Universidade de Vigo

Testing critical points of regression curves. An application to the management of aquatic living resources.

Marta Sestelo and Javier Roca-Pardiñas

Report 12/06

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://webs.uvigo.es/depc05/
E-mail: depc05@uvigo.es

Universidade de Vigo

Testing critical points of regression curves. An application to the management of aquatic living resources.

Marta Sestelo and Javier Roca-Pardiñas

Report 12/06

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita:

Universidade de Vigo
Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

Testing critical points of regression curves. An application to the management of aquatic living resources.

Marta Sestelo[a,*], Javier Roca-Pardiñas[a]

[a]*Department of Statistics and Operations Research, University of Vigo, C/ Torrecedeira 86, E-36280 Vigo, Spain.*

June 2012

Abstract

This paper was focused on regression models incorporating the so-called factor-by-curve interaction, where the effect of a continuous covariate on response varies across groups defined by levels of a categorical variable. This study sought to compare regression curves and their derivatives that may vary across groups defined by different experimental conditions. The goals of this paper ware a) to provide a global test to detect significant features of regression curves through the study of their derivatives, and b) to draw inferences about critical points (such as maxima or change points) linked to the derivative curves. The regression curves were estimated using local polynomial kernel smoothers. Such nonparametric regression models allow for a more flexible fit of real data than do the parametric regression techniques usually used. Similarly, they make it possible for the derivatives of the regression curve to be calculated. Bootstrap methods were used to draw inferences from the derivative curves, and binning techniques were applied to speed up computation in the estimation and testing processes. A simulation study was conducted to assess the validity of these bootstrap-based tests. This methodology was applied to study the relative growth of barnacles, in particular, in the estimation of the minimum size of capture of this species.

Keywords: bootstrap, derivatives, factor-by-curve, kernel smoothing

*Corresponding author. Tel.: +34 986813948; fax: +34 986813644

Email address: sestelo@uvigo.es (Marta Sestelo)

1. Introduction

In many practical situations, the target response, Y, depends on a continuous covariate, X. In this regression framework, consideration might well be given to the nonparametric regression model

$$Y = m(X) + \varepsilon \tag{1}$$

where m is a smooth unknown function and ε is the error, which is assumed to be independent of the covariate X. By studying m we can establish the functional relationship between the mean response and the covariate X. Additionally, it might be interesting to make inferences about critical points of m, called x_0 (e.g. minima, maxima or inflection points which signal the change in the sign of curvature) studying for this purpose the derivatives of m. For instance, in the application to real data shown below, it is necessary to determine which point maximizes the first derivative of the regression curve m. Explicitly, the cited point, x_0, is defined as follows

$$x_0 = \arg\max_x m^1(x)$$

being $m^1(x)$ the first derivative of m at the point x. In some circumstances, the relationship between Y and X can vary among subsets defined by levels $1 \ldots, M$ of a categorical covariate F resulting in a regression model with factor-by-curve interactions. In this framework, we will denote x_{0l} as the critical point specific to the l level of F. For instance, x_{0l} can be defined as

$$x_{0l} = \arg\max_x m_l^1(x)$$

being $m_l(X) = E(Y|X = x, F = l)$. At this stage, it is of interest to test the null hypothesis $H_0 : x_{01} = \ldots = x_{0M}$.

It is important to highlight that it is possible to observe that the specific critical points could coincide even if the derivative regression curves of m_l are different. One example of this can be observed in the application to real data of this paper. In this section, we will apply this methodology in studying the growth of the stalked barnacle, *Pollicipes pollicipes* (Gmelin, 1789), in particular, in the estimation of the minimum size of capture of this species. The study of derivatives is extremely useful when it comes to establishing this ideal size. In particular, in this work we propose that the minimum size corresponds to the point (or size) where the first derivative reaches the maximum. From this point onwards, weight gain from one size

to the next decreases, so that the yield obtained ceases to be profitable. At this point, the proposed methodology lets us compare the derivatives of the growth curves between the two years of study. Also, it shows how the point which maximizes the first derivative is the same in both years.

The possibility of incorporating the factor-by-curve interactions in nonparametric regression models has already been discussed by Hastie and Tibshirani (1990). Also, Ruppert and Wand (1994) presented an algorithm based on penalized splines (P-splines), which would enable these types of interactions to be incorporated into these types of models. Recently, Cadarso-Suárez et al. (2006) and Roca-Pardiñas et al. (2006) have successfully applied these types of interactions for estimating neuron firing rates.

Additionally, a question that tends to arise in this type of model is to know if the estimated curves are equal to each other. This problem - testing for the equality of nonparametric regression curves - has also been widely treated in the statistical literature. Relevant papers on this topic are Bowman and Young (1996); Dette and Neumeyer (2001); Hardle and Marron (1990); Pardo-Fernández et al. (2007); Kulasekera (1995); Neumeyer and Dette (2003); Srihera and Stute (2010); Young and Bowman (1995) among others. However, unlike the above references where are given global tests to detect significant differences between curves, with this paper we add, over existing approach, the possibility of testing locally curves, allowing to draw inference about critical points.

The main goals of this paper are (1) to provide a global test to detect globally significant features of regression curves by studying their derivatives and (2) to propose a new methodology that can be used to draw inferences about critical points (such as maxima or change points) linked to the derivative curves. To this end, a computational algorithm was developed and implemented, based on local kernel polynomial smoothers, which allowed for nonparametric estimation of the curves. The bootstrap method was used for practical implementation of tests capable of detecting the significance of these curves.

The layout of this paper is as follows. The estimation algorithm based on kernel smoothers is presented in Section 2. In Section 3, we propose bootstrap methods for the implementation of different tests designed to detect significant features of the curves based on the study of their derivatives. Practical questions, such as bandwidth selection and computational acceleration based on binning techniques are addressed in Section 4. In Section 5, we present the results regarding the numerical performance of the different

test statistics under review. In Section 6, the above-described methodology was applied in the evaluation of the relative growth of *P. pollicipes*. Finally, we conclude the paper with a discussion section.

2. Nonparametric Estimation Procedures

In this paper the following nonparametric regression model including factor-by-curve interactions was considered

$$Y = f_0(X) + \begin{cases} f_1(X) + \varepsilon_1 & \text{if} \quad F = 1 \\ \vdots & \\ f_M(X) + \varepsilon_M & \text{if} \quad F = M \end{cases} \tag{2}$$

where $\varepsilon_1, \ldots, \varepsilon_M$ are the mean zero errors for each factor's levels, f_0 represents the global effect of X on the response, and f_l is the specific effect of X associated with the l^{th} level of the factor F. Note that under model (2) the regression curves $m_l(x) = E(Y|X = x, F = l)$ are given by

$$m_l(X) = f_0(X) + f_l(X) \qquad \text{for} \quad l = 1, \ldots, M$$

In order to avoid different combinations of $f_0, f_1, \ldots, f_M$ that lead to the same model, we assume that the sum of the specific effects across the levels must be zero. That is, for each x, $\sum_{l=1}^{M} f_l(x) = 0$ is satisfied. Note that this condition does not represent restrictions on our model because it can be modified to conform to the said identifiable condition.

