A Data-driven Approach to Detecting Change Points in Linear Regression Models

Vyacheslav Lyubchich*[1], Tatiana V. Lebedeva[2], and Jeremy M. Testa[1]

[1]Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD, USA
[2]Department of Statistics and Econometrics, Orenburg State University, Orenburg, Russia

Abstract

Change points appear in various types of environmental data—from univariate time series to multivariate data structures—and need to be accounted for in proper analysis and inference. The analysis of change points is challenging when no exact information about their number and locations is available, and statistical tests developed under such conditions often have low power identifying the change points. This paper provides a powerful, data-driven procedure for detecting at-most-m change points in linear regression models by adapting a sieve bootstrap approach for a modified cumulative sum statistic. The new procedure does not assume a particular dependence structure nor a particular distribution of regression residuals. It employs a data-driven selection of the order of an autoregressive model and a robust estimation of the model coefficients. Our simulation studies show a competitive performance of the new bootstrap-based procedure compared with its asymptotic counterpart. We apply the new testing procedure to address an important environmental problem in Chesapeake Bay—severe oxygen depletion—and detect two change points in the relationship between the volume of low-oxygen waters and nutrient inputs to the bay during 1985–2017.

*Corresponding author. E-mail: lyubchich@umces.edu.

Keywords: sieve bootstrap, time series, hypothesis test, regime shift, CUSUM, Chesapeake Bay anoxia.

1 Introduction

It is difficult to underestimate the influence of change points on the results of statistical analysis because change points may affect the structural stability of regression models, conclusions about shapes and significance of temporal trends, and other inferences. The complexity of change point detection problems ranges from the task of testing one or several specific change points to completely relaxing the assumptions about the number and locations (i.e., the time of change) of possible change points in a given sample. There is an increasing interest in methods targeting the latter, less structured problems, because the rampant expansion of available datasets, complexity of analyses, and growing computing power demand and allow for development of powerful data-driven solutions.

Change point techniques have been applied in freshwater and marine ecosystems to quantify the timing of 'regime shifts'. Regime shifts have been identified as the abrupt and dramatic alternation between different steady states, which can include shifts from benthic to pelagic-dominated states in lakes associated with nutrient loading (Scheffer and Jeppesen, 2007) or food-web shifts driven by switches between large-scale climate cycles (Hare and Mantua, 2000). Regime shifts, given their abrupt nature, are often associated with the crossing of an environmental threshold, and techniques have been articulated to identify thresholds in coastal ecosystems (Andersen et al, 2009). In Chesapeake Bay, such threshold techniques have indicated regime shifts associated with changes to nutrient cycling associated with eutrophication-induced oxygen, O_2, depletion (Testa and Kemp, 2012).

Oxygen depletion, and the associated development of hypoxia (O_2 concentration $<$ 2 mg/L) and anoxia (oxygen absent), is a primary societal concern in many aquatic ecosystems, such as lakes, estuaries, and the open ocean (Kemp et al, 2009; Scavia et al, 2014;

Breitburg et al, 2018). In Chesapeake Bay, the largest estuary in the United States, large hypoxic and anoxic volumes develop each summer and have been associated with food web disruptions and altered biogeochemical cycling (Kemp et al, 2009; Sturdivant et al, 2013). As a consequence, large socio-economic commitments have been made to relieve hypoxic conditions via the reduction of watershed nutrients that stimulate algal biomass production and subsequent oxygen depletion via respiration. Given these commitments, there is a continuing need for new understanding of the temporal variations in low-oxygen volumes in response to biological, chemical, and physical controls.

Early approaches to the problem of changing parameters in time-dependent linear regression have used a likelihood ratio test (Quandt, 1958, 1960), theory of Markov processes (e.g., Vaman, 1985), and Bayes-type statistics (Jandhyala and MacNeill, 1989, 1991). Most of the recent studies, however, are based on cumulative sum (CUSUM) statistics calculated on regression residuals (e.g., see Horváth et al, 2004; Aue et al, 2006; Gombay, 2010; Horváth et al, 2017); see more references in the reviews by Jandhyala et al (1999), Reeves et al (2007), and Horváth and Rice (2014). Motivated by the applied questions from our Chesapeake Bay study, we favored the flexible framework of Horváth et al (2017) for detecting at-most-m change points in a linear regression model with potentially autocorrelated errors. The results by Horváth et al (2017), however, involve kernel-based estimation of the long-run variance function—an approach that can be sensitive to dominantly positive or negative autocorrelations. In practice, the estimation also requires selecting a kernel function and optimal bandwidth. Also, the general problem of CUSUM-based test statistics is their slow convergence to an asymptotic distribution with the increase of sample size (e.g., see Gombay, 2010).

To overcome the problem of estimating the long-run variance function and to enhance the performance of the test in small samples, we propose to use a non-parametric sieve bootstrap approach for the modified CUSUM statistic. Bootstrap approaches have been used for CUSUM statistics before (Kirch, 2006, 2007; Chatterjee and Qiu, 2009; Gandy

and Kvaløy, 2013; Zhao and Driscoll, 2016), as well as in other approaches to change point detection (e.g., see Antoch et al, 1995; Gombay and Horváth, 1999; Kirch, 2007; Hušková and Kirch, 2008; Seijo and Sen, 2011; Hlávka et al, 2016). However, to the best of our knowledge, this is the first time the bootstrap is applied to a CUSUM statistic for detecting at-most-m change points in linear regression coefficients (we describe how our approach relates to other methods in more detail in Section 2.1, after introducing the testing procedure). We apply robust estimation of autoregressive coefficients (Hall and Van Keilegom, 2003) and automatic selection of the autoregressive order based on the Bayesian information criterion to make the testing procedure fully data-driven and convenient for fast implementation by the user. We show that the new bootstrapped procedure maintains the size of the test similar to its asymptotic counterpart, but has higher power in detecting the changes. Finally, we describe several techniques for pre-selecting locations of change points in real data and compare the techniques in a case study.

The remainder of the article is organized as follows. Section 2 presents our data-driven testing technique. Section 3 demonstrates the performance of the method in finite simulated samples. In Section 4, the method is applied to a long-term study of low-oxygen waters in Chesapeake Bay. Concluding remarks are given in Section 5.

2 Methods

We are interested in detecting and testing for changes in coefficients in a time-dependent linear regression model

$$\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $t = 1, \ldots, T$; T is the sample size (i.e., length of the time series); $\mathbf{Y}_t$ is the dependent variable, $\mathbf{X}_t$ is the design matrix with d regressors; $\boldsymbol{\beta}_t$ is a vector of regression coefficients,

and $\boldsymbol{\varepsilon}_t$ are regression errors:

$$\mathbf{Y}_t = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \quad \mathbf{X}_t = \begin{bmatrix} x_{10} & x_{11} & \dots & x_{1d} \\ x_{20} & x_{21} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T0} & x_{T1} & \dots & x_{Td} \end{bmatrix}, \quad \boldsymbol{\beta}_t = \begin{bmatrix} \beta_{0t} \\ \beta_{1t} \\ \vdots \\ \beta_{dt} \end{bmatrix}, \quad \boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix},$$

and $x_{t0} = 1$ for all t. The changes in coefficients shall be reflected in the series of residuals

$$\hat{\boldsymbol{\varepsilon}}_t = \mathbf{Y}_t - \mathbf{X}_t \hat{\boldsymbol{\beta}}_t, \tag{2}$$

where $\hat{\boldsymbol{\beta}}_t$ are estimated using the method of ordinary least squares (OLS). We assume possible autoregressive dependence in the errors (under which the OLS estimates are still unbiased; see Lee and Lund, 2004 and references therein on asymptotic efficiency of the OLS estimates under common cases of autocorrelated regression errors), such as:

$$\varepsilon_t = \sum_{i=1}^{p} \phi_i \varepsilon_{t-i} + v_t, \tag{3}$$

where $v_t \sim \mathrm{WN}(0, \sigma^2)$; $\phi(\lambda) = 1 - \phi_1 \lambda - \ldots - \phi_p \lambda^p \neq 0$, $\forall |\lambda| \leqslant 1$; ϕ_i are autoregressive coefficients, and p is the autoregressive order.

2.1 The testing procedure

Our testing procedure is based on the flexible framework for detecting at-most-m changes, by Horváth et al (2017):

$$\begin{aligned} &H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_T \\ &H_a : \boldsymbol{\beta}^{(j)} \neq \boldsymbol{\beta}^{(l)} \text{ for some } 1 \leqslant j, l \leqslant m+1, \end{aligned} \tag{4}$$

where $\boldsymbol{\beta}^{(i)}$ $(i = 1, \ldots, m+1)$ are regression coefficients in the ith subperiod; $k_1, \ldots, k_m$ are the locations of m change points $(1 \leqslant k_1 \leqslant k_2 \leqslant \cdots \leqslant k_m < T)$; $k_{i-1} < t \leqslant k_i$ ($k_0 = 0$ and $k_{m+1} = T$). The H_0 of no changes in the coefficients is rejected in all cases from at least one change point to at most m change points being present, where the m change point locations are specified before applying the test.

