available at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.elsevier.com/locate/ecocom

ELSEVIER

ECOLOGICAL COMPLEXITY

Identification of regime shifts in time series using neighborhood statistics

Christopher W. Pawlowski [1], *Heriberto Cabezas* [*]

US EPA Office of Research and Development, 26 W. Martin Luther King Dr., Cincinnati, OH 45268, United States

ARTICLE INFO

Article history:
Received 17 July 2007
Accepted 27 July 2007
Published on line 14 September 2007

Keywords:
Nonlinear dynamics
Alternative dynamic regimes
Detection
Thresholds
Regime shifts
Time series
Statistical methods

ABSTRACT

Sudden and significant changes in biotic and abiotic variables have been observed across a variety of systems. The identification of such regime shifts in time series includes both model-fitting and statistical approaches. We introduce two methods that use state- or measurement-space neighborhood statistics to pick out regime shifts. Analysis of simulated and real data sets shows that these methods can be an effective means of identifying regime shifts for single variable as well as multivariable time series. In addition, these methods can be used on systems with non-equilibrium steady states. However, care must be taken in interpreting results as these methods do respond to changes in time series that are not consistent with the regime shift concept.

1. Introduction

Sudden and significant changes in biotic and abiotic measures have been identified across a variety of system types from the distant past to the present (Scheffer et al., 2001; Burkett et al., 2005). While evidence for such 'regime shifts' is cited in a number of disciplines, the characterization and understanding of the phenomenon is not straightforward. In ecology, early theoretical work on alternative stable states (Lewontin, 1969; Holling, 1973; May, 1977) set the stage for recent theoretical and empirical studies explaining large changes in observable variables in response to small changes in conditions in terms of shifts between system attractors and their basins of attraction brought about by bifurcations (Scheffer et al., 1993; Carpenter et al., 1999; Anderies et al., 2002; Scheffer and Carpenter, 2003). Bifurcations can involve the appearance (or disappearance) of attractors as well as changes in the boundary between basins of attraction of existing system attractors (Vandermeer and Yodzis, 1999).

The regime shift concept in marine systems has been more controversial. The difficulty in identifying and explaining regime shifts stems from the fact that these are large, open systems for which the concept of state is not well defined, and the underlying mechanisms for change are poorly understood (de Young et al., 2004). Some have argued that observed changes in system variables need not be anything but normal statistical deviations (Wunsch, 1999; Rudnick and Davis, 2003). There is evidence, however, that deviations in at least some biotic variables are better explained by low dimensional, deterministic nonlinear processes characteristic of alternative stable states than simply linear stochastic processes (Hsieh et al., 2005). The difficulty in studying these large systems has led to a pragmatic approach to the

* *Corresponding author.* Tel.: +1 513 569 7350; fax: +1 513 487 2511.
E-mail address: cabezas.heriberto@epa.gov (H. Cabezas).

[1] Metcalf & Eddy, United States.

1476-945X/$ – see front matter
doi:10.1016/j.ecocom.2007.07.005

characterization of regime shifts. On this view a regime shift is a response to changes in environmental forcing that results in changes in ecosystem status and function. Most importantly, this transient response occurs over a short-time scale relative to the time the ecosystem is in a non-transient status (de Young et al., 2004).

A variety of tools to explore data for regime shifts or alternative stable states are available. The emphasis in ecology on nonlinear dynamics and alternative stable states has led to the use of model regressions to test for multiple equilibria in time series data (Carpenter and Pace, 1997; Solow et al., 2003). In the oceanographic literature the approach has been to identify points of significant change in the statistics of the data – mainly changes in the mean – rather than underlying stable states. Easterling and Peterson (1995) review these methods and propose their own using linear regressions with Student's *t*-test and multiresponse permutation procedures to test for significance. Lanzante (1996) proposes a method based on a Wilcoxon–Mann–Whitney type test for the equality of medians. These methods suffer from deterioration of the statistics toward the ends of the time series. Rodionov (2004) proposes a sequential data processing technique based on Student's *t*-test to circumvent this problem. It compares favorably with the method of Lanzante (1996).