The factor-by-curve regression model in (2) was estimated using local polynomial kernel smoothers (Wand and Jones, 1995). Given a sample $\{(X_i, F_i, Y_i)\}_{i=1}^{n}$ the estimate of the f_0 at a point x is given by $\hat{f}_0(x) = \hat{\alpha}^0(x)$ where $\left(\hat{\alpha}^0(x), \hat{\alpha}^1(x), \ldots, \hat{\alpha}^R(x)\right)$ is the minimizer of

$$\sum\nolimits_{i=1}^{n} \left\{ Y_i - \sum\nolimits_{r=0}^{R} \alpha^r(x) (X_i - x)^r \right\}^2 \cdot K\left(\frac{X_i - x}{h}\right)$$

where $K(x) = \exp(-0.5x^2)$ is the kernel function, h is the smooth bandwidth and R is the degree of the polynomial. Once obtained the estimation of $\hat{f}_0$, for $l = 1, \ldots, M$, we obtain the estimates $\hat{f}_l(x) = \hat{\alpha}_l^0(x)$ with $\left(\hat{\alpha}_l^0(x), \hat{\alpha}_l^1(x), \ldots, \hat{\alpha}_l^R(x)\right)$ minimizing

$$\sum\nolimits_{i=1}^{n} \left\{ Y_i^l - \sum\nolimits_{r=0}^{R} \alpha_l^r(x) (X_i - x)^r \right\}^2 \cdot K\left(\frac{X_i - x}{h}\right) I_{\{F_i = l\}}$$

with $Y_i^l = Y_i - \hat{f}_0(X_i)$. Note that it is not necessary for the obtained estimates to have satisfied the imposed identification condition. To do so, the following procedure is used. For each x, calculate the mean of the specific effects of each level, $S(x) = M^{-1}\sum_{l=1}^{M}\hat{f}_l(x)$, and replace the original $\hat{f}(x)$ and $\hat{f}_l(x)$ respectively by $\hat{f}_l(x) - S(x)$ and $\hat{f}_0(x) + S(x)$. Finally, the estimated curves for each level at point x are given by

$$\hat{m}_l(x) = \hat{f}_0(x) + \hat{f}_l(x) \quad \text{for} \quad l = 1, \ldots, M$$

Moreover, the estimated $r^{th}(r \leq R)$ derivative of $\hat{m}_l(x)$ is given by $\hat{m}_l^r(x) = \hat{f}_0^r(x) + \hat{f}_l^r(x)$ where $\hat{f}_0^r(x) = r!\hat{\alpha}_0^r(x)$ and $\hat{f}_l^r(x) = r!\hat{\alpha}_l^r(x)$.

3. Inferences

When a factor-by-curve interaction is detected in model (2), it might be interesting to make inferences about some critical points of curves (such as minima, maxima or inflection points) studying for this purpose the derivatives. In general, the critical point x_{0l} referring to the l level will be obtained, for some r, from the derivative curve $m_l^r(x)$. For example, in the application to real data, we will be interested in determining, for each l level, which point x_{0l} maximizes the first derivative of the regression curve $m_l^1(x)$.

The proposed procedure in this section allows us to test the hypothesis that the critical points among the levels of the factor are equal. Based on this goal, we first need to propose a global test that assumes the following null hypothesis

$$H_0^r : m_1^r(\cdot) = \ldots = m_M^r(\cdot) \tag{3}$$

Note that if H_0^r is not rejected, then the equality of critical points $x_{01}, \ldots, x_{0M}$ will be also accepted. By contrast, if H_0^r is rejected, the conclusion about these critical points should be postponed, and it will be necessary to use the local test that we propose below.

3.1. Global test

Here we propose a bootstrap procedure that allows us to test the null hypothesis (3) based on the model (2). Note that this hypothesis is equivalent to $f_1^r(\cdot) = \ldots = f_M^r(\cdot) = 0$ and, therefore, $f_l(x) = \sum_{j=0}^{r-1} a_l^j X^j$ is satisfied for $l = 1, \ldots, M$. Accordingly, the null regression model is given by

$$Y = f_0(X) + \begin{cases} \sum_{j=0}^{r-1} a_1^j X^j + \varepsilon_1 & \text{if} \quad F = 1 \\ \vdots & \\ \sum_{j=0}^{r-1} a_M^j X^j + \varepsilon_M & \text{if} \quad F = M \end{cases} \tag{4}$$

To test H_0^r we propose the use of a statistic based on direct nonparametric estimates of f_l^r curves. This statistic test is as follows:

$$T = \sum_{l=1}^{M} \sum_{i=1}^{n} |\hat{f}_l^r(X_i)|$$

Note that if H_0^r is verified, the T value should be close to zero, but generally greater. The test rule based on T consists of rejecting the null hypothesis if $T > T^{1-\alpha}$, where T^p is the empirical $p-$percentile of T under H_0. Nevertheless, it is well known that, within a nonparametric regression context, the asymptotic theory for determining such percentiles is not closed, and resampling methods such as the bootstrap introduced by Efron (1979) (see also Efron and Tibshirani, 1993; Härdle and Mammen, 1993; Kauermann and Opsomer, 2003) can be applied instead. The testing procedure used here involves the following steps:

Step 1. Compute the T value from the sample as explained above.

Step 2. Estimate the null regression model in (4) and obtain for $i = 1, \ldots, n$ the pilot estimates

$$\hat{m}_{F_i}(X_i) = \hat{f}_0(X_i) + \sum\nolimits_{j=0}^{r-1} \hat{a}_{F_i}^j X_i^j$$

Step 3. For $b = 1 \ldots B$ (e.g. B=1000), generate bootstrap samples $\left\{X_i, F_i, Y_i^{\bullet b}\right\}_{i=1}^n$ with $Y_i^{\bullet b} = \hat{m}_{F_i}(X_i) + \varepsilon_i^{\bullet b}$ being

$$\hat{\epsilon}_i^{\bullet b} = \begin{cases} \hat{\epsilon}_i \cdot \frac{(1-\sqrt{5})}{2} & \text{with probability } p = \frac{5+\sqrt{5}}{10} \\ \hat{\epsilon}_i \cdot \frac{(1+\sqrt{5})}{2} & \text{with probability } p = \frac{5-\sqrt{5}}{10} \end{cases}$$

where $\hat{\epsilon}_i = Y_i - \hat{m}_{F_i}(X_i)$ are the errors under the H_0, and compute $T^{\bullet b}$ as in **Step 1**.