To test the null hypothesis, compute the modified CUSUM

$$M(k_1, \ldots, k_m) = |M_1(k_1)| + |M_2(k_1, k_2)| + \ldots + |M_m(k_{m-1}, k_m)| + |M_{m+1}(k_m)|, \qquad (5)$$

where

$$\begin{aligned}
M_1(k_1) &= \frac{1}{\sqrt{k_1}} \left(\sum_{t=1}^{k_1} \hat{\varepsilon}_t - \frac{k_1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t \right), \\
M_i(k_{i-1}, k_i) &= \frac{1}{\sqrt{T}} \left(\sum_{t=k_{i-1}+1}^{k_i} \hat{\varepsilon}_t - \frac{k_i - k_{i-1}}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t \right), \quad 2 \leqslant i \leqslant m, \\
M_{m+1}(k_m) &= \frac{1}{\sqrt{T - k_m}} \left(\sum_{k_m+1}^{T} \hat{\varepsilon}_t - \frac{T - k_m}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t \right).
\end{aligned}$$

Then, obtain the statistic

$$M_T = \max_{\mathfrak{M}} M(k_1, \ldots, k_m), \text{ where } \mathfrak{M} = \{1 \leqslant k_1 \leqslant k_2 \leqslant \cdots \leqslant k_m < T\}. \qquad (6)$$

The statistic M_T is the maximum of statistics M calculated over all combinations of one to m change points $k_1, \ldots, k_m$ in $\mathfrak{M}$. For example, for given $m = 3$ and respective change point locations $\mathfrak{M} = \{1 \leqslant k_1 \leqslant k_2 \leqslant k_3 < T\}$, the seven possible combinations of the change point locations are explored:

$$\begin{aligned}
\mathfrak{M}^* &= \{(k_1), (k_2), (k_3), (k_1, k_2), (k_1, k_3), (k_2, k_3), (k_1, k_2, k_3)\}, \\
M_T &= \max\{M(\mathfrak{M}_1^*), \ldots, M(\mathfrak{M}_7^*)\},
\end{aligned}$$

and the final change point locations are those corresponding to the M_T:

$$\arg\max_{\mathfrak{M}^*}\{M(\mathfrak{M}_1^*),\ldots,M(\mathfrak{M}_7^*)\}.$$

The assumptions for the test include stationarity of $\mathbf{x}_t$ and ε_t and that the sequence is a Bernoulli shift, which can be approximated with finitely dependent time series (see Horváth et al, 2017 for more details).

Remark 1 An alternative to trying all combinations of change points in $\mathfrak{M}^*$ could be an approach utilizing penalized likelihood, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), or risk inflation criterion (RIC); see Stine (2004) and references therein for a comparative assessment of these criteria in variable selection. The approach using criterion-based stepwise elimination of terms, starting from the total of $m+1$ terms (one for each subperiod), may be faster than using the $\mathfrak{M}^*$, however, both of these approaches require the candidate change point locations $k_1,\ldots,k_m$ to start with. If $k_1,\ldots,k_m$ are unknown, they can be pre-selected from the data, which usually constitutes a more computationally demanding task (see Section 2.2) than exploring the combinations of m terms in $\mathfrak{M}^*$.

Here we propose to replace the asymptotic approximations based on the long-run variance function with a data-driven bootstrap approach for approximating the distribution of M_T under the null hypothesis and calculating bootstrap-based p-values. Our modified testing procedure consists of the following steps:

1. Calculate the observed statistic (6), $M_T^{(obs)}$, based on $\hat{\varepsilon}_t$.

2. Select autoregressive order p using a data-driven information criterion, such as BIC, and get the estimates $\hat{\phi}_1,\ldots,\hat{\phi}_p$ in (3). Similarly to Lyubchich and Gel (2016), here we employ the robust difference-based estimator of autocorrelation functions by Hall and Van Keilegom (2003) within the Yule–Walker equations to get the estimates of the

autoregressive coefficients.

3. Calculate the residuals $\hat{v}_t$ in (3). If the optimal autoregressive order selected at the previous step is zero, let $\hat{v}_t = \hat{\varepsilon}_t$.

4. Let $\hat{v}_t^*$ be a sample with replacement (i.e., bootstrap sample) from $\hat{v}_t$. Alternatively, $\hat{v}_t^*$ can be generated from a kernel smoothing of the $\hat{v}_t$.

5. Generate a new autoregressive series $\hat{\varepsilon}_t^*$ using the estimated coefficients $\hat{\phi}_1, \ldots, \hat{\phi}_p$ and innovations $\hat{v}_t^*$.

6. Calculate the bootstrapped statistic (6), $M_T^{(*1)}$, based on $\hat{\varepsilon}_t^*$.

7. Repeat steps 4–6 a large number of times to get a distribution of bootstrapped statistics $\left\{M_T^{(*b)}\right\}_{b=1}^{B}$, where B is the total number of bootstrap repetitions.

8. The bootstrap p-value for testing (4) is the proportion of $\left\{M_T^{(*b)}\right\}_{b=1}^{B}$ that exceed $M_T^{(obs)}$.

An implementation of the above procedure is available from `mcusum.test` function in R package *funtimes* (Lyubchich and Gel, 2019).

Remark 2 Consider a simple linear regression without intercept term, which is a reduced version of model (1):

$$y_t = \beta_t x_t + \varepsilon_t.$$

The coefficient β_t is estimated using OLS as a weighted ratio of y_t/x_t:

$$\hat{\beta}_t = \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} x_t^2} = \frac{\sum_{t=1}^{T} x_t^2 \frac{y_t}{x_t}}{\sum_{t=1}^{T} x_t^2},$$

where x_t^2 $(t = 1, \ldots, T)$ are the weights. When the weights are almost equal and are not zero, the non-weighted average ratio, $T^{-1} \sum_{t=1}^{T} (y_t/x_t)$, can serve as an intuitive estimate of

β_t. Therefore, the task of detecting changes in β_t can be performed using the series y_t/x_t instead of $\hat{\varepsilon}_t$.

Relevance to other methods Without the goal of providing a comprehensive review on the topic of change point analysis, here we outline how our testing procedure is positioned among other methods in the field.

Our procedure is developed for linear parametric regression models with stationary errors. For results on change point detection in non-linear models, we refer to Aue et al (2008), Jandhyala and Al-Saleh (1999), Jandhyala et al (1999), and references therein. Other approaches examine cases of non-stationarity, such as periodicity, stochastic and deterministic trends in regressors (Bai et al, 1998; Hanson, 2002; Lund et al, 2007; Kejriwal and Perron, 2008; Gallagher et al, 2013), changes in persistence, and heteroscedasticity (Busetti and Taylor, 2004; Cavaliere and Taylor, 2008; Górecki et al, 2018).

Our test is more general than a number of previously developed procedures for testing at-most-one change (AMOC; for example, see Quandt, 1960; Gombay and Horváth, 1999; Jandhyala et al, 1999; Jarušková, 2003; Kirch, 2006, 2007; Reeves et al, 2007; Aue et al, 2008, and references therein), but it shares the common property of decreasing power when a change point moves away from the center of a time series.

One of the inputs used by our testing procedure is the suggested set of change point locations $k_1, \ldots, k_m$ for which the test is applied. The locations can be given beforehand (the so-called case of 'documented' change points), but in many applications both the number of change points and their locations need to be determined from the analysis. An exhaustive search over all possible positions of m change points in a sample is feasible only when m and sample size T are small, because of the heavy computational burden. We apply phase analysis, regression trees, and alternating conditional expectations to locate possible change points. Other available approaches use the principle of dynamic programming (see Zeileis et al, 2003; Antoch and Jarušková, 2013, and references therein) and genetic algorithms

(e.g., Li and Lund, 2012) to find a solution optimizing some likelihood function, which is usually the residual sum of squares, AIC, or BIC. Li and Lund (2015) and Li et al (2017) use Bayesian techniques to account for partial information about change points documented in metadata. A separate group of testing approaches that often do not require to set the upper limit m for the number of change points are developed for monitoring tasks (e.g., see Horváth et al, 2004; Aue et al, 2006; Eichinger and Kirch, 2018).

The residuals $\hat{v}_t$ calculated at Step 3 of our procedure are the differences between the potentially autocorrelated observed values $\hat{\varepsilon}_t$ and the one-step-ahead predictions obtained using the autoregressive model (3). That is, $\hat{v}_t$ are one-step-ahead prediction residuals. Robbins et al (2011) provide a summary of using such residuals in CUSUM methods when considering a problem of detecting a single mean shift in an autoregressive moving average (ARMA) process. Robbins et al (2011) note that under certain regularity conditions the CUSUM-based inferences are asymptotically the same, whether based on the raw ARMA series or the uncorrelated one-step-ahead prediction residuals. In our sieve bootstrap approach, however, the bootstrap counterparts of residuals $\hat{v}_t$ are plugged back into the autoregressive model to generate new autoregressive series $\hat{\varepsilon}_t^*$.

Our bootstrap procedure is most similar to Hušková et al (2008) and Hušková and Kirch (2012) who apply two types of bootstrapping while considering change points in coefficients of autoregressive models and sequential change point testing in linear models. In the paired bootstrap, the resampling of paired y_t and $\mathbf{x}_t$ (Hušková and Kirch, 2012) or of paired y_t and regression residuals $\tilde{e}_t$ (Hušková et al, 2008) is performed. In the regression bootstrap, just the residuals are bootstrapped. The difference is that the residuals $\tilde{e}_t$ are defined after a hypothesized change point has been incorporated into the regression model (cf. the residuals (2) that are estimated without assuming a change point). Hušková and Kirch (2012) note that their procedure is most suitable for independent and identically distributed (i.i.d.) vectors, and point in the direction of block bootstrap for dependent series.

Thus, our method stands out by automatically accommodating possible serial depen-

dence, using the version of sieve bootstrap. We also bypass standardization of the test statistic with a long-run variance estimate. Such an estimate is used, for example, in the $V^{(3)}$ version of the test statistic by Horváth et al (2017), but can lead to a number of complications in analysis of real data, especially when the sample size is small (see more details in Horváth et al, 2017; Eichinger and Kirch, 2018).

