1-1-2006

Modelling fish IBI with agricultural stress gradient and estimation of threshold effects.

Jabed Hossain Tomal
University of Windsor

Modelling Fish IBI with Agricultural Stress Gradient and Estimation of Threshold Effects

by

Jabed Hossain Tomal

A Major Paper
Submitted to the Faculty of Graduate Studies and Research
through Mathematics and Statistics
in Partial Fulfillment of the Requirements of
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada

2006

Your file Votre référence
ISBN: 978-0-494-42341-7
Our file Notre référence
ISBN: 978-0-494-42341-7

Abstract

To assess the health of aquatic habitats, Uzarski et al. (2005) and Bhagat (2005) developed a multimetric index of biotic integrity (IBI) through assessing fish community composition at thirty sites with dominant *Scirpus* vegetation across the entire US Great Lakes coastline. Danz et al. (2005) derived an agricultural stress gradient to characterize the degradation of nature using GIS based data. The IBI-agricultural stress relationship resulting from the combined data set suggests that threshold relationships rather than a linear relationship describe how the fish IBI changes as agricultural stress increases. The main objective in this paper is to estimate the threshold effects of the agricultural stressor to the health of aquatic habitats as represented by the fish IBI. First, we employed four tests for bivariate randomness hypothesis among 'Fish IBI' and 'Agricultural Stress Gradient'. All of the four tests gave results rejecting the null hypothesis of bivariate randomness. A simulation study was performed to verify the power of those tests against our specific alternative in our data. Various regression techniques including linear, non-linear, nonparametric, and piecewise linear regression, were used to regress 'Fish IBI' with 'Agricultural Stress Gradient'. Among all these regression techniques, only the piecewise linear regression model was able to estimate the threshold effects. We applied quantile regression techniques to identify the prediction band of the data as well as to estimate the threshold effects. Based upon the estimates of threshold parameters we identified the "undegraded" zone, the "transition" zone, and the "degraded" zone in terms of agricultural stress. We defined "undegraded" zone as the collection of sites with minimum level of agricultural stress, indicating a habitat suitable for fish communities typical of unimpaired

locations at the US Great Lakes Coastal margins and "degraded" zone as the collection of sites with high level of agricultural stress, indicating degradation of natural habitat for fish communities at this Great Lakes Coastal margin. We also defined "transition" zone as the collection of sites that fall between the "undegraded" zone and "degraded" zone with rapid decline in fish communities with the increase of agricultural stress. The results showed that seven sites among the selected thirty sites in the US Great Lakes coastal margin are in the "degraded" zone. So, care should be taken to get rid of the degradation of natural habitat.

Finally, we recommended LOESS, piecewise linear regression and quantile regression techniques to model data with potential breakpoints and to estimate the threshold(s) as well. We also made recommendation for further research in terms of sampling scheme.

Dedicated

to

My Family

Acknowledgements

My reverent gratefulness goes to my co-supervisor Dr. Karen Fung, Department of Mathematics and Statistics, University of Windsor, and co-supervisor Dr. Jan J.H. Ciborowski, Department of Biological Sciences, University of Windsor, for their beneficial guidance and imperative supervision.

I would like to thank Dr. S. Nkurunziza, Department of Mathematics and Statistics, University of Windsor, for providing his valuable time as Department Reader.

I am grateful to the Department of Mathematics and Statistics for providing financial support in terms of graduate assistantship throughout my study. The research assistantship from Dr. Karen Fung and Dr. Jan J.H. Ciborowski is also appreciated.

I gratefully acknowledge the assistance that I have received from the Ph.D. students, specifically, from Md. Shakhawat Hossain, Lihua An, and from my friends to complete this paper.

Finally, I am grateful to my family for their continuous support and encouragement. I remember with gratification of their love, affection and inspiration throughout my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Objectives

Many recent studies in ecology have been devoted to estimation of critical thresholds associated with human-induced natural habitat fragmentation (e.g., Andren 1994, Fahrig 2001). Critical thresholds occur when the response of a species or ecological process to habitat loss is not linear, but changes abruptly at some threshold level of loss (Toms and Lesperance 2003). Abrupt changes in ecological processes can also occur in other systems. Plant and animal communities change within a threshold distance of habitat edges (edge effects; Wales 1972, Gates and Mosher 1981). Changes in management regimes may have threshold-type effects if processes are viewed through time. Human produced disturbance from agriculture is the major cause of natural habitat loss for fish population in lakes and rivers. In this paper, most of the statistical analyses have been devoted for estimating threshold effects of agricultural stressor on fish population in the US Great Lakes coastal margins.

The effects of human-induced disturbances affecting the Great Lakes basin have been of major concern to managers and researchers alike. Since the passing of the US

Clean Water Act of 1972 (PL 92 – 500), there has been much focus on developing biological indicators of anthropogenic stress to measure 'biotic integrity' of a habitat. Biotic integrity is one of the primary objectives set forth by the Clean Water Act and is defined as "the ability of a habitat to support and maintain a balanced, integrated, adaptive community of organisms having a composition, diversity and functional organization comparable to that of a natural habitat" (Frey 1977). A natural habitat as such, refers to an area with minimal levels of anthropogenic stress, also referred to as a reference condition area (Host et al. 2005).

Many groups of taxa have been used to develop indicators of stress, ranging from diatoms (Ferguson et al. 2003) and benthic invertebrates (Burton et al. 1999) to aquatic plants (Gallatowitsch et al. 1999), amphibians (Grabas et al. 2004) and more commonly, fish (e.g., Simon 1991). Fishes serve as a good indicators of stress because they are often philopatric, represent a broad spectrum of community tolerance to manifestations of anthropogenic disturbance from very sensitive to highly tolerant, and respond to physical, chemical and biological degradation (Plafkin et al. 1989). Fish communities include species representing various trophic levels (piscivores, omnivores, insectivores, herbivores), and their position in relation to diatoms and benthic invertebrates helps to provide an integrative approach to habitat assessment (Karr 1981).

The index of biotic integrity (IBI) is the most commonly used multimetric approach for stress, particularly in assessing streams and rivers in the US (Karr 1981, Lyons and Wang 1996, Mundahl and Simon 1999, Emery et al. 2003). Theoretically, the IBI reflects the degree to which the local environment influences the fish community.

Development of an IBI depends on the a priori identification of reference sites (locations that are subject to minimal levels of anthropogenic stress) typically delineated by watershed boundaries (Karr 1981, Fausch et al. 1984) or ecoregions (Hughes and Larson 1988, Omernik 1995). Thus, each ecological zone could theoretically require its own IBI consisting of metrics that represent measures of a healthy community, specific to that ecoregion or watershed. One of the main advantages of using the IBI approach is that it takes into account a variety of attributes that represent the fish community in a site (Simon 1991). The IBI produces a single score that can be compared to a sampling distribution of expected scores based on scores of minimally disturbed sites in the region. Thus, the IBI reflects the fish community response to relative degree of disturbance at a particular site (Karr 1981). In this way the multimetric index of biotic integrity (IBI) becomes the popular approach to assess the health of an aquatic habitat.

Uzarski et al. (2005) used correspondence analysis to determine that the primary correlate in coastal wetland fish community composition is emergent plant zonation. They developed an IBI for sites dominated by (> 50% cover) *Scirpus* (bulrush) and *Typha* (cattail). Uzarski et al. (2005) proposed that the IBIs they developed would serve as good indicators of overall habitat integrity as long as the sites sampled and tested were dominated by either *Typha* and *Scirpus*. Danz et al. (2005), the Great Lakes Environmental Indicators group, sampled fish at a total of 82 coastal wetlands along US Great Lakes coastal margins in 2002 and 2003, according to a design that balanced effort across lakes, hydrogeomorphic wetland type, and degree of local anthropogenic disturbance. Of those 82 coastal wetlands, Bhagat et al. (in review)

found 36 sites with dominant *Typha* and *Scirpus* vegetation, more specifically 23 sites with dominant *Typha* vegetation and 13 sites with dominant *Scirpus* vegetation across the entire US Great Lakes shoreline. In their study, vegetation density and cover were noted at the fish-capture net level rather than the site level and the criterion for dominant cover was lowered to 30% or greater. So, they calculated Uzarski et al.'s IBI scores for 13 Great Lakes wetland sites with dominant *Scirpus* vegetation 23 Great Lakes wetland sites with dominant *Typha* vegetation, using their data from overnight sets fyke nets.

Most researchers to date have quantified primarily agricultural and land use as disturbance measures affecting fish communities (Brazner and Beals, 1977, Crosbie and Chow-Fraser 1999). General patterns of human activity and land use in the US Great Lakes basin differ between ecoprovinces, with most agricultural activities occurring in the southern portion of the basin, while the northern portion of the basin remains largely forested. The southern portion of the US Great Lakes basin contains deeper, more permeable, and more highly buffered soils than that of the northern portion. Metropolitan areas are more common in the southern basin.

Measures of Anthropogenic Stress

Danz et al. (2005) calculated measures five of anthropogenic stresses in the US Great Lakes Basin. For many regions of the continental US, there is a wealth of spatially explicit data from monitoring and reporting programs related to human activities; these variables can be used to represent stress if appropriate scales of impact and interactions with related factors can be identified. To integrate spatial stress data they used different methods ranging from relatively simple rank or scoring (Bryce

et al., 1999) to multivariate statistical techniques in combination with a geographic information system (GIS) (Tran et al., 2003; Tran et al., 2004).

Danz et al. (2005) developed a geographic information system (GIS) database with 149 spatial variables previously used to distribute sampling effort across a range of environmental conditions in the Great Lakes basin. A preliminary multivariate analysis had been used to classify the variables into five categories of anthropogenic stress that are prominent in the Great Lakes basin. To calculate agricultural stress gradient they employed 21 variables characteristic of the major types of stresses associated with agricultural activities, including nutrient run off, pesticides, and erosion. They used principal component analysis (PCA) to integrate the information within each of the five categories of stress variables into a smaller number of stress measures. PCA is a multivariate statistical technique that creates a set of novel orthogonal variables (PCs) that are linear combinations of the original variables (Rencher, 1995).

To assess the relationship between GLEI stressors and the fish data collected by Uzarski et al. (2005), Bhagat (2005) overlaid geographic co-ordinates for each site sampled by Uzarski et al. (2005) on a map of their GLEI segment-shed delineations and determined its corresponding GLEI stressor score. Since the stressor information was only available for sites on the US coastline, they were only able to get Uzarski et al. (2005) data and the GLEI data for *Scirpus* dominant sites.

Fish Sampling

Fish communities were sampled using 2 large fyke net arrays (1.25 cm mesh) and 2 small fyke net arrays (0.5 cm mesh) set overnight at each site. Each fyke net array was placed lead-to-lead (leads parallel to shore), with the wings set at 45^0 angles

(Brazner and Beals 1997). One set of large and one set of small nets were placed near each of the two dominant shoreline habitats at a site. Fish community composition (number of individuals of each species) and condition of up to 25 fish per species (total length, incidence of damage or disease) were measured at each fyke net the next day. Unidentifiable fish were euthanized in clove oil, preserved in 9 : 1 v/v ethanol: formalin mixture, and taken to the lab for identification. Physicochemical variables (temperature, dissolved oxygen concentration, conductivity pH) were measured at each net using a multi-probe meter (YSI 556 MPS). Water clarity at each net was measured using a Secchi disk and turbidity tube. Dominant and subdominant genera of emergent, subemergent and floating vegetation (cover and density) were also noted at each net per site.

IBI Development

The Fish IBI that we used in this paper was developed by Uzarski et al. (2005). The IBI was developed according to Uzarski et al.'s (2005) method, and Table 1 in the Appendix contains the final set of IBI metrics for *Scirpus* zones and Table 2 in the Appendix contains the data with observations from Uzarski et al. (2005) sites and GLEI sites. Bhagat (2005) concluded that *Typha* IBI was significantly negatively correlated to agricultural stress. In contrast, *Scirpus* IBI scores reflected a threshold effect (Figure 1.1) at agricultural stress scores of -0.5 or less, whereas at stress scores > 0, there were no IBI score greater than 45. The pattern was more consistent with a threshold effect rather than a linear response.

We defined, **undegraded Zone**: A zone with the collection of sites with minimum level of agricultural stress, indicating a habitat suitable for fish communities in the

US Great Lakes Coastal margins. **Degraded Zone**: A zone with the collection of sites having high level of agricultural stress, indicating degradation of natural habitat for fish communities in this Great Lakes Coastal margin. **Transition Zone**: A zone with the collection of sites that fall between the "undegraded" zone and "degraded" zone with rapid decline in fish communities with the increase of agricultural stress.

Before beginning to track the pattern between the random variables 'Fish IBI' and 'Agricultural Stress Gradient', it is wise to test the randomness between them. A random process is a repeating process whose outcomes follow no describable pattern. Chapter 2 talks about the tests for bivariate randomness. To test the hypothesis of bivariate randomness of the two random variables, we used (1) the test based on the mean nearest-neighbor distance proposed by Clark and Evans (1954) that takes into account dependencies amongst the nearest-neighbor distances and incorporates a correction for edge effects; (2) the test based on the cumulative R-spectrum proposed by Mugglestone and Renshaw (1990) known as spectral test; (3) the test based on the reduced second-order moment function by Ripley (1976); and (4) the test based on the bivariate Cramer-von Mises statistic proposed by Zimmerman (1993). The reason for choosing these tests is that they were shown to be powerful against different types of alternatives especially against a clustered alternative, which is what we observe in our data. We also performed a simulation study to compare the power of those chosen tests.

In Chapter 3 we employed regression techniques to regress 'Fish IBI' by 'Agricultural Stress Gradient'. Some nonparametric regression techniques (e.g., Kernel smoothing, LOESS) have been used to identify the pattern that is present in the data. Among

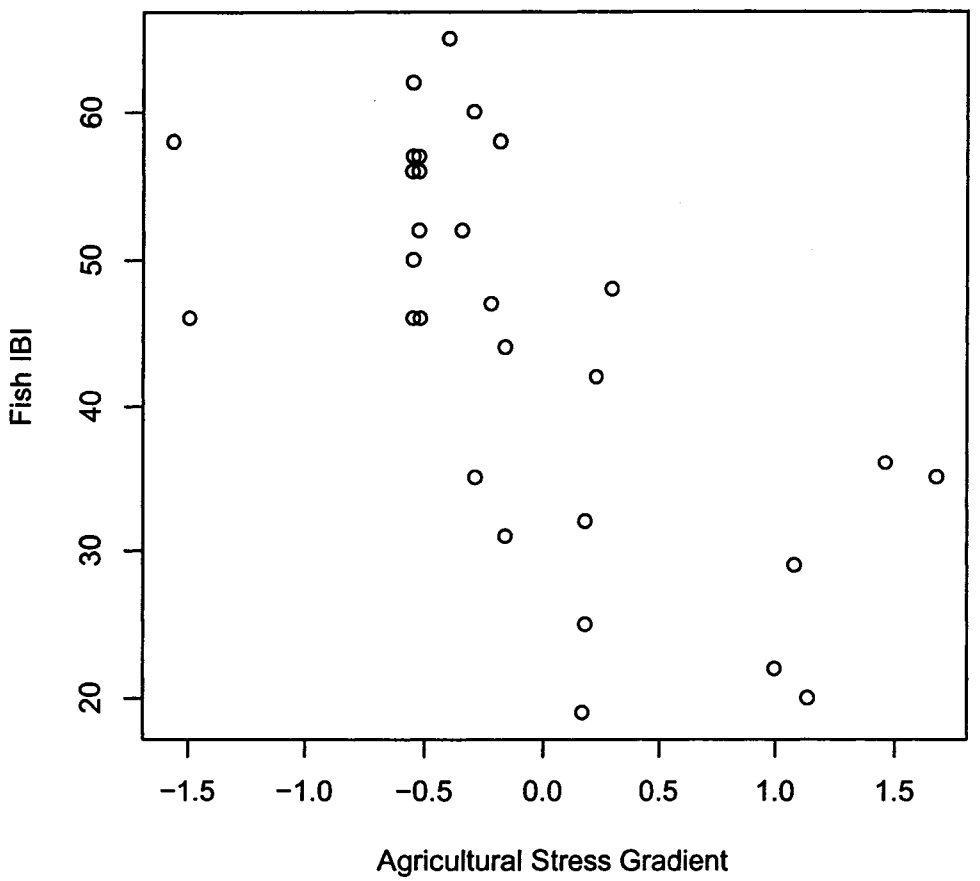

Figure 1.1: Plot of Fish IBI of GLEI and Uzarski sites against agricultural stress gradient (from Bhagat et al. in review)

the parametric regression procedures, we used (1) linear, (2) nonlinear: logistic curves and (3) piecewise linear regression techniques and compared them in terms of residual sum of squares. Toms and Lesperance (2003), used piecewise linear regression models to estimate break points and used these as estimates of thresholds. We followed Toms and Lesperance (2003) techniques to fit piecewise linear model for two breakpoints and hence to estimate two thresholds. We specified the interval between two breakpoints as the "transition" zone and the zone beyond the second break point as the "degraded" zone.

Classical least squares linear regression technique gives the regression line through the mean of the conditional distribution of the dependent variable y. As the mean alone gives an incomplete picture about the distribution of a random variable, similarly mean regression gives an incomplete picture about regression line, especially when the data are sparse and heterogeneous. By complementing the exclusive focus of classical least-squares regression on the conditional mean, quantile regression offers a systematic strategy for examining how covariates influence the location, scale, and shape of the entire response distribution (Koenker 2005). We fitted linear regression, nonlinear regression (logistic curve) and piecewise linear regression through different quantiles. Among those, piecewise linear quantile regression gives us a set of values for the first break point (threshold) and another set of values for the second break point (threshold), from which we can get a comprehensive view of the "transition" zone and hence we were able to determine the "degraded" zone in terms of agricultural stress gradient. Quantile regression techniques have permitted us to get an idea about prediction band of the data. So, Chapter 4 contains the methodology of

quantile regression and its application to our data.

Finally Chapter 5 contains the summary and conclusion of this paper.

Objectives

The overall objective of this paper is to examine how values of the Uzarski et al. (2005) and Bhagat (2005) *Scirpus* IBI varied across the agricultural anthropogenic stressor gradient derived from GIS-recorded data measured at the segment shed scale. More specifically,

- Testing bivariate randomness in 'Fish IBI' and 'Agricultural Stress Gradient'.
- Regressing 'Fish IBI' with 'Agricultural Stress Gradient' and selecting an appropriate model.
- Regression through different quantiles and estimation of prediction interval.
- Estimation of threshold parameters and determination of the "undegraded" zone, the "transition" zone, and the "degraded" zone in terms of 'Agricultural Stress Gradient'.

The statistical analyses and graphical procedures have been done by **R** (version 2.3.1) and **S-Plus** (version 7.0) programming environment and the programs have been included in the Appendix.

Chapter 2

Test for Bivariate Randomness

2.1 Introduction

The objective of this chapter is to test bivariate randomness of the random variables 'Fish IBI' and 'Agricultural Stress Gradient'. If they are bivariate random, then it is foolhardy to go further to try to identify any pattern between them. There is a huge collection of literature to test for bivariate randomness. Among them we have chosen four popular and powerful tests to serve for our objective in this chapter. Mathematically, we can express the null hypothesis as:

$$H_0 : ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)) i.i.d. \sim u([x, y]) \tag{2.1}$$

This implies that the two random variables x and y are uniformly and independently distributed. The alternative hypothesis can be expressed as:

$$H_A : ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)) i.i.d. \nsim u([x, y]) \tag{2.2}$$

This implies that there is some pattern between the two random variables.

