

# Flexible quantile regression models: application to the study of the purple sea urchin

Isabel Martínez-Silva<sup>1</sup>, Javier Roca-Pardiñas<sup>2</sup>, Vicente Lustres-Pérez<sup>3</sup>  
Altea Lorenzo-Arribas<sup>1</sup>, Carmen Cadarso-Suárez<sup>1</sup>

---

## Abstract

In many applications, it is often of interest to assess the possible relationships between covariates and quantiles of a response variable through a regression model. In some instances, the effects of continuous covariates on the outcome are highly nonlinear. Consequently, appropriate modelling has to take such flexible smooth effects into account. In this work, various flexible quantile regression techniques were reviewed and compared by simulation. Finally, all the techniques were used to construct the overall zone specific reference curves of morphologic measures of sea urchin *Paracentrotus lividus* (Lamarck, 1816) located in NW Spain.

---

*MSC:* 62G08, 62J02, 62P10.

*Keywords:* Boosting, quantile regression, reference curves, smoothness.

## 1. Introduction

Quantile regression is a statistical technique which allows, among other applications, to calculate growth curves and reference values, and is extremely useful in various fields of application, such as Ecology, Economy and Medicine, examples of which can be seen in Brian (2003), Koenker (2001), González-Barcala (2008), respectively. In the applied field, the need arises to extend the classic parametric approach by using smoothing techniques in regression to capture all the variations that occur in population quantile curves in response to a set of covariates.

---

<sup>1</sup> Departamento de Estatística e I. O., Universidade de Santiago de Compostela (USC), Spain.  
isabelmaria.martinez@usc.es

<sup>2</sup> Departamento de Estatística e I. O., Universidade de Vigo (UVigo), Spain.

<sup>3</sup> Departamento de Zoología y Antropología Física, USC, Spain.

Received: December 2011

Accepted: November 2012



**Figure 1:** Shown at left is a colony of *P. lividus* sea urchins from the intertidal site. Shown at right in orange are the gonads (the part that is marketed).

Quantile regression is used in cases where a study seeks to estimate the different percentiles (e.g., the median) of a population of interest. One advantage of using quantile regression to estimate the median rather than using ordinary least squares regression (to estimate the mean), is that the former is less sensitive to the presence of atypical values. When it comes to using different measures of central trend and dispersion, quantile regression can be regarded as a natural analogue in regression analysis for ensuring a more complete and robust data analysis. A further advantage of this type of regression lies in the possibility of estimating any quantile and thus being able to ascertain what occurs in the case of extreme population values.

In practice there are different methodologies – with freeware implementations developed by the R Development Core Team (2011) – which address quantile regression. To our knowledge, while no general comparative analysis has targeted all of these methodologies, one such analysis has reportedly been conducted by Fenske (2011) on two of them.

Our principal aim was to conduct a comparative study, using simulation and application to real data, to carry out a brief review of a number of currently used flexible quantile regression techniques implemented in R software. Specifically, the following were reviewed: i) Koenker and Basset's methodology in Koenker (1978), using the `quantreg` package; ii) Cole (1988)'s least means squares (LMS) method, represented here in the form of a vector generalised additive model as proposed by (Yee (1996)), using the `VGAM` package; iii) the method based on generalised additive models for location, scale and shape proposed by Rigby (2001) and implemented in the `gamlss` package; and, iv) a new approach to quantile regression using the boosting process described by Buehlmann (2007), with the `mboost` package.

This study is structured as follows: Section 2.1 takes classic quantile regression and extends it to the non-parametric case; Section 2.2 outlines four current methods of non-parametric quantile regression; and Section 3 then makes a comparative study of the different techniques reviewed. The simulation study envisages a non-parametric

scenario that allows for the respective results yielded by the above-mentioned quantile regression techniques to be compared. Lastly, section 4 takes two of the four.

## 2. Quantile regression

### 2.1. Overview

Let  $(x_1, y_1), \dots, (x_n, y_n)$  be a random sample with variable response  $y$  and covariate  $x$ . The problem of parametric quantile regression is thus defined as

$$y_i = \beta_{0\tau} + \beta_{1\tau} x_i + \varepsilon_{i,\tau} \quad \forall i \in \{1, \dots, n\} \quad (1)$$

with  $\beta_{0\tau}, \beta_{1\tau} \in R$  and  $\varepsilon_{i,\tau} \sim H_\tau$  verifying  $H_\tau(0) = \tau$ . The estimated  $\hat{\beta}_{0\tau}$  and  $\hat{\beta}_{1\tau}$  are obtained by solving

$$(\hat{\beta}_{0\tau}, \hat{\beta}_{1\tau}) = \arg \min_{(\beta_{0\tau}, \beta_{1\tau}) \in R^2} \left\{ \sum_{y_i \geq A} \tau |y_i - \beta_{0\tau} - \beta_{1\tau} x_i| + \sum_{y_i < A} (1 - \tau) |y_i - \beta_{0\tau} - \beta_{1\tau} x_i| \right\} \quad (2)$$