Hare and Mantua (2000) use principal component analysis (PCA) to detect regime shifts in multivariable data sets of the North Pacific, but this approach requires additional methods to test for significance. The average standard deviates compositing approach of Ebbesmeyer et al. (1991) requires *a priori* specification of a candidate shift. The analysis demonstrated by Noakes (1986) and the method proposed by Solow and Beet (2005) use time series models to identify times at which regime shifts have occurred, but these approaches assume a single regime shift. Fath et al. (2003) propose a method using Fisher Information as a summary statistic, and Mayer et al. (2006) apply it to several data sets. However, it was developed in the context of cyclical systems and does not give any guidance as to significance of regime shifts. See Mantua (2004) for a recent review of these methods.

Here we describe two methods that identify potential regime shifts through changes in neighborhood statistics. Both methods can be applied to multivariable data as well as to non-equilibrium systems. Furthermore, the nearest neighbor method gives objective estimates of significance. These methods contrast with current approaches in that they are not based on changes in the mean or median, and do not assume equilibrium dynamics. However, the methods presented here are not meant to take the place of detailed modeling, experiments and comparative analyses (Carpenter, 2001). Rather, they are proposed as a means to quickly identify shifts in time series that merit further study.

Throughout we assume that time series involving one or several variables of interest is available, and that over the measurement period the system has experienced one or more shifts in dynamic regime. The problem we address is the identification of the time point(s) at which likely regime shifts have occurred.

2. Nearest neighbor method

We have adapted an approach by Kennel (1997) to the detection of multiple 'change points', points in time at which a change in system parameters has occurred. We have found the method effective in identifying regime shifts.

Consider a sequence of measurements (data) taken at uniform intervals in time and indexed by N integers $\mathbf{x}(1), \mathbf{x}(2), \mathbf{x}(3), \ldots, \mathbf{x}(N)$. Assume further that over the time period for which data have been collected a single regime shift has occurred. For each point in the data sequence a nearest neighbor (NN) (in measurement- or state-space) over the entire time series is found. We then calculate the proportion of NN pairs that lie on the same side (in time) of a candidate change point and compare this to the proportion expected under the null hypothesis of stationarity. When the candidate change point coincides with the time of a regime shift (see Fig. 2) we expect the proportion of NN pairs on the same side to be high relative to the proportion under the null. Instead of performing statistics on nearest neighbors directly, these are collected in sets, termed strands, to eliminate correlations. For each NN pair, made up of a reference point and its nearest neighbor, the difference in time indices is determined, $\Delta(\mathbf{x}) = T(\mathbf{x}^{NN}) - T(\mathbf{x})$. Here $T(\cdot)$ gives the time index of the point. Nearest neighbors are found such that $|\Delta(\mathbf{x})| \geq W$, where W is a characteristic autocorrelation time.

Moving along the time series, every $\mathbf{x}(i)$ and its nearest neighbor that share the same Δ for $\mathbf{x}(i-k)$, $k \in [1, W]$, are appended to the strand associated with $\mathbf{x}(i-k)$. Otherwise, a new strand is started. This corrects for serial correlation (in which iterates of nearest neighbors remain nearest neighbors). Strands then are collections of pairs of points, a reference and its nearest neighbor, that are Δ apart in time. The reference points of sequential pairs in a given strand are at most W apart in time. As a final correction, if any two strands share underlying points, in either the reference or neighbor part, one strand is randomly deleted until no remaining strands share any points. The resulting strands represent nearly independent nearest neighbors and help in dealing with noise. A correction for oversampling can be introduced by defining two NN pairs as having the same Δ when $|\Delta(\mathbf{x}(i)) - \Delta(\mathbf{x}(j))| \leq F$, where F (constant) acts to 'fuzzify' time.

A statistical test is then applied to the strands. For a candidate change point ($\alpha, 1 < \alpha < N$) we apply the indicator function:

$$I_\alpha(\mathbf{x}(i), \mathbf{x}(j)) = \begin{cases} 1 & \text{if } i, j \leq \alpha \text{ or } i, j > \alpha \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

to the first NN pair in each strand to determine the proportion of strands (ρ) whose reference and nearest neighbor are on the same side of the change point.

Under the hypothesis of stationarity, the time index of a nearest neighbor can be anywhere in the data set with uniform probability except in the interval W time steps before the beginning and W time steps after the end of the reference portion of a strand. To estimate the proportion (ρ_0) of reference–neighbor pairs on the same side of the candidate change point (α) under the null hypothesis, we keep the reference portion of the first NN pair in each strand as it is and