect caused by undersampling is much stronger. A compromise between financial and logistical resources and sample representativeness must be reached to maximize the information obtained from the observations taken at the chosen sampling frequency.

© International Society of Limnology 2016

DOI: 10.5268/IW-6.3.948
Acknowledgements

R. Aguilera benefited from a Short-Term Scientific Mission at the Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, funded by the ESSEM COST Action ES1201 NETLAKE (Networking Lake Observatories in Europe). D.M. Livingstone benefited from time spent as guest scientist at the Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, funded by the IGB Fellowship Program in Freshwater Science. R. Adrian was supported through the German Science Foundation project LakeRisk (AD 91/13-1).

References


DOI: 10.5268/IW-6.3.948