2.2 Identifying candidate change points

There are a number of ways we can select the candidate change points $k_1, k_2, \ldots, k_m$ for applying the testing procedure. For example, we might consider metadata, expert opinions, or perform a quantitative and exhaustive search over all combinations. The exhaustive search over all possible locations will require calculating the value in (5) $K_{T,m} = \sum_{i=1}^{m} \frac{(T-1)!}{i!(T-1-i)!}$ times, which is still computationally feasible for small T and m, but is impractical for larger problems. For example, $K_{33,3} = 5{,}488$ possible locations can be explored in our case study (Section 4), but there are too many possible change point locations for larger T and m ($K_{100,3} = 161{,}799$; $K_{100,5} = 75{,}449{,}319$, etc.). Below we present several other methods for identifying the change points.

Phase analysis The main idea of the method of phase analysis is iterative cleaning of a time series from low-power (random) fluctuations. *Phases* are periods of positive or negative fluctuations (Lukashin, 2003)—similar to runs in the test for time series randomness. In the filtering process, the phases are aggregated to incorporate or smooth out smaller phases.

The fluctuations can be computed as deviations of time series from the mean, previous values (i.e., as consecutive differences), zero or some other target value, or the temporal trend. In this application to the problem of studying changes in regression coefficients, we define fluctuations as the regression residuals $\hat{\varepsilon}_t$.

Let $p_i = \sum_{t=l_{i-1}+1}^{l_i} |\hat{\varepsilon}_t|$ be the power of ith phase that includes time points from $l_{i-1}+1$ to l_i, where $l_0 = 0$. Thus, the overall power is $P = \sum_i p_i$. The phase analysis is performed

in the following steps:

1. Set a stopping criterion for the aggregation process. This could be, for example, percent of the total power we choose to lose during the aggregation, or pre-defined number or average length of final phases.

2. Find the least powerful phase

$$j = \arg\min_i p_i$$

and aggregate it with its neighboring phase(s). If the jth phase is located not at the beginning or at the end of the time series, replace the three phases, $(j-1)$th, jth, and $(j+1)$th, with one phase which power is $p_{j-1} - p_j + p_{j+1}$. Hence, the number of phases decreases by two, while the power P decreases by $2p_j$. If the least powerful phase is the first (or the last) in the sequence, join it only with the one after, $(j+1)$th (or the one before, $(j-1)$th).

3. Repeat the previous step until the stopping criterion is met.

The time points $l_1, l_2, \ldots$ at which the final phases end (except the ending time of the last phase, which is T) can serve as an approximation for the change points to be tested.

CART Classification and regression trees (CART; Breiman et al, 1984) is a machine learning method that aims to learn a sequence of splitting rules that can split the observations of a response variable into relatively small homogeneous groups of observations, using associated values of predictors (also termed as splitting variables). When the response variable is numeric, the heterogeneity is quantified by the response within-group sum of squares.

For our task of identifying locations of possible change points, we set $\hat{\varepsilon}_t$ as the response, and use time index $t = 1, \ldots, T$ as the splitting variable. In this case, CART solutions identify periods of distinctively homogeneous $\hat{\varepsilon}_t$.

The algorithm proceeds starting with the whole data set of size T (Hastie et al, 2009):

1. Using an exhaustive search over all available t, select a value s that separates the response into two groups, $R_1(s)$ and $R_2(s)$, by solving

$$\min_s \left[\sum_{t \in R_1(s)} (\hat{\varepsilon}_t - \hat{c}_1)^2 + \sum_{t \in R_2(s)} (\hat{\varepsilon}_t - \hat{c}_2)^2 \right],$$

where $\hat{c}_1$ and $\hat{c}_2$ are average values of $\hat{\varepsilon}_t$ within the subperiods $R_1(s)$ and $R_2(s)$, respectively.

2. Repeat Step 1 on each of the two regions, $R_1(s)$ and $R_2(s)$.

3. Repeat Step 1 on all of the resulting regions until a stopping criterion is met.

Growing big trees with many splits can reduce the in-sample errors drastically (perfect goodness-of-fit), however, such 'deep' trees are unstable (adding or removing few observations may change the tree dramatically) and hard to interpret. Hence, a number of tuning parameters are used to define conditions when the next split should be attempted. Such parameters (e.g., minimal number of observations per group, *minbucket*, and maximal depth of the final tree, *maxdepth*) protect against over-fitting the data, but their values should be chosen adaptively from the data. Hastie et al (2009) propose a cost-complexity pruning technique that is based on growing a big tree until reaching, for example, some minimal node size. Then, the cost complexity criterion $C_\lambda(\Theta)$ is applied to prune the tree:

$$C_\lambda(\Theta) = \sum_{i=1}^{|\Theta|} N_i Q_i(\Theta) + \lambda |\Theta|, \tag{7}$$

where $|\Theta|$ is the number of terminal nodes in tree Θ; N_i is the number of observations in ith terminal node R_i (i.e., N_i is the node size); $Q_i(\Theta) = N_i^{-1} \sum_{t \in R_i} (\hat{\varepsilon}_t - \hat{c}_i)^2$ is the heterogeneity of ith node, and $\lambda \geqslant 0$ is the penalty parameter for the tradeoff between tree goodness-of-fit and complexity. Overall, (7) resembles the form of penalized regression estimators, such as the least absolute shrinkage and selection operator (LASSO), and, similarly, Hastie et al

(2009) suggest to select the optimal value of λ using a ten-fold cross-validation.

ACE Alternating conditional expectations (ACE; Breiman and Friedman, 1985) is an optimization procedure for suggesting smooth non-parametric transformations of regressors (and response) to maximize the proportion of explained variation.

Similar to CART, ACE can be employed to find breakpoints in regression residuals $\hat{\varepsilon}_t$ using the sequence $t = 1, \ldots, T$ as the predictor. The transformations $\hat{f}(\cdot)$ are chosen in the process of minimizing $\mathrm{E}\left\{\hat{\varepsilon}_t - f(t)\right\}^2$. Unlike phase analysis or CART, ACE do not provide clearly defined change point locations, and an additional step should be taken for inferring the candidate change points from ACE transformations (for example, using visual analysis of a plot of $\hat{f}(t)$ against t—see Soliman et al, 2015 and Lyubchich and Gel, 2017).

3 Simulation experiments

We apply the new sieve bootstrap-based testing procedure (Section 2.1) in a series of simulation experiments similar to of Horváth et al (2017). In particular, we simulate three types of error processes and incorporate them in two models for the change points (Model I with one change point and Model II with two change points) with varying size of the change, δ. We use the nominal significance level $\alpha = 0.05$, number of bootstrap replications $B = 1000$, and 5000 Monte Carlo runs for each combination. We use sample sizes $T = 100$ and 400 same as Horváth et al (2017), and additionally use $T = 30$ as it is close to the sample size in our case study (Section 4). The proportion of Monte Carlo runs when the null hypothesis (4) is rejected represents the empirical size of the test when $\delta = 0$, and power of the test when $\delta \neq 0$ (for both Model I and Model II).

Model I Assume $m = 1$ and

$$\mathbf{Y}_t = \begin{cases} \mathbf{X}_t \boldsymbol{\beta}^{(1)} + \boldsymbol{\varepsilon}_t, & \text{if } 1 \leqslant t \leqslant k_1^*, \\ \mathbf{X}_t \boldsymbol{\beta}^{(2)} + \boldsymbol{\varepsilon}_t, & \text{if } k_1^* + 1 \leqslant t \leqslant T, \end{cases}$$

where $\boldsymbol{\beta}^{(1)} = (0,1)^\top$, $\boldsymbol{\beta}^{(2)} = (0,1+\delta)^\top$; $\delta = -2, -1.8, \ldots, 1.8, 2$. Thus, Model I is a simple linear regression through the origin, with a change δ added to its only parameter after time k_1^*. With $k_1^* = \lfloor T\theta \rfloor$ and $\theta = \{0.2, 0.5, 0.9\}$, the cases of the change early, in the middle, and late are considered.

Model II Assume $m = 2$ and

$$\mathbf{Y}_t = \begin{cases} \mathbf{X}_t\boldsymbol{\beta}^{(1)} + \boldsymbol{\varepsilon}_t, & \text{if } 1 \leqslant t \leqslant k_1^*, \\ \mathbf{X}_t\boldsymbol{\beta}^{(2)} + \boldsymbol{\varepsilon}_t, & \text{if } k_1^* + 1 \leqslant t \leqslant k_2^*, \\ \mathbf{X}_t\boldsymbol{\beta}^{(3)} + \boldsymbol{\varepsilon}_t, & \text{if } k_2^* + 1 \leqslant t \leqslant T, \end{cases}$$

where $\boldsymbol{\beta}^{(1)} = (0,1)^\top$, $\boldsymbol{\beta}^{(2)} = (0,1+\delta)^\top$, $\boldsymbol{\beta}^{(3)} = (0,1-2\delta)^\top$; $\delta = -3, -2.8, \ldots, 2.8, 3$ (i.e., the definition of $\boldsymbol{\beta}^{(3)}$ and granularity of δ are different from Horváth et al, 2017). The change points $k_1^* = \lfloor T\theta_1 \rfloor$ and $k_2^* = \lfloor T\theta_2 \rfloor$ are set to split the series into equal periods with $(\theta_1, \theta_2) = (1/3, 2/3)$ or unequal, with $(\theta_1, \theta_2) = (0.2, 0.5)$ and $(\theta_1, \theta_2) = (0.5, 0.9)$.

In both models, $\mathbf{X}_t = (1, x_{t1})$, where x_{t1} are i.i.d. $N(1,1)$ random variables. The regression errors, ε_t, are obtained from the following three processes:

1. Independent standard normal: ε_t are i.i.d. $N(0,1)$;
2. GARCH(1,1): $\varepsilon_t = \sigma_t u_t$, $\sigma_t^2 = 0.25 + 0.25\varepsilon_{t-1}^2 + 0.5\sigma_{t-1}^2$, where u_t are i.i.d. $N(0,1)$;
3. AR(1): $\varepsilon_t = 0.5\varepsilon_{t-1} + u_t$, where u_t are i.i.d. $N(0,1)$.

