# Frontiers in Ecology and the Environment


## Using SiZer to detect thresholds in ecological data


**Derek L Sonderegger, Haonan Wang, William H Clements, and Barry R Noon**



*Front Ecol Environ* 2009; 7, doi:10.1890/070179

# Using SiZer to detect thresholds in ecological data

**Derek L Sonderegger[1*], Haonan Wang[1], William H Clements[2], and Barry R Noon[2]**

**Ecological systems can change substantially in response to small shifts in environmental conditions. Such changes are characterized by a non-linear relationship between the value of the response variable and one or more explanatory variables. Documenting the magnitude of change and the environmental conditions that give rise to these threshold responses is important for both the scientific community and the agencies charged with ecosystem management. A threshold is defined as a substantial change in a response variable, given a marginal change in environmental conditions. Here, we demonstrate the usefulness of a derivative-based method for detecting ecological thresholds along a single explanatory variable. The "significant zero crossings" (SiZer) approach uses a non-parametric method to approximate the response function and its derivatives and then examines how those functions change across the range of the explanatory variable. SiZer makes fewer assumptions than conventional threshold models and explores a full range of smoothing functions. We believe SiZer is a useful technique for the exploratory analysis of many ecological datasets.**

*Front Ecol Environ* 2009; 7, doi:10.1890/070179

Theoretical and empirical studies suggest that some ecosystems may show abrupt, non-linear changes in one or more response variables in response to environmental drivers (May 1977; Connell and Sousa 1983; Knowlton 1992; Estes and Duggins 1995; Groffman *et al.* 2006). Shifts to alternative stable states have been reported in a variety of ecosystems, including lakes, coral reefs, deserts, and oceans (Scheffer *et al.* 2001). These shifts can be triggered by natural disturbance, such as fire or flooding, or anthropogenic factors, such as climate change, nutrient accumulation, exotic species, and toxic chemicals. Although communities may recover from natural disturbance through successional processes, human-induced disturbances are often unprecedented and move ecological systems to novel, alternative states (Holling 1986; Folke *et al.* 2002). In addition, if ecosystems are chronically stressed due to natural or anthropogenic disturbances, such systems may move to alternative states that remain stable, even when the stressors are removed (eg Carpenter 2001; Scheffer *et al.* 2001; van Nes *et al.* 2002; Scheffer and Carpenter 2003).

One can consider thresholds as ecological non-linearities, where substantial changes in an ecological state variable are a consequence of small, continuous changes in an independent (stressor) variable (Muraian 2001). The point or region at which rapid change initially occurs defines the threshold. Near this point, small changes in stressor intensity produce large effects on state variables. Unfortunately, there can be an inherent arbitrariness to the threshold concept, because it does not take into account whether the change in the value of the state variable is ecologically relevant. Statistical models have been developed in other disciplines, to detect breakpoints in non-linear response functions, but it is not always clear which models are appropriate for a particular ecological dataset.

Here, we demonstrate a method by Chaudhuri and Marron (1999) that makes few model assumptions and is therefore suitable for a broad range of ecological problems. Their method, "significant zero crossings" (SiZer), applies a non-parametric smoother to the stressor–response data, and then examines the derivatives of the smoothed curve to identify the existence of a threshold. To illustrate this method, we consider benthic macroinvertebrate data collected on the Arkansas River, a metal-polluted stream in Colorado. We use SiZer to examine the nature of the threshold(s) and to select between two competing threshold models. We then use SiZer in a multivariate setting by examining the first axis of the canonical discriminant analysis for the same dataset. Similar to principal components analysis, this axis is the linear combination of the 20 dominate taxa, that contains the most yearly variation.