Clark and Evans (1954) proposed the 'Distance to nearest neighbor' method as a mea-

sure of spatial relationships. One mentionable feature of that proposed methodology is the simplicity of the procedure. Bartlett (1964) extended his proposed spectral analysis from a one-dimensional point process to a two-dimensional stationary point processes. Various distance-based methods of testing for randomness in a population of spatially distributed events were described by Diggle et al. (1976). They concluded that the method of T-square sampling can help to provide quick and informative results and suited to large populations.

A rigorous foundation for the second-order analysis of stationary point processes on general spaces was provided by B.D.Ripley in 1976. In that paper, the main tool was the decomposition of moment measures pioneered by Krickeberg and Vere-Jones and the Ripley's K function was proposed. Diggle (1977) proposed a two-stage procedure for the detection of random-heterogeneity and applied it in plant populations. Ripley (1979) incorporated the methods of edge-correction in tests of "randomness" and investigated the asymptotic distribution theory and power of tests based on the nearest-neighbour distances and estimates of the variance function. Diggle (1979) discussed the objectives of spatial point pattern analysis, for the mapped data and reviewed the available models, discussed the role of preliminary testing and outlined a procedure for fitting a parametric model.

Renshaw and Ford (1984) described spatial point pattern using two-dimensional spectral analysis. Four functions were described: the autocorrelation function; the periodogram; and, the R- and Θ-spectra, which respectively summarize the periodogram in terms of scale and directional components of patterns. Zimmerman(1993) examined the randomness of a mapped spatial pattern of events in a rectangle D in R^2 using

a bivariate Cramer-von Mises type statistic that is based on the 'distance' between the bivariate empirical distribution function of the events' Cartesian co-ordinates and the uniform distribution function. In a simulation study, he showed that the proposed test was superior to existing tests for detecting heterogeneous alternatives to spatial randomness but inferior for detecting regular or aggregated alternatives. The feature of this test is its simplicity.

A test based on the angle between the vectors joining each sample point to its two nearest neighbors was derived by Renato in 1994. He proposed that the test statistic can be useful in forestry and ecology studies in regions with terrains that make distance measurements difficult. Zimmerman (1994) derived and tabulated the limiting null distribution for an origin-invariant bivariate Cramer-von Mises-type statistic using the principal component decomposition method. Mugglestone and Renshaw (1996) showed how spectral analysis can be used as a tool for the exploratory analysis of spatial point patterns. They compared the interpretations obtained using spectral analysis with those derived from analysis using the reduced second-order moment function. A test for the complete spatial randomness hypothesis of a point pattern in R^2, based on functions of the spacings between x-ordinates and the spacings between y-ordinates was proposed by Cucala et al. in 2000. They showed in a simulation study that the proposed test was inferior to existing tests for detecting regularity or clustering but more powerful for detecting certain types of heterogeneity. Prayag et al. (2000) proposed a test of randomness based on Eberhardt's index and empirically obtained its distribution. The test is powerful under various degress of aggregation and regularity.

Mugglestone and Renshaw (2001) developed a spectral framework for testing the hypothesis of complete spatial randomness (CSR) for a spatial point pattern. They compared five spectral tests and performed a simulation study to compare the power of those tests with the widely used tests for CSR specifically with the test based on the mean nearest-neighbour distance, test based on the reduced second-order moment function, and the test based on the bivariate Cramer-von Mises statistic. They showed that test based on the scaled cumulative R-spectrum was more powerful than the widely used tests for detecting clustered alternatives, especially when the number of events was small.

In this chapter, four widely used tests for CSR are investigated. They are tests based on mean nearest-neighbor distance, scaled cumulative R-spectrum, the reduced second-order moment function (K-function), and bivariate Cramer-von Mises statistic. The reason for choosing these tests is that they were shown to be powerful against different types of alternatives especially against clustered alternative which is what we observe in our data.

2.2 Test Statistics

In this section the description of the four test statistics that has been chosen are narrated briefly.

2.2.1 Test based on the mean nearest-neighbour distance

This test was first proposed by Clark and Evans (1954). This is the simplest test among the four chosen tests. Let an observed pattern consists of N events (points)

in a rectangle with sides of length l_x and l_y. The first step in calculating the test statistic is to scale the coordinates of the events to the unit square. Let $(x_i, y_i), i = 1, 2, ..., N$ denote the coordinates of the events relative to the unit square. For N events observed in the unit square, the mean nearest-neighbour distance is given by $\bar{d} = \sum_{i=1}^{N} d_i/N$, where d_i (i=1,2,...,N) represents the distance from the ith event to its nearest neighbour. Under the null hypothesis, the distribution of $\bar{d}$ is approximately normal with mean and variance given by $\mu_{\bar{d}} = 0.500N^{-1} + 0.164N^{-3/2}$ and $\sigma^2_{\bar{d}} = 0.070N^{-2} + 0.148N^{-5/2}$, respectively (Donnelly, 1978). For cluster processes, $\bar{d}$ tends to be less than $\mu_{\bar{d}}$ whilst for inhibition processes, it tends to be greater than $\mu_{\bar{d}}$. Thus a two-sided test for CSR is obtained by comparing

$$T = (\bar{d} - \mu_{\bar{d}})/\sigma_{\bar{d}} \sim N(0, 1) \tag{2.3}$$

with critical values of the standard normal distribution.

2.2.2 Test based on the cumulative R-spectrum

The latest version of this test is due to Mugglestone and Renshaw(1996). The calculation of the periodogram depends on the coordinates of the events to the unit square; this reduces the bias in the periodogram at low frequencies (Bartlett, 1964). Let $(x_i, y_i), i = 1, 2, ..., N$, denote the coordinates of the events relative to the unit square. The periodogram is given by:

$$\hat{f}(w_p, w_q) = \{\sum_{i=1}^{N} cos\{N(w_p x_i + w_q y_i)\}\}^2 + \{\sum_{i=1}^{N} sin\{N(w_p x_i + w_q y_i)\}\}^2 \tag{2.4}$$

which is calculated for the frequencies $(w_p, w_q) = (2\pi p/N, 2\pi q/N)$, where $p = 0, 1, ..., p_{max}$ and $q = -q_{max}, -q_{max} + 1, ..., q_{max} - 1$ for suitable values of p_{max} and q_{max}. The suitable values of p_{max} and q_{max} depend on the optimum selection of periodogram values

which in turn depends on the number of events suggested by Mugglestone and Renshaw (1996). The values for $p_{max} = 5$ and $q_{max} = 5$ are capable of capturing the suggestion of Mugglestone and Renshaw.

Calculation of the periodogram amounts to a transformation from coordinate data to a $(p_{max} + 1) \times 2q_{max}$ matrix of periodogram values. The ordinates (elements) of the periodogram hold information about the strength of periodicities and the intensity of the point pattern. The term $\hat{f}(w_p, w_q)$ corresponds to a periodic pattern with p repeats in the x direction and q repeats in the y direction.

Completely random spatial point processes are characterized by "flat" periodograms; the values are roughly constant at all frequencies because no one frequency dominates the process. The R-spectrum, $\hat{f}_R(r)$, summarizes average periodogram values for ordinates with similar values of $\acute{r} = \sqrt{(p^2 + q^2)}$; it is used to investigate scales of pattern. If a spatial pattern is isotropic (that is, if it does not contain any directional structure), then the R-spectrum fully captures the second-order structure of the pattern. Formally

$$\hat{f}_R(r) = [\sum_{\acute{r}} \sum_{\theta} \hat{f}(w_p, w_q) I_{\{r-1<\acute{r}\leq r\}}(\acute{r}) I_{\{0^0\leq\theta<180^0\}}(\theta)]/n_r \qquad (2.5)$$

where $\theta = tan^{-1}(p/q)$. Here n_r denotes the number of periodogram ordinates for which $r - 1 < \acute{r} \leq r$ and $0^0 \leq \theta < 180^0$.

Note that the ordinate for $p = q = 0$ is excluded from the averaging procedure (since it has different sampling properties to the rest of the periodogram). Ordinates corresponding to positive values of q on the row for $p = 0$ are also excluded since these are exact repeats of ordinates for negative values of q. These ordinates are discarded before calculating n_r.

Tests based on the precise bands of frequency magnitudes might be more powerful than the other spectral tests, since certain frequency ranges will be more affected by departures from CSR than others. One way of achieving this more sensitive form of the test is to determine appropriate "cut-off" levels $r_1 \leq r \leq r_2$ in the R-spectrum that are especially tailored for testing against specific alternatives to CSR.

A natural extension to the R-spectrum is the cumulative R-spectrum proposed by Mugglestone (1990)

$$\hat{f}_{CR}(r) = \sum_{s=1}^{r} n_s \hat{f}_R(s)/N_r \tag{2.6}$$

where $N_r = \sum_{s=1}^{r} n_s$. This indicates the total power present in the spectrum up to a scale of pattern, r. The scaled cumulative R-spectrum has the distribution

$$\hat{f}_{CR}(r)/N \sim (2N_r)^{-1}\chi^2_{2N_r} \tag{2.7}$$

Invoking the above equation for a given value of r allows us to calculate critical values of $\hat{f}_{CR}(r)/N$ for a test of CSR. Two-sided critical region from the χ^2 distribution is necessary to capture the specific alternative.

The simulation experiments of Mugglestone (1990) and Mugglestone and Renshaw (1996) suggest that tests of CSR should be based on ordinates of the cumulative R-spectrum for which $r \leq 5$. According to their experiments the ordinates of the cumulative R-spectrum for $r \leq 2, 3, 4, 5$ are based on $N_r = 6, 14, 24, 40$ periodogram values, these form appropriate test statistics for patterns containing at least $N = 12, 28, 48, 80$ events, respectively.

2.2.3 Test based on the reduced second-order moment function

The reduced second-order moment function, or K-function, represents the expected number of events within distance t of a randomly chosen event. Under the null hypothesis, K(t) is equal to πt^2, whereas for cluster processes it exceeds πt^2 and for an inhibition processes it is smaller than πt^2. A test of CSR is typically based on

$$L_m = \sup_{t \leq t_0} |\sqrt{\{\hat{K}(t)/\pi\}} - t| \tag{2.8}$$

where $\hat{K}(t)$ is Ripley's (1976) estimator for $K(t)$ defined as $\hat{K}(t) = N^{-2} \sum k(x, y)$, here $1/k(x, y)$ is the proportion of the circumference of the circle x passing through y within an area and t_0 is some maximum distance of interest. With reference to Mugglestone and Renshow(2001), we shall use $t_0 = 1.25/\sqrt{N}$, so that as N increases the test concentrates on smaller inter-event distances.

According to Zimmerman (1993), the critical value of the test based on L_m is obtained by Monte Carlo simulation each and every time as follows. Let L_{1m} denote the value of L_m for the sample pattern, and calculate L_{im}(i=2,3,...,j), the values of L_m for (j-1) realizations of CSR with intensity N. The attained significance level of the test of CSR is given by $rank(L_{1m})/j$. We shall use $j - 1 = 999$ realizations of CSR, these being obtained by simulating from the null distribution.

2.2.4 Test based on the bivariate Cramer-von Mises statistic

This test, which is due to Zimmerman(1993), measures deviations between the empirical distribution function of the two-dimensional coordinates of a point pattern and

Table 2.1: Selected percentiles of the limiting null distribution of $\bar{w}^2$.

$Pr(\bar{w}^2 \leq x)$	x	$Pr(\bar{w}^2 \leq x)$	x
0.01	0.0433	0.75	0.1713
0.02	0.0487	0.85	0.2062
0.05	0.0569	0.90	0.2337
0.10	0.0664	0.95	0.2806
0.15	0.0747	0.98	0.3425
0.25	0.0881	0.99	0.3892
0.50	0.1219		

the bivariate uniform distribution. For data in the unit square, the test statistic is

$$\bar{w}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}(1-|x_i-x_j|)(1-|y_i-y_j|)/(4N) - \sum_{i=1}^{N}(x_i^2-x_i-0.5)(y_i^2-y_i-0.5)/2 + N/9 \tag{2.9}$$

A two-sided alternative is necessary to capture the specific alternative described earlier. The limiting distribution of $\bar{w}^2$ under CSR is that of an infinite sum of χ_1^2 random variables, and is tabulated by Zimmerman(1994) and is given in Table 2.1.

2.3 Simulation

In order to evaluate the performance of the four chosen tests for our type of alternatives, we generated data from the null situation as well as from the alternatives and converted to the unit square using the origin and scale transformations. We used the mean and variances of the 'Agricultural Stress Gradient' and 'Fish IBI' variables

given in Table 2 in the Appendix. The mean and variance of the stress variable calculated from the data are -0.1033 and 0.5678 respectively. For the 'Fish IBI' variable the mean and variance are 44.4 and 174.2483 respectively. The estimated correlation coefficient between the two variables is -0.6474. We evaluated the four test statistics under several situations by generating data from:

1. null distribution to check the empirical size of those tests. We calculated the standard error of the empirical size using the formula $\sqrt{\hat{p}\,(1-\hat{p})/n}$.

2. bivariate normal distribution with means and variances of the fish data and correlation coefficients ranging from -0.30 to -0.80.

A nominal level of 4% was used. One thousand samples with these parameter values were generated. The empirical power of each test is given by the proportion of the 1000 replicates for which the bivariate randomness null hypothesis was rejected.

Results of empirical α level for the four tests are given in Table 2.2. All of the four tests show slight fluctuations around 0.04 in their size for smaller sample sizes. For larger sample sizes, the empirical level of significance for all of the four tests stabilizes with a little bit of conservativeness.

The overall performance of these four tests (mean nearest-neighbour distance, cumulative R-spectrum, reduced second-order moment function, bivariate Cramer-von mises statistic) were evaluated under bivariate normal distributions with specific means and variances but for different correlation coefficients (ρ= -0.30, -0.40, -0.50, -0.60, -0.64737, -0.70, -0.80). Results of empirical power are displayed in Table 2.3.

Mean nearest-neighbour distance (T): The performance of this test for smaller sample sizes ($N \leq 25$) and for weaker associations ($-0.70 \leq \rho \leq -0.30$) is not satisfactory

Table 2.2: Estimated size (standard error) of nominal 4% tests of bivariate randomness

	N				
Test statistic	12	25	50	75	100
T	0.035 (0.0058)	0.046 (0.0066)	0.032 (0.0056)	0.032 (0.0056)	0.038 (0.0060)
$\hat{f}_{CR}(r)/N$	0.040 (0.0062)	0.034 (0.0057)	0.044 (0.0065)	0.042 (0.0063)	0.037 (0.0060)
L_m	0.044 (0.0065)	0.032 (0.0056)	0.050 (0.0016)	0.036 (0.0059)	0.039 (0.0061)
$\bar{w}^2$	0.036 (0.0059)	0.026 (0.0050)	0.045 (0.0065)	0.037 (0.0060)	0.037 (0.0060)

comparing to the other three tests. It showed better power performance for larger sample sizes ($N \geq 30$) in comparison with smaller sample sizes for weaker correlations ($-0.70 \leq \rho \leq -0.30$). In stronger correlation situations ($-0.80 \leq \rho \leq -0.70$), the power of this test is better in comparison to weaker correlations but showed the lowest power of rejecting the null hypothesis among the four chosen tests.

Cumulative R-spectrum ($\hat{f}_{CR}(r)/N$): This test showed better power performance compared to T irrespective to every situation. The performance of this test and of the test L_m are almost the same for weaker correlations. But for higher correlation situations ($-0.80 \leq \rho \leq -0.60$), its performance is below L_m. For the weaker correlation situations, the performance of this test is far below than that of the test based on the bivariate Cramer-von mises statistic ($\bar{w}^2$) but for stronger associations ($-0.80 \leq \rho \leq -0.60$), the performances of the tests based on cumulative R-spectrum ($\hat{f}_{CR}(r)/N$) and bivariate Cramer-von mises statistic ($\bar{w}^2$) are almost the same. This

Table 2.3: Power of four tests at 4% level. Data generated from bivariate normal with $\mu_x = -0.1033$, $\sigma_x^2 = 0.5678$, $\mu_y = 44.4$, $\sigma_y^2 = 174.2483$ and different correlation coefficients ρ

Corr. coef.		N				
ρ	Test statistic	15	20	25	30	35
-0.30	T	0.035	0.095	0.157	0.243	0.300
	$\hat{f}_{CR}(r)/N$	0.216	0.448	0.684	0.661	0.802
	L_m	0.228	0.391	0.644	0.794	0.901
	$\bar{w}^2$	0.447	0.583	0.770	0.871	0.950
-0.40	T	0.058	0.126	0.215	0.284	0.386
	$\hat{f}_{CR}(r)/N$	0.259	0.495	0.744	0.735	0.870
	L_m	0.261	0.496	0.725	0.855	0.950
	$\bar{w}^2$	0.452	0.616	0.801	0.896	0.969
-0.50	T	0.064	0.162	0.250	0.383	0.504
	$\hat{f}_{CR}(r)/N$	0.297	0.593	0.820	0.764	0.887
	L_m	0.315	0.606	0.791	0.919	0.971
	$\bar{w}^2$	0.491	0.643	0.817	0.915	0.966
-0.60	T	0.115	0.234	0.398	0.531	0.653
	$\hat{f}_{CR}(r)/N$	0.345	0.644	0.889	0.879	0.940
	L_m	0.394	0.702	0.901	0.971	0.992
	$\bar{w}^2$	0.496	0.680	0.853	0.958	0.987
-0.64737	T	0.147	0.294	0.453	0.613	0.722
	$\hat{f}_{CR}(r)/N$	0.388	0.705	0.917	0.890	0.976
	L_m	0.418	0.743	0.938	0.987	0.998
	$\bar{w}^2$	0.509	0.706	0.893	0.971	0.992
-0.70	T	0.191	0.341	0.567	0.710	0.817
	$\hat{f}_{CR}(r)/N$	0.439	0.774	0.940	0.928	0.984
	L_m	0.540	0.813	0.963	0.996	0.999
	$\bar{w}^2$	0.522	0.745	0.911	0.973	0.999
-0.80	T	0.338	0.634	0.815	0.908	0.968
	$\hat{f}_{CR}(r)/N$	0.546	0.910	0.983	0.988	0.999
	L_m	0.651	0.943	0.993	1.000	1.000
	$\bar{w}^2$	0.550	0.786	0.934	0.996	0.999

test showed lower power performance for the sample size 30 than for the sample size of 25. This might be due to the definition of initial values proposed by Mugglestone and Renshaw (1996). According to them, for sample sizes of 25 and 30, the test statistic uses 6 and 14 periodogram values respectively. So, there is further scope for redefinition of the initial values for this test.

Reduced second-order moment function (L_m): This test has the highest power among the four chosen tests in stronger correlation situations ($-0.80 \leq \rho \leq -0.60$) and showed good power (power > 0.40) performances for larger ($N \geq 20$) sample sizes.

Bivariate Cramer-von mises statistic ($\bar{w}^2$): This test showed the best power performances among the four chosen tests in the situations of weaker correlations ($-0.50 \leq \rho \leq -0.30$) irrespective to small and large sample sizes, and the test based on the reduced second-order moment function L_m showed second highest power performance. But for strong correlation situations ($-0.80 \leq \rho \leq -0.60$) and for the sample sizes ($N \geq 20$), the performance of ($\bar{w}^2$) is slightly lower than the test based on the reduced second order moment function (L_m).

Table 2.3 also gives the estimated power of the four tests for our specific alternative situation (correlation coefficient, $\rho = -0.64737$). We see that test based on the mean nearest-neighbour distance (T) gives the lowest power among the four chosen tests for all of the chosen sample sizes. The test based on the reduced second-order moment function (L_m) gives the highest power among the four chosen tests when the sample sizes are large, i.e., ≥ 20. The test based on the bivariate Cramer-von Mises statistic ($\bar{w}^2$) gives the highest power when the sample size is small. This is extremely important in ecology, because often it is difficult to obtain large sample sizes. The test

based on the cumulative R-spectrum is in the third position among the four tests in terms of power. Except for the test based on mean nearest-neighbour distance (T), all tests showed good power performance in that situation.