Finally, the test rule based on T consists of rejecting the null hypothesis if $T > T^{1-\alpha}$, where T^p is the empirical $p-$percentile of values $T^{\bullet b}(b = 1, ..., B)$ obtained before.

3.2. Local test

As we mentioned before, if the previous test is significative and therefore we reject the equality of the m_l^r curves ($l = 1, \ldots, M$), we will be interested in testing the null hypothesis of equality of critical points. Note that, it is possible that these points can be equal, even if the curves and/or their derivatives are different.

For instance, taking into account the maxima of the first derivatives, the interest lies in testing the following null hypothesis

$$H_0 : x_{01} = \ldots = x_{0M}$$

The cited hypothesis is true if $D = x_{0j} - x_{0k} = 0$ being

$$(j, k) = \quad \arg\max \quad {}_{1 \leq l < m \leq M} |x_{0j} - x_{0k}|$$

otherwise, H_0 is false. It is important to highlight that, in practice, the true x_{0j} are not known and consequently neither is D, so an estimate $\hat{D} = \hat{x}_{0j} - \hat{x}_{0k}$ is used, where, in general, $\hat{x}_{0l}$ are the estimates of x_{0l} based on the estimated curves $\hat{m}_l$.

In our application, we need to know the point which maximizes the first derivative of the m_l curves. Accordingly, we have defined this point, x_{0l} for each l level, as

$$x_{0l} = \arg\max_x m_l^1(x)$$

A natural estimator of the cited x_{0l} can be obtained as the maximizer of

$$\hat{m}_l^1(k_1), \ldots, \hat{m}_l^1(k_N)$$

with $k_1, \ldots, k_N$ being a grid of N equidistant points in a range of the X values.

Of course, since $\hat{D}$ is only an estimate of the true D, the sampling uncertainty of these estimates need to be acknowledged. Hence, a confidence interval (a, b) is created for D for a specific level of confidence (e.g., 95%). Based on this, the null hypothesis is rejected if a zero value is not within the interval.

The steps for construction of the bootstrap confidence interval for the true D are as follows:

Step 1. Obtain from the sample data $\{(X_i, F_i, Y_i)\}_{i=1}^n$ the estimates of x_{0l} based on the model in (2) and consequently retrieve the $\hat{D}$ value.

Step 2. For $b = 1, \ldots, B$ (e.g. B=1000), generate bootstrap samples $\left\{(X_i^{\bullet b}, F_i^{\bullet b}, Y_i^{\bullet b})\right\}_{i=1}^n$ by randomly sampling the n items from the original data set with replacement (that is, each individual value (X_i, F_i, Y_i) has a probability n^{-1} of occurring), and compute $\hat{D}^{\bullet b}$ as in **Step 1**.

Finally, the $100(1-\alpha)\%$ limits for the confidence interval of D are given by

$$I = \left(\hat{D}^{\alpha/2}, \hat{D}^{1-\alpha/2}\right)$$

where $\hat{D}^p$ represents the p-percentile of $\hat{D}^{\bullet 1}, \ldots, \hat{D}^{\bullet B}$.

4. Bandwidth Selection and Computational Aspects

Bandwidth Selection

It is well known that the nonparametric estimates $\hat{m}_l^r(X)$ depend heavily on the bandwidths $h_0, h_1, \ldots, h_M$ used in the kernel-based algorithm for the estimation of the partial functions $f_0, f_1, \ldots, f_M$. Various methods for an optimal selection have been suggested, such as Generalized Cross Validation (GCV) (Golub et al., 1979) or plug-in methods (see e.g. Ruppert et al., 1995). For a nice overview on this topic, we recommend the reading of Wand and Jones (1995). However, given the difficulty of asymptotic theory, optimal bandwidth selection is still a challenging problem. Furthermore, the results obtained from the tests presented in Section 3 depend heavily on the smoothing parameter, and a distinction should be drawn between the bandwidth choice for estimation and for testing.

As a practical solution, in the first step of the estimation algorithm, bandwidth h_0 is selected automatically by minimizing the following cross-validation criterion:

$$CV_0 = \sum\nolimits_{i=1}^n \left(Y_i - \hat{f}_0^{(-i)}(X_i)\right)^2 \tag{5}$$

where $\hat{f}_0^{(-i)}(X)$ indicates the fit at X leaving out the i-th data point. Likewise, windows h_j $(j = 1, \ldots, M)$ are selected by minimizing

$$CV_l = \sum\nolimits_{i=1}^n I_{\{F_i=l\}} \left(Y_i - \hat{f}_0(X_i) - \hat{f}_l^{(-i)}(X_i)\right)^2 \tag{6}$$

Computational Aspects

Bootstrap resampling techniques are time-consuming processes because it is necessary to estimate the model many times. Moreover, the use of the cross-validation technique for the choice of the bandwidths used in the nonparametric estimates implies a high computational cost, inasmuch as it is necessary to repeat the estimation operations several times to select the optimal bandwidths.

Additionally, in the application to real data, we have a large amount of data ($n = 16562$). Consequently, recourse to some computational acceleration technique is fundamental to ensure that the problem can be addressed adequately in practical situations.

To speed up this process, in this paper we have used binning techniques. A detailed explanation of this technique can be found in Fan and Marron (1994). There now follows a brief description of the procedure that we used to construct the binning versions of the estimators $\hat{f}_0(x)$ and $\hat{f}_l(x)$ given in Section 2.

In the first step of the algorithm, we consider a grid of N equidistant points $X_1^\bullet < \ldots < X_N^\bullet$ and construct the binned sample $\left\{X_j^\bullet, Y_j^\bullet\right\}_{j=1}^N$ with weights $\left\{W_j^\bullet\right\}_{j=1}^N$ where

$$Y_j^\bullet = \sum\nolimits_{i=1}^n \left(1 - \left|X_i - X_j^\bullet\right|/\delta\right)_+ Y_i \quad \text{and}$$

$$W_j^\bullet = \sum\nolimits_{i=1}^n \left(1 - \left|X_i - X_j^\bullet\right|/\delta\right)_+$$

with $X_+ = \max\{0, X\}$ and δ denoting the distance between two neighboring knots. The binning approximations $\hat{f}_0(x)$ in the first step of the estimation algorithm are obtained by minimizing

$$\sum\nolimits_{i=1}^N \left\{Y_i^\bullet - \sum\nolimits_{r=0}^R \alpha^r \left(X_i^\bullet - X\right)^r \cdot K\left(\frac{X_i^\bullet - X}{h}\right) W_i^\bullet\right\}$$

Similarly, the approximations $\hat{f}_l(x)$ in the second step of the algorithm are obtained by minimizing

$$\sum\nolimits_{i=1}^N \left\{Y_i^{\bullet l} - \sum\nolimits_{r=0}^R \alpha_l^r \left(X_i^\bullet - X\right)^r\right\}^2 \cdot K\left(\frac{X_i^\bullet - X}{h}\right) W_i^{\bullet l}$$

where $Y_i^{\bullet l} = Y_i^\bullet - \hat{f}_0\left(X_i^\bullet\right)$ and $W_i^{\bullet l} = W_i^\bullet I_{\{F_i = l\}}$.