## Data

The Arkansas River (Figure 1) is located in the southern Rocky Mountain ecoregion of Colorado. Mining operations in this watershed have had a major impact since the mid-1800s, when gold was discovered near Leadville, CO. Concentrations of heavy metals, particularly cadmium (Cd), copper (Cu), and zinc (Zn), are greatly elevated downstream from Leadville and often exceed acutely toxic levels (Clements 2004). Over the past 18 years (1989–2006), physiochemical characteristics, habitat quality, heavy metal concentrations, and the

[1]*Department of Statistics, Colorado State University, Fort Collins, CO* [*]*(sondereg@stat.colostate.edu);* [2]*Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO*

 

***Figure 1.*** *The Arkansas River after restoration efforts, 200 meters downstream of the confluence with the California Gulch.*

responses of macroinvertebrate communities were quantified at several stations in the Upper Arkansas River Basin. In 1993, 4 years after this research program began, state and federal agencies initiated a large-scale restoration program designed to improve water quality in the Arkansas River. To quantify recovery, we examined temporal changes in the abundance of metal-sensitive mayflies (Ephemeroptera: Heptageniidae) collected in the fall from 1989 to 2006. During each of the 18 years, five replicate samples were taken. The mayfly counts were transformed (square root) to stabilize the variance. Recovery was defined as the threshold where mayfly abundance became asymptotic. While the study was primarily concerned with the effect of heavy metal pollution, this paper uses time as the independent variable for clearer illustration of the method.

## SiZer approach

### *Derivative definition of thresholds*

An intuitive way of defining a threshold for a state variable that is a continuous function of an environmental driver is to consider where the function's derivatives change significantly. Non-parametric smoothers provide a method for finding a smooth response function that is data driven and requires only weak assumptions. Smoothing splines (Green and Silverman 1993; Wahba 1975), LOESS (Cleveland and Devlin 1988) and locally weighted polynomial regression (Fan and Gijbels 1996) are well known. These techniques result in an estimated smooth response function, the estimated derivative(s), and confidence intervals (CIs) for the functions and derivatives. The SiZer methodology can be implemented using any of these techniques, but we have restricted our discussion to locally weighted polynomials. All SiZer CIs in this paper are reported at the 95% level, based on Hannig and Marron (2006) row-wise intervals.

In the Arkansas River data, both a piecewise linear (PL; Barrowman and Myers 2000; Toms and Lesperance 2003) or bent-cable (BC; Chiu *et al.* 2006) model would fit the data. Both models assume a linear relationship with a single threshold. The difference is that the PL model assumes an abrupt transition between the linear sections, whereas the BC model assumes a quadratic bend connecting the two linear pieces. The PL model is a simple case of the BC model where the half-width of the bend is zero.

Traditional model selection methods, such as Akaike's information criterion (Burnham and Anderson 2002), do not decisively rate one model over the other. The difference between AIC values is 1.82 in favor of the PL model. However, an inverted likelihood ratio test (Seber and Wild 1989) resulted in a 95% CI for the half-width of the quadratic bend that does not contain zero (0.04, 5.52). Point estimates of the threshold were similar for the PL and BC models (1996.7, 1996.1), but 95% bootstrap CIs for the threshold were (1994.6, 1997.5) and (1989.0, 2000.0), respectively.

Using a non-parametric smoother, researchers can classify every point along the independent axis into one of three states: the estimated slope is positive (ie the CI of the first derivative contains only positive values), possibly zero (the CI contains zero), or negative (the CI contains only negative values). Each point could be similarly classified by the estimated second (or higher order) derivative.

Many interesting relationships can be found by examining the state changes of the derivatives. By noting how many times the state of the first derivative changes, inference about where the true relationship is increasing or decreasing can be made. The Arkansas River data show that abundance of mayflies is clearly increasing, then flattens out and seems to decrease slightly near $x = 2002$. The second derivative contains information about the curvature of the data. At small $x$ values, the second derivative could be zero, indicating that there is little or no curvature. However, between 1994 and 2000, the second derivative is significantly negative, indicating that the function is concave down and providing support in the data for the BC model.

There is no mathematical reason to partition the curve by where the derivative is different than zero. A similar procedure could be implemented to partition the $x$ axis into segments that have a derivative different than 5, for example. However, the choice to partition on $\hat{f}'(x) = 0$ is appropriate for many ecological problems in which the increase or decrease of the response variable is of interest. Moreover, detecting a change in the rate of increase (or decrease) can be made by second derivative, because a change of rate causes curvature.