Table 1 shows the empirical size of the new bootstrapped testing procedure when one or two change points (in Model I and Model II, respectively) split the series into equal parts[1]. When the error process is i.i.d. $N(0,1)$, the results are the most satisfactory as the observed rejection probabilities are close to the nominal $\alpha = 0.05$ even for small samples of size $T = 30$. With GARCH(1,1) errors and small sample size, the empirical size of the test is slightly higher than it should be under $\alpha = 0.05$, however, it quickly approaches the

[1]The results for other θ's—i.e., $\theta = 0.2$ and $\theta = 0.9$ in Model I and $(\theta_1, \theta_2) = (0.2, 0.5)$ and $(\theta_1, \theta_2) = (0.5, 0.9)$ in Model II—differ trivially from those shown in Table 1 (see Figures 1 and 2 when $\delta = 0$).

nominal level as the sample size increases (e.g., the size of the test under Model I is 0.059 for $T = 30$, and 0.049 for $T = 400$). With autoregressive dependence in the regression errors, as the sample size T increases, the empirical size of the test approaches the nominal α slower than in the uncorrelated case (see the last column of Table 1).

Overall, Table 1 shows that the performance of the data-driven bootstrapped procedure matches the parametric approach of Horváth et al (2017). It also shows that our non-parametric bootstrap approach can be implemented for small T when the regression errors are uncorrelated.

Figures 1 and 2 correspond to Model I and Model II and show the power of the new testing procedure when the change $\delta \neq 0$. With δ moving away from zero, the power of the test quickly approaches 1 (i.e., 100%), especially in large samples of $T = 100$ and $T = 400$. This growth of power is faster when the simulated regression errors are uncorrelated (the first two columns of plots in Figures 1 and 2) than when the errors are autocorrelated (the third column of plots in Figures 1 and 2). There is no substantial difference between the power curves for the i.i.d. $N(0, 1)$ and GARCH(1,1) errors (i.e., between the first and second columns of the plots).

By studying the curves in each of the plots, we observe that power is higher when the periods both before and after change points are longer—it corresponds to the cases of $\theta = 0.5$ and $(\theta_1, \theta_2) = (1/3, 2/3)$ that split the series into equal parts (the cases for which the size of the test is presented in Table 1). In other words, it is easier to detect change points if a single change point appears in the middle of the series, or if several change points stand far apart from each other and from the beginning and the end of the series.

Based on the power curves by Horváth et al (2017), our data-driven procedure is more powerful (see Figures 1d, 1e, and 1i), thus, should be preferred in real data applications. However, if autocorrelated regression errors are observed in a small sample, the user should be aware that both procedures may over-reject the null hypothesis.

4 Low-oxygen water in Chesapeake Bay

The frequency and severity of low oxygen volumes (hypoxia and anoxia) have been highly variable in past decades and have persisted as a significant environmental problem in Chesapeake Bay. Hypoxia is the dissolved oxygen concentration, often defined as < 2 mg/L, at which many aquatic organisms are physically stressed; while anoxia corresponds to complete depletion of oxygen, operationally defined as < 0.2 mg/L (Diaz and Rosenberg, 2008; Testa and Kemp, 2014; Wang et al, 2016). Reduced oxygen in the bottom water of the Chesapeake occurs naturally due to biological processes, but the extent and severity of hypoxia and anoxia has increased in the past as a result of elevated nitrogen loading into the bay resulting from anthropogenic activities in the watershed (for example, see Kemp et al, 1992; Hagy et al, 2004; Kemp et al, 2005; Scully, 2010; Testa and Kemp, 2014; Li et al, 2016 and references therein). Various physical and statistical models have been used to study the dynamics of oxygen-depletion events towards a better prediction of hypoxia (and anoxia) and more complete understanding the bay ecosystem as a whole (Murphy et al, 2011; Zhou et al, 2014). Although some relationships between hypoxia and its controlling variables are linear in nature, Conley et al (2009) considered regime shifts associated with the degradation of ecosystem buffers when hypoxic events are particularly severe, allowing the ecosystem to become more susceptible to future hypoxic events. Hagy et al (2004) and Testa and Kemp (2012) illustrated how Chesapeake Bay may have become more susceptible to nitrogen inputs over the period of 1950–2010. In the most recent four decades, sufficiently resolved oxygen measurements over space and time allow for more detailed statistical analyses of changes in Chesapeake Bay hypoxia.

Using high-quality data for the period 1985–2017 ($T = 33$ years), we investigate the relationships between anoxic events and the correlated factors in two regression models used to predict Chesapeake Bay early summer anoxic volumes (y_{1t}) and late summer anoxic volumes (y_{2t}) for public releases by the National Oceanic and Atmospheric Administration

(NOAA) of the USA[2] and originally developed in Murphy et al (2011). The anoxic volumes are calculated using bay-wide statistical interpolation of dissolved oxygen concentrations sampled from the main stem of Chesapeake Bay by the Chesapeake Bay Program (Murphy et al, 2011; Testa and Kemp, 2012). The sampling is conducted in May through first two weeks of July to estimate early summer volumes, and in the second two weeks of July through September for the late summer volumes. The following factors were investigated as potentially correlated with the anoxic volumes:

- total nitrogen load from Susquehanna and Potomac Rivers during January–April, $JanAprTNLoad_t$ (kg/day),
- total nitrogen load from Susquehanna River during January–May, $JanMayTNLoad_t$ (kg/day),
- freshwater discharge from Susquehanna and Potomac Rivers in May, $MayFlow_t$ (m^3/s),
- mean sea level, MSL_t (m), and
- fraction of hours with southeastern wind over Chesapeake Bay during March–May.

The final regression models, obtained with step-wise model selection based on AIC and analysis of statistical significance of the coefficients, retain only $JanAprTNLoad_t$ as a predictor for early summer anoxic volumes, and $JanMayTNLoad_t$ for late summer anoxic volumes:

$$\hat{y}_{1t} = -\underset{(0.405)}{0.980} + \underset{(1.069\cdot 10^{-6})}{6.903 \cdot 10^{-6}} JanAprTNLoad_t, \tag{8}$$

$$\hat{y}_{2t} = -\underset{(0.426)}{0.217} + \underset{(1.360\cdot 10^{-6})}{5.596 \cdot 10^{-6}} JanMayTNLoad_t, \tag{9}$$

[2]E.g., see http://www.noaa.gov/media-release/noaa-usgs-and-partners-predict-larger-summer-dead-zone-for-chesapeake-bay and http://ian.umces.edu/ecocheck/forecast/chesapeake-bay/2017/.

where standard errors of the coefficients are given in parentheses. Residuals $\hat{\varepsilon}_{1t}$ and $\hat{\varepsilon}_{2t}$ of the respective models show no strong patterns or significant autocorrelations (Figure 3). Hence, the observed residuals are more similar to the first type of errors explored in the simulations in Section 3—independent normal—rather than to the GARCH or AR types. At the next step, we identify possible change points using the methods outlined in Section 2.2, then apply the data-driven testing procedure that is described in Section 2.1. All test results are combined in Table 2.

We applied an exhaustive search of up to three change points (we do not expect more than three change points in the relatively short time series of $T = 33$ observations)

$$\hat{k}_1, \ldots, \hat{k}_m = \arg\max\{M(k_1, k_2, k_3),\ 1 \leqslant k_1 \leqslant k_2 \leqslant k_3 < T\},$$

and in both cases (for early and late summer anoxic volumes) two candidate change points were found that were further tested using the suggested sieve bootstrap approach (Table 2).

In the phase analysis of residuals from each regression model, the phase aggregation was stopped to retain 50% of the total power. This gave two change points for the model of early summer anoxia (1992 and 2013) and one change point, at 2001, for the model of late summer anoxia (Table 2).

In the CART analysis, the data can be generally split into $2^{maxdepth}$ groups, and the value $maxdepth = 2$ gives us potentially up to four groups (i.e., up to three change points). Additionally, we set $minbucket = \lceil 0.1T \rceil = 4$, i.e., 10% of the available data, rounded up (the CART algorithm implementation by Therneau and Atkinson, 2018 was used in this analysis). The results in Figure 4 give us two change points in each case (with up to three possible under the current settings of the tuning parameters). Notice that based on the definitions in Section 2, the change points $k_1, \ldots, k_m$ split the data set into periods $[1, k_1], [k_1 + 1, k_2], \ldots, [k_m + 1, T]$ (i.e., each change point is the last point in the respective subseries), hence, the change points from CART analysis are 1988 and 2013 for the early

summer anoxia model, and 2001 and 2009 for the late summer anoxia model (Figure 4).

The obtained ACE transformations (using the algorithm by Spector et al, 2016) allowed us to select easily at least one change point, closer to the end of the series (i.e., the change point at 2012 in Figure 5a and at 2013 in Figure 5b when the rapid changes start to occur). While the rest of the transformed series look non-linear, it is less obvious where additional change points occurred. For the model of early summer anoxia, we hypothesized an additional change point at 2002 as it is the time when $\hat{f}(Year)$ in Figure 5a start declining. The year of 1992 was selected as another candidate change point for late summer anoxic volume regression model for the same reason (Figure 5b).

Thus, all the employed methods (exhaustive search, phase analysis, CART, and ACE) suggest roughly two change points in the model coefficients during the analyzed period (Table 2). The methods, however, sometimes disagree on where the change points are located. We observed a good correspondence between the exhaustive search and CART: CART-selected change points are the closest to exhaustive search in terms of their location and p-values. Hence, CART can be the preferred method for identifying change points when exhaustive search is not computationally feasible. The change points selected based on phase analysis led to the test p-value at the border of statistical significance for the early summer anoxia model (p-value of 0.0527). For the late summer model, the phase analysis identified only one change point, whereas all other methods identified two. The ACE results were the least consistent with other methods—the ACE-selected thresholds were located at very different years, what could lead to completely missing the change points in the early summer anoxia model (the p-value for ACE-selected change points is 0.3306, whereas CART and exhaustive search could identify significant change points).