As in the estimation with the binning technique, the cross-validation errors CV in 5 and 6 can be respectively approximated by

$$CV_0 \approx \sum\nolimits_{i=1}^{N} W_i^{\bullet} \left(\frac{Y_i^{\bullet(-i)}}{W_i^{\bullet}} - \hat{f}_0^{(-i)}\left(X_i^{\bullet}\right) \right)^2 \text{ and}$$

$$CV_l \approx \sum\nolimits_{i=1}^{N} W_i^{\bullet l} \left(\frac{Y_i^{\bullet} - \hat{f}_0^{(-i)}\left(X_i^{\bullet}\right)}{W_i^{\bullet l}} - \hat{f}_l^{(-i)}\left(X_i^{\bullet}\right) \right)^2$$

where the estimates $\hat{f}_0^{(-i)}$ and $\hat{f}_l^{(-i)}$ $(l = 1, \ldots, M)$ are obtained by leaving out the ith grid point.

These approximations substantially reduce computing time because the calculation of CV_l is only necessary to evaluate kernel K at a maximum of N different points for each choice of bandwidth. Needless to say, the finer the grid of points selected, the better the approximation. The choice of the number of grid points is a compromise between approximation error and computational speed. In practice, depending on the sample size n and on the distribution of the covariates, a larger amount of grid points might be more appropriate.

A detailed study of the compromise between the computational time and the error of the binning approximations can be seen in De Uña Álvarez and Roca Pardiñas (2009). The conclusion to be drawn from this study is that, as N increases, the errors of the estimates decrease, but waiting time may increase substantially.

5. Simulation Study

This section reports the results of two simulation studies to assess the validity of both (a) global derivative factor-by-curve interaction tests and (b) critical points tests. In both cases we consider a factor-by-curve unidimensional regression problem where the explanatory covariate X was drawn from a uniform $U[-2, 2]$ distribution, the factor F was chosen in accordance with $F \sim \quad \text{Bernoulli} \quad (0.5) + 1$, and the outcome variable Y was generated according to

$$Y = \begin{cases} m_1(X) + N(0, \sigma_1(x)) & \text{if} \quad F = 1 \\ m_2(X) + N(0, \sigma_2(x)) & \text{if} \quad F = 2 \end{cases} \tag{7}$$

with $\sigma_j(x) = 0.2 + |0.25m_j(x)|$ for $j = 1, 2$. One thousand independent samples $\{X_i, F_i, Y_i\}_{i=1}^n$ were generated from the model (7).

In the first study we assessed the validity of the global first derivative factor-by-curve interaction test. In particular, we considered the null hypothesis $H_0 : m_1^1(\cdot) = m_2^1(\cdot)$ under the model in (7) with $m_1(x) = (2 - 3x^2)$ and $m_2(x) = 1-(1-a)3x^2$ being a a constant. Consequently, the first derivative regression curves are given by $m_1^1(x) = -6x$ and $m_2^1(x) = -6(1 - a)x$. Note that the regression curves m_1 and m_2 are always different, yet the constant, a, governs the first degree factor-by-curve interaction of the model. The value $a = 0$ corresponds to the hypothesis H_0 and as the value of a rises, so does the degree of interaction of the model.

To determine the critical values of the global test statistic, we applied the bootstrap method as described above in Subsection 3.1. Specifically, this entailed 400 bootstrap samples for calculating type 1 errors and 200 bootstrap samples for calculating the power under the alternative. Both type 1 errors and power were calculated on the basis of 1000 simulation runs.

Table 1 displays the estimated type 1 errors for the method at different significance levels and for different sample sizes. As can be seen from this table, all tests performed reasonably well, with almost all holding the level and several coming quite close to the nominal level.

Level	n=100	n=200	n=500
1	0.9	0.6	0.9
5	5.1	4.4	5.4
10	11.2	8.6	11.1
15	16.7	14.3	15.4
20	21.0	19.8	20.8

Table 1: Estimated type 1 error (in percent) for the global test.

We then studied power performance for the alternatives as a function of a. Power results are shown in Figure 1. The test produces satisfactory power curves, with the probability of rejection rising in response to any increase in the value of the constant a.

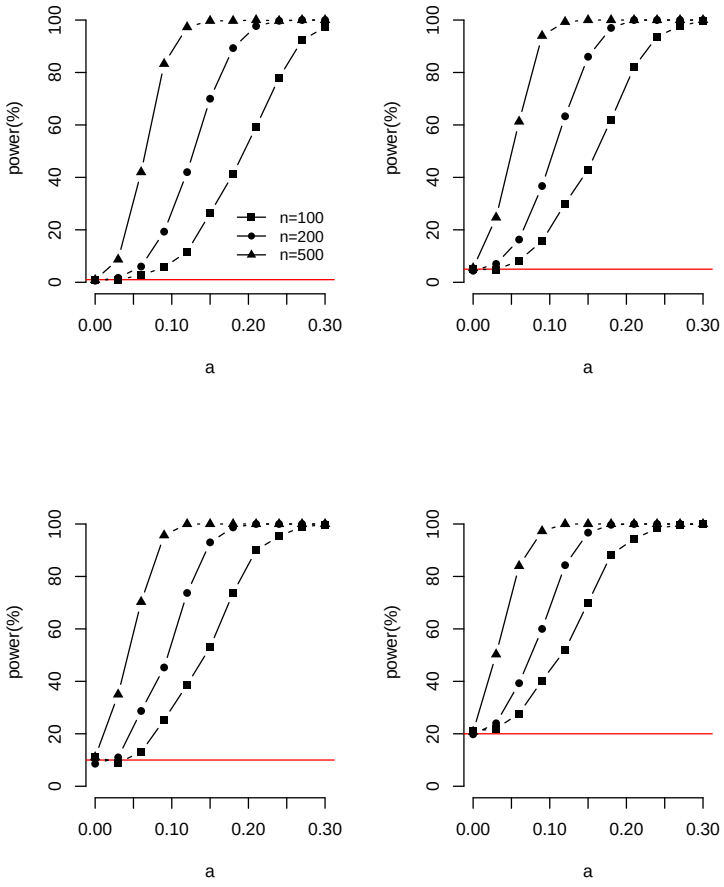

Figure 1: Percentage rejection for global test on increasing a for nominal levels of 1, 5, 10, and 20 percent and sample sizes of $n = 100$, $n = 200$ and $n = 500$.