Given a small number of models that are thought to quantify a researcher's beliefs and hypotheses about a system, traditional model selection methods such as AIC, AICc, Mallows $C_p$, and goodness-of-fit tests (Kutner *et al.* 2005) fail to directly explain why one model is selected over another. By using a derivative-based method in the process of model selection, the researcher can investigate how the data support one model over the other, or what assumptions in a model are being violated. The non-parametric implementation allows the researcher to examine a broad range of questions, including the number of thresholds in a system.

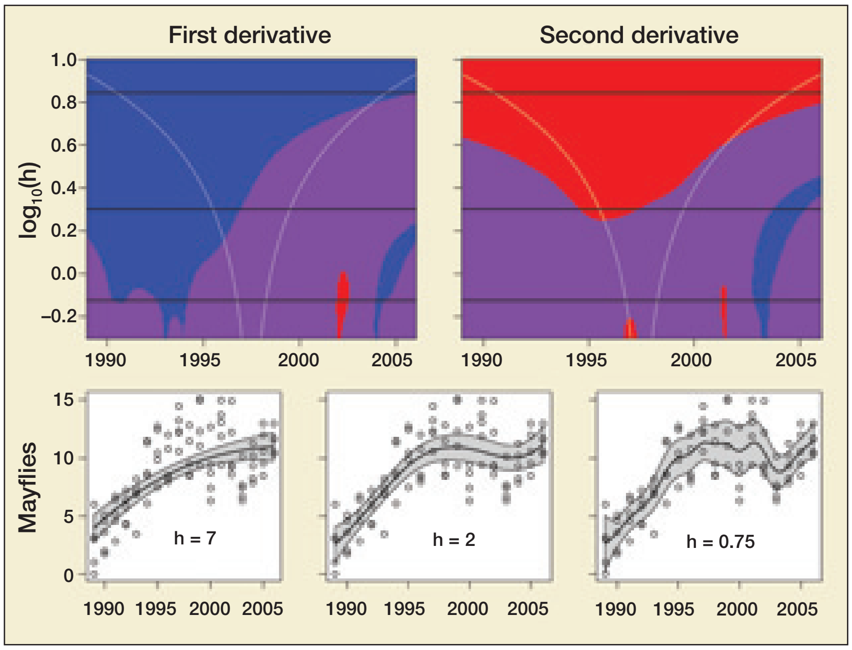


***Figure 2.*** *SiZer maps of the Arkansas River data and associated smoothing functions at three different bandwidths. SiZer maps categorize the derivative as positive (blue), negative (red), or possibly zero (purple). The black lines in the SiZer maps show bandwidth parameter corresponding to the three smoothing functions. The bandwidth* h *= 7 is clearly over-smoothing the data and does not capture the flatness (or decrease) in the second half of the data. The bandwidth* h *= 0.75 is under-smoothing the data and is being affected by random perturbations in the data.*

## Smoothing bandwidths

One complication of the derivative approach is the estimation of the smoothed function and its derivative(s). Most non-parametric smoothing algorithms, including smoothing splines and locally weighted polynomial regression, have a tuning parameter that controls the smoothness of the resulting curve. By manipulating this parameter, the resulting smoothing function can range from a simple linear regression to perfectly (over)fitting the data. There are several methods for selecting the tuning parameter (Fan and Gijbels 1995; Ruppert *et al.* 1995; Hengartner *et al.* 2002), but none are uniformly superior.

SiZer, as proposed and implemented by Chaudhuri and Marron (1999) and Hannig and Marron (2006), uses the idea of locally weighted polynomial regression (Fan and Gijbels 1996; Loader 1999). When the weight function (also called a kernel function) is the normal density curve, the level of smoothing is controlled by the standard deviation of the kernel. For a given tuning parameter $h$, and a given point $x_0$, $\hat{f}'(x_0)$ is obtained by weighting the data points according to a normal curve centered at $x_0$, with a standard deviation $\sigma = h$. This means that data close to $x_0$ (eg within $\pm h$) have a major influence over the smoothing function, data between $h$ and $2h$ away are less influential, and data farther than $2h$ from $x_0$ have only a slight influence. In locally weighted polynomial regression, the tuning parameter $h$ is the width parameter of the kernel function and is commonly referred to as the bandwidth. Other common choices for the kernel function include the uniform density and triangle density functions. Here, we use the normal kernel.