Due to the assumption of linearity in the covariate, the above model can be very restrictive in some instances. This constraint can be avoided by replacing the linear index  $\beta_{0\tau} + \beta_{1\tau} \cdot x_i$  with a non-parametric structure. Accordingly, a generalisation of the model in (1) is given by

$$y_i = f_\tau(x_i) + \varepsilon_{i,\tau} \quad \forall i \in \{1, \dots, n\} \quad (3)$$

with  $f_\tau$  being an unknown smooth function and  $\tau \in (0, 1)$ . Moreover, the  $\tau$ -th quantile of the error  $\varepsilon$  conditional on the covariate  $x$  is assumed to be zero, namely,  $Q_\tau(\varepsilon_{i,\tau}|x) = 0$ . Given the sample  $(x_1, y_1), \dots, (x_n, y_n)$  the estimation of  $f_\tau$  is obtained by using some smoother of the form

$$\hat{f}_\tau(x) = \sum_{i=1}^n \omega_{\lambda,\tau}(x_i) y_i \quad (4)$$

where  $\lambda$  is the smoothing parameter and  $\omega_{\lambda,\tau}$  is the function of weights (kernel type, splines, etc.). Some of these methods are now reviewed below.

## 2.2. Methods reviewed

A number of techniques for calculating population growth curves are described in the current literature. Four techniques displaying different approaches and implemented in R software developed by the R Development Core Team (2011), are further discussed below.

### 2.2.1. Linear-programming-based technique

As its starting point, the linear-programming-based (LP-based) approach to the calculation of the quantile reference curve  $\tau$  deems estimations of penalised quantile regression splines to be solutions to the minimisation of:

$$\sum_{i=1}^n \rho_{\tau}\{y_i - g(x_i)\} + \lambda \int \{g''(x)\}^2 dx \quad (5)$$

where  $\rho_{\tau}(u) = u\{\tau - I(u < 0)\}$  is the function *check* proposed by Koenker (1978) and  $\lambda$  is the smoothing parameter of the resulting cubic spline, which generalises the classic approach of least squares smoothing splines pioneered by Wahba (1990). Since the minimisation problem posed entails a high computational cost, in expression (5)  $\{g''(x)\}^2$  is usually replaced by  $|g''(x)|$  (Koenker, 1994). Indeed, this is the approach used in the `quantreg` package. In our study, the `rqss` function was used to estimate the quantile curves, with smoothing being added in the non-parametric case via the `qss` function. No specifications were laid down as to the monotonicity of the data. This was due to the fact that, since the work scenarios encountered by us are not always monotonic, we felt this was something that should be borne in mind when it came to fitting the model.

### 2.2.2. Cole's least means squares method

In this case, the percentile reference curves are calculated on the basis of the distribution of the data. Hence, based on the LMS technique described in Cole (1988), the calculation of the  $\tau$ -th percentile uses Box-Cox family power transformations  $\lambda$  to obtain the pertinent estimates for the mean and standard deviation. In this procedure, one obtains the  $\tau$ -th percentile curve given by the equation

$$Q_{y_i}(\tau|x_i) = M(x_i)[1 + L(x_i)S(x_i)z_{\tau}]^{1/L(x_i)} \quad (6)$$

with  $z_{\tau}$  being the normal equivalent deviate for tail area  $\tau$  and  $L(x)$ ,  $M(x)$  and  $S(x)$  being functions that, as shown in (Cole, 1988), relate to the parameters  $\lambda$ ,  $\mu$  and  $\sigma$  of the distribution of the original simple data. These functions are estimated using vector generalised additive models (VGAM) proposed by Yee (1996) and based on smoothing

splines (Hastie, 1990). To implement this method, we used the VGAM-library `vgam` function.

### 2.2.3. Methodology of generalised linear models for location, scale and shape

The generalised linear models for location, scale and shape (GAMLSS) methodology proposed by Rigby (2005) assumes the structure

$$Q_{y_i}(\tau|x_i) = f_\tau(x_i) + \exp(g_\tau(x_i))z_\tau = \mu(x_i) + \sigma(x_i)z_\tau \quad (7)$$

with  $z_\tau$  as being defined previously and where smoothing is introduced into the estimation of the data-distribution parameters,  $\mu(x)$  and  $\sigma(x)$ , via the functions  $f_{\tau j}$  and  $g_{\tau j}$  using regression B-splines described in Boor (1978). Computational implementation was performed using `gamlss` belonging to the package of the same name. The resulting estimations,  $\hat{\mu}$  and  $\hat{\sigma}$ , are based on B-Spline regression.