In the second simulation study, we considered the local hypothesis H_0 : $x_{01} = x_{02}$ being $x_{0j} = \arg\max_x m_j^1(x)$. In this study we consider again the model (7) with $m_1(x) = 2+x-x^3$ and $m_2(x) = 1+2x-(x-a)^3$, being again a a constant. In this case, the first derivatives are given by $m_1^1(x) = 1-3x^2$ and $m_2^1(x) = 2-3(x-a)^2$ and they are always different. However, it is important to highlight that if $a = 0$ then the null hypothesis will be true.

Graphical average results are displayed in Figure 2. The left panel plots the data generating function and their mean estimates with their 100 simulation replicates for estimate m_2 with $a = 0$. The good performance of the resulting estimates is evident, recovering the functional forms of the corresponding true curve very successfully. In the right panel we can observe the estimate of their first derivative, and their simulation replicates. Note that this estimate opens at the limits of the curve resulting from the intrinsic features of the kernel estimator. Ticks on the horizontal axis of this right panel represent the estimated a value for each simulation run, which corresponds with the maximizers of the first derivative. It is important to highlight that the estimation of this point should be close to zero because in the data generating function the true a value was forced to be zero. The Box-Plots for $\hat{a}$, according to the sample size n, can be seen in Figure 3. As expected, the interquartile range decreased as the sample size n increased.

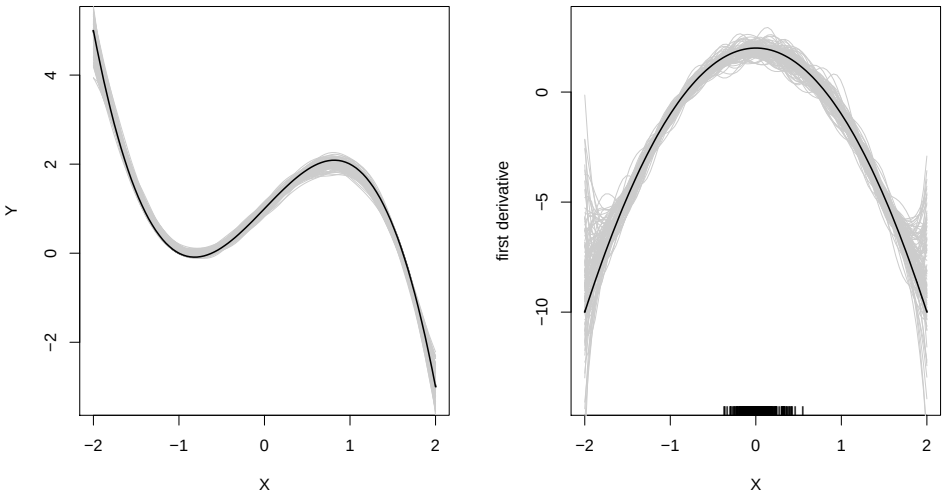

Figure 2: True estimation and first derivative (solid broad lines) with their 100 simulation runs (grey lines) with $a = 0$ and $n = 500$. Ticks on the horizontal axis represent the estimated a value for each simulation trial.

To determine the confidence interval for the statistic D, we applied the bootstrap method as described in subsection 3.2 with a total of 1000 bootstrap samples. Type 1 errors and power were calculated as the proportions of rejections of H_0 in 1000 runs. Test size and power were determined for different levels (1, 5, 10, and 20 percent) and for different sample sizes n (n=100, 200, 500).

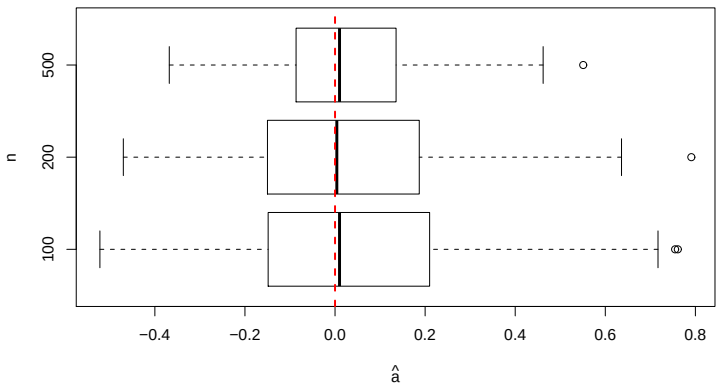

Figure 3: Box-Plot for the estimated a with different sample sizes (n = 100, 200 and 500).

Table 2 shows the results obtained according to type 1 errors. As we can see, this test performed well in general, with type 1 errors proving to be relatively close to nominal errors. Moreover, the differences between nominal levels and type 1 errors decreased as the sample size increased.

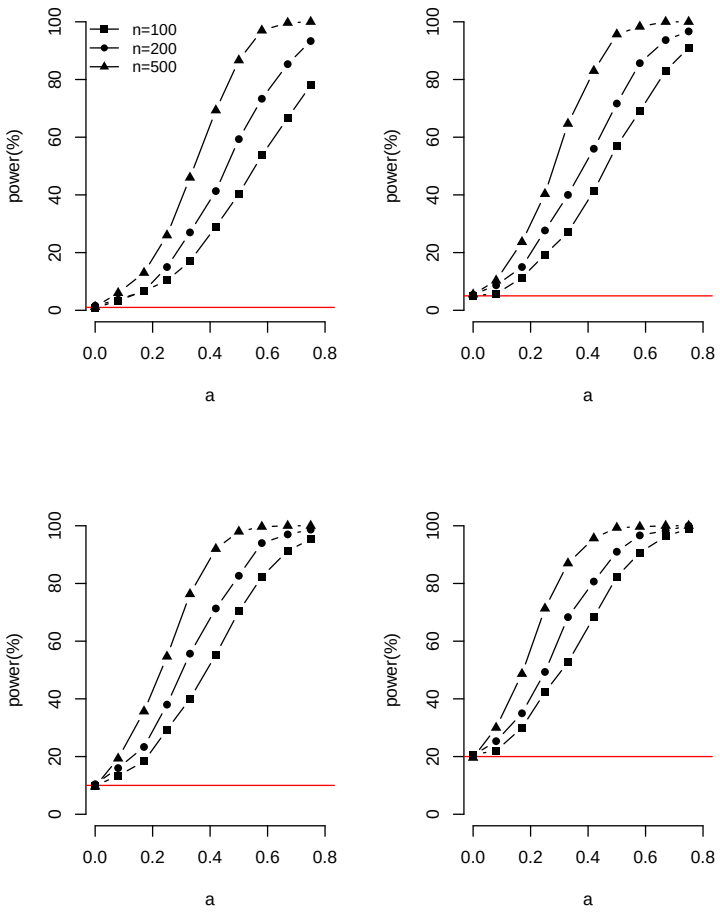

Figure 4: Percentage rejection for local test on increasing a for nominal levels of 1, 5, 10, and 20 percent and sample sizes of $n = 100$, $n = 200$ and $n = 500$.