The novel aspect of SiZer (Chaudhuri and Marron 2000) is that it considers all reasonable bandwidth values and exploits the notion that different values provide different information about the data. The SiZer approach explores how the derivative changes along the independent axis, as well as across the range of bandwidth values, and displays this information in one image (Figure 2). To read the SiZer map, first notice that the $y$ axis represents the bandwidth parameter $h$, displayed in units of $\log(h)$ for visual clarity. Wherever the map is blue, the derivative is significantly increasing; wherever it is purple, the derivative is possibly zero; and wherever it is red, the derivative is significantly decreasing. At very small bandwidths, $\hat{f}'(x_0)$ is influenced by a small number of data points, and gray areas in the SiZer map indicate that the estimated effective sample size (Chaudhuri and Marron 1999) is less than five. The white lines give a visual representation of the size of the bandwidth. The horizontal distance between the lines is drawn to be $2h$, indicating the effective width of the locally weighted polynomial.

To demonstrate the effect of bandwidth on the smoothing function, Figure 2 displays Arkansas River data with three different choices of bandwidth $h$, each highlighted by the horizontal black line in the adjacent SiZer maps.

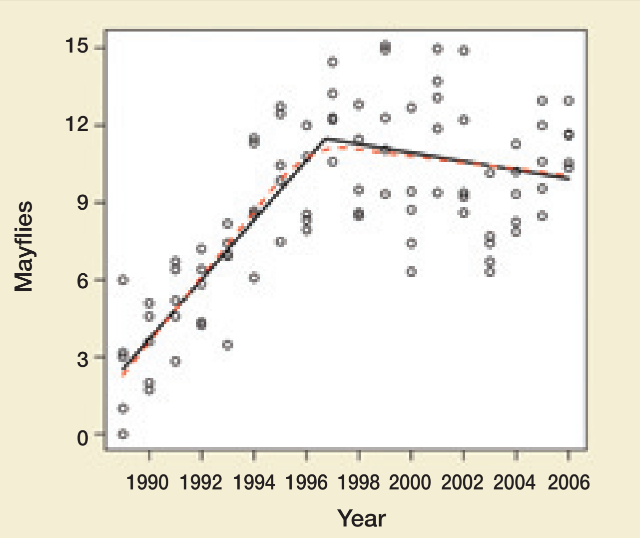


***Figure 3.*** *The Arkansas River mayfly data fit by the piecewise linear (black, solid, threshold = 1996.7, CI = 1994.6, 1997.5) and bent-cable (red, dashed, threshold = 1996.1, CI = 1989.0, 2000.0) models.*

The bottom left graph represents a smoothing parameter that is too large ($h = 7$) and has over-smoothed the data and fails to detect the transition from an increasing to a flat (or possibly decreasing) function. At this scale of view, the first derivative SiZer row is completely blue, suggesting an increasing function with no threshold. The second derivative is negative (red), indicating that there is downward curvature in the function. At an intermediate level of smoothing ($h = 2$), the smoother captures the initial increasing section and transition to a flat function, as indicated by the first derivative SiZer map. The second derivative map shows that the function is reasonably linear, except for a region of concavity in the middle and a second region of convexity near 2004. The piecewise linear and bent-cable models presented in Figure 3 capture features that are visible at this scale. At a very low level of smoothing ($h = 0.75$), at any given point, the smoother is being influenced by a very small number of data points (estimated effective sample size ~15). Consequently, the power of testing if the derivative(s) are not equal to zero is low. In this study, researchers were not particularly interested in annual variation, but wanted to detect trends occurring over multiple years associated with recovery of this system; therefore, bandwidths $h > 1$ ($\log_{10} h > 0$) should be considered. After considering each of these bandwidths, particularly $h$ near 2, the data support the bent-cable model over the piecewise linear model, because of the curvature near the threshold at the intermediate bandwidths. However, the SiZer analysis suggests considering a model with two threshold points to account for the decrease near 2003.