### 2.2.4. Boosting algorithms for quantile regression

Calculation of percentile curves based on boosting algorithms (BOOSTING) for quantile regression evolved from boosting algorithms for classification, the best known of which is the AdaBoost described in Freund (1997). Over the following two years, this algorithm was propounded by Breiman (1998, 1999), as a backward stepwise algorithm, known as the functional gradient descent FGD algorithm. Friedman, Hastie and Tibshirani (2000) and Friedman (2001) then carried out statistical developments which enabled the FGD algorithm to be applied to estimating functions, including regression. Subsequently Buehlmann (2007) developed boosting methods for estimation in quantile regression, and more recently, Fenske (2009) propounded the functional gradient boosting algorithm for additive quantile regression. In this approach, the  $\tau$ -th percentile is given by

$$Q_{y_i}(\tau|x_i) = f_\tau(x_i) \quad (8)$$

where the non-linear term of equation (8) introduces smoothing function,  $f_\tau$ , for continuous non-linear covariate  $x$ . In this paper, we fitted this model by means of smoothing P-splines with B-spline bases, using the `mboost` package `gamboost` function for the purpose.

## 3. Simulations

A simulation study was conducted to compare the behaviour of the different quantile regression techniques reviewed. To this end, samples were generated in accordance with the model

$$y = 2 + 1.5 \log(x) + 0.5x\epsilon \quad (9)$$

with errors  $\epsilon_i$  independently and identically distributed, and the covariate  $x$  was generated following a uniform distribution  $U(0,3)$ . One hundred ( $m = 100$ ) independent samples  $\{(x_i, y_i)\}_{i=1}^n$  of size  $n = 400$  were generated from the model (9) with independent random variables distributed following these different scenarios: **Scenario A** Normal standard distribution; **Scenario B** Student t distribution; and **Scenario C** Gamma distribution.

The mean squared error (MSE) and the mean absolute deviation error (MADE) were calculated for the quantile curves corresponding to  $\tau \sim 0.3, 0.5, 0.7$ . These errors are given by the equations (10) and (11) respectively,

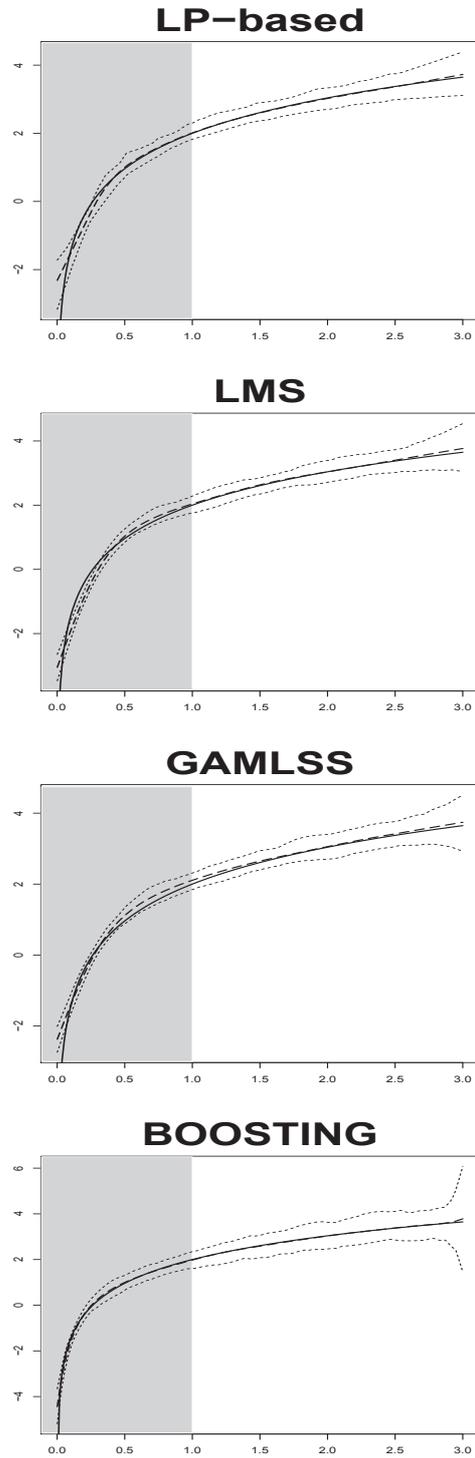
$$MSE = \frac{1}{100} \sum_{j=1}^{100} (\widehat{Q}_\tau^{(j)}(x) - Q_\tau^{(j)}(x))^2 \quad (10)$$

$$MADE = \frac{1}{100} \sum_{j=1}^{100} |\widehat{Q}_\tau^{(j)}(x) - Q_\tau^{(j)}(x)| \quad (11)$$