The power curves shown in Figure 4 display the expected behavior pattern. For $a = 0$, the probability of rejection was approximately at the nominal level, reaching a value of 1 when the sample size grew. Moreover, the test showed an improvement in power as the sample size grew.

Level	n=100	n=200	n=500
1	0.8	1.6	1.2
5	5.1	5.4	5.5
10	9.9	10.4	9.6
15	15.7	15.8	14.6
20	20.7	20.4	19.6

Table 2: Estimated type 1 errors (in percent) for local test.

The bandwidth used in this simulation study were obtained using the CV mechanism explained in Section 3.1. While this choice of bandwidths may be far from optimal, as stated before, the complete testing procedures seemed to perform reasonably well in this simulation study.

6. Application to real data

Our methodology was used to determine the ideal size of capture of the stalked barnacle (Sestelo and Roca-Pardiñas, 2011) . This species, *Pollicipes pollicipes* (Gmelin, 1789), is a strictly littoral and essentially intertidal pedunculate cirripede which form dense aggregates or clumps on the exposed rocky shores of Algeria, France, Spain, Morocco, Portugal and Senegal (Barnes, 1996; Cruz, 2000; Darwin, 1851). The commercial interest of this species resides in their muscular peduncle, the edible stalk of the barnacle, which commands high prices on the market (Goldberg, 1984). In Spain and Portugal, where harvesting of *P. pollicipes* is highest, the phenomenon of overfishing has affected this species to differing degrees (Bernard, 1988; Cardoso and Yule, 1995; Cruz, 2000; Molares and Freire, 2003).

Accordingly, we sought to determine a sustainable, harvestable specimen size, i.e., a size that would ensure high commercial yield while simultaneously guaranteeing the regeneration and conservation of the population.

To this end, specimens were collected from five sites along an intertidal zone that are representative of the region's Atlantic coastline and correspond to the stretches of coast where this species is harvested (Figure 5, first row). The study was conducted over two years, from January 2006 to December 2007, during which we sought to maintain a monthly sampling periodicity.

The following biometric variables of each specimen were measured: rostro-carinal length (RC; maximum distance across the capitulum between the ends of the rostral and carinal plates; the variable that best represents the growth of the species (Cruz, 1993, 2000)) (Figure 5, second row); and dry weight (DW), obtained on the basis of drying individuals in a forced air oven for 24 hours at 100 ^{o}C (Montero-Torreiro and Martínez, 2003). All measurements were made using a digital caliper with a precision of 0.1 mm, and a 0.01 g precision balance. A total of 16562 specimens were measured.

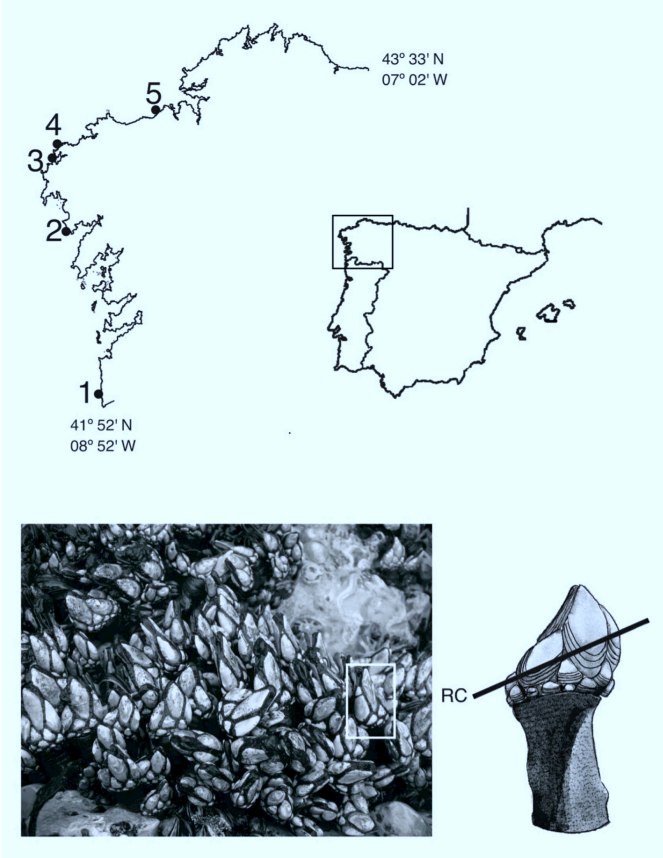

Figure 5: First row: Sampling sites. Second row: Picture of *P. pollicipes* on the rock and sketch depicting longitudinal variable measured.

In order to ascertain individual weight gain versus size, and be able to relate this evolution with the optimal size of capture for the species, the following regression model was used

$$DW = m(RC) + \varepsilon \tag{8}$$

where m is a smooth function and ε is the error that is assumed to have mean zero and variance as a function of the covariate RC.

Figure 6 shows the estimated regression curve of the previous model and its first derivative. As we can see, the regression curve m is a monotone increasing function, and the value of DW thus increases with the values of RC. Yet the increase in weight per unit of RC (given by the first derivative of m) registers a maximum at a given size, that we named rc_0, beyond which this weight gain declines (or at least remains constant). Consequently, this study proposes that the minimum size of capture should never be less than rc_0. In this overall study, this rc_0 corresponded to an RC of 21.5 mm (broken vertical line).

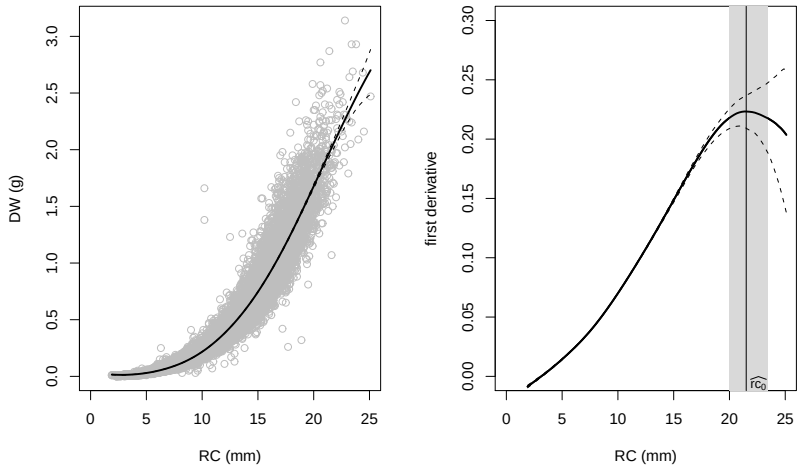

Figure 6: Regression curve and first derivative (solid lines) with bootstrap-based 95% confidence intervals (broken lines) for dry weight and rostro-carinal length (overall study). Solid vertical line: estimated rc_0. Grey area: confidence interval constructed for $\widehat{rc_0}$.