The mechanisms behind the timing of the identified change points cannot be definitively explained, but there are clear associations with these time points and periods of environmental change in Chesapeake Bay. The fact that the change points were slightly different for the two periods is not surprising, given that the temporal trend in the early and late sum-

mer low-oxygen volumes have been shown to be in different directions (Murphy et al, 2011). The ACE and exhaustive search methods generally suggest a change point in the 2013–2015 period where the model residuals were negative, which coincides with a period of relatively low eutrophic conditions in Chesapeake Bay. From 2013 to 2017, Susquehanna River discharge was consistently average or low, and metrics of dissolved oxygen (Zhang et al, 2018), nutrient availability (Testa et al, 2018), and submerged aquatic vegetation (Orth et al, 2017) all indicated that eutrophication effects were weak. Thus, it appears that internal processes that control a wide variety of ecosystem properties were better than expected from nutrient loading from the Susquehanna, which is consistent with lower-than-predicted volumes of anoxia. Sampling difficulties in 2016 and 2017 may also have biased the observed volumes to be lower[3]. The CART, ACE, and exhaustive search methods also indicated a change point in 2001 and 2002 for anoxic volume in both periods. This time period corresponds with the end of a prolonged drought in the Bay watershed (1999–2002), which is associated with the resurgence of a large submerged aquatic vegetation bed in the upper Chesapeake Bay (Gurbisz and Kemp, 2014) and a landward shift in the peak winter-spring phytoplankton biomass from the lower to the upper Bay (Testa et al, 2018). Both of these changes would be expected to lead to elevated nutrient retention in the upper regions of Chesapeake Bay, thus reducing the fraction of watershed nitrogen load that reaches seaward waters where it can eventually support oxygen depletion. Future research can better resolve the mechanisms responsible for these change points by examining volumes of other oxygen thresholds (e.g., 1 mg/L) and including other forcing variables (e.g., summer wind speeds). For example, a preliminary analysis of the linear relationship between late summer anoxic volumes and wind speeds (m/s) in 1986–2015 detected two change points: 1998 and 2005. In the 1998–2005 period, both the level and variability of wind speeds were high, and regression analysis showed a significant linear relationship for these variables ($\hat{y}_{2t} = 9.740 - 2.401 WindSpeed$; p-value 0.033), whereas no such relationship was detected for the years outside 1998–2005.

[3] https://news.maryland.gov/dnr/2017/10/26/summer-2017-hypoxia/

Thus, this initial analysis reveals that wind speed has a potentially significant, but secondary impact on anoxic volume.

5 Conclusions

In this paper we propose a data-driven non-parametric sieve bootstrap framework for testing at-most-m change points in coefficients of a linear regression model. The test statistic has the form of modified CUSUM by Horváth et al (2017). Our simulation studies indicate that the new testing procedure outperforms its asymptotic approximation counterpart.

We illustrate the new approach by applying it to the data on Chesapeake Bay ecosystem, where annual re-occurrence of 'dead zones' with extremely low concentration of oxygen in the water has been a long-lasting problem. Our results for 1985–2017 show two statistically significant changes in the coefficients of simple linear regression models quantifying the relationships between anoxic water volumes and nitrogen loadings entering the bay.

Possible lines of future research include providing theoretical proofs of consistency of the bootstrap procedure, identifying if other processes linked to anoxia in Chesapeake Bay had change points co-occurring with the ones identified, and expanding the testing approach to other types of data beyond time series.

The code for the method is available from `mcusum.test` function in R package *funtimes* (Lyubchich and Gel, 2019); the data that support the findings of this study are available from the corresponding author upon request.

Acknowledgements

We thank Yulia R. Gel for several valuable suggestions in the preparation of this paper, and Robert Jarrett for the help with Chesapeake Bay data analysis at the initial steps of the project. The work of V. L. and J. M. T. was supported in part by the funds from the U.S. National Oceanic and Atmospheric Administration (NOAA) grant #NA12OAR4320071. This

is UMCES contribution No XXXX.

References

Andersen T, Carstensen J, Hernandez-Garcia E, Duarte CM (2009) Ecological thresholds and regime shifts: approaches to identification. Trends in Ecology & Evolution 24(1):49–57, DOI 10.1016/j.tree.2008.07.014

Antoch J, Jarušková D (2013) Testing for multiple change points. Computational Statistics 28(5):2161–2183, DOI 10.1007/s00180-013-0401-1

Antoch J, Hušková M, Veraverbeke N (1995) Change-point problem and bootstrap. Journal of Nonparametric Statistics 5(2):123–144, DOI 10.1080/10485259508832639

Aue A, Horváth L, Hušková M, Kokoszka P (2006) Change-point monitoring in linear models. The Econometrics Journal 9:373–403, DOI 10.1111/j.1368-423X.2006.00190.x

Aue A, Horváth L, Hušková M, Kokoszka P (2008) Testing for changes in polynomial regression. Bernoulli 14(3):637–660, DOI 10.3150/08-BEJ122

Bai J, Lumsdaine RL, Stock JH (1998) Testing for and dating common breaks in multivariate time series. The Review of Economic Studies 65(3):395–432, DOI 10.1111/1467-937X.00051

Breiman L, Friedman JH (1985) Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association 80:580–598, DOI 10.1080/01621459.1985.10478157

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees. Taylor & Francis, New York

Breitburg D, Levin LA, Oschlies A, Grégoire M, Chavez FP, Conley DJ, Garçon V, Gilbert D, Gutiérrez D, Isensee K, Jacinto GS, Limburg KE, Montes I, Naqvi SWA, Pitcher GC, Rabalais NN, Roman MR, Rose KA, Seibel BA, Telszewski M, Yasuhara M, Zhang J (2018) Declining oxygen in the global ocean and coastal waters. Science 359(6371):eaam7240,

DOI 10.1126/science.aam7240

Busetti F, Taylor AMR (2004) Tests of stationarity against a change in persistence. Journal of Econometrics 123(1):33–66, DOI 10.1016/j.jeconom.2003.10.028

Cavaliere G, Taylor AMR (2008) Testing for a change in persistence in the presence of non-stationary volatility. Journal of Econometrics 147(1):84–98, DOI 10.1016/j.jeconom.2008.09.004

Chatterjee S, Qiu P (2009) Distribution-free cumulative sum control charts using bootstrap-based control limits. The Annals of Applied Statistics 3(1):349–369, DOI 10.1214/08-AOAS197

Conley DJ, Carstensen J, Vaquer-Sunyer R, Duarte CM (2009) Ecosystem thresholds with hypoxia. Hydrobiologia 629(1):21–29, DOI 10.1007/s10750-009-9764-2

Diaz RJ, Rosenberg R (2008) Spreading dead zones and consequences for marine ecosystems. Science 321(5891):926–929, DOI 10.1126/science.1156401

Eichinger B, Kirch C (2018) A mosum procedure for the estimation of multiple random change points. Bernoulli 24(1):526–564, DOI 10.3150/16-BEJ887

Gallagher C, Lund RB, Robbins M (2013) Changepoint detection in climate time series with long-term trends. Journal of Climate 26(14):4994–5006, DOI 10.1175/JCLI-D-12-00704.1

Gandy A, Kvaløy JT (2013) Guaranteed conditional performance of control charts via bootstrap methods. Scandinavian Journal of Statistics 40:647–668, DOI 10.1002/sjos.12006

Gombay E (2010) Change detection in linear regression with time series errors. Canadian Journal of Statistics 38(1):65–79, DOI 10.1002/cjs.10043

Gombay E, Horváth L (1999) Change-points and bootstrap. Environmetrics 10:725–736, DOI 10.1002/(SICI)1099-095X(199911/12)10:6⟨725::AID-ENV387⟩3.0.CO;2-K

Górecki T, Horváth L, Kokoszka P (2018) Change point detection in heteroscedastic time series. Econometrics and Statistics 7:63–88, DOI 10.1016/j.ecosta.2017.07.005

Gurbisz C, Kemp WM (2014) Unexpected resurgence of a large submersed plant bed in Chesapeake Bay: Analysis of time series data. Limnology and Oceanography 59(2):482–

494, DOI 10.4319/lo.2014.59.2.0482

Hagy JD, Boynton WR, Keefe CW, Wood KV (2004) Hypoxia in Chesapeake Bay, 1950–2001: long-term change in relation to nutrient loading and river flow. Estuaries 27(4):634–658, DOI 10.1007/BF02907650

Hall P, Van Keilegom I (2003) Using difference-based methods for inference in nonparametric regression with time series errors. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65(2):443–456, DOI 10.1111/1467-9868.00395

Hanson BE (2002) Tests for parameter instability in regressions with I(1) processes. Journal of Business & Economic Statistics 20(1):45–59, DOI 10.1198/073500102753410381

Hare SR, Mantua NJ (2000) Empirical evidence for North Pacific regime shifts in 1977 and 1989. Progress in Oceanography 47(2–4):103–145, DOI 10.1016/S0079-6611(00)00033-1

Hastie TJ, Tibshirani RJ, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York, DOI 10.1007/978-0-387-84858-7

Hlávka Z, Hušková M, Kirch C, Meintanis SG (2016) Bootstrap procedures for online monitoring of changes in autoregressive models. Communications in Statistics – Simulation and Computation 45(7):2471–2490, DOI 10.1080/03610918.2014.904346