2.4 Data Analyses

Table 2.4 shows the results after the application of four chosen tests to our data set. All of the tests resulted in rejection of the null hypothesis of bivariate randomness at 4% level of significance. The test based on the reduced second order moment function (L_m) shows p-value of < 0.001, indicating the strongest evidence against the null hypothesis.

Hence, we conclude that there is some non-random pattern present between "Fish IBI" and "Agricultural Stress Gradient". So it is worthwhile to go and find the pattern between the two random variables.

Table 2.4: Application of the tests to the data (level of significance= 0.04)

Test statistic	Calculated Value	LCV	UCV	P-val	Decision
T	-3.606	-2.054	2.054	—	Rejected
$\hat{f}_{CR}(r)/N$	3.314	0.530	1.622	—	Rejected
L_m	0.123	—	—	<0.001	Rejected
$\bar{w}^2$	0.797	0.0487	0.3425	—	Rejected

LCV: Lower Critical Value UCV: Upper Critical Value

2.5 Summary

In this chapter we have chosen the four most powerful tests available, which have frequently been used as tests for CSR. We performed a simulation study to compare their empirical power and size. All of the tests showed some fluctuations in size for small sample sizes but stabilized for large sample sizes and showed some conservativeness for rejecting the null hypothesis. The test based on reduced second-order moment function (L_m) showed the highest power for larger sample sizes and for stronger correlations while the test based on bivariate Cramer-von Mises statistic ($\bar{w}^2$) showed the highest power for small sample size and for weaker correlations. The test based on the mean-nearest neighbor distance showed least tendency to reject the null hypothesis when the alternative is true among the four chosen tests. All of the four chosen tests were applied to our data set and showed strong evidence against the null hypothesis, indicating some pattern in the data.

Chapter 3

Regression and Estimation of Thresholds

3.1 Introduction

Regression is the basic statistical technique to model statistical data. Linear regression is the most popular and widely used regression technique among others. There are other techniques also, such as non-linear regression, nonparametric regression etc. Recently, the piecewise regression technique has become popular for modelling environmental threshold parameters in ecology. In ecology thresholds occur when the response of a species or ecological process to an independent variable is not linear, but changes abruptly at some threshold level of loss (Toms and Lesperance 2003). Piecewise regression models are "broken-stick" models, where two or more lines are joined at some unknown points, called "break-points". "Breakpoints" can be used as estimates of thresholds and are used to determine the width of edge effects. Piecewise

linear regression allows transitions from one linear regime to another linear regime. The aim is to estimate the position of the transition points between the linear regimes and to take these as estimates of threshold parameters.

Bacon and Watts (1971) defined a model that includes two intersecting straight lines. They used a Bayesian estimation procedure to determine the plausibility of different parameter values. Watts and Bacon (1974) extended their proposed model using the hyperbola as a transition model to fit two regime straight line data. Tishler and Zang (1981) proposed a maximum likelihood algorithm for piecewise regression model. Chiu (2002) incorporated a bent-cable model to estimate ecological thresholds. Toms and Lesperance (2003) used all of the previously defined models to estimate ecological threshold parameters. They also showed how the piecewise regression model can be fitted using a standard non-linear least squares algorithm. All of the previous works were used to define and estimate one breakpoint (i.e., threshold). These models incorporated two straight lines joined at the threshold. In this paper, we fitted a piecewise linear regression model to our data to capture two breakpoints incorporating three linear regression lines joined at two threshold points (Seber and Wild 1989). We also incorporated the non-linear least squares method to estimate the model parameters proposed by Toms and Lesperance in 2003.

3.2 Model fitting

Different types of regression models can be fitted to our data given in Table 2 in the Appendix. The most widely used linear regression model was fitted first. As an improvement over the linear model, the non-linear logistic model was fitted also. Then

we tried to guess the position of breakpoints using some nonparametric regression techniques (local mean, local linear regression and LOESS). Finally, the piecewise linear regression model was fitted to the data to estimate the threshold parameters.

3.2.1 Linear Regression

Methods

Denote the variable 'Fish IBI' as y, and variable 'Agricultural Stress Gradient' as x, and the linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{3.1}$$

where ε is the error term independently and identically distributed with $N(0, \sigma^2)$. Our estimation procedure is the least square method where we get the estimates by minimizing the sum of squares of deviations from the true line

$$S = \Sigma_i^n \varepsilon_i^2 = \Sigma_i^n (y_i - \beta_0 - \beta_1 x_i)^2. \tag{3.2}$$

We can determine b_0 and b_1 (estimates of β_0 and β_1 respectively) by differentiating equation (3.2) with respect to β_0 and β_1 and setting the results equal to zero.

Results

From the estimates of the coefficients of the simple linear regression equation presented in Table 3.1, we can see that the intercept and slope parameters are significantly different from zero. The sign of the estimated slope coefficient is -11.340, implying that 'Fish IBI' decreases as 'Agricultural Stress Gradient' increases. From the analysis of variance table (Table 3.2), we can see the F-value is very high with small p-value, indicating that the slope coefficient is significantly different from zero.

Table 3.1: Estimate of the parameters of the linear regression line

Coefficients	Estimate	Std. Error	t-value	$P(>\|t\|)$
β_0	43.229	1.887	22.903	< 0.001
β_1	-11.340	2.523	-4.494	0.00011

Table 3.2: Analysis of variance for the straight line relationship

Source	df	SS	MS	Calculated F-value	$Pr(> F)$
Regression	1	2117.74	2117.74	20.2	0.00011
Residual	28	2935.46	104.84		
Total(corrected)	29	5053.2			

$R^2 = 0.4191$

The model explained only 41.91% variation of the total variation in 'Fish IBI' which is rather a poor performance. The residual sum of squares for this linear regression model is 2935.46. Figure 3.1 shows the fitted linear regression line and the data. We can see that the linear model overestimates the value of Fish IBI at smaller Stress, underestimates Fish IBI when the Stress Gradient is around -0.5, and overestimates IBI for the Stress around 0.08. For the purpose of improvement in regression in the next section of this chapter we are going to fit non-linear regression model.

3.2.2 Non-linear Regression: Logistic Curves

We attempted to improve the fit by using logistic curves. Again, we denote the 'Fish IBI' as y and 'Agricultural Stress Gradient' as x. Our first logistic curve with the

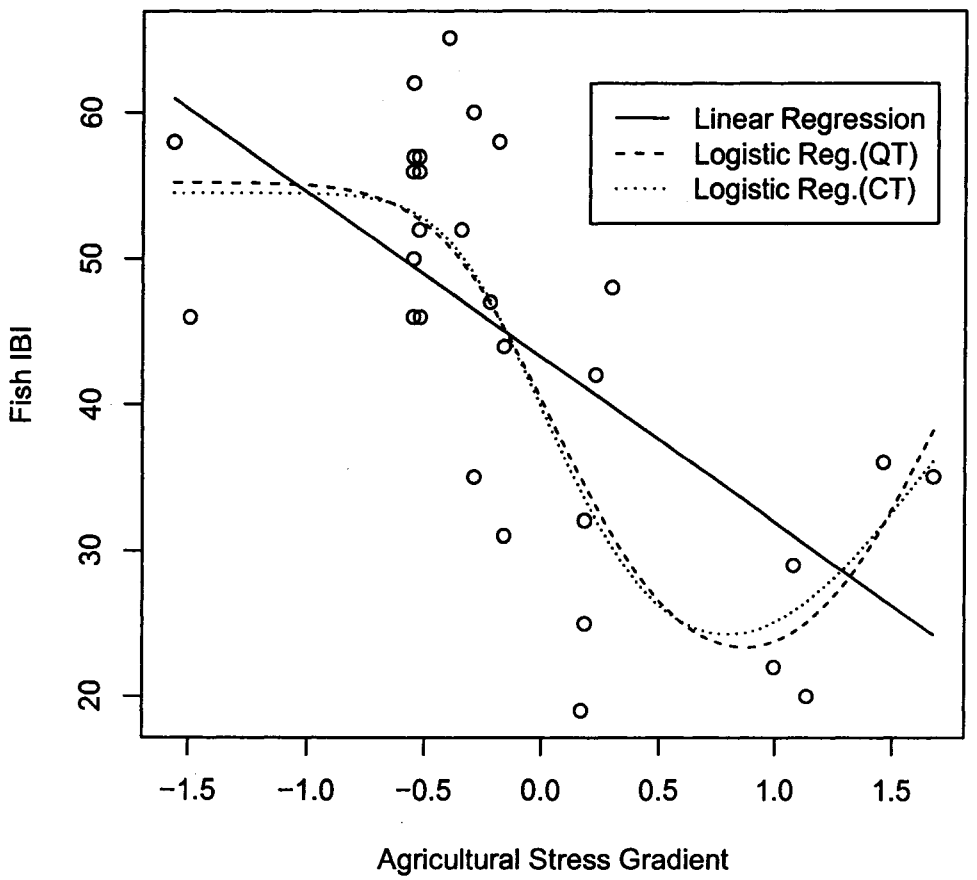

Figure 3.1: Fitted linear regression line and logistic curves one with quadratic term and the other with cubic term superimposed to the data

quadratic term is given by:

$$E(y|x) = A/(1 + \exp(B + Cx + Dx^2)) \tag{3.3}$$

where A, B, C, D are the model parameters. Our second logistic curve including a cubic term is given by:

$$E(y|x) = A/(1 + \exp(B + Cx + Dx^2 + Ex^3)) \tag{3.4}$$

where A, B, C, D , E are the model parameters. The estimation procedure is the Gauss-Newton non-linear least squares and the parameter estimates are displayed in the result section at Tables 3.3 and 3.4. To fit the models with non-linear least squares we used the software package R version 2.3.1.

Model Selection

Usually, there is no true model. Instead, a model only approximates reality. The question then is to find which model would best approximate reality given the data. In other words, we are trying to minimize the loss of information. Kullback and Leibler (1951) addressed such issues and developed a measure, the Kullback-Leibler information, to represent the information lost when approximating reality. A few decades later, Akaike (1973) used Kulback-Leibler information for model selection. He established a relationship between the maximum likelihood and the Kullback-Leibler information. In essence, he developed an information criterion to estimate the Kullback-Leibler information, resulting in Akaike's information criterion (AIC)

$$AIC = -2\{\log(likelihood)\} + 2K \tag{3.5}$$

where K is the number of parameters included in the model. This reflects the overall fit of the model with small values indicating good fit.

In cases where analyses are based on more conventional least squares regression for normally distributed errors, one can compute the AIC with the following formula

$$AIC = n\log(\hat{\sigma}^2) + 2K \tag{3.6}$$

where, $\hat{\sigma}^2$ = Residual Sum of Squares/n, and n is the sample size.

Results

The parameter estimates of the logistic curves with quadratic and cubic terms are presented in Tables 3.3 and 3.4 respectively. From the parameter estimates of the logistic curve with quadratic term, we see that all the parameters are significant at the 5% level. The residual sum of squares is given by 2037.855 which is far below the residual sum of squares of simple linear regression fit. The AIC for this model is 219.69. On the other hand the estimates of the parameters D and E of the logistic curve with cubic term are not significant. Though the residual sum of squares decreases to 2018.729, the AIC increases to 221.407 indicating that there is more loss of information in the later model comparing to the former. So among the two logistic

Table 3.3: Estimates of the parameters of the logistic curve with quadratic term

Parameters	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
A	55.257	4.274	12.928	< 0.001
B	-0.985	0.478	-2.061	0.049
C	2.988	1.250	2.391	0.024
D	-1.723	0.812	-2.122	0.044

$RSS = 2037.855, \quad AIC = 219.69$

curves, the model with quadratic term is better. The fitted logistic curve is presented

in Figure 3.1, showing it fits the data well. It shows better performance of residual sum of squares comparing to the linear counterpart. However, one result of this fitted line is against our expectation. It shows the increase of Fish IBI with increase of Stress Gradient when the Stress is ≥ 1.00. Our expectation of decreasing Fish IBI with increasing Stress is violated. So, our next effort is to find a way to overcome this difficulty.

Table 3.4: Estimates of the parameters of the logistic curve with cubic term

Parameters	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
A	54.541	4.338	12.573	< 0.001
B	-0.982	0.480	-2.046	0.051
C	3.318	1.851	1.792	0.085
D	-2.632	2.696	-0.976	0.338
E	0.455	1.125	0.404	0.689

$RSS = 2018.729, \quad AIC = 221.407$

3.2.3 Nonparametric Regression

There are cases where linear models fit poorly because of intrinsic nonlinearity in the data. Nonparametric regression aims to provide a means of modelling such nonlinearity in the data. Even where the suitability of linear models has not yet been brought into question, smoothing techniques are still useful by enhancing scatterplots to display the underlying structure of the data, without reference to a parametric model. In this section, the main goal of nonparametric regression is to gather

knowledge about the pattern of the data. A suitable model can be described as

$$y = m(x) + \varepsilon \tag{3.7}$$

where y denotes the response variable, x the covariate, $m(x)$ is the model of interest and ε denotes an independent error term with mean 0 and variance σ^2.

Kernel Smoothig: Local Mean

A simple kernel approach is to construct the local mean estimator

$$\tilde{m}(x) = \frac{\sum_{i=1}^{n} w(x_i - x; h) y_i}{\sum_{i=1}^{n} w(x_i - x; h)} \tag{3.8}$$

which was first proposed by Nadaraya (1964) and Watson (1964). The kernel function $w(z; h)$ is generally a smooth positive function which peaks at 0 and decreases monotonically as z increases in size. This ensures that most weight is given to the observations whose covariate values x_i lie close to the point of interest x. For convenience, a normal density function, with standard deviation h, is commonly used as the kernel. But in our problem we used the kernel 'box' because of the sparseness of the data. The kernel 'box' is a square with point of interest x as the center having $2h$ length with each sides. The smoothing parameter h controls the width of the kernel function, and hence the degree of smoothing applied to the data. When h is the standard deviation of a normal density, observations over an effective range of $4h$ in the covariate axis contribute to the estimate. As the smoothing parameter increases, the resulting estimator misses some details in the curvature of the data. As the smoothing parameter decreases too much, the estimator begins to track the data closely and ends up interpolating the observed points. Clearly, some effective compromise is required. In order to do that, we choose a collection of values of h

from 0.3 to 1.0 with interval 0.01 and calculate $RSS = \sum_{i=1}^{n}(y_i - \tilde{m}(x_i))^2$ like statistic. For each h, we choose that h which gives the smallest RSS which is $h = 0.47$ for our data set.

Local Linear Regression

An alternative approach to the construction of a local mean for the data is to fit a local linear regression. This involves solving the least squares problem

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 w(x_i - x; h) \tag{3.9}$$

and taking as the estimate at x the value of $\hat{\alpha}$, as this defines the position of the local regression line at the point x. Again, it is the role of the kernel weights to ensure that observations close to x have the most weight in determining the estimate. The local linear estimator can be given an explicit formula

$$\hat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{s_2(x; h) - s_1(x; h)(x_i - x)\} w(x_i - x; h) y_i}{s_2(x; h) s_0(x; h) - s_1(x; h)^2} \tag{3.10}$$

where $s_r(x; h) = \{\sum (x_i - x)^r w(x_i - x; h)\}/n$. The local mean estimator described above can be derived in a similar way by removing the $\beta(x_i - x)$ term from the formulation of the least squares problem in equation (3.9). The advantage of the local linear regression estimator over the local mean estimator is its superior behavior near the edges of the covariate space. The local mean estimator is biased at the edges of the covariate space. In our local linear regression estimator, we used the normal kernel function.

One of the properties of nonparametric regression estimators is that bias increases with the size of the smoothing parameter while variance decreases. In order to define a suitable level of smoothing it is therefore necessary to find some compromise between

these properties. A simple strategy is to define at each point of x the mean squared error $E\{\hat{m}(x) - m(x)\}^2$, which is the sum of the squared bias and variance terms. To prepare the way for asymptotic analysis, it is natural to consider the mean integrated squared error given by

$$MISE(h) = \int E\{\hat{m}(x) - m(x)\}^2 f(x)dx \tag{3.11}$$

where $f(x)$ represents the density of observed design points. The MISE is a function of the smoothing parameter h. An optimal value h_{opt} can be defined as the value which minimizes this quantity.

Cross-validation has provided a popular means of selecting smoothing parameters by constructing an estimate of MISE and minimizing this over h. The philosophy of cross-validation is to attempt to predict each response value y_i from the remainder of the data. For the value y_i, this prediction can be denoted by $\hat{m}_{-i}(x_i)$, where the subscript $-i$ denotes omission of the observation (x_i, y_i). The cross-validation function is then defined as

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{m}_{-i}(x_i)\}^2 \tag{3.12}$$

Some simple algebra shows that

$$E\{CV(h)\} = \frac{1}{n}\sum E\{\hat{m}_{-i}(x_i) - m(x_i)\}^2 + \sigma^2 \tag{3.13}$$

The averaging over the design points x_i provides a discrete analogue of the integral and factor $f(x)$ in the MISE curve defined above, and so CV(h) provides a simple estimator for MISE(h), apart from the unimportant additive constant σ^2(For more details see Bowman and Azzalini 1997). For our data, the value of h that minimizes

the cross-validation curve is $h_{cv} = 0.465$. We used a routine named *'sm'* in *R*, and the program has been included in the Appendix.

Variable Bandwidths: LOESS

The estimators described in the previous sections have used the same smoothing parameter h in the weights attached to each observation (x_i, y_i). In some situations it can be advantageous to use different smoothing parameters for different covariate values. Sometimes it is appealing to use a large smoothing parameter where the data are sparse and to use a small smoothing parameter where the data are dense. A simple way to implement this is to employ a variable bandwidth that reflects the density of the design points through a nearest neighbor distance. The bandwidth used in the kernel function for estimation at the point x could be defined by $h_i = h \times d_k(x)/\bar{d}$, where $d_k(x)$ denotes the distance to the kth nearest neighbor of the covariate value x_i and $\bar{d}$ denotes the geometric mean of the $d_k(x)$. In this way, the overall bandwidth h is scaled to increase the degree of smoothing applied in regions where data are sparse, and to decrease the degree of smoothing where the data are dense.