In biological studies, and specifically in population dynamics and stock assessment, it is relevant to ascertain whether this size remained constant across time and was not altered by any possible annual variability in the growth of this species. Therefore, the study was repeated including the factor-by-curve interaction. The first and second rows of Figure 7 thus refer to 2006 and 2007, respectively. As with the overall study, in all cases the initial regression curves show the way in which smaller-sized individuals increased in weight exponentially whereas larger-sized individuals increased in weight proportionally. To summarize, Table 3 shows the values estimated for rc_0 by each of the studies conducted.

Study	$\widehat{rc_0}$	95% IC
Global	21.50	(19.96,23.42)
2006	21.18	(19.75,23.56)
2007	21.10	(19.60,22.89)

Table 3: Size, $\widehat{rc_0}$, which maximizes the first derivative of the regression curves, with 95% confidence interval, for each of the studies conducted.

Once obtained the regression curves m_{2006} and m_{2007}, the following step is to determine whether the year factor produces an effect on the response and we are really dealing with a true interaction or, by contrast, the previous regression curves are equal. For this, we have applied the global test explained above. The p-value obtained, both initial regression curves and first derivatives, is less than 0.01 and therefore both the null hypothesis of equality of curves and equality of first derivatives are rejected.

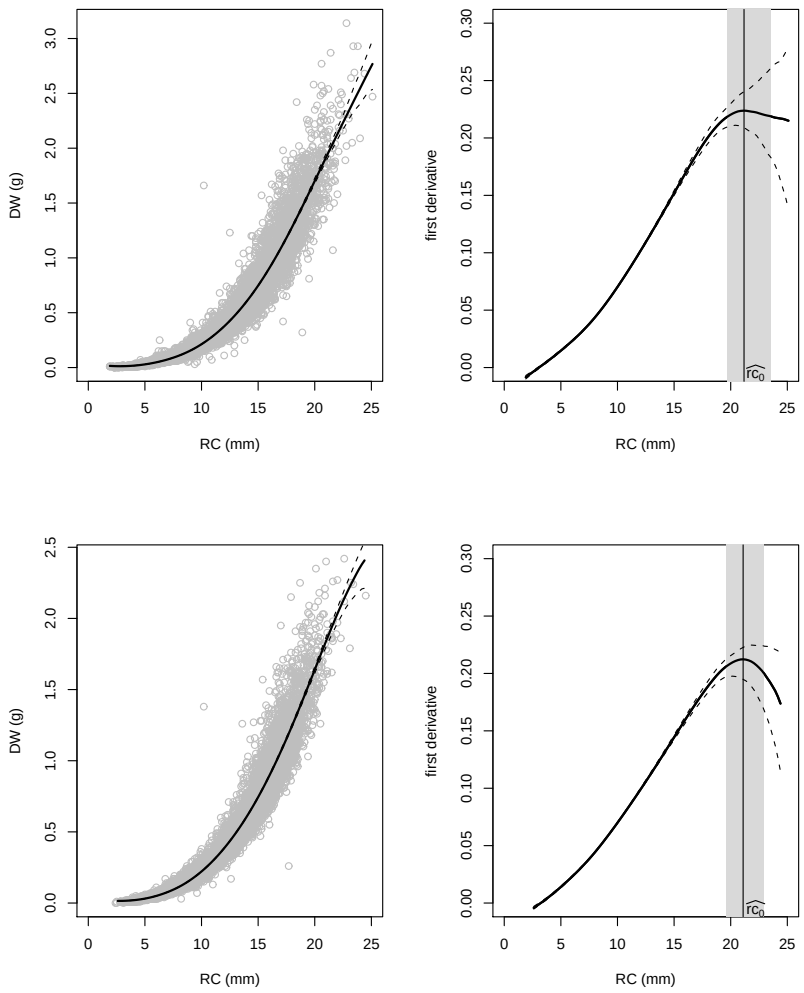

Figure 7: Regression curve and first derivative (solid lines) with bootstrap-based 95% confidence intervals (broken lines) for dry weight and rostro-carinal length. First row: year 2006; second row: year 2007. Solid vertical line: estimated rc_0. Grey area: confidence interval constructed for $\widehat{rc_0}$.

As we mentioned before, in this application it will be useful to prove whether the size sought (rc_0), which maximizes the first derivative of both years, is equal for the two levels. Therefore, the local test was applied result-

ing a D value of 0.0812 (-3.2264, 3.1562). This confidence interval indicates that, although the effects of RC on the response depends on the factor and consequently the curves and their derivatives are different for each level, the size where this barnacle reaches its maximum yield is significantly equal.

In this application we are also interested in testing if the optimal size varies depending on the geographic distribution of the species. Therefore, we have selected two sites where we collected the sample and we have fitted a model with factor-by curve interaction but, in this case, the site has been considered as the factor. The estimated rc_0 was 20.93 (19.97, 22.37) and 17.35 (16.82, 18.13) for site 1 and site 2, respectively. The p-value obtained with the global test is less than 0.01 and the local test applied to this context results in a D value of -3.573 (-5.025,-2.531). This confidence interval indicates significant differences in size between sites, suggesting the existence of a possible geographical differentiation of growth of *P. pollicipes*.

7. Discussion

In this paper, local polynomial kernel smoothers were used to obtain nonparametric estimates of regression curves and their derivatives, based on regression models with factor-by-curve interactions. The main goals of this paper were to provide a global test to detect significant features of the regression curves and their derivatives, and to draw inferences about critical points linked to the derivative curves. We have also shown here the application of this methodology in a real study, in order to obtain and compare the size of capture of a biological species.

In the application to real data, to draw inferences about the critical point rc_0 it is essential to know the confidence interval of the maximizer of the first derivative. To this end, we have used bootstrap techniques. To construct this punctual confidence interval and to determine the critical values of the T statistic (global test), we have used the wild bootstrap. This resampling method is valid for heteroscedastic models where variance of ε is a function of the covariate X. By contrast, according to the local test, where we must not generate bootstrap samples under H_0, the chosen bootstrap was a "simple bootstrap" where replicates have been generated by randomly sampling the n items from the original data set with replacements (that is, each individual value has a probability n^{-1} of occurring).