Because the SiZer map contains information at many different scales, there is seldom a "best" bandwidth to examine. Therefore, we recommend an evaluation of the derivative at different resolutions of the data. Just as when viewing a tree from a distance (ie at large bandwidths), only gross features are discernible, so, as the observer gets closer to the tree (ie as the bandwidth decreases), the overall pattern cannot be seen, but smaller features come into focus. Only by examining the function across a range of bandwidths can a researcher gain a clear understanding of the data.

SiZer cannot, however, always estimate the location of the threshold. Because $\hat{f}$ is calculated from nearby values, if $f$ has a threshold at $x = \alpha$, then $\hat{f}$ is not necessarily first affected by the threshold at $x = \alpha$. Furthermore, where $\hat{f}$ is affected by the threshold changes with the bandwidth. This phenomenon can be seen in the Arkansas River example: the threshold from an increasing to a flat function drifts from near 1995 to 2004 as the bandwidth increases.

### *Using SiZer to identify multiple thresholds*

Multivariate analysis of macroinvertebrate data collected from the Arkansas River provided an opportunity to investigate ecological thresholds in community composition over the duration of the monitoring project. In this example, canonical discriminant analysis was used to examine differences among years, based on abundance of the 20 dominant taxa. A threshold response in this example represents an abrupt shift in community composition from one year to the next. Ignoring the issue of non-independence of observations in multivariate space, we applied SiZer to the first canonical axis, which explained 58.2% of the total variation (Figure 4). The first derivative shows a generally increasing function, but there is a sharp decreasing trend between 1995 and 1997. These results reflect macroinvertebrate community responses to changes in water quality from 1989 to 2006. Heavy metal concentrations declined from 1989–1994, increased abruptly in 1995 and 1996, and then declined again as a result of ongoing restoration in the Arkansas River (Clements 2004).

The second derivative from the SiZer plot shows two distinct thresholds. The first, near 1996, is a change from concave down to concave up. The second, near 2000, is a change from concave up to concave down. At intermediate smoothing levels near 1996, the 95% CI goes from being less than zero, to containing zero, to being greater than zero. For the second threshold, near 2000, which is a result of macroinvertebrate community recovery after improvements in water quality, the 95% CI changes from being greater than zero, to containing zero, to being less than zero.

## Discussion

Fitting mathematical models to observed data is difficult, in part because of the uncertainty in model selection. Traditional model selection methods tend to encourage examination of vast numbers of models. Burnham and Anderson (2002) address this issue by differentiating exploratory studies from confirmatory ones. When this is not possible, inference must be made carefully to avoid

 