One of the earliest, and still very popular, approaches to nonparametric regression uses nearest neighbor distances in a particularly simple and appealing way. This was described by Cleverland (1979) and is referred to as the lowess estimator, or LOESS after its more general S-Plus implementation. The key component is that in the local linear estimator defined through the least squares criterion (3.10), the kernel function for estimation at the point x is $w(x_i - x; d_k(x))$. This achieves an approximately variable pattern of smoothing without the need for an overall smoothing parameter h. To provide weights which are easily evaluated but adequately smooth, Cleverland (1979)

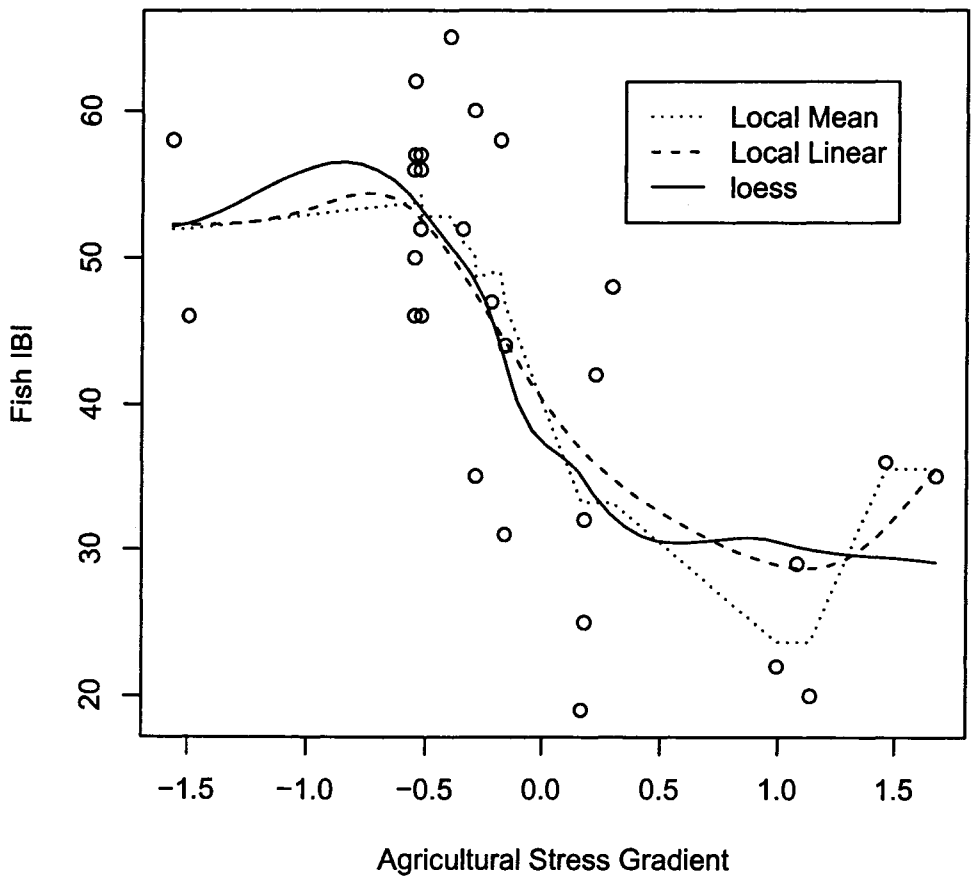

Figure 3.2: Nonparametric regression lines (local mean, local linear regression and LOESS) are superimposed to the data

used the tricube kernel function defined by $w(z;h) = (1-(|z|/h)^3)^3$, for $z\epsilon[-h,h]$. A further appealing feature of this formulation is that the degree of smoothing applied to the data can be expressed in the parameter k/n, which describes the proportion of the sample which contributes positive weight to each local linear regression. This is referred to as the span of the estimator and it has the attraction of a clear and simple interpretation. Here, in our LOESS estimator, we are using the default span of 0.5. The LOESS estimator proposed by Cleverland (1979) incorporates robustness in the fitting procedure, to prevent unusual observations from exerting large influence on the fitted curve. We used an S-plus package named *loess* and the program has been included in the Appendix.

The bootstrap and nonparametric regression

The bootstrap method can often provide a very useful means of deriving the properties of estimators, and of constructing confidence intervals. In order to apply the bootstrap in the context of nonparametric regression, Härdle and Bowman (1988) proposed the following algorithm. For convenience, the notation $\hat{m}(x;h)$ is adopted for the estimator in order to make the values of the smoothing parameter explicit.

1. Construct residuals $\hat{\varepsilon}_i = y_i - \hat{m}(x_i;h_p)$ through a pilot estimator $\hat{m}(x;h_p)$.

2. Create a set of normalized residuals $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \frac{1}{n}\sum_j \hat{\varepsilon}_j$, with mean 0.

3. Repeatedly create bootstrap observations $y_i^* = \hat{m}(x_i;h_p) + \varepsilon_i^*$, through the pilot estimator $\hat{m}(x_i;h_p)$ and random sampling of ε_i^* from $\{\tilde{\varepsilon}_j\}$. We created 500 sets of bootstrap observations.

4. Repeatedly create bootstrap estimators $\hat{m}^*(x;h)$ by smoothing the observations (x_i, y_i^*).

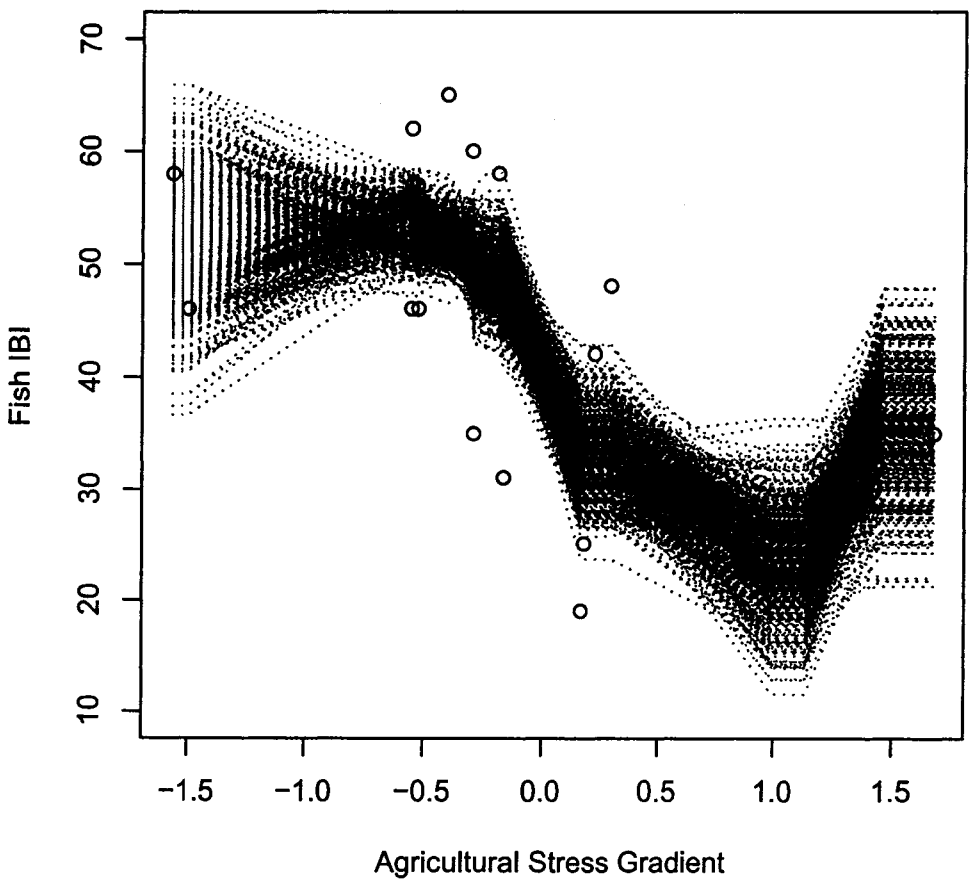

Figure 3.3: Five hundred bootstrap estimates of the local mean regression (kernel smooth)

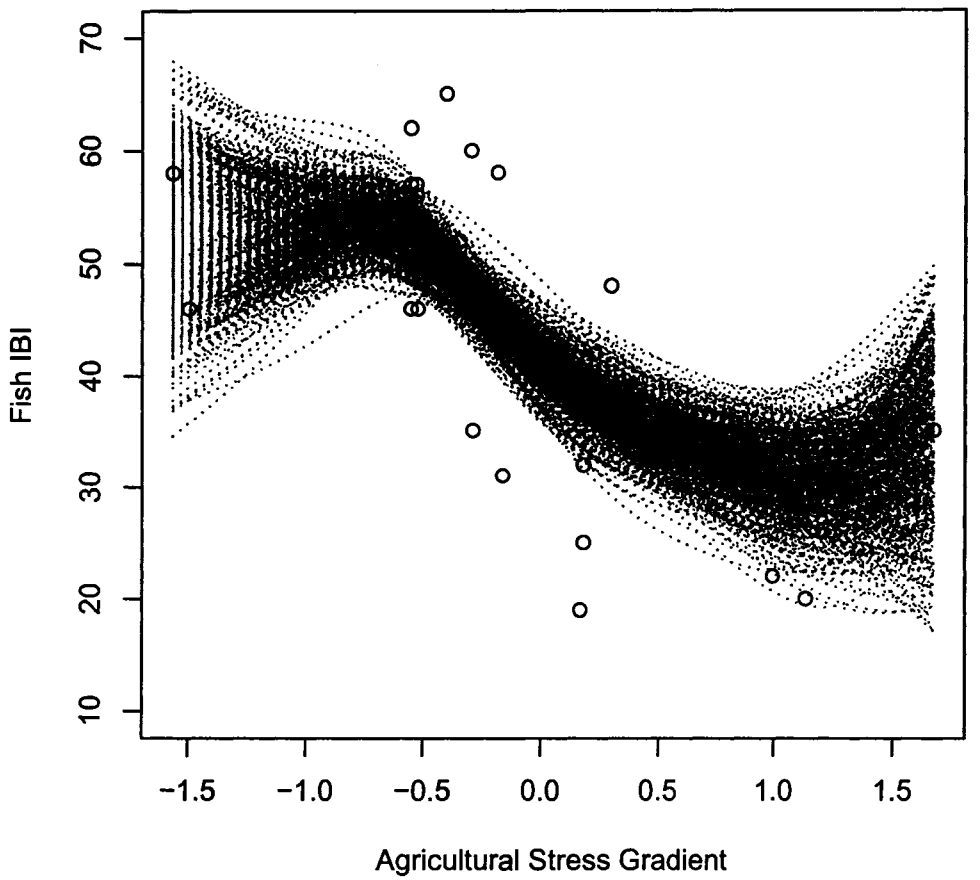

Figure 3.4: Five hundred bootstrap estimates of the local linear regression

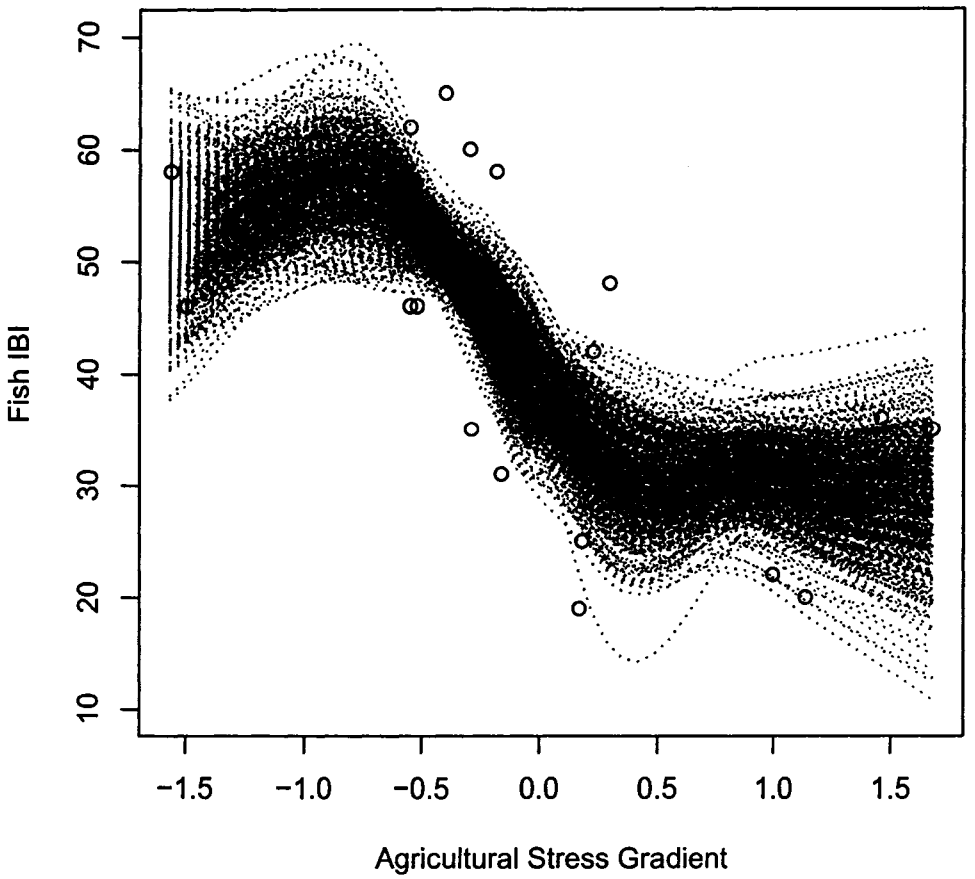

Figure 3.5: Five hundred bootstrap estimates of LOESS

Results

Among the three nonparametric smoothing techniques that have been applied to the data, the 'local mean' is the least appealing and the 'LOESS' is the most appealing of the techniques. The bandwidth for LOESS is wider where the data are sparser and narrower where the data are denser. For our data, there are few observations below stress level -0.5 and above stress level 0.3, which advocates in favor of LOESS technique. Figure 3.2 displays the three types of smoothing curves; local mean, local linear and LOESS. The local mean curve shows that the Fish IBI is stable up to stress level -0.5, then it decreases with the increase of stress level and reaches minimum at stress level around 1.0. The most disappointing thing is that it indicates the increase of Fish IBI with the increase of stress after stress level 1.0. Figure 3.3 gives us the idea of the band of local mean curve. The most disappointing performance of local mean estimator is its indication of positive response of Fish IBI at the stress level ≥ 1.0. Also it shows some sharp changes of the curve at some points. The local linear regression estimator is comparatively stable and does not show sharp changes, but it also shows the positive change of Fish IBI with agricultural stress level beyond 1.0 and gives an upward biased result at this area. Seemingly local linear regression is stable in response to the increase of the stress variable. Both of these techniques (local mean regression and local linear regression) show wider bands at the edge of the stressor. Perhaps the cause behind that is, they use fewer observations in smoothing at the edges.

On the other hand, LOESS gives results consistent with our expectation. It shows some increment in Fish IBI up to stress level -0.7. This might be for very few ob-

servations lying in that region. After stress level -0.7, it is showing a decreasing trend with increment in stress, and reaches its minimum at stress level around 0.4 and stabilizes after that. Figure 3.5 gives an idea of the band of LOESS from 500 bootstrap estimates. The band for LOESS at the edges are narrower than the edges given by the other two nonparametric methods, because at the edges where the data are sparser the LOESS uses wider bandwidth. The five hundred bootstrap band of LOESS indicates that there might be two break points (i.e., thresholds), one is around stress level -0.7 and the other is around stress level 0.4.

3.2.4 Piecewise Linear Regression and Thresholds

Many recent studies have looked for critical thresholds associated with habitat fragmentation (e.g., Andren 1994, Fahring 2001). Critical thresholds occur when the response of a species or ecological process to habitat loss is not linear, but changes abruptly at some threshold level of loss. Abrupt changes in ecological processes can also occur in other systems. Plant and animal communities change within a threshold distance of habitat edges (edge effects; Wales 1972, Gates and Mosher 1981). In this section we described a piecewise linear regression model that is effective in modelling abrupt thresholds. Piecewise regression models are "broken-stick" models, where two or more lines are joined at some unknown point(s), called "breakpoints(s)", representing the threshold(s). We describe a model where the segments are straight lines. We have parameterized the model in such a way that the standard nonlinear models can be fitted and the break point estimates can be obtained as well.

Methods

The fitted LOESS model gives us the indication that there might be two break points in the data. We assume straight line models among the segments separated by thresholds. The piecewise regression model (Seber and Wild 1989) that joins three straight lines sharply at the breakpoints is as follows:

$$y_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i & \text{for } x_i \leq \alpha_1 \\ \beta_0 + \beta_1 x_i + \beta_2 (x_i - \alpha_1) + \varepsilon_i & \text{for } \alpha_1 < x_i \leq \alpha_2 \\ \beta_0 + \beta_1 x_i + \beta_2 (x_i - \alpha_1) + \beta_3 (x_i - \alpha_2) + \varepsilon_i & \text{for } x_i > \alpha_2 \end{cases} \tag{3.14}$$

where y_i is the value for the ith observation, x_i is the corresponding value for the independent variable, α_1 and α_2 are the breakpoints (the two thresholds), and ε_i are assumed to be normal and independent errors with mean zero, constant variance, and finite absolute moment for some *order* > 2 (Seber and Wild 1989). The slopes of the lines are β_1, $\beta_1 + \beta_2$ and $\beta_1 + \beta_2 + \beta_3$, so β_2 can be interpreted as the difference in slopes between lines in first segment and second segment and β_3 can be interpreted as the difference in slopes between lines in second segment and in third segment. Parametrizing the model in this way forces continuity at the breakpoints. We assume that there are abrupt transition at the breakpoints. One drawback of the model is that there may be convergence problems when fitting, especially when the data are sparse, so a careful study of the residual sum of squares surface is needed.

Results

Using non-linear least squares algorithm, we estimated the parameters of the piecewise linear regression model. We wrote the program statements in *R* version 2.3.1. Table 3.5 contains the estimates of the parameters of the model for our data set. According

to the fitted model, the estimates of the first break point (i.e., first threshold) is -0.397 and of the second break point (i.e., second threshold) is 0.087. The residual sum of squares of the fitted model is $RSS = 1976.205$ which is a great reduction in residual sum of squares comparing to the logistic curve that gives $RSS = 2037.855$. The chosen model is very sensitive to the initial values and might converge to a local minimum rather than the global minimum especially when the data are sparse. To check whether the model converged to a global minimum, we produced the contour plot of the residual sum of squares surface (Figure 3.6). The contour plot shows that the convergence of the model in global minimum. Figure 3.7 contains the fitted piecewise linear regression model and depicts the breakpoints (i.e., thresholds). It gives a similar pattern as LOESS though there is some deviation in the first segment between them.