Horváth L, Rice G (2014) Extensions of some classical methods in change point analysis. Test 23:219–255, DOI 10.1007/s11749-014-0368-4

Horváth L, Hušková M, Kokoszka P, Steinebach J (2004) Monitoring changes in linear models. Journal of Statistical Planning and Inference 126(1):225–251, DOI 10.1016/j.jspi.2003.07.014

Horváth L, Pouliot W, Wang S (2017) Detecting at-most-m changes in linear regression models. Journal of Time Series Analysis 38:552–590, DOI 10.1111/jtsa.12228

Hušková M, Kirch C (2008) Bootstrapping confidence intervals for the change-point of time series. Journal of Time Series Analysis 29(6):947–972, DOI 10.1111/j.1467-9892.2008.00589.x

596 Hušková M, Kirch C (2012) Bootstrapping sequential change-point tests for linear regression.
597 Metrika 75(5):673–708, DOI 10.1007/s00184-011-0347-7
598 Hušková M, Kirch C, Prášková Z, Steinebach J (2008) On the detection of changes in au-
599 toregressive time series, II. Resampling procedures. Journal of Statistical Planning and
600 Inference 138(6):1697–1721, DOI 10.1016/j.jspi.2007.06.029
601 Jandhyala VK, Al-Saleh JA (1999) Parameter changes at unknown times in non-linear re-
602 gression. Environmetrics 10:711–724, DOI 10.1002/(SICI)1099-095X(199911/12)10:6⟨711::
603 AID-ENV398⟩3.0.CO;2-T
604 Jandhyala VK, MacNeill IB (1989) Residual partial sum limit process for regression models
605 with applications to detecting parameter changes at unknown times. Stochastic Processes
606 and their Applications 33(2):309–323, DOI 10.1016/0304-4149(89)90045-8
607 Jandhyala VK, MacNeill IB (1991) Tests for parameter changes at unknown times in linear
608 regression models. Journal of Statistical Planning and Inference 27(3):291–316, DOI 10.
609 1016/0378-3758(91)90043-E
610 Jandhyala VK, Zacks S, El-Shaarawi AH (1999) Change-point methods and their applica-
611 tions: contributions of Ian MacNeill. Environmetrics 10:657–676, DOI 10.1002/(SICI)109
612 9-095X(199911/12)10:6⟨657::AID-ENV390⟩3.0.CO;2-Y
613 Jarušková D (2003) Asymptotic distribution of a statistic testing a change in simple linear
614 regression with equidistant design. Statistics & Probability Letters 64(1):89–95, DOI
615 10.1016/S0167-7152(03)00143-3
616 Kejriwal M, Perron P (2008) The limit distribution of the estimates in cointegrated regression
617 models with multiple structural changes. Journal of Econometrics 146(1):59–73, DOI
618 10.1016/j.jeconom.2008.07.001
619 Kemp WM, Sampou PA, Garber J, Tuttle J, Boynton WR (1992) Seasonal depletion of
620 oxygen from bottom waters of Chesapeake Bay: roles of benthic and planktonic respiration
621 and physical exchange processes. Marine Ecology Progress Series 85(1–2):137–152, DOI
622 10.3354/meps085137

623 Kemp WM, Boynton WR, Adolf JE, Boesch DF, Boicourt WC, Brush G, Cornwell JC,
624 Fisher TR, Glibert PM, Hagy JD, Harding LW, Houde ED, Kimmel DG, Miller WD,
625 Newell RIE, Roman MR, Smith EM, Stevenson JC (2005) Eutrophication of Chesapeake
626 Bay: historical trends and ecological interactions. Marine Ecology Progress Series 303:1–
627 29, DOI 10.3354/meps303001
628 Kemp WM, Testa JM, Conley DJ, Gilbert D, Hagy JD (2009) Temporal responses of coastal
629 hypoxia to nutrient loading and physical controls. Biogeosciences 6(12):2985–3008, DOI
630 10.5194/bg-6-2985-2009
631 Kirch C (2006) Resampling methods for the change analysis of dependent data. PhD thesis,
632 Universität zu Köln
633 Kirch C (2007) Resampling in the frequency domain of time series to determine critical values
634 for change-point tests. Statistics & Decisions 25(3):237–261, DOI 10.1524/stnd.2007.0902
635 Lee J, Lund RB (2004) Revisiting simple linear regression with autocorrelated errors.
636 Biometrika 91(1):240–245, DOI 10.1093/biomet/91.1.240
637 Li M, Lee YJ, Testa JM, Li Y, Ni W, Kemp WM, Di Toro DM (2016) What drives interan-
638 nual variability of hypoxia in Chesapeake Bay: Climate forcing versus nutrient loading?
639 Geophysical Research Letters 43(5):2127–2134, DOI 10.1002/2015GL067334
640 Li S, Lund RB (2012) Multiple changepoint detection via genetic algorithms. Journal of
641 Climate 25(2):674–686, DOI 10.1175/2011JCLI4055.1
642 Li Y, Lund RB (2015) Multiple changepoint detection using metadata. Journal of Climate
643 28(10):4199–4216, DOI 10.1175/JCLI-D-14-00442.1
644 Li Y, Lund RB, Hewaarachchi A (2017) Multiple changepoint detection with partial infor-
645 mation on changepoint times. arXiv preprint arXiv:151107238v3
646 Lukashin YP (2003) Adaptive Methods of Short-term Forecasting of Time Series. Finances
647 and Statistics, Moscow
648 Lund RB, Wang XL, Lu QQ, Reeves J, Gallagher C, Feng Y (2007) Changepoint detection
649 in periodic and autocorrelated time series. Journal of Climate 20(20):5178–5190, DOI

10.1175/JCLI4291.1

Lyubchich V, Gel YR (2016) A local factor nonparametric test for trend synchronism in multiple time series. Journal of Multivariate Analysis 150:91–104, DOI 10.1016/j.jmva.2016.05.004

Lyubchich V, Gel YR (2017) Can we weather proof our insurance? Environmetrics 28(2):e2433, DOI 10.1002/env.2433

Lyubchich V, Gel YR (2019) funtimes: Functions for Time Series Analysis. URL https://CRAN.R-project.org/package=funtimes, R package version 6.1

Murphy RR, Kemp WM, Ball WP (2011) Long-term trends in Chesapeake Bay seasonal hypoxia, stratification, and nutrient loading. Estuaries and Coasts 34(6):1293–1309, DOI 10.1007/s12237-011-9413-7

Orth RJ, Dennison WC, Lefcheck JS, Gurbisz C, Hannam M, Keisman J, Landry JB, Moore KA, Murphy RR, Patrick CJ, Testa JM, Weller DE, Wilcox DJ (2017) Submersed aquatic vegetation in Chesapeake Bay: sentinel species in a changing world. Bioscience 67(8):698–712, DOI 10.1093/biosci/bix058

Quandt RE (1958) The estimation of the parameters of a linear regression system obeying two separate regimes. Journal of the American Statistical Association 53(284):873–880, DOI 10.2307/2281957

Quandt RE (1960) Tests of the hypothesis that a linear regression system obeys two separate regimes. Journal of the American Statistical Association 55(290):324–330, DOI 10.1080/01621459.1960.10482067

Reeves J, Chen J, Wang XL, Lund RB, Lu QQ (2007) A review and comparison of change-point detection techniques for climate data. Journal of Applied Meteorology and Climatology 46(6):900–915, DOI 10.1175/JAM2493.1

Robbins M, Gallagher C, Lund RB, Aue A (2011) Mean shift testing in correlated data. Journal of Time Series Analysis 32(5):498–511, DOI 10.1111/j.1467-9892.2010.00707.x

Scavia D, Allan JD, Arend KK, Bartell S, Beletsky D, Bosch NS, Brandt SB, Briland RD,

677 Daloğlu I, DePinto JV, Dolan DM, Evans MA, Farmer TM, Goto D, Han H, Höök TO,
678 Knight R, Ludsin SA, Mason D, Michalak AM, Richards RP, Roberts JJ, Rucinski DK,
679 Rutherford E, Schwab DJ, Sesterhenn TM, Zhang H, Zhou Y (2014) Assessing and ad-
680 dressing the re-eutrophication of Lake Erie: Central basin hypoxia. Journal of Great Lakes
681 Research 40(2):226–246, DOI 10.1016/j.jglr.2014.02.004
682 Scheffer M, Jeppesen E (2007) Regime shifts in shallow lakes. Ecosystems 10(1):1–3, DOI
683 10.1007/s10021-006-9002-y
684 Scully ME (2010) The importance of climate variability to wind-driven modulation of hypoxia
685 in Chesapeake Bay. Journal of Physical Oceanography 40(6):1435–1440, DOI 10.1175/201
686 0JPO4321.1
687 Seijo E, Sen B (2011) Change-point in stochastic design regression and the bootstrap. The
688 Annals of Statistics 39(3):1580–1607, DOI 10.1214/11-AOS874
689 Soliman M, Lyubchich V, Gel YR, Naser D, Esterby S (2015) Evaluating the impact of
690 climate change on dynamics of house insurance claims, Springer, Switzerland, chap 16, pp
691 175–183. DOI 10.1007/978-3-319-17220-0_16
692 Spector P, Friedman J, Tibshirani R, Lumley T, Garbett S, Baron J (2016) acepack: ACE
693 and AVAS for Selecting Multiple Regression Transformations. URL https://CRAN.R-pro
694 ject.org/package=acepack, R package version 1.4.1
695 Stine RA (2004) Model selection using information theory and the MDL principle. Sociolog-
696 ical Methods & Research 33(2):230–260, DOI 10.1177/0049124103262064
697 Sturdivant SK, Brush MJ, Diaz RJ (2013) Modeling the effect of hypoxia on macroben-
698 thos production in the lower Rappahannock River, Chesapeake Bay, USA. PloS One
699 8(12):e84140, DOI 10.1371/journal.pone.0084140
700 Testa JM, Kemp WM (2012) Hypoxia-induced shifts in nitrogen and phosphorus cycling in
701 Chesapeake Bay. Limnology and Oceanography 57(3):835–850, DOI 10.4319/lo.2012.57.3
702 .0835
703 Testa JM, Kemp WM (2014) Spatial and temporal patterns of winter–spring oxygen de-

pletion in Chesapeake Bay bottom water. Estuaries and Coasts 37(6):1432–1448, DOI 10.1007/s12237-014-9775-8