It is well known that the use of bootstrap resampling techniques may entail a high computational burden. In our particular database, this bur-

den increased even further, since the sample size was large ($n = 16600$). The computational cost that is involved can be considerably reduced by using binning-type acceleration techniques. With the use of these types of techniques, we can considerably reduce computation time and render our procedures operational in practical situations.

Finally, the behavior of the proposed statistical methodology was verified with biological data obtained from a crustacean. In terms of weight gain, in the case of the overall study, individuals were estimated to grow exponentially and thus ensure a high commercial yield until they reached an RC of 21.50 mm. This cutoff point ensures that any barnacle under this size has not yet attained its maximum yield in weight and, in accordance with FAO guidelines (Sparre and Venema, 1997), should therefore not be captured. From this threshold onwards, the accumulated weight of individual specimens will continue to rise with size but the increase in weight from one size to the next will be progressively less, so that the yield obtained ceases to be profitable when seen against the time that the barnacle remains in place without being harvested.

According to this, our testing methods reveal that: (a) Stalked barnacles reach a maximum commercial yield with a size of 21.50 mm; (b) this point or size (rc_0) is the same in both years of the study; and (c) this point or size (rc_0) is different between sites.

Software implementing the nonparametric model estimation (with binning), and the bootstrap-based tests proposed in this paper can be obtained by contacting the first author at sestelo@uvigo.es.

Acknowledgements

The authors gratefully acknowledge the financial support from the project MTM2011-23204 of the Spanish Ministry of Science and Innovation (FEDER support included) and Xunta de Galicia (10 PXIB 300 068 PR).

References

Barnes, M., 1996. Pedunculate cirripedes of the genus *Pollicipes*. Oceanography and Marine Biology: An Annual Review 34, 303–394.

Bernard, F.R., 1988. Potential fishery for the gooseneck barnacle *Pollicipes polymerus* (Sowerby, 1833) in British Columbia. Fisheries Research 6, 287–298.

Bowman, A., Young, S., 1996. Graphical comparison of nonparametric curves. Journal of the Royal Statistical Society. Series C (Applied Statistics) 45, pp. 83–98.

Cadarso-Suárez, C., Roca-Pardiñas, J., Molenberghs, G., Faes, C., Nácher, V., Ojeda, S., Acuña, C., 2006. Flexible modelling of neuron firing rates across different experimental conditions: an application to neural activity in the prefrontal cortex during a discrimination task. Journal Of The Royal Statistical Society Series C 55, 431–447.

Cardoso, A.C., Yule, A.B., 1995. Aspects of the reproductive biology of *Pollicipes pollicipes* (Cirripedia; Lepadomorpha) from the southwest coast of Portugal. Netherlands Journal of Aquatic Ecology 29, 391–396.

Cruz, T., 1993. Growth of *Pollicipes pollicipes* (Gmelin, 1790) (Cirripedia, Lepadomorpha) on the SW coast of Portugal. Crustaceana 65, 151–158.

Cruz, T., 2000. Biologia e ecologia do percebe, *Pollicipes pollicipes* (Gmelin, 1790), no litoral sudoeste português. Ph.D. thesis. Universidad de Évora.

Darwin, C., 1851. A monograph on the subclass *Cirripedia*, with figures of all the species. The *Lepadidae*; or, pedunculated cirripedes. London: The Ray Society.

De Uña Álvarez, J., Roca Pardiñas, J., 2009. Additive models in censored regression. Computational Statistics & Data Analysis 53, 3490–3501.

Dette, H., Neumeyer, N., 2001. Nonparametric analysis of covariance. The Annals of Statistics 29, pp. 1361–1400.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1–26.

Efron, E., Tibshirani, R.J., 1993. An introduction to the Bootstrap. Chapman and Hall, London.

Fan, J., Marron, J., 1994. Fast implementation of nonparametric curve estimators. Journal of Computational and Graphical Statistics 3, 35–56.

Goldberg, H., 1984. Posibilidades de cultivo de percebe, *Pollicipes cornucopia* Leach, en sistemas flotantes. Informes Técnicos del Instituto Español de Oceanografía 11, 1–13.

Golub, G., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215–223.

Härdle, W., Mammen, E., 1993. Comparing nonparametric versus parametric regression fits. The Annals of Statistics 21, pp. 1926–1947.

Hardle, W., Marron, J.S., 1990. Semiparametric comparison of regression curves. The Annals of Statistics 18, pp. 63–89.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. London: Chapman and Hall.

Kauermann, G., Opsomer, J., 2003. Local Likelihood Estimation in Generalized Additive Models. Scandinavian Journal of Statistics 30, 317–337.

Kulasekera, K.B., 1995. Comparison of regression curves using quasi-residuals. Journal of the American Statistical Association 90, pp. 1085–1093.

Molares, J., Freire, J., 2003. Development and perspectives for community-based management of the goose barnacle (*Pollicipes pollicipes*) fisheries in Galicia (NW Spain). Fisheries Research 65, 485–492.

Montero-Torreiro, M.F., Martínez, P.G., 2003. Seasonal changes in the biochemical composition of body components of the sea urchin, *Paracentrotus lividus*, in Lorbe (Galicia, north-western Spain). Journal of the Marine Biological Association of the United Kingdom 83, 575–581.

Neumeyer, N., Dette, H., 2003. Nonparametric comparison of regression curves: An empirical process approach. The Annals of Statistics 31, pp. 880–920.

Pardo-Fernández, J., Van Keilegom, I., González-Manteiga, W., 2007. Testing for the equality of k regression curves. Statistica Sinica 17, 1115–1137.

Roca-Pardiñas, J., Cadarso-Suárez, C., Nácher, V., Acuña, C., 2006. Bootstrap-based methods for testing factor-by-curve interactions in generalized additive models: assessing prefrontal cortex neural activity related to decision-making. Statistics in Medicine 25, 2483–2501.

Ruppert, D., Sheather, S.J., Wand, M.P., 1995. An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association 90, pp. 1257–1270.

Ruppert, D., Wand, M.P., 1994. Multivariate locally weighted least squares regression. The Annals of Statistics 22, pp. 1346–1370.

Sestelo, M., Roca-Pardiñas, J., 2011. A new approach to estimation of length-weight relationship of *Pollicipes pollicipes* (Gmelin, 1789) on the Atlantic coast of Galicia (Northwest Spain): some aspects of its biology and management. Journal of Shellfish Research 30, 939–948.

Sparre, P., Venema, S., 1997. Introduction to tropical fish stock assessment. Part 1. Manual. FAO Fisheries Technical Paper Rev. 2, 420 pp.

Srihera, R., Stute, W., 2010. Nonparametric comparison of regression functions. Journal of Multivariate Analysis 101, 2039–2059.

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman & Hall: London.

Young, S.G., Bowman, A.W., 1995. Non-parametric analysis of covariance. Biometrics 51, pp. 920–931.