Table 3.5: Estimates of the parameters of the piecewise linear regression model

Parameters	Estimate	Std. Error
α_1-TH1	-0.397	0.190
α_2-TH2	0.087	0.376
β_0	55.323	5.337
β_1	2.508	6.913
β_2	-49.564	51.109
β_3	45.932	50.898

$RSS = 1976.205$

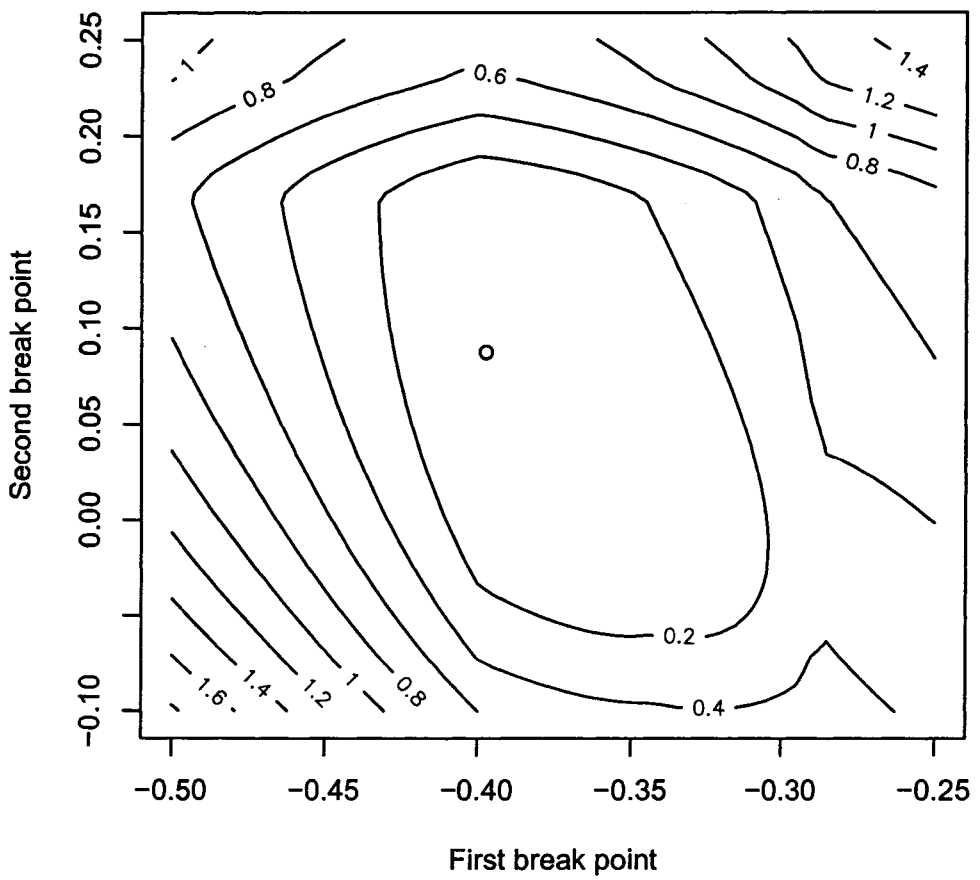

Figure 3.6: Contour plot of the residual sum of squares surface of the piecewise linear regression model

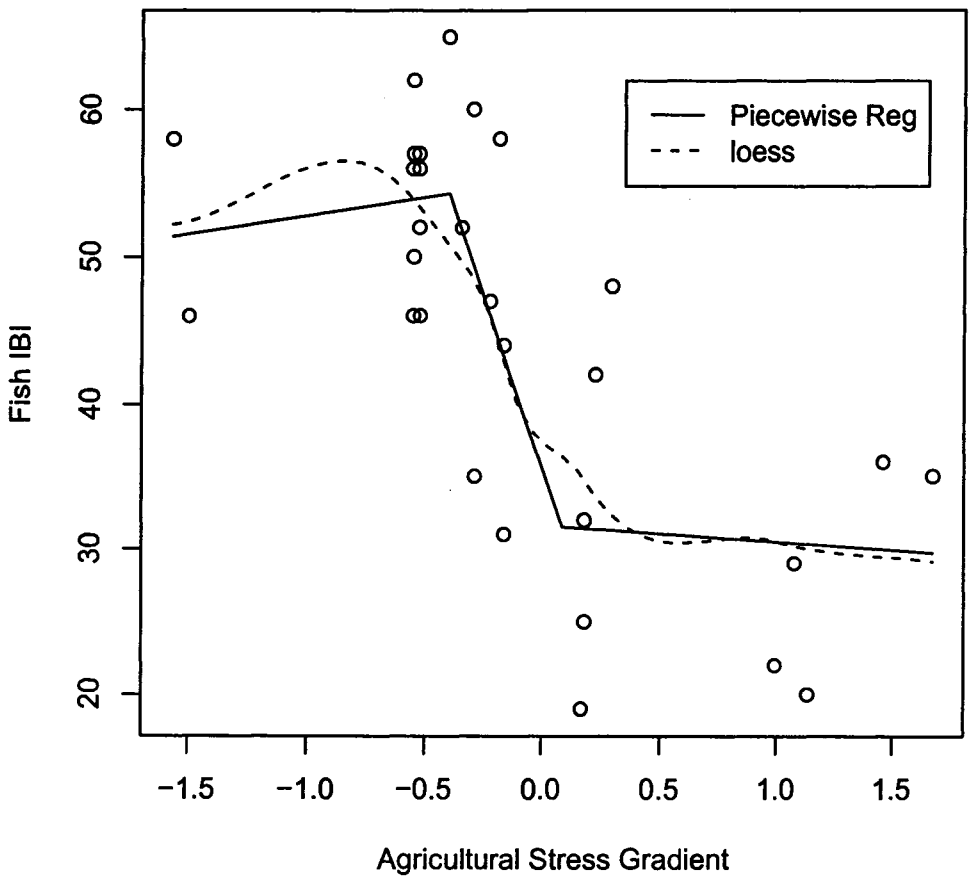

Figure 3.7: Estimated piecewise linear regression model and LOESS superimposed to the data

3.3 Summary

This chapter deals with modelling 'Fish IBI' variable with 'Agricultural Stress Gradient' variable. The linear regression model tells us that the 'Fish IBI' responds negatively with an increase of 'Agricultural Stress Gradient'. Among the non-linear models, we tried the logistic curve with quadratic term fits the data very well. One of the least appealing findings of non-linear logistic curve is its indication of a positive response of 'Fish IBI' with 'Agricultural Stressor' for the region of stress level ≥ 1.00. Three nonparametric regression techniques were employed to investigate the pattern of relationship between the two variables. LOESS gave the most promising results among them. It gives us the idea that the 'Fish IBI' and 'Agricultural Stressor' are negatively related with two thresholds. Finally, the piecewise linear regression model with two breakpoints was fitted using non-linear least squares technique. It gives two thresholds in stresses; one with stress level -0.397 and the other with stress level 0.087. The residuals of the fitted model were calculated and analyzed and indicated consistency with the assumptions behind the estimation technique. To extract a clear view about the "transition" zone (area between two thresholds) and to identify the "degraded" zone (area beyond the second breakpoint), a sophisticated statistical technique will be employed in the next chapter.

Chapter 4

Quantile Regression and Thresholds

4.1 Introduction

Regression is the basic technique in statistics to identify relationships between variables. Usually, a dependent variable y is some function of independent variables x, such as $y = f(x)$. Regression techniques mainly focus on estimating rates of change in the mean of the distribution of the dependent variable as some function of a set of independent variables; in other words, the technique deals with the expected value of y conditional on x, i.e., $E(y|x)$. But it is possible to fit regression curves to other parts of the distribution of the response variable, which is rarely done, and therefore most regression analysis gives an incomplete picture of the relationships between variables (Cade and Noon 2003). Just the mean gives an incomplete picture of a single distribution, so the classical regression gives correspondingly incomplete pic-

ture for a set of distributions (Mosteller and Tukey 1977). Alternatively, quantile regression methods offer a mechanism for modelling the conditional median function, and the full range of other conditional quantile functions. By supplementing the estimation of conditional mean functions with techniques for estimating an entire family of conditional quantile functions, quantile regression is capable of providing a more complete statistical analysis of the stochastic relationships among random variables. (*http* : *//www.econ.uiuc.edu/* $\sim$ *roger/research/rq/rq.html*).

A regression model with heterogeneous variances implies that there is not a single rate of change that characterizes changes in the probability distributions. Focusing exclusively on changes in the means may underestimate, overestimate, or fail to distinguish real nonzero changes in heterogeneous distributions (Terrell et al. 1996; Cade et al. 1999). Unequal variation implies that there is more than a single slope (rate of change) describing the relationship between a response variable and predictor variables measured on a subset of these factors. Quantile regression estimates multiple rates of change (slopes) from the minimum to maximum response, providing a more complete picture of the relationships between variables missed by the other regression methods. According to Cade and Noon (2003) "Quantile regression is a method for estimating functional relationships between variables for all portions of a probability distribution".

When the predictor variables x exert both a change in means and a change in variance on the distribution of y, we have a regression model with unequal variances a "location-scale model" in statistical terminology. Consequently, changes in the quantiles of y across x cannot be the same for all quantiles. Slope estimates differ across

quantiles since the variance in y changes as a function of x. In this situation, ordinary least squares regression for the mean is commonly modified by incorporating weights in inverse proportion to the variance function (Neter et al. 1996). To improve the estimates of the sampling variation for the estimated mean function, the use of weighted least squares is done. Estimating prediction intervals based on weighted least squares estimates implicitly recognize the unequal rates of change in the quantity of y (Cunia 1987). Generalized linear models offer an alternative way to link changes in the variances of y with changes in the mean based on assuming some specific distributional form in the exponential family for example, Poisson, negative binomial, or gamma (McCullagh and Nelder 1989). The purpose is to provide better estimates of rates of change in the mean of y rather than estimates in the changes in the quantiles of y that must occur when variances are heterogeneous. Estimating prediction interval for a generalized linear model would implicitly recognize that rates of change in the conditional distribution of y given x cannot be the same for all quantiles, and these interval estimates would be linked to and sensitive to violations of the assumed error distribution.

An advantage of using quantile regression, to model heterogeneous variation in response distributions, is that no specification of how variance changes are linked to the mean is required, nor is there any restriction to the exponential family of distributions. Furthermore, one can also detect changes in the shape of the distribution of y across the predictor variables (Koenker and Machado 1999). Complicated changes in central tendency, variance, and shape of distributions are common in statistical analysis. Quantile regression is able to effectively address those statistical problems and

precisely answer the research questions. Regression quantile estimates can be used to construct prediction and tolerance intervals without assuming any parametric error distribution and without specifying how variance heterogeneity is linked to changes in means.

Scharf et al.(1998) described methodologies to quantify the boundaries of scatter diagrams. They tested regression techniques based on least squares and least absolute values models using several independent data sets on prey length and predator length for piscivorous fishes and compared estimated slopes for consistency. They observed that least squares regression techniques were particularly sensitive to outlying y values and irregularities in the distribution of observations, and that they frequently produced inconsistent estimates of slopes for upper and lower bounds. In contrast, quantile regression techniques based on least absolute values models appeared robust to outlying y values and spareness within data sets, while providing consistent estimates of upper and lower bound slopes. They recommended quantile regression as an improvement to currently available techniques used to examine potential ecological relationships dependent upon quantitative information on the boundaries of polygonal relationships.

This chapter contains the results from the application of linear quantile regression, nonlinear quantile regression and piecewise linear quantile regression to our data set to regress the 'Fish IBI' by 'Agricultural Stress Gradient'. Among them we emphasize on piecewise linear quantile regression to estimate the thresholds that are present in our data set. In the previous chapter, we compared those regression techniques in terms of residual sum of squares. We found that linear regression is the least appeal-

ing and piecewise linear regression is the most appealing technique among the three techniques for our particular data set. In this chapter, we are evaluating the efficiency of those techniques in terms of quantiles. So, section 4.2 contains the linear quantile regression, section 4.3 the nonlinear quantile regression and section 4.4 the piecewise linear quantile regression methodologies and the results.

4.2 Linear Quantile Regression

The linear quantile regression is simply the linear regression techniques through different quantiles of the conditional distribution of y given x though the parameter estimation techniques is different. Before defining the quantile linear regression function, it is necessary to define quantile.

Quantile

Any real-valued random variable Z may be characterized by its distribution function $F(z) = P(Z \leq z)$, whereas for any $0 < \tau < 1$,

$$F^{-1}(\tau) = \inf\{z : F(z) \geq \tau\} \tag{4.1}$$

is called the τth quantile of Z. The median, which usually plays the central role, is denoted by $F^{-1}(1/2)$. It is possible to define an optimization problem to estimate quantiles of a random variable Z using linear programming techniques (for details see Koenker 2005). The τth sample quantile, $\hat{\alpha}(\tau)$, could be estimated by solving the objective function

$$\min_{\alpha \varepsilon \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha) \tag{4.2}$$

where ρ_τ is the loss function defined as:

$$\rho_\tau(u) = u(\tau - I(u < 0)) \tag{4.3}$$

for some $\tau\epsilon(0,1)$, and u is the residual or error. The above equation led to specifying the τth conditional quantile function as: $Q_y(\tau|x) = x^T\beta(\tau)$, and to consideration of $\hat{\beta}(\tau)$ by solving

$$\min_{\beta\epsilon\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta) \tag{4.4}$$

This is the main idea to estimate linear quantile regression elaborated by Koenker and Bassett (1978). They also described how to convert this optimization problem into linear programming techniques to solve for the parameters of the model through different quantiles. The techniques of linear programming selects a subset of elements, called the basic solutions, by minimizing the loss function, equation 4.3, for some specific quantile $\tau\epsilon(0,1)$. Quantile regression then interpolates the selected observations. In this section we specified the τth conditional quantile function as:

$$Q_y(\tau|x) = \beta_0(\tau) + \beta_1(\tau)x \tag{4.5}$$

where y denotes the dependent variable 'Fish IBI' and x denotes the independent variable 'Agricultural Stress Gradient'. Hence, the corresponding objective function to estimate for the parameters is:

$$\min_{\beta\epsilon\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \beta_1 x_i) \tag{4.6}$$

We used the "quantreg" package in R for quantile regression, more specifically the statement "rq" for the linear quantile regression.

Results

Table 4.1 contains the parameter estimates (intercept and slope) of the linear quantile regression models. Figure 4.2 contains the parameter plot (intercept and slope) against different quantiles. We can see from the table and from the intercept plot (Figure 4.2 left) that the intercept is increasing with the increase of quantile.

Table 4.1: Estimates of parameters of the linear quantile regression model

τ	$b_0(\tau)$	$b_1(\tau)$	τ	$b_0(\tau)$	$b_1(\tau)$
0.05	27.268	−12.537	0.55	45.671	−7.881
0.10	29.194	−11.248	0.60	47.772	−8.058
0.15	31.236	−9.881	0.65	48.955	−8.868
0.20	34.271	−12.550	0.70	50.715	−10.073
0.25	36.002	−14.062	0.75	50.888	−9.743
0.30	37.532	−15.418	0.80	51.563	−9.898
0.35	40.025	−11.491	0.85	55.328	−12.148
0.40	40.843	−10.968	0.90	55.743	−12.396
0.45	42.399	−9.973	0.95	56.264	−12.708
0.50	44.344	−10.297			

The relationship between intercept and quantile is linear and the direction is positive. For the $50th$ quantile, the estimate of the intercept is 44.344. On the other hand, all of the slope estimates are negative which indicates that with the increase of 'Agriculture Stressor' the 'Fish IBI' decreases. For the $50th$ quantile the estimate of slope is −10.297. But estimates of slope show some oscillation (Figure 4.2 right) with different quantiles.

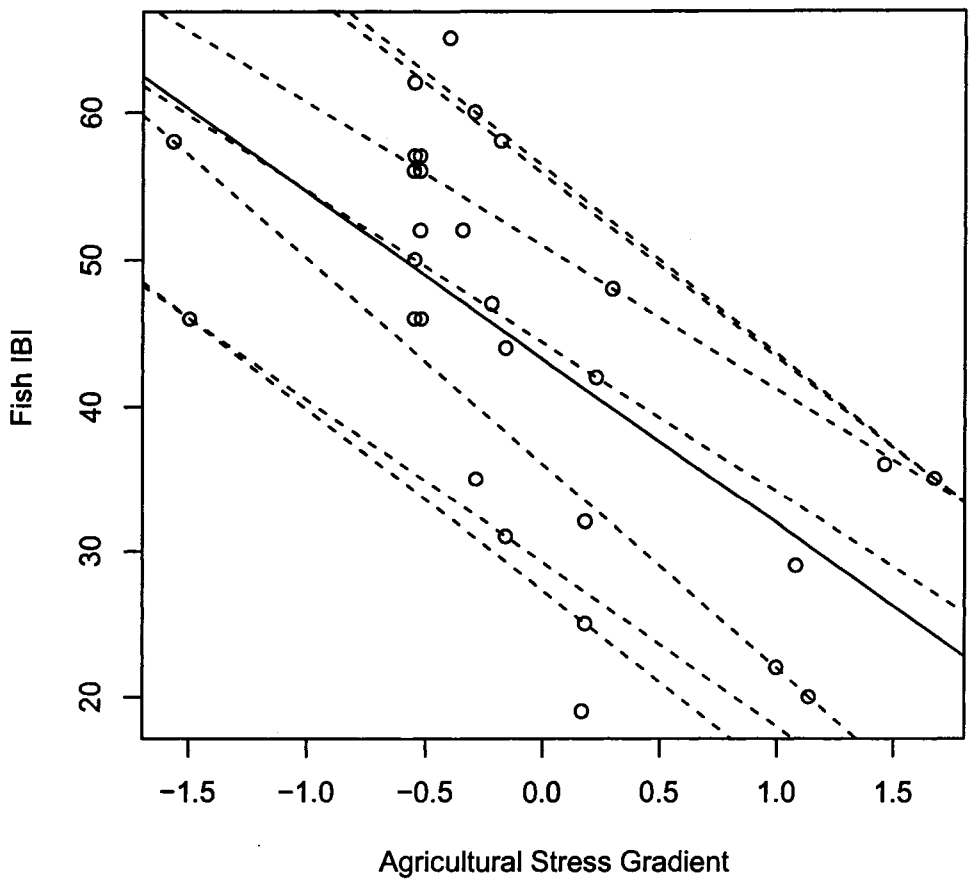

Figure 4.1: Fitted linear quantile regression functions (tau=(0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95)) from bottom to top and the least squares linear regression (solid line) superimposed to the data

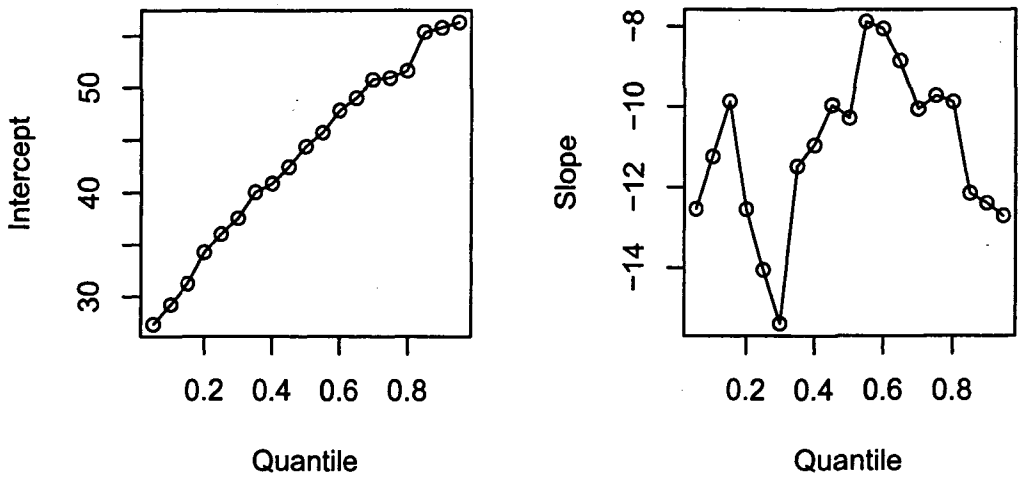

Figure 4.2: Parameter plot of the linear regression model against different quantiles.

The minimum slope is for the $30th$ quantile which is -15.418 and the maximum slope estimate is for the $55th$ quantile which is -7.881. Figure 4.1 show the estimated quantile regression lines and the fitted least squares regression line. The mean regression line is steeper than the median regression line. Figure 4.1 contains the $5th$ and $95th$ quantile regression lines, the bottom one and top one respectively, from which we get an idea about 90% prediction interval of the data.

4.3 Non-Linear Quantile Regression

The asymptotic behavior of the nonlinear quantile regression estimator closely parallels the well-established theory for nonlinear least squares, and so the inference apparatus for nonlinear quantile regression can be adapted directly from existing methods (Koenker 2005). In this section we fit the logistic regression curves through

the different quantiles. Our chosen logistic regression curve is:

$$Q_y(\tau|x) = A(\tau)/(1 + \exp(B(\tau) + C(\tau)x + D(\tau)x^2)) \tag{4.7}$$

We are not going to fit the logistic curve with a cubic term in this model, since the cubic term appeared to be insignificant in fitting the nonlinear curve through the mean of distribution of y, fitted in the previous chapter of this paper. To estimate the parameters of the nonlinear quantile regression model the nonlinear quantile regression technique solve the objective function:

$$\min_{\theta \epsilon \mathbb{R}_p} \sum_{i=1}^{n} \rho_\tau(g_i(\theta)) \tag{4.8}$$

where, we have $g_i(\theta) = (y_i - Q_{y_i}(\tau|x_i))$. Here, ρ_τ is the loss function defined in terms of residual in equation 4.3. The function g_i is differentiable with respect to θ. It is possible to convert that optimization problem into linear programming techniques and estimates the parameter vector θ for different quantiles. For further details see Koenker (2005). We used the statement "nlrq" in the package "quantreg" to fit the non-linear quantile regression models.