Testa JM, Murphy RR, Brady DC, Kemp WM (2018) Nutrient- and climate-induced shifts in the phenology of linked biogeochemical cycles in a temperate estuary. Frontiers in Marine Science 5:114, DOI 10.3389/fmars.2018.00114

Therneau T, Atkinson B (2018) rpart: Recursive Partitioning and Regression Trees. URL https://CRAN.R-project.org/package=rpart, R package version 4.1-13

Vaman HJ (1985) Optimal online detection of parameter changes in two linear models. Stochastic Processes and Their Applications 20(2):343–351, DOI 10.1016/0304-4149(85)90221-2

Wang P, Wang H, Linker L, Hinson K (2016) Influence of wind strength and duration on relative hypoxia reductions by opposite wind directions in an estuary with an asymmetric channel. Journal of Marine Science and Engineering 4(3):62–85, DOI 10.3390/jmse4030062

Zeileis A, Kleiber C, Krämer W, Hornik K (2003) Testing and dating of structural changes in practice. Computational Statistics & Data Analysis 44(1-2):109–123, DOI 10.1016/S0167-9473(03)00030-6

Zhang Q, Murphy RR, Tian R, Forsyth MK, Trentacoste EM, Keisman J, Tango PJ (2018) Chesapeake Bay's water quality condition has been recovering: Insights from a multimetric indicator assessment of thirty years of tidal monitoring data. Science of The Total Environment 637:1617–1625, DOI 10.1016/j.scitotenv.2018.05.025

Zhao MJ, Driscoll AR (2016) The c-chart with bootstrap adjusted control limits to improve conditional performance. Quality and Reliability Engineering International 32(8):2871–2881, DOI 10.1002/qre.1971

Zhou Y, Scavia D, Michalak AM (2014) Nutrient loading and meteorological conditions explain interannual variability of hypoxia in Chesapeake Bay. Limnology and Oceanography 59(2):373–384, DOI 10.4319/lo.2014.59.2.0373

Table 1: Empirical size (in both models, $\delta = 0$) of the bootstrapped test under different specifications of the models and error processes

		Error process ε_t		
Model	T	i.i.d. $N(0,1)$	GARCH(1,1)	AR(1)
Model I, $m = 1$ and $\theta = 0.5$	30	0.049	0.059	0.185
	100	0.057	0.055	0.075
	400	0.056	0.049	0.059
Model II, $m = 2$ and $(\theta_1, \theta_2) = (1/3, 2/3)$	30	0.051	0.063	0.219
	100	0.053	0.052	0.083
	400	0.053	0.052	0.062

Table 2: Results of identifying and testing ($B = 10^4$) for change points in regression coefficients of the regression models for anoxic volumes in Chesapeake Bay

Regression model	Method of selecting change points	Change points $\hat{k}_1, \ldots, \hat{k}_m$	Bootstrapped p-value
Early summer (8)	Exhaustive search	1987, 2014	0.0039
	Phase analysis	1992, 2013	0.0527
	CART	1988, 2013	0.0238
	ACE	2002, 2012	0.3306
Late summer (9)	Exhaustive search	2001, 2015	0.0017
	Phase analysis	2001	0.0125
	CART	2001, 2009	0.0077
	ACE	1992, 2013	0.0078

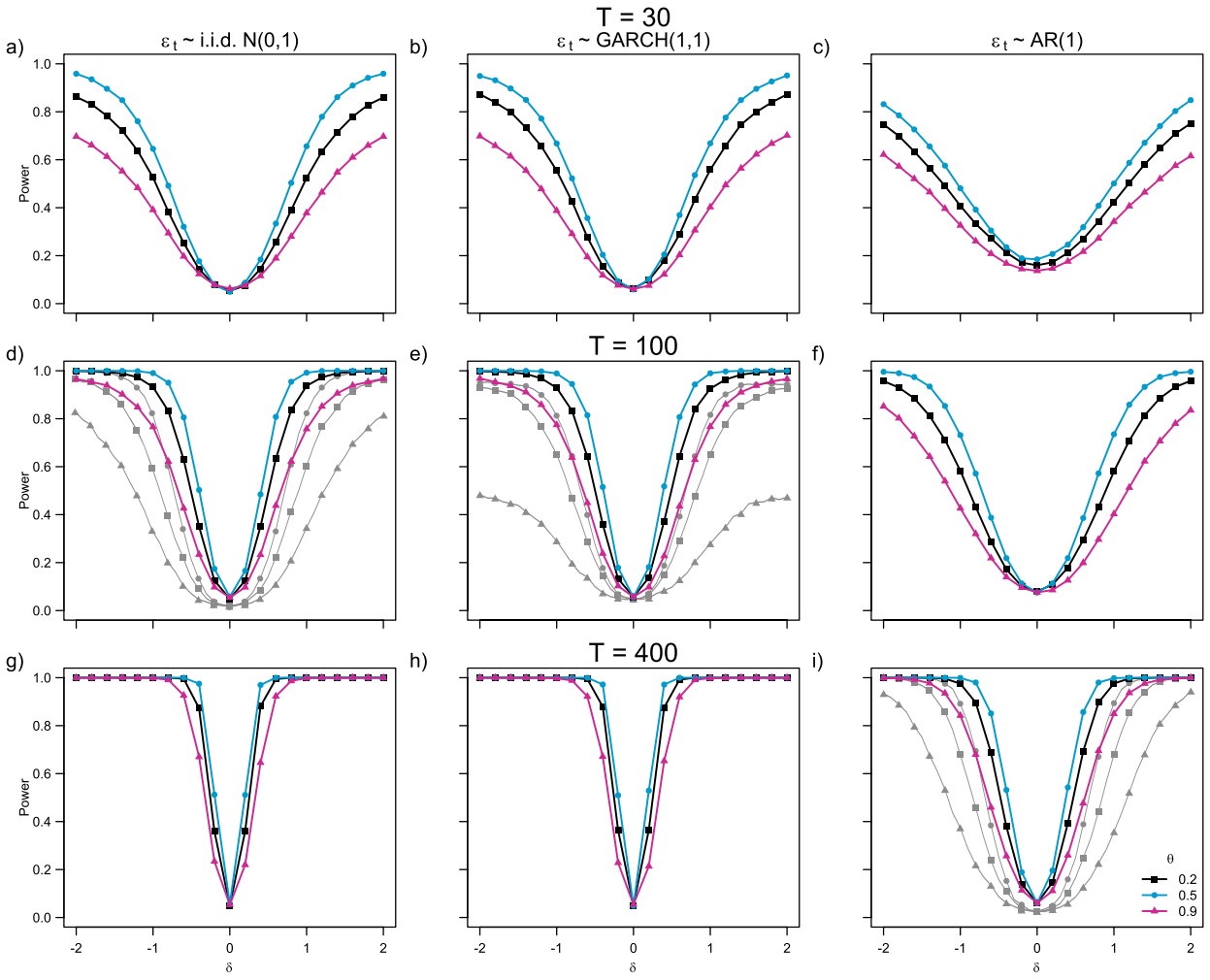

Figure 1: The empirical power functions of the bootstrapped test for testing the null hypothesis of at-most-one change point at $k_1^* = \lfloor T\theta \rfloor$, when the data are simulated using Model I. Grey lines in (d), (e), and (i) correspond to $V^{(3)}$ curves from respective Figures 4, 5, and 6 of Horváth et al (2017).