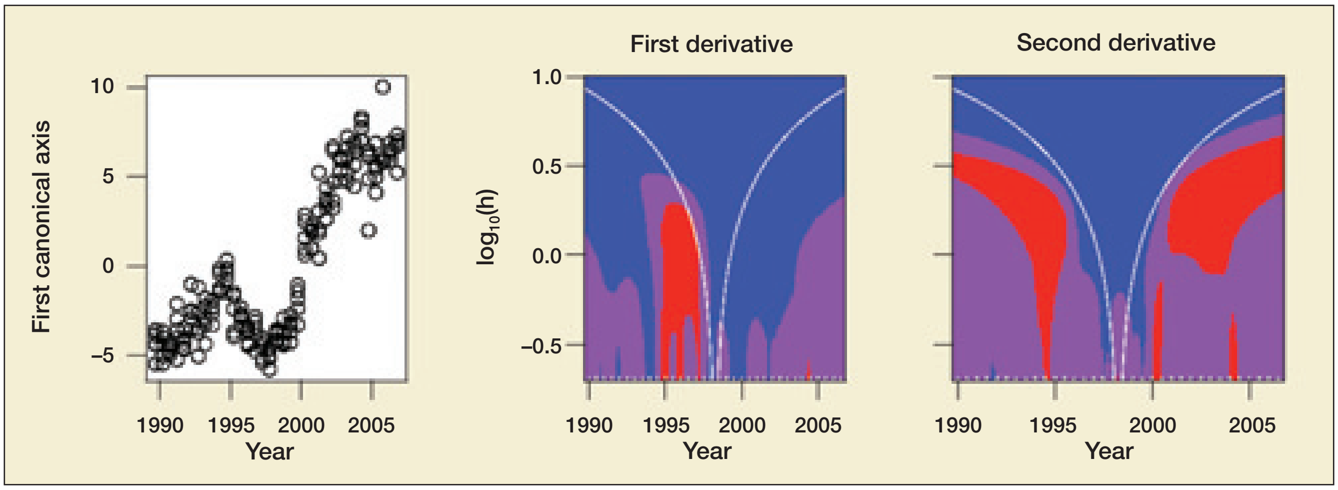


***Figure 4.*** *A scatterplot of the first axis of a canonical discriminate analysis of the Arkansas River data versus time, along with its first and second derivative SiZer maps.*

inflated Type I errors (including a variable in the model when, in fact, it has no effect on the response), due to picking the "best" model for the data. As presented here, SiZer is most naturally used in exploratory studies, but if sufficient data are available, part of the data may be used for exploratory model selection and the other part used for inference. Burnham and Anderson (2002) also strongly advocate only examining models that have sound scientific explanations. Since SiZer encourages the practitioner to create an appropriate response function from the SiZer map, the ecological justification and empirical evidence for a particular form of the response function can be combined. This should result in small numbers of models, to be used in subsequent model selection steps or model averaging. SiZer also separates the question of statistical significance from ecological significance, by showing the statistically significant features at each bandwidth and then allowing the researcher to decide which features are important.

Fitting threshold models is particularly difficult, because researchers often assume that the existence and number of thresholds are known. For example, a piecewise linear analysis will always find a single threshold, regardless of whether the true functional relationship contains no threshold or multiple thresholds. SiZer can provide insight into the number of thresholds and general form of the relationship. Unfortunately, SiZer cannot provide estimates and confidence regions for the thresholds it detects. SiZer can only be used to select a model; a model fitting procedure such as maximum likelihood estimation must be used subsequently. Furthermore, by definition, SiZer can only address thresholds in the context of changes in state of the derivative. Uniform or gradual changes that lead to irreversible state changes are not detectable by SiZer.

The mathematics that SiZer employs can be readily extended to multiple dimensions by using the multidimensional gradient rather than the one-dimensional derivative (Godtliebsen *et al.* 2002); however, SiZer's main strength, its graphical presentation, cannot be easily extended past two dimensions. Covariates can be accounted for in an additive fashion, by using SiZer on the non-parametric portion of a generalized additive model (GAM). Finally, one direction for future research is to extend SiZer to work with local quantile regression instead of local mean regression. Pointwise estimates of the quantile function should not be difficult to obtain, but appropriate row-wise CIs might be.

A Matlab implementation of SiZer has been made available by S Marron (www.stat.unc.edu/faculty/marron/marron_software.html). An R package of SiZer, along with code for the piecewise linear and bent-cable models, is available on the Comprehensive R Archive Network (www.cran.r-project.org).

## Acknowledgements

This research was supported by the US Environmental Protection Agency (EPA) STAR Program, grant #R832441, the National Science Foundation (NSF), IGERT grant #DGE-0221595, and NSF grant #DMS-0706761. This manuscript has not been formally reviewed by the EPA or NSF. The views expressed here are solely those of the authors. The EPA does not endorse any products or commercial services mentioned in this report.

## References

Barrowman NJ and Myers RA. 2000. Still more spawner-recruitment curves; the hockey stick and its generalizations. *Can J Fish Aquat Sci* **57**: 665–76.

Burnham KP and Anderson DR. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. New York, NY: Springer-Verlag.

Carpenter SR. 2001. Alternative states of ecosystems: evidence and its implications for environmental decisions. In: Press MC, Huntley NJ, and Levin S (Eds). Ecology: achievement and challenge. Cornwall, UK: Cambridge University Press.