Results

The esimates of the parameters of the nonlinear logistic curve through different quantiles are given in Table 4.3. The parameter A increases with the increase of τ but remains stable for quantiles from 0.45 to 0.80 (Figure 4.4 top left). The parameter B decreases almost in linear fashion with the increase of τ (Figure 4.4 top right). The parameter C remains stable for quantiles from 0.05 to quantile 0.75 and after that it increases for the larger quantiles (Figure 4.4 bottom left). Parameter D remains stable for quantiles from 0.05 to 0.75 and it is around -2.0 but after quantile 0.75

it decreases. At the quantile 0.80 the parameter B is minimum, C is maximum and again D is minimum (Table 4.3).

Table 4.2: Estimates of the parameters of the non-linear quantile regression model

τ	$A(\tau)$	$B(\tau)$	$C(\tau)$	$D(\tau)$
0.05	46.003	−0.098	3.049	−2.200
0.10	46.002	−0.136	3.298	−2.336
0.15	50.802	0.062	3.056	−2.133
0.20	49.217	−0.493	2.885	−1.870
0.25	49.171	−0.500	2.921	−1.890
0.30	49.271	−0.560	3.009	−1.919
0.35	51.514	−1.047	3.402	−1.927
0.40	54.433	−0.877	3.212	−1.816
0.45	58.016	−0.674	2.685	−1.514
0.50	58.008	−0.758	2.805	−1.556
0.55	58.001	−0.748	3.375	−2.138
0.60	58.082	−0.739	3.349	−2.176
0.65	58.007	−1.523	2.777	−1.266
0.70	57.716	−1.730	3.732	−1.981
0.75	57.837	−1.720	3.720	−1.975
0.80	58.000	−3.068	5.896	−2.830
0.85	59.016	−2.918	5.689	−2.737
0.90	62.133	−2.538	5.183	−2.509
0.95	62.133	−2.538	5.183	−2.509

For the quantiles 0.90 and 0.95, all the estimates of parameter values are the same, hence the $90th$ and $95th$ quantile regression lines fall close to each other (Figure 4.2). For the $50th$ quantile the parameter values for A, B, C, and D are 58.008, −0.758, 2.805 and −1.556 respectively. Figure 4.3 displays the fitted logistic regression curve with different quantiles. The shape of the right tail of those curves is against our expectation of decreasing 'Fish IBI' with increasing 'Agricultural Stress Gradient'.

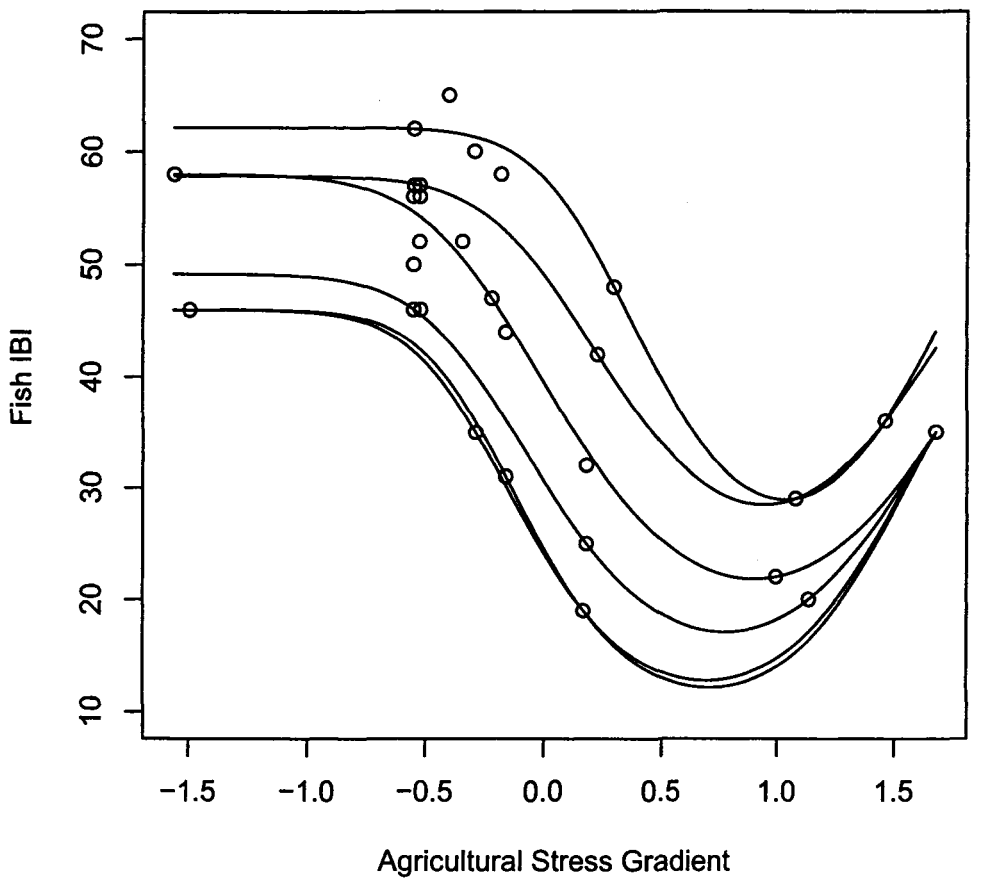

Figure 4.3: Fitted non-linear quantile regression functions (tau=(0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95)) from bottom to top superimposed to the data

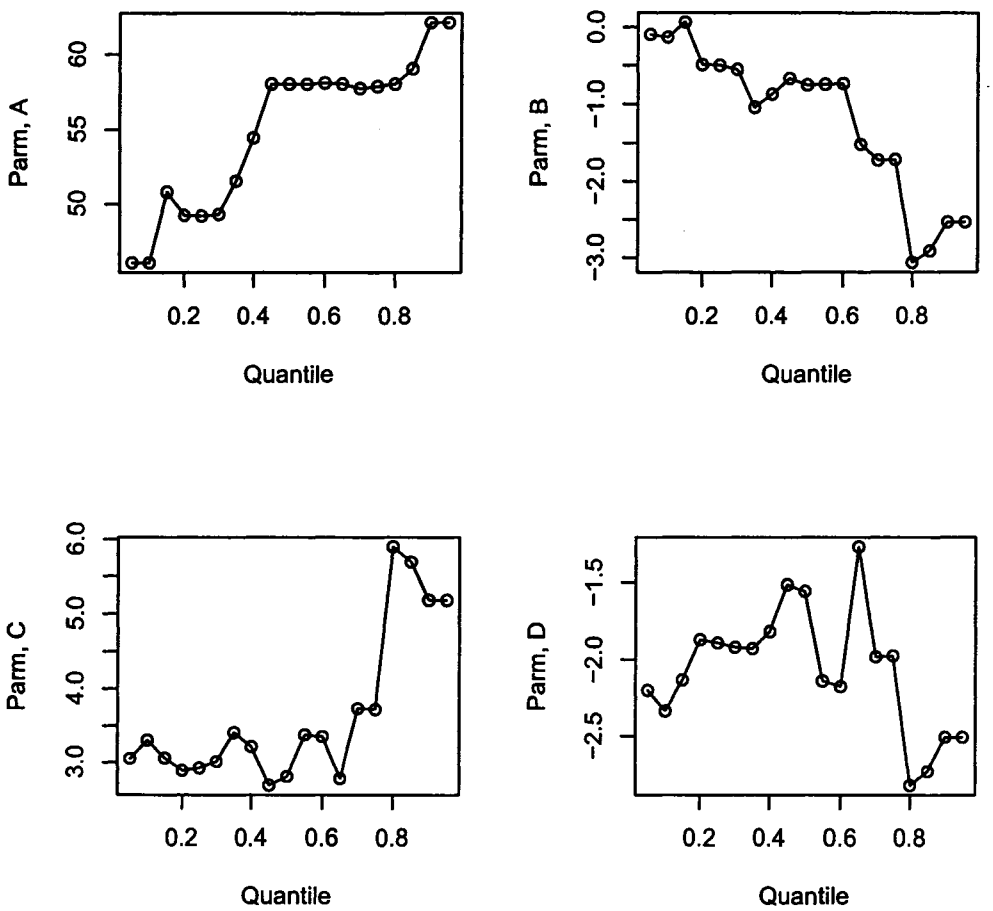

Figure 4.4: Plots of parameters of the non-linear least squares regression against different quantiles.

Those curves fit the data very well except for the right tail. The curves through quantiles 0.05 and 0.95 give us the 90% prediction interval of the data. Those curves indicate that the 'Fish IBI' is stable up to stressor level around -0.5 then it decreases and reaches its minimum at the stressor level around 0.6. There are two weaknesses of the fitted logistic curves through different quantiles. The first one is its right tail after stressor level ≥ 1.0, which is against our expectation. The second one is its inability to give us the threshold estimates. So we propose to overcome these difficulties using piecewise linear quantile regression techniques and applied to our data.

4.4 Piecewise Linear Quantile Regression

Piecewise linear regression model is a broken stick model where two or more linear regression lines join each other at the break points. Those break points could be considered as thresholds. In this section we are going to fit the piecewise linear regression lines through different quantiles. The piecewise linear regression function through the different quantiles is defined as:

$$Q_y(\tau|x) = \begin{cases} \beta_0(\tau) + \beta_1(\tau)x & \text{for } x \leq \alpha_1(\tau) \\ \beta_0(\tau) + \beta_1(\tau)x + \beta_2(\tau)\{x - \alpha_1(\tau)\} & \text{for } \alpha_1(\tau) < x \leq \alpha_2(\tau) \\ \beta_0(\tau) + \beta_1(\tau)x + \beta_2(\tau)\{x - \alpha_1(\tau)\} & \\ \quad +\beta_3(\tau)\{x - \alpha_2(\tau)\} & \text{for } x > \alpha_2(\tau) \end{cases} \tag{4.9}$$

where y is the value for the dependent variable 'Fish IBI', x is the value for the independent variable 'Agricultural Stress Gradient', $\alpha_1(\tau)$ and $\alpha_2(\tau)$ are the breakpoints (the two thresholds) at quantile τ. For parameter estimation the objective function is

the same as for nonlinear quantile regression described in previous section. We used the statement "nlrq" in the package "quantreg" to fit the piecewise linear quantile regression.

Results

Among the three parametric regression techniques fitted in our data, piecewise linear regression came out with our expected results, gaving minimum residual sum of squares and allowed us to estimate the thresholds. We saw that linear and nonlinear quantile regression techniques gave different estimates of the parameter values for different quantiles. This might be due to the sparseness of our data. So, in this section we fitted piecewise linear regression through different quantiles to capture the thresholds in our data. We got different estimates of thresholds for different quantiles (Table 4.3). Figure 4.6 gives the threshold plots against different quantiles. From the left plot in Figure 4.6 of the first threshold against different quantiles and from the Table 4.3, we can see that there is little difference of the first threshold estimates around the value -0.39 but the minimum value is -0.557 which is for the quantile 0.65. For the $50th$ quantile, the estimate of the first threshold parameter is -0.390. From the plot for the second threshold parameter (Figure 4.6 right) and Table 4.3, we see that there is little variation of the second threshold estimates around the value 0.05 but most of the values are around 0.15. Quantile regression based on $85th$ quantile did not give the estimate for the second threshold parameter. It gives a straight line after the first threshold. The maximum value of the second threshold parameter is 0.220 which corresponds to the quantile 0.95. So, from a conservative point of view we took the minimum of the first threshold estimates for different quantiles as our

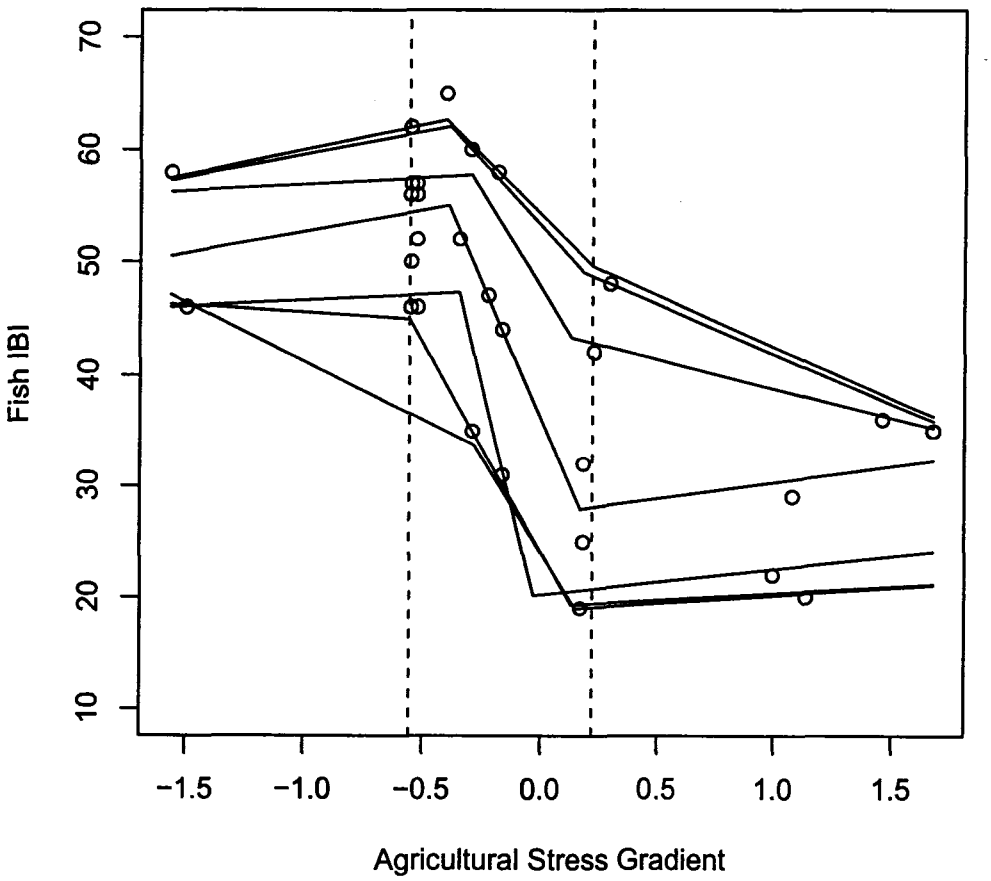

Figure 4.5: Fitted piecewise linear quantile regression functions (tau=(0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95)) from bottom to top superimposed to the data

Table 4.3: Estimates of the parameters of the piecewise linear quantile regression model

τ	$\alpha_1(\tau)$	$\alpha_2(\tau)$	τ	$\alpha_1(\tau)$	$\alpha_2(\tau)$
0.05	−0.281	0.146	0.55	−0.399	0.103
0.10	−0.555	0.128	0.60	−0.399	0.101
0.15	−0.552	0.139	0.65	**-0.557**	0.126
0.20	−0.399	0.032	0.70	−0.294	−0.109
0.25	−0.343	−0.031	0.75	−0.294	0.130
0.30	−0.343	−0.096	0.80	−0.344	0.168
0.35	−0.359	0.168	0.85	−0.547	NA
0.40	−0.364	0.168	0.90	−0.399	0.189
0.45	−0.368	0.168	0.95	−0.399	**0.220**
0.50	−0.390	0.168			

first threshold (−0.557) and the maximum of the second threshold estimates for different quantiles as our second threshold (0.220). We defined the region of agricultural stressor between −0.557 to 0.220 as the "transition" zone. We also defined the region of stressor value greater than 0.220 as the "degraded" zone. Figure 4.5 gives us the three zones in terms of agricultural stressor: the "undegraded" zone, the "transition" zone, and the "degraded" zone. We can see from the figure that 7 of the sites among total selected sites are in "degraded" zone in terms of agricultural stress gradient.

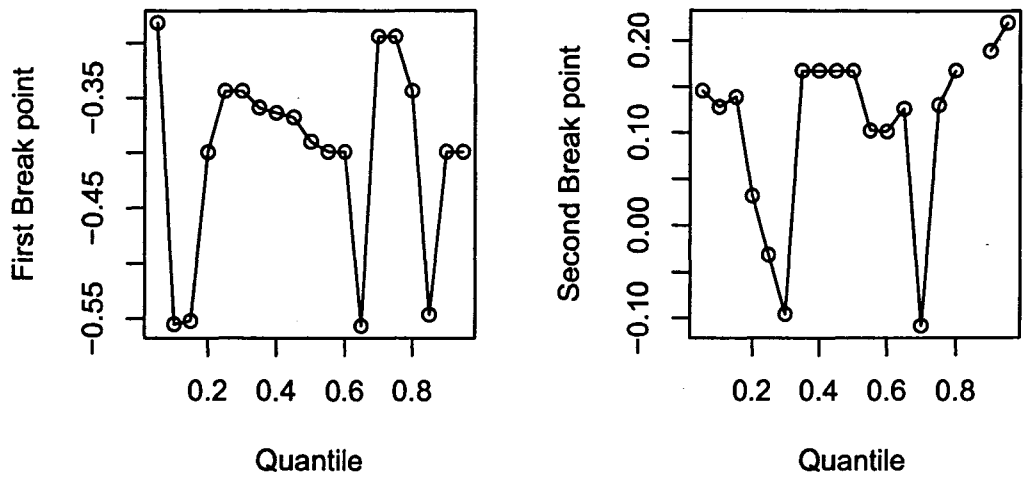

Figure 4.6: Parameter plot of the piecewise linear regression against different quantiles.

4.5 Summary

We fitted linear regression lines and nonlinear logistic curves through different quantiles. The slope parameter in the linear model and the parameters in the nonlinear logistic curve showed fluctuations for different quantiles. The fluctuations among the parameter estimates for different quantiles might be due to sparseness of our data. If there is sparseness in the data, the estimate of breakpoints from piecewise linear model through the mean fitted in chapter 3 might not give the complete picture. So, we calculated the breakpoints through different quantiles and took the minimum of the first break point estimates and the maximum of the second break point estimates to define the "transition" zone. According to our analysis the "transition" zone is between the Agricultural stressor values from −0.557 to 0.220. From Figure 4.5 we

can see that seven sites are in the "degraded" zone and three sites are in the boundary of "transition" zone and "degraded" zone. We also see that the Fish IBIs are very low among the sites in "degraded" zone. One of the sites in the "degraded" zone is showing moderate Fish IBI. This may be due to the fact that other types of stressor might have less impact on this particular site.

Chapter 5

Summary and Conclusion

With the advancement of the science and technology, there is an increase of agricultural activities in the US Great Lakes coastal region that causes the degradation of nature and loss of aquatic habitat of fish in the US Great Lakes coastal margins. The fish communities across lakes and rivers serve as a good indicators of stress (Karr 1981). The fish IBI reflects the fish community response to relative degree of disturbance at a particular site. Uzarski et al.(2005) and Bhagat (2005) calculated fish IBI scores at 17 and 13 sites respectively that were dominated by *Scirpus* vegetation across the US Great Lakes coastal margins. We got agricultural stress gradient for those 30 sites developed by Danz et al. (2005). In this paper, we have applied several statistical techniques to analyze the fish data that contains only two variables 'Agricultural Stress Gradient' and 'Fish IBI'. Bhagat (2005) established that the fish IBI responded negatively to agricultural stress gradient but not in a linear fashion. Instead it exhibited some threshold effects. Our goal of this paper was to estimate those threshold effects. We observed that there were two threshold points, one at the

agricultural stress level around -0.50 where the fish IBI starts falling suddenly and the other one is at the stress level around 0.20 where it reaches its minimum IBI level. We applied several statistical techniques to arive at our objectives. First, four statistical tests were used to test the null hypothesis of bivariate randomness among 'Fish IBI' and 'Agricultural Stress Gradient'. Our chosen tests were based on the mean nearest-neighbour distance, cumulative R-spectrum, reduced second-order moment function and the bivariate Cramer-von Mises statistic. The reason for choosing those tests was that they had been found to be powerful against Poisson clustered processes which was close to our specific alternative. We did a simulation study to check the performance of these tests by generating bivariate normal random sample with the specific means and variances for our data and for different negative correlation coefficients since the two variables were negatively correlated. We applied these tests to our data to verify whether they are bivariate random. We used the 4% level of significance since Zimmerman (1993) tabulated the value for the distribution of the bivariate Cramer-von Mises statistic for 4% significance level.