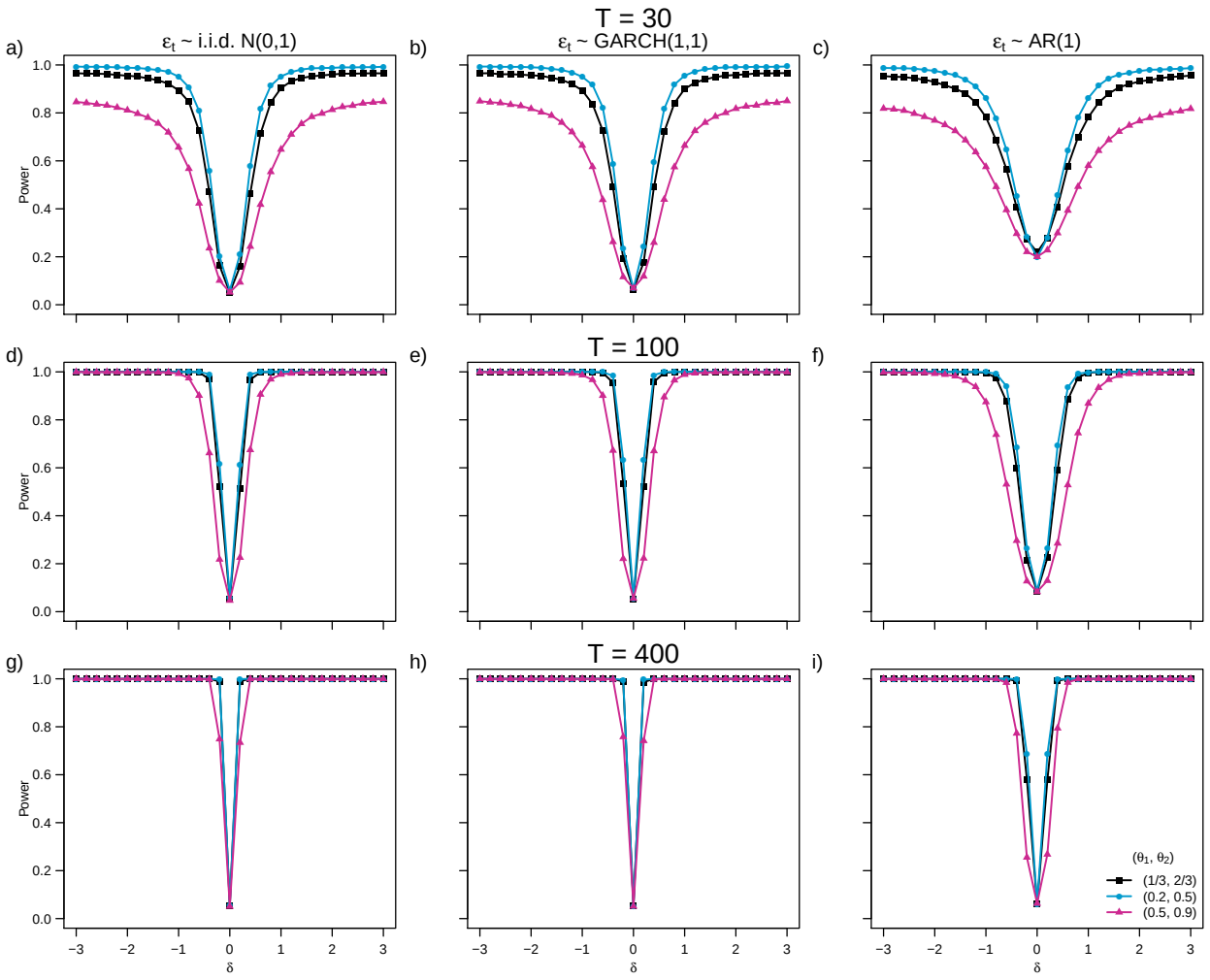

Figure 2: The empirical power functions of the bootstrapped test for testing the null hypothesis of at-most-two change points at $k_1^* = \lfloor T\theta_1 \rfloor$ and $k_2^* = \lfloor T\theta_2 \rfloor$, when the data are simulated using Model II.

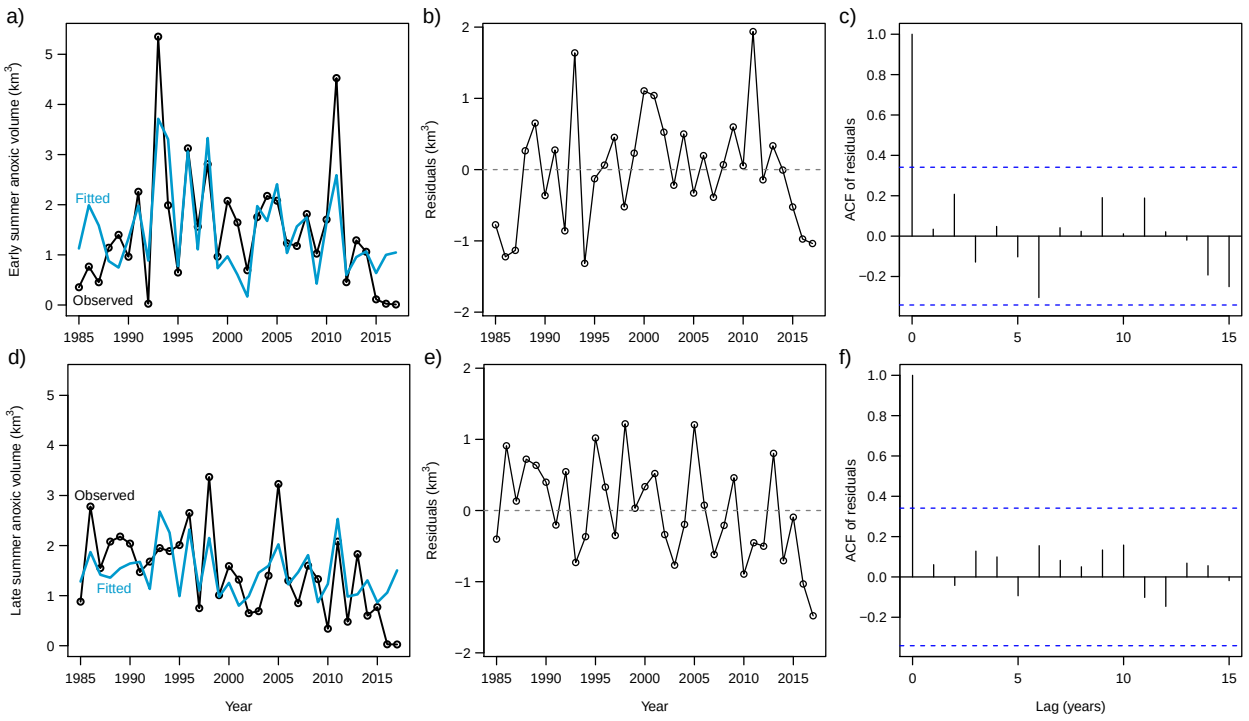

Figure 3: Regression models of anoxic volumes in Chesapeake Bay: a) anoxic volumes in early summer and fitted with (8); d) anoxic volumes in late summer and fitted with (9); b) and e) respective residuals; c) and f) sample autocorrelation functions, ACFs, of the residuals.

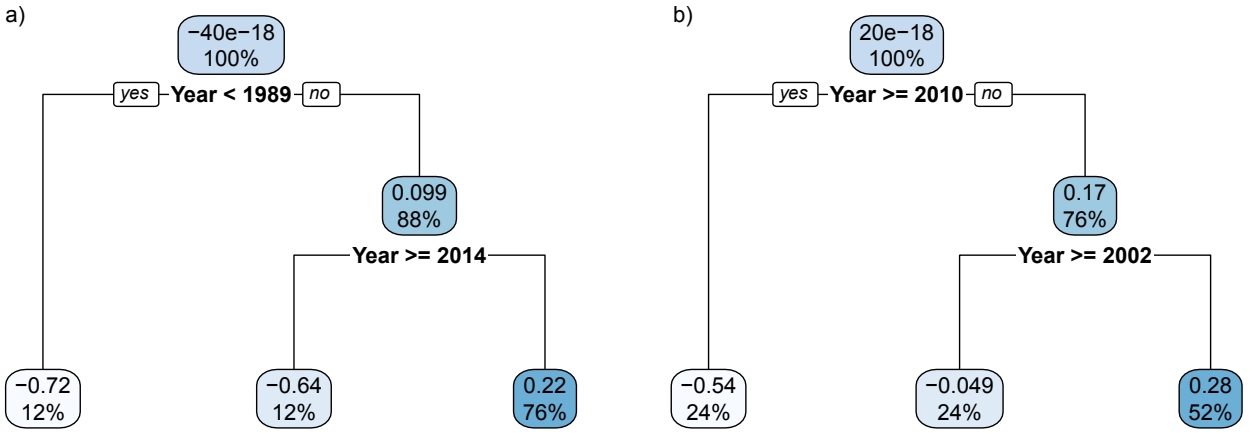

Figure 4: Classification and regression trees (CART) applied to residuals of: a) model (8) for anoxic volumes in early summer, and b) model (9) for anoxic volumes in late summer. For each node, average value of the residuals is reported along with the node size expressed as percentage of the total sample size $T = 33$.

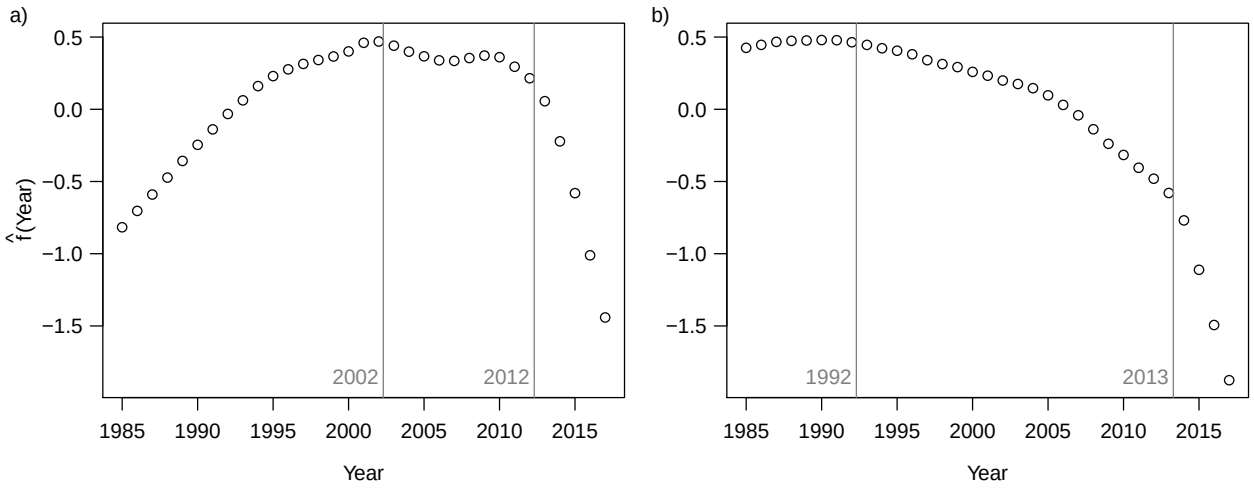

Figure 5: Transformations $\hat{f}(t)$ estimated using alternating conditional expectations for the residuals of: a) model (8) for anoxic volumes in early summer, and b) model (9) for anoxic volumes in late summer. The labeled vertical lines denote visually identified change points.