Chaudhuri P and Marron JS. 1999. SiZer for exploration of structures in curves. *J Am Stat Assoc* **94**: 807–23.

Chaudhuri P and Marron JS. 2000. Scale space view of curve estimation. *Ann Stat* **28**: 408–28.

 

Chiu GS, Lockhart R, and Routledge R. 2006. Bent-cable regression theory and applications. *J Am Stat Assoc* **101**: 542–53.

Clements WH. 2004. Small-scale experiments support causal relationships between metal contamination and macroinvertebrate community responses. *Ecol Appl* **14**: 954–67.

Cleveland WS and Devlin SJ. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc* **83**: 596–610.

Connell JH and Sousa WP. 1983. On the evidence needed to judge ecological stability or persistence. *Am Nat* **121**: 789–824.

Estes JA and Duggins DO. 1995. Sea otters and kelp forests in Alaska: generality and variation in a community ecological paradigm. *Ecol Monogr* **65**: 75–100.

Fan J and Gijbels I. 1995. Data-driven bandwidth selection in local polynomial fitting – variable bandwidth and spatial adaptation. *J Roy Stat Soc B Met* **57**: 371–94.

Fan J and Gijbels I. 1996. Local polynomial modeling and its applications. London, UK: Chapman and Hall.

Folke C, Carpenter S, Elmqvist T, *et al.* 2002. Resilience and sustainable development: building adaptive capacity in a world of transformations. *Ambio* **31**: 437–40.

Godtliebsen F, Marron JS, and Chaudhuri P. 2002. Significance in scale space for bivariate density estimation. *J Comput Graph Stat* **11**: 1–21.

Green PJ and Silverman BW. 1993. Nonparametric regression and generalized linear models: a roughness penalty approach. New York, NY: Chapman and Hall.

Groffman PM, Baron JS, Blett T, *et al.* 2006. Ecological thresholds: the key to successful environmental management or an important concept with no practical application? *Ecosystems* **9**: 1–13.

Hannig J and Marron JS. 2006. Advanced distribution theory for SiZer. *J Am Stat Assoc* **101**: 484–99.

Hengartner NW, Wegkamp MW, and Matzer-Lober E. 2002. Bandwidth selection for local linear regression smoothers. *J Roy Stat Soc B Met* **64**: 791–804.

Holling H. 1986. A classification of trivariate regression-models. *Biometrical J* **28**: 783–90.

Knowlton N. 1992. Thresholds and multiple stable states in coral-reef community dynamics. *Am Zool* **32**: 674–82.

Kutner MH, Nachtsheim CJ, Neter J, and Li W. 2005. Applied linear statistical models, 5th edn. New York, NY: McGraw-Hill/Irwin.

Loader C. 1999. Local regression and likelihood. New York, NY: Springer-Verlag.

May RM. 1977. Thresholds and breakpoints in ecosystems with a multiplicity of stable states. *Nature* **269**: 471–77.

Muraian R. 2001. Ecological thresholds: a survey. *Ecol Econ* **38**: 7–24.

Ruppert D, Sheather SJ, and Wand MP. 1995. An effective bandwidth selector for local least squares regression. *J Am Stat Assoc* **90**: 1257–70.

Scheffer M and Carpenter SR. 2003. Catastrophic regime shifts in ecosystems: linking theory to observation. *Trends Ecol Evol* **18**: 648–56.

Scheffer M, Straile D, van Nes EH, and Hosper H. 2001. Climatic warming causes regime shifts in lake food webs. *Limnol Oceanogr* **46**: 1780–83.

Seber GAF and Wild CJ. 1989. Nonlinear regression. New York, NY: John Wiley and Sons.

Toms JD and Lesperance ML. 2003. Piecewise regression: a tool for identifying ecological thresholds. *Ecology* **84**: 2034–41.

van Nes EH, Scheffer M, van den Berg MS, and Coops H. 2002. Dominance of charophytes in eutrophic shallow lakes – when should we expect it to be an alternative stable state? *Aquat Bot* **72**: 275–96.

Wahba G. 1975. Smoothing noisy data with spline functions. *Numer Math* **24**: 383–93.

 