Regression techniques to regress 'Fish IBI' by 'Agricultural Stress Gradient' were applied in chapter 3. We used linear, non-linear and piecewise linear regression models and compared them in terms of residual sum of squares. Some nonparametric techniques were applied also and bootstrap techniques were adopted to figure out the band of the regression line. A piecewise linear regression model with three linear regimes was incorporated to estimate two threshold effects. Finally, we employed quantile regression techniques to identify the prediction band of the data and to estimate the threshold parameters. To accomplish these analysis, sometimes we wrote our own

programs and sometimes we used some packages in R and $S-Plus$.

We applied those above mentioned statistical techniques to our fish and stressor data. The null hypothesis of bivariate randomness was rejected by all of the four chosen tests, this indicates that there is association, more specifically there are some patterns between 'Fish IBI' and 'Agricultural Stress Gradient'. From the nonparametric regression techniques, specifically from LOESS, we got the idea about the position of the two breakpoints, and piecewise linear regression allowed us to estimate the threshold parameters. Estimates of the breakpoints, i.e., threshold parameters, from the piecewise linear regression lines were -0.397 and 0.087.

We used quantile regression techniques for different regression models. From the piecewise linear quantile regression model, we got a set of estimates for the first and second threshold parameters. We took the minimum of the estimates of the first threshold parameter (-0.557) and the maximum of the estimates of the second threshold parameter (0.220) to define the "transition" zone. We defined the zone that was beyond the maximum of the second threshold estimates (0.220) as the "degraded" zone. Figure 4.5 displayed that seven of the sites among the thirty selected sites in the US Great Lakes coastal margin were in "degraded" zone. The sites were Rapid River, Venderbilt Park, Wigwam Bay, Wildfowl Bay, Menominee River, L. Pickerel Creek, and Black River. Three of the sites were in the boundary between the "transition" zone and the "degraded" zone. The sites were Bradleyvile Rd., Pinconning[1], and Pinconning[2]. One important point to note here is that only two sites among the thirty selected sites were in the "undegraded" zone and most of the sites were in the "transition" zone. From these findings, we could say that the natural habitat for fish

in the US Great Lakes coastal margins has been degraded for agricultural activities. So, care should be taken to get rid of the degradation of natural habitat. Finally, we are recommending LOESS, piecewise linear regression and quantile regression techniques to model data with potential breakpoints and to estimate the threshold(s) as well.

Limitations and Scope of Further Research

The data set in this paper is created by merging two data sets. So, one might question the randomness of the observations in our data. This is the major limitation of this study. Secondly, this study employed only thirty sites in the US Great Lakes coastal margins with dominant *Scirpus* vegetation. So, an extension of this research work is to incorporate more sites from the US and Canada Great Lakes coastal margins and calculating the different stressors for those regions and study their relationships with fish IBI and to find the threshold effect if there is any.

Recommendation for Further Study

(1) Determine agricultural stress gradient across the US and Canada Great Lakes Coastline. (2) Define three strata in terms of the agricultural stress gradient according to the findings of this research. (3) Randomly select sites from the different strata. If possible, use stratified random sampling with proportional allocation. (4) Select at least twenty sites or more from each stratum depending on the cost of your study and the precision you want. (5) Calculate fish IBI for those randomly selected sites. (6) Use LOESS, piecewise linear regression and quantile regression techniques to model the data and to estimate the threshold effects.

Bibliography

[1] Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), Second international symposium on information theory (pp. 267 – 281). Budapest: Academiai Kiado.

[2] Andren, H. 1994. Effects of habitat fragmentation on birds and mammals in landscapes with different proportions of suitable habitat-a review. *Oikos* **71** : 355–366.

[3] Bacon, D.W., and Watts, D.J. 1971. Estimating the transition between two intersecting straight lines. *Biometrika* **58** : 525 – 534.

[4] Bartlett, M.S. 1964. The spectral analysis of two-dimensional point processes. *Biometrica* **51** : 299 – 311.

[5] Bhagat, Y. 2005. Fish Indicators of Anthropogenic Stress at Great Lakes Coastal Margins: Multimetric and Multivariate Approach. Dissertation. University of Windsor, Windsor, Ontario, Canada.

[6] Bowman, A.W. and Azzalini, A. 1997. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-plus Illustrations.* Oxford, Oxford University Press.

[7] Brazner, J.C., and Beals, E.W. 1997. Patterns in fish assemblages from coastal wetland and beach habitats in Green Bay, Lake Michigan: a multivariate analysis of abiotic and biotic forcing factors. *Canadian Journal of Fisheries and Aquatic Sciences* **54** : 1743 – 1761.

[8] Bryce, S.A., Larsen, D.P., Hughes, R.M., and Kaufmann, P. 1999. Assessing relative risks to aquatic ecosystems: a mid-Appalachian case study. *Journal of the American Water Resources Association* **35**(1) : 23 – 36.

[9] Burton, T.M., Uzarski, D.G., Gathman, J.P., Genet, J.A., Keas, B.E. and Stricker, C.A. 1999. Development of a preliminary invertebrate index of biotic integrity for Lake Huron coastal wetlands. *Wetlands* **19** : 869 – 882.

[10] Cade, B.S., J.W. Terrell, and R.L. Schroeder. 1999. Estimating effects of limiting factors with regression quantiles. *Ecology* **80** : 311 – 323.

[11] Cade, B.S., and Noon, B.R. 2003. A gentle introduction to quantile regression for ecologists. *Front Ecol Environ*, **1**(8) : 412 – 420.

[12] Chiu, D.S. 2002. Bent-cable regression for assessing abruptness of change. Dissertation. Simon Fraser University, Vancouver, British Columbia, Canada.

[13] Clark, P.J., and Evans, F.C. 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* **35** : 23 – 30.

[14] Cleverland, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74** : 829 – 836.

[15] Crosbie, B., and Chow-Fraser, P. 1999. Percentage land use in the watershed determines the water and sediment quality of 22 marshes in the Great Lakes basin. *Canadian Journal of Fisheries and Aquatic Science* **56** : 1781 – 1791.

[16] Cucala, L. and Thomas-Agnan, C. 2004. Seminaire Statistique Mathematique et applications.

[17] Cunia, T. 1987. Construction of tree biomass tables by linear regression techniques. In: Estimating tree biomass regressions and their error. USDA Forest Service, General Technical Report NE-GTR-117, p 27 – 36.

[18] Danz, N.P., Regal, R.R., Niemi, G.J., Brady, V.J., Hollenhorst, T., Johnson, L.B., Host, G.E., Hanowski, J.M., Johnston, C.A., Brown, T., Kingson, J., and Kelly, J.R. 2005. Environmentally stratified sampling design for the development of great lakes environmental indicators. *Environmental Monitoring and Assessment* **102** : 41 – 65.

[19] Diggle, P.J., Besag, J., and Gleaves, J.T. 1976. Statistical analysis of spatial point patterns by means of distance methods. *Biometrics* **32** : 659 – 667.

[20] Diggle, P.J. 1977. The detection of random heterogeneity in plant populations. *Biometrics* **33**(2) : 390 – 394.

[21] Diggle, P.J. 1979. On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics* **35**(2) : 87 – 101.

[22] Donnelly, K. 1978. Simulations to determine the variance and edge-effect of total nearest-neighbor distance. *In Simulation Studies in Archaeology* Hodder I. (ed), Cambridge University Press, pp. 91 – 95.

[23] Draper, N.R., and Smith, H. 1998. *Applied Regression Analysis.* 3rd edition. Wiley series in probability and statistics. John Wiley & Sons. New York. USA.

[24] Emery, E.B., Simon, T.P., McCormick, F.H., Angermeier, P.L., Deshion, J.E., Yoder, C.O., Sanders, R.E., Pearson, W.D., Hickman, G.D., Reash, R.J., and Thomas, J.A. 2003. Developmnet of a multimetric index for assessing the biological condition of the Ohio river. *Transactions of the American Fisheries Society* **132** : 791 – 808.

[25] Fahrig, L. 2001. How much habitat is enough? *Biological Conservation* **100** : 65 – 74.

[26] Fausch, K.D., Karr, J.R., and Yant, P.R. 1984. Regional Application of an Index of Biotic Integrity Based on Stream Fish Communities. *Transactions of the American Fisheries Society* **113** : 39 – 55.

[27] Feder, P. 1975. The log likelihood ratio in segmented regression. *Annals of Statistics*, **3** : 84 – 97.

[28] Frey, D.G. 1977. Biological integrity of water-an historic approach. Pages 127 – 140 in R.K. Ballantine and L.J.Guarraia, editors. The Integrity of Water. Proceedings of a Symposium, U.S. Environmental Protection Agency, Washington, D.C.

[29] Galatowitsch, S.M., Anderson, N.O., and Ascher, P.D. 1999. Invasiveness in wetland plants in temperate North America. *Wetlands* **19** : 733 – 755.

[30] Gates, J.E., and Mosher, J.A. 1981. A functional approach to estimating habitat edge width for birds. *American Midland Naturalist* **105** : 189 – 192.

[31] Grabas, G., Pemanen, A.D., Galloway, M., and Holmes, K. 2004. Durham Region Coastal Wetland Monitoring Project: Year 2 Technical Report. ECB-OR. Environment Canada and Central Lake Ontario Conservation Authority, Downsview, Ontarion, Canada.

[32] Hardle, W., and Bowman, A.W. 1988. Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *Journal of the American Statistical Association*, **83** : 102 – 110.

[33] Host, G.E., Schuldt, J., Ciborowski, J.J.H., Johnson, L.B., Hollenhorst, T., and Richards, C. 2005. Use of GIS and remotely sensed data for a priori identification of reference areas for Great Lakes coastal ecosystems. *International Journal of remote sensing* **26** : 1 – 18.

[34] Hughes, R.M., and Larson, D.P. 1988. Ecoregions: an approach to surface water protection. *Journal of the Water Pollution Control Federation* **60** : 486 – 493.

[35] Karr, J.R. 1981. Assessment of Biotic Integrity Using Fish Communities. *Fisheries* **6** : 21 – 27.

[36] Koenker, R., and Bassatt, G. 1978. Regression quantiles. *Econometrica* **46** : 33 – 50.

[37] Koenker, R., and Machado, J.A.F. 1999. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* **94** : 1296 – 1310.

[38] Koenker, R. 2005. Quantile Regression. Cambridge University Press, New York, USA.

[39] Kullback, S., and Leibler, R.A. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, **22** : 79 – 86.

[40] Lyons, J. and Wang, L., and Simonson, T.D. 1996. Development and validation of an index of biotic integrity for coldwater streams in Wisconsin. *North American Journal of Fisheries Management* **16** : 241 – 256.

[41] McCullagh, P., and Nelder, J.A. 1989. *Generalized linear models.* New York: Chapman and Hall.

[42] Mosteller, F., and Tukey, J.W. 1977. *Data analysis and regression.* New York: Addison-Wiley.

[43] Mugglestone, M.A. 1990. *Spectral Analysis of Spatial Processes.* Ph.D. thesis, University of Edinburgh.

[44] Mugglestone, M.A., and Renshaw, E. 1996*a*. A practical guide to the spectral analysis of spatial point processes. *Computational Statistics and Data Analysis* **21** : 43 – 65.

[45] Mugglestone, M.A., and Renshaw, E. 1996*b*. The exploratory analysis of bivariate spatial point patterns using cross-spectra. *Environmetrics* **7** : 361 – 377.

[46] Mugglestone, M.A., and Renshaw, E. 2001. Spectral tests of randomness for spatial point patterns. *Environmental and Ecological Statistics* **8** : 237 – 251.

[47] Mundahl, N.D., and Simon, T.P. 1999. Development and application of an index of biotic integrity for coldwater streams of the upper midwestern Unites States. Pages 383 – 415 in T.P. Simon, editor. Assessing the Sustainability and Biological Integrity of Water Resources Using Fish Communities. CRC Press, Boca Raton, Florida.

[48] Myers, R.H. 1990. *Classical and modern regression with applications.* Second edition. Duxbury Press, Belmont, California, USA.

[49] Nadaraya, E.A. 1964. On estimating regression. *Theory of Probability Applied,* **10** : 186 – 190.

[50] Neter, J.M., Kutner, H., Nachtsheim, C.J., and Wasserman, W. 1996. *Applied linear statistical models.* Chicago, IL: Irwin.

[51] Plafkin, J.L., Barbour, M.T., Porter, K.D., Gross, S.K., and Hughes, R.M. 1989. Rapid Bioassessment Protocols for Use in Streams and Rivers. Office of Water Regulations and Standards, U.S. Environmental Protection Agency, Washington, D.C.

[52] Prayag, V.R, and Deshmukh, S.R. 2000. Testing randomness of spatial pattern using Eberhardt's index. *Environmetrics* **11** : 571 – 582.

[53] Renato, A. 1994. Testing spatial randomness by means of angles. *Biometrics* **50** : 531 – 537.

[54] Rencher, A.C. 1995. Methods of Multivariate Analysis. John Wiley & Sons, New York, NY.

[55] Renshaw, E., and Ford, E.D. 1983. The interpretation of process from pattern using two-dimensional spectral analysis: methods and problems of interpretation. *Applied Statistics* **32** : 51 – 63.

[56] Renshaw, E., and Ford, E.D. 1984. The description of spatial pattern using two-dimensional spectral analysis. *Vegetatio* **56** : 75 – 85.

[57] Ripley, B.D. 1976. The second-order analysis of stationary point processes. *Journal of Applied Probability* **13** : 255 – 266.

[58] Ripley, B.D. 1979. Tests of "randomness" for spatial point patterns. *Journal of the Royal Statistical Society, Series B* **41** : 368 – 374.

[59] Scharf, F.S., Juanes, F., and Sutherland, M. 1998. Inferring ecological relationships from the edges of scatter diagrams: Comparison of regression techniques. *Ecology* **79**(2) : 448 – 460.

[60] Seber, G.A.F., and Wild, C.J. 1989. *Nonlinear regression.* John Wiley and Sons, New York, USA.

[61] Simon, T.P. 1991. Development of Ecoregion Expectations for the Index of Biotic Integrity. I. Central Corn Belt Plain. U.S. Environmental Protection Agency, Region 5, Chicago, Illinois.

[62] Terrell, J.W., Cade, B.S., Carpenter, J., and Thompson, J.M. 1996. Modeling stream fish habitat limitations from wedged-shaped patterns of variation in standing stock. *Trans Am Fish Soc* **125** : 104 – 117.

[63] Tishler, A., and Zang, I. 1981. A new maximum-likelihood algorithm for piecewise regression. *Journal of the American Statistical Association* **76** : 980 – 987.

[64] Toms, J.D., and Lesperance, M.L. 2003. Piecewise regression: A tool for identifying ecological thresholds. *Ecology* **84** : 2034 – 2041.

[65] Tran, L.T., Knight, C.G., O'Neill, R.V., Smith, E.R., and O'Connell, M. 2003. Self-organizing maps for integrated environmental assessment of the mid-Atlantic region. *Environmental Management* **31**(6) : 822 – 825.

[66] Tran, L.T., Knight, C.G., O'Neill, R.V., and Smith, E.R. 2004. Integrated environmental assessment of the mid-Atlantic region with analytical network process. *Environmental Monitoring and Assessment* **94** : 263 – 277.

[67] Uzarski, D.G., Burton, T.M., Cooper, M.J., Ingram, J.W., and Timmermans, S. 2005. Fish habitat use within and across wetland classes in coastal wetlands of the five Great Lakes: development of a fish-based index of biotic integrity. *Journal of Great Lakes Research* **31** : 171 – 187.

[68] Wales, B.A. 1972. Vegetation analysis of north and south edges in a mature oak-hickory forest. *Ecological Monographs* **42** : 451 – 471.

[69] Watson, G.S. 1964. Smooth regression analysis. *Sankhya*, Ser. A, **26** : 259 – 372.

[70] Watts, D.J., and Bacon, D.W. 1974. Using a hyperbola as a transition model to fit two-regime straight-line data. *Technometrics* **16** : 369 – 373.

[71] Zimmerman, D.L. 1993. A bivariate Cramer-von Mises type of test for spatial randomness. *Applied Statistics* **42** : 43 – 54.

[72] Zimmerman, D.L. 1994. On the limiting distribution of and critical values for an origin-invariant bivariate Cramer-von Mises-type statistic. *Statistics and Probability Letters* **20** : 189 – 195.

Appendix

Table 1: Preliminary fish-based index of biotic integrity metrics for Great Lakes coastal wetlands.

Scirpus Zone:
1. Mean catch per net-night: < 10 score=0 $10 - 30$ score= 3 > 30 score=5
2. Total richness: < 5 score=0 5 to < 10 score=3 10 to 14 score =5 > 14 score=7
3. Percent non-native richness: $> 12\%$ score=0 7 to 12% score=3 $< 7\%$ score=5
4. Percent omnivore abundance: $> 70\%$ score=0 50 to 70% score=3 $< 50\%$ score=5
5. Percent piscivore richness: $< 15\%$ score=0 15 to 25% score=3 $> 25\%$ score=5
6. Percent insectivore abundance: $< 20\%$ score=0 $20 - 30\%$ score=3 $> 30\%$ score=5
7. Percent insectivorous Cyprinidae abundance: $< 1\%$ score=0 $1 - 2\%$ score=3 $> 2\%$ score=5
8. Percent carnivore (insectivore+piscivore+zooplanktivore) richness: $< 60\%$ score=0 $60 - 70\%$ score=3 $> 70\%$ score=5
9. White sucker (Catostomus commersoni) mean abundance per net-night: 0 score=0 > 0 to 0.4 score=3 > 0.4 score=5
10. Black bullhead (Ictalusrus melas) mean catch per net-night: 0 score=0 > 0 to 3 score=3 > 3 score=5
11. Rock bass (Ambloplites rupestries) mean catch per net-night: 0 score=0 > 0 to 4 score=3 > 4 score=5
12. Alewife (Alosa psuedoharengus) mean catch per net-night: > 11 score=0 1 to 11 score=3 < 1 score=5
13. Smallmouth bass (Micropterus dolomieu) mean catch per net-night: 0 score=0 > 0 to 5 score=3 > 5 score=5
14. Pugnose shiner(Notropis anogenus) mean catch per net-night: 0 score=0 > 0 to 5 score=3 > 5 score=5

Table 2: Site locations and IBI scores for Scirpus dominant sites sampled through the GLEI project and by Uzarski et al. (2005)

Site Name	Latit	Longit	Lake	F. IBI	Ag Chem	Project
Big Fishdam	45.893	−86.585	Michigan	65	−0.39931	Uzarski
Bradleyvile Rd.	43.622	−83.635	Huron	19	0.16783	Uzarski
Cedarvile	45.997	−84.363	Huron	57	−0.54924	Uzarski
Escanaba	45.818	−87.052	Michigan	47	−0.22034	Uzarski
Garden Bay	45.997	−86.573	Michigan	52	−0.34341	Uzarski
Hessel Bay	46.005	−84.434	Huron	56	−0.5247	Uzarski
Hill Island	45.982	−84.317	Huron	57	−0.52569	Uzarski
Mackinac Bay	46.001	−84.409	Huron	52	−0.5247	Uzarski
Moscoe Channel	45.992	−84.314	Huron	62	−0.54924	Uzarski
Ogontz Bay	45.832	−86.782	Michigan	60	−0.29399	Uzarski
Pinconning1	43.859	−83.913	Huron	25	0.18094	Uzarski
Rapid River	45.914	−86.966	Michigan	48	0.29638	Uzarski
Shephads Bay	45.984	−84.364	Huron	50	−0.54924	Uzarski
St. Ignace	45.845	−84.739	Michigan	56	−0.55164	Uzarski
Vanderbilt Park	43.601	−83.661	Huron	22	0.99574	Uzarski
Wigwam Bay	43.961	−83.859	Huron	35	1.67332	Uzarski
Wildfowl Bay	43.802	−83.463	Huron	20	1.13714	Uzarski
Middle River	46.68	−91.82	Superior	58	−0.18207	GLEI
Clover	46.88	−91.17	Superior	35	−0.28764	GLEI
McKay Creek	45.99	−84.34	Huron	46	−0.54924	GLEI
Menominee River	45.09	−87.59	Michigan	42	0.22765	GLEI
Pinconning2	43.85	−83.92	Huron	32	0.18094	GLEI
L. Pickerel Creek	41.46	−82.79	Erie	36	1.4608	GLEI
Sterling Creek	43.35	−76.68	Ontario	44	−0.16053	GLEI
Deer Tick Creek	43.61	−76.19	Ontario	58	−1.56431	GLEI
Skinner Creek	43.67	−76.18	Ontario	46	−1.49408	GLEI
Black River	43.99	−76.06	Ontario	29	1.07982	GLEI
Sterling Creek	43.35	−76.68	Ontario	31	−0.16053	GLEI
Bear Lake	45.97	−84.16	Huron	46	−0.51999	GLEI
McKay Creek	45.99	−84.34	Huron	46	−0.54924	GLEI

R and S-plus Programs

 # Function Used in Simulation for the four tests

"brantest" < − function ()

{ # nnn gives no of iteration

nnn < − 1

install.packages("spatstat")

library(spatstat)

install.packages("MASS")

library(MASS)

P < − rep(0 , 4)

repeat{

Generation of Bivariate random sample

covxy < − -0.80* sqrt (0.567825)* sqrt (174.2483)

Sigma < − matrix(c(0.567825, covxy, covxy, 174.2483),ncol=2)

data2 < − mvrnorm(35, c(-0.1032857,44.4), Sigma)

x2 < − data2[,1]

y2 < − data2[,2]

mix2 < − min(x2)

miy2 < − min(y2)

ifelse((mix2 ≤ 0), x1 < − x2 + abs(mix2), x1 < − x2)

mx1 < − max(x1)

ifelse((miy2 ≤ 0), y1 < − y2 + abs(miy2), y1 < − y2)

my1 < − max(y1)

```
ifelse((mx1 > 1),x < − x1/mx1, x < − x1)

ifelse((my1 > 1), y < − y1/my1, y < − y1)

# Generation of Uniform random sample

# x < − runif(100,0,1)

# y < − runif(100,0,1)

data < − cbind(x,y)

# No of Data points

n < − length(x)

# Test Based on the Cumulative R-spectrum

if(n < 12){r=1}

if(n >= 12 && n<28){r=2}

if(n>=28 && n<48){r=3}

if(n>=48 && n<80){r=4}

if(n>=80){r=5}

count < − 0

CR < − 0

f< −matrix(rep(0,60), ncol=10)

for (p in 0:5){

i< −p+1

for (q in -5:4){

j< −q+6

r1< − sqrt(p∧2 + q∧2)

if((p ! = 0 || q ! = 0) && (p ! = 0 || q <=0) && (r1 <= r)){
```

```
f[i,j]< −(sum(cos(n*((2*pi*p/n)*x+(2*pi*q/n)*y))))∧ 2
+ (sum(sin(n*((2*pi*p/n)*x+(2*pi*q/n)*y))))∧ 2
if(f[i,j] > 0 ){
count< −count+1
CR< −CR+f[i,j] } } } }
CRF< −CR/count
# Test statistic
T < −CRF/n
df< −2*count
# Lower and Upper critical region
L< −qchisq(0.02,df)/df
U< −qchisq(0.98,df)/df
ifelse((T <= L || T >= U),P[1]< −P[1]+1,P[1]< −P[1]+0)
# Test Based on the Mean Nearest-neighbor Distance
dm< −pairdist(x,y)
d< −rep(0,n)
for(i in 1:n){
d[i]< −min(dm[i,][dm[i,]!=0]) }
dbar< −mean(d)
mu < − 0.50/sqrt(n) + 0.206/n + 0.164/(n∧(3/2))
sig < − 0.070/(n ∧ 2) + 0.148/(n ∧ (5/2))
# Test Statistic
z< −(dbar-mu)/(sqrt(sig))
```

```
ifelse(z <= qnorm(0.02) || z >= qnorm(0.98), P[2] < - P[2]+1, P[2]< -P[2]+0)

# Test Based on the Reduced second-order Moment Function

if(nnn == 1){

LM < - rep(0,1000) }

m < - as.ppp(data, c(0,1,0,1))

t0 < - 1.25/sqrt(n)

# Calling the function LRM.

LM[1] < - LRM(m,t0)

if(nnn == 1){

for(i in 2:1000){

m1 < - as.ppp(matrix(c(runif(n,0,1),runif(n,0,1)),ncol=2),c(0,1,0,1))

LM[i] < - LRM(m1,t0) } }

RLM < - rank(LM)

# P-value of this test

pval < - 1 - RLM[1]/1000

ifelse(pval <= 0.04, P[3] < - P[3]+1, P[3] < - P[3]+0)

# Test Based on the Bivariate Cramer-von Mises Statistic

A < - matrix(rep(0,n*n),ncol=n)

B < - rep(0,n)

for(i in 1:n){

B[i] < - (x[i]^2 - x[i] - 0.5)* (y[i]^2 - y[i] - 0.5)

for(j in 1:n){

A[i,j] < - (1-abs(x[i]-x[j]))*(1-abs(y[i]-y[j])) } }
```

```
A1 < − sum(A)

B1 < − sum(B)

# Test statistic

wbar < − A1/(4*n) - B1/2 + n/9

ifelse((wbar <= 0.0487 || wbar >= 0.3425), P[4]< −P[4]+1, P[4]< −P[4]+0)

# Controls the total number of simulated sample

if(nnn >= 1000) break

nnn < − nnn+1

}

cat("No of iteration =")

cat(nnn)

cat("\ n")

powr < − P/nnn

method < − c("Cumulative R-spectrum:","Mean Nearest-neighbor Distance:","Reduced

second-order Moment Fuction:","Bivariate Cramer-von Mises Statistic:")

power < − list(Method = method, Power = powr)

return(power) }
```

 # Supporting Program: LRM

```
function (data,t0) {

# Estimate Ripley's K-function

kt < − Kest(data,correction="Ripley")

k < − kt$iso

t < − kt$r
```

```
l < - abs(sqrt(k/pi)-t)
# Estimate Lm
lm < - max(l[t <= t0])
return(lm) }

# Reading the data file from the location "c:/data.txt".
data< -matrix(scan("c:/data.txt"),ncol=2,byrow=TRUE)
data < - data[sort.list(data[,1]),]
x < - data[,1]
y < - data[,2]
# Making Postscript file.
postscript("j:/Major Paper/plot1.ps",horizontal=FALSE, width=6,height=6)
plot(x,y,xlab="Agricultural Stress Gradient",ylab="Fish IBI")
dev.off()
# Linear regression fit
plot(x,y,xlab="Agricultural Stress Gradient",ylab="Fish IBI")
pm< -glm(y~x,family=gaussian)
lines(x,pm$fitted.values, lty=1)
# Logistic regression function
fm< -nls (y~A/(1+exp (B+C*x+D*x^2)) ,start=list (A=55,B=-0.98,
C= 2.98,D=-1.72))
summary(fm)
AIC(fm)
```

```
fm1< −nls(y∼A/(1+exp (B+C*x+D*x∧2+E*x∧3)) ,start=list(A=54,
B=-0.98, C=3.31,D=-2.63,E=0.45))
# Plot of linear, and nonlinear Logistic regression
fm< −nls(y∼A/(1+exp(B+C*x+ D*x∧2)),start=list
(A=55 ,B=-0.98,C=2.98,D=-1.72))
fm1< −nls(y∼A/(1+exp(B+C*x+ D*x∧2+E*x∧3)) ,start=list
(A=54,B=-0.98, C=3.31,D=-2.63,E=0.45))
tt < − seq(-1.564, 1.674, length = 100000)
pm < − glm(y∼x,family=gaussian) lines(x,pm$fitted.values,lty=1)
lines(tt, predict(fm, list(x = tt)),lty=2)
lines(tt, predict(fm1, list(x = tt)),lty=3)
# Bandwidth selection for Kernel smooth (box).
bd < − seq(0.3, 1.0, 0.001)
RSS < − rep (0,length(bd))
for (i in 1:length(bd)){
model < − ksmooth(x,y,kernel="box", bandwidth=bd[i], x.points=x)
RSS[i] < − sum((y-model$y)∧2) }
# cbind(bd,RSS)
plot(bd, RSS, xlab="bandwidth", ylab="Residual SS")
# Plotting the Kernel smooth.
plot(x,y,xlab="Agricultural Stress Gradient",ylab="Fish IBI")
lines(ksmooth(x,y,kernel="box", bandwidth=0.47, x.points=x),lty=3)
# fitting local linear regression
```

```
install.packages("sm")

library(sm)

sm.regression(x,y,h=0.4646703,add=T,lty=2)

# fitting loess.

lines(loess.smooth(x,y),lty=1)

legend(0.35,62,c("Local Mean","Local Linear","loess"),lty=c(3:1))

# 500 Bootstrap lines of the local mean regression

plot(x,y,xlab="Agricultural Stress Gradient",ylab="Fish IBI",ylim=c(10,70))

model < - ksmooth(x,y,kernel="box",bandwidth=0.47, x.points=x)

mhat < - model$y

r < - y-mhat

rr < - r-mean(r)

for(i in 1:500)

lines(ksmooth(x,mhat+ sample(rr,length(x),replace=TRUE),

bandwidth=0.47, x.points=x),lty=3,col=2)

# 500 Bootstrap lines of the local linear regression

plot(x,y,xlab="Agricultural Stress Gradient",ylab="Fish IBI",ylim=c(10,70))

smodel < - sm.regression(x,y,h=0.4646703,eval.points=x,display="none")

mhat < - smodel$estimate

r < - y-mhat

rr < - r-mean(r)

for(i in 1:500)

sm.regression(x, mhat+sample(rr,replace=T),h=0.4646703, add=T , col=2 ,lty=3)
```

```
# 500 Bootstrap lines of the of loess
plot(x,y,xlab="Agricultural Stress Gradient",ylab="Fish IBI",ylim=c(10,70))
model < − loess(y x)
mhat < − fitted(model)
r < − y-mhat
rr < − r-mean(r)
for(i in 1:500)
lines(loess.smooth(x,mhat+ sample(rr,length(x),replace=TRUE)),lty=3,col=2)
# Fit of Piecewise Linear Regression
fit21< −nls(y ∼ cbind(1,x,ifelse((x>br1),x-br1,0),
ifelse(x>br2,x-br2,0)), start=list(br1=-0.53,br2=0.00),control=list(maxiter=200,
toler=0.00001,minFactor=0.00001,minscale=0.000001),algorithm= "plinear",trace=T)
tt < − seq(min(x), max(x), length = 10000)
lines(tt, predict(fit21, list(x = tt)),lty=1)
lines(loess.smooth(x,y),lty=2)
legend(0.35,62,c("Piecewise Reg","loess"),lty=c(1:2))
# Contour Plot for the Non Linear Least squares algorithm
data.sharp.wt1< −llsurface.sharp(x,y,seq(-0.50,-0.25,
0.005), seq(-0.1,0.25,0.005),c(-0.39,0.08))
contour(data.sharp.wt1$br1,data.sharp.wt1$br2,data.sharp.wt1$loglik,xlab="First break
point",ylab="Second break point",cex=0.8)
points(-0.39727286,0.08711336)
"loopfn.sharp" < − function(br,data) {
```

```
datadf < − data
br1 < − br[1]
br2 < − br[2]
loglik < − summary(lm(datadf$y ∼ cbind(x,ifelse(x > br1, x-br1,0),ifelse(x >br2,x-
br2,0))))$sigma^2
return(loglik) }
"llsurface.sharp" < − function(dx, dy, br1,br2, br.start) {
# Calculates a linear fit for each grid point (grid of possible br1 and br2 values).
# Calculates a RSS surface across the grid.
# For the sharp piecewise regression model.
y < − dy
x < − dx
datadf < − data.frame(x = dx, y = dy)
br01 < − br.start[1]
br02 < − br.start[2]
nlsmax < − nls(y ∼ cbind(1, x, ifelse(x > br1, x - br1, 0),ifelse (x > br2, x - br2, 0)),
start = list(br1 = br01,br2=br02), control = list(maxiter = 50, tolerance = 0.000001,
minscale = 0.000001,minFactor=1/2048), algorithm = "plinear", data=datadf)
nlsum < − summary(nlsmax)
ds2 < − nlsum$sigma^2
error.df < − nlsum$df[2]
out < − data.frame(br1 = br1, br2 = br2, loglik = rep(NA, length(br1)* length(br2)))
out < − NA
```

```
out$loglik < − matrix(nrow = length(br1), ncol = length(br2))
for(i in 1:length(br1)) {
for(j in 1:length(br2)) {
out$loglik[i, j] < − loopfn.sharp(c(br1[i],br2[j]), datadf) }}
out$loglik < − (out$loglik * (error.df + 2) - ds2 * error.df)/(ds2* error.df)
out$br1 < − br1
out$br2 < − br2
return(out) }
# Linear Quantile Regression(Quantiles =0.05,0.10,0.25,0.50,0.75,0.90,0.95)
plot(x,y,xlab="Agricultural Stress Gradient",ylab="Fish IBI")
taus < − c(0.05,0.10,0.25,0.50,0.75,0.90,0.95)
abline(lm(y ~ x),lty=1)
for(i in 1:length(taus)){
abline(rq(y ~ x,tau=taus[i]),lty=2) }
# 5th Linear Quantile Regression
res05 < − rq(y ~ x,tau=0.05)
fit1 < − summary(rq(y~x, tau=10:90/100))
plot(fit1,nrow=1,ncol=2)
# Nonlinear Quantile Regression
tt < − seq(min(x), max(x), length = 100000)
plot(x,y,xlab="Agricultural Stress Gradient",ylab="Fish IBI",ylim=c(10,70))
us < − c(0.05,0.10,0.25,0.50,0.75,0.90,0.95)
TModel < − function(x,A,B,C,D){
```

```
z < − A/(1+exp(B+C*x+D*x^2)) }
for(i in 1:length(us)){
tau < −us[i]
fit1 < − nlrq(y ~ TModel(x,A,B,C,D),tau=tau,start=
list (A=55,B=-0.98,C=2.98,D=-1.72))
lines(tt, predict(fit1, list(x = tt)),lty=1) }
TModel < − function(x,A,B,C,D){
z < − A/(1+exp(B+C*x+D*x^2)) }
# 5th nonlinear quantile regression
fit5 < − nlrq(y ~ TModel(x,A,B,C,D),tau=0.05,start=
list (A=55,B=-0.98,C=2.98,D=-1.72))
# Plot of the par for non-linear quant. reg.
qt< − seq(0.05,0.95,0.05)
A < − c(46.00322757,46.0015831,50.80178101, 49.2167128,49.170653,49.2710882,
51.514157, 54.433383,58.0164279,58.0075565,58.0013685, 58.0817096,
58.007417,57.715818, 57.836875,58.000,59.016324,62.133323,
62.133324)
B < − c(-0.09818163,-0.1362395,0.06224087, -0.4926758,-0.500369,-0.5596914,
-1.047112, -0.876823,-0.6738135,-0.7579338,-0.7482654, -0.7393034,
-1.522860,-1.729951, -1.719833,-3.067618,-2.917635, -2.538496,-2.538496)
C < − c(3.04874793,3.2978630,3.05614167, 2.8854126,2.920869,3.0090278,
3.402148,3.211561, 2.6852760,2.8052632,3.3750957, 3.3495226,
2.777200,3.732109,3.719965, 5.896347,5.689467,5.183207,5.183207)
```

```
D < − c(-2.20018181,-2.3355178,-2.13263174, -1.8696890,-1.889770,-1.9187520,
-1.927464,-1.816251,-1.5141582,-1.5559036 ,-2.1381157,-2.1755281,
-1.265649, -1.981030,-1.974853,-2.829626, -2.737126,-2.508716,-2.508715)
par(mfrow=c(2,2))
plot(qt,A,type="o",xlab="Quantile",ylab="Parm, A")
plot(qt,B,type="o",xlab="Quantile",ylab="Parm, B")
plot(qt,C,type="o",xlab="Quantile",ylab="Parm, C")
plot(qt,D,type="o",xlab="Quantile",ylab="Parm, D")
# Plot of the par for linear and piecewise linear quant. reg.
qt< −seq(0.05,0.95,0.05)
B0< −c(27.26847,29.19433,31.236476,34.27074,36.00232,37.53202,
40.02493,40.84313,42.399023,44.34422,45.67131,47.771745,
48.954590, 50.71468,50.887709,51.563350,55.32775,55.74300,56.26406)
B1< −c(-12.53716,-11.24817,-9.881348,-12.54968,-14.06223,-15.41764,
-11.49073,-10.96769,-9.973073,-10.29747,-7.88123,-8.058424,-8.868147,
-10.07303,-9.743265,-9.898495,-12.14815,-12.39632,-12.70771)
br1 < − c(-0.2814259,-0.5552924,-0.5522797,-0.39935516,-0.34341016,-0.34339834,
-0.3588400,-0.3636134,-0.3677486,-0.3901717,-0.3993249,-0.3993144, -0.5568753,
-0.2939816,-0.2939890,-0.3438695,-0.5466294,-0.3993366,-0.3993374)
br2< −c(0.1455510,0.1278622,0.1388161,0.03187916,-0.03131366,-0.09572109,
0.1678362,0.1678544,0.1678288,0.1678487,0.1025280,0.1014321,0.1261136,
-0.1085645,0.1299741,0.1678324,NA,0.1887980,0.2197467)
plot(qt,B0,type="o",xlab="Quantile",ylab="Linear Intercept")
```

```
plot(qt,B1,type="o",xlab="Quantile",ylab="Linear Slope")

plot(qt,br1,type="o",xlab="Quantile",ylab="First Break point")

plot(qt,br2,type="o",xlab="Quantile",ylab="Second Break point")

# Piecewise Linear Quantile Regression

tt < − seq(min(x), max(x), length = 100000)

plot(x,y,xlab="Agricultural Stress Gradient",ylab="Fish IBI",ylim=c(10,70))

us < − c(0.05,0.10,0.25,0.50,0.76,0.90,0.95)

for(i in 1:length(us)){

tau < −us[i]

fit30< −nlrq(y~ A+B*x+C*ifelse((x>br1),x-br1,0)+D*ifelse(x>br2,x-br2,0),

start=list(A=55,B=2.5,C=-49.56,br1=-0.39,D=45.93,br2=0.08),

tau=tau,trace=FALSE)

lines(tt, predict(fit30, list(x = tt)),lty=1) }

abline(v=c(-0.5552924,0.2197467),lty=2,col=1)

# 5 th piecewise linear quantile regression

fit5< −nlrq(y ~ A+B*x+C*ifelse((x>br1),x-br1,0)+D*ifelse(x>br2,x-br2,0),start=list

(A=55,B=2.5,C=-49.56,br1=-0.39,D=45.93,br2=0.08),tau=0.05,trace= FALSE)
```

Vita Auctoris

Name:	Jabed Hossain Tomal
Place of Birth:	Mymensingh, Bangladesh
Year of Birth:	1977
Education:	B.Sc. (Honours) in Statistics University of Dhaka, Dhaka, Bangladesh 1999
	M.Sc. in Statistics University of Dhaka, Dhaka, Bangladesh 2001
	M.Sc. in Statistics University of Windsor, Windsor, Ontario 2006