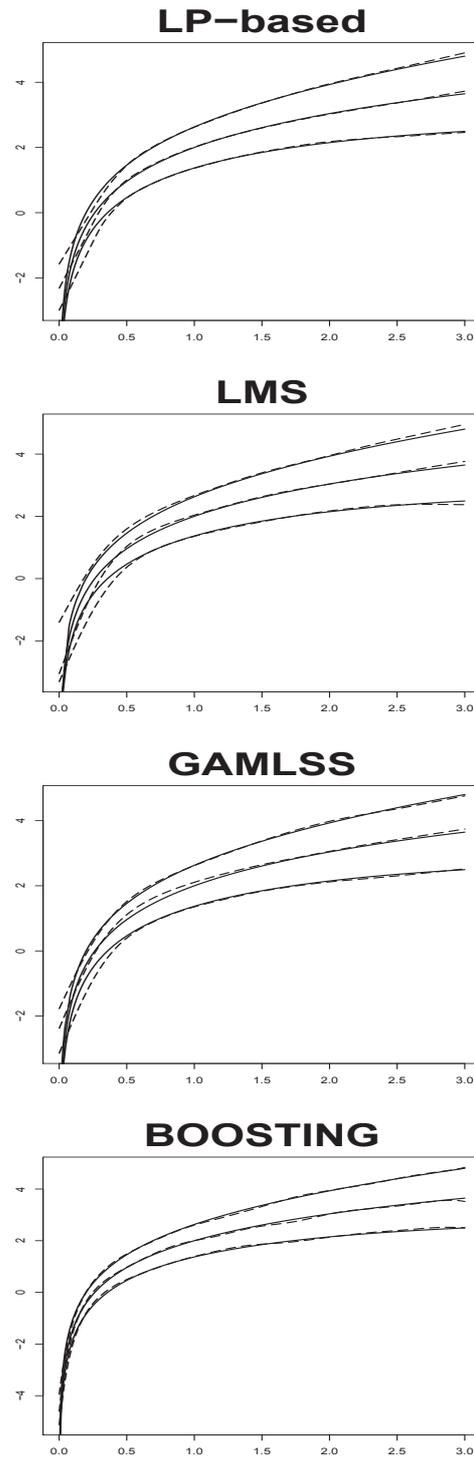
where  $\widehat{Q}_\tau^{(j)}(x)$  is the estimation of the  $\tau$ -th percentile for  $x_i$ ,  $Q_\tau^{(j)}(x)$  is the real value of the  $\tau$ -th percentile for  $x$ .

**Table 1:** This table shows the mean (standard deviation) of the MSE and the MADE for the different methodologies and scenarios in the simulation sample.

Scenario	$\tau$		LP-based	LMS	GAMLSS	BOOSTING
A	0.3	MSE	0.115(0.133)	0.146(0.028)	0.143(0.017)	0.126(0.041)
		MADE	0.254(0.172)	0.204(0.067)	0.201(0.060)	0.241(0.068)
	0.5	MSE	0.110(0.092)	0.151(0.027)	0.171(0.028)	0.119(0.030)
		MADE	0.259(0.165)	0.184(0.061)	0.192(0.055)	0.224(0.056)
	0.7	MSE	0.134(0.115)	0.174(0.028)	0.171(0.031)	0.137(0.048)
		MADE	0.284(0.182)	0.188(0.060)	0.182(0.064)	0.240(0.062)
B	0.3	MSE	0.230(0.064)	0.359(0.043)	0.264(0.049)	0.188(0.069)
		MADE	0.296(0.078)	0.267(0.055)	0.276(0.065)	0.298(0.066)
	0.5	MSE	0.295(0.052)	0.169(0.038)	0.202(0.040)	0.144(0.057)
		MADE	0.256(0.064)	0.219(0.057)	0.235(0.050)	0.248(0.057)
	0.7	MSE	0.376(0.086)	0.227(0.050)	0.226(0.057)	0.193(0.078)
		MADE	0.288(0.077)	0.251(0.050)	0.239(0.064)	0.293(0.064)
C	0.3	MSE	0.191(0.036)	0.515(0.077)	0.771(0.069)	0.120(0.044)
		MADE	0.195(0.045)	0.581(0.080)	0.624(0.076)	0.222(0.060)
	0.5	MSE	0.390(0.086)	0.645(0.071)	0.458(0.084)	0.171(0.067)
		MADE	0.309(0.077)	0.588(0.088)	0.540(0.090)	0.285(0.066)
	0.7	MSE	0.524(0.120)	0.611(0.169)	0.690(0.157)	0.231(0.089)
		MADE	0.404(0.094)	0.981(0.139)	0.942(0.085)	0.340(0.081)



**Figure 2:** The solid line shows the theoretical median curve and the dashed line shows the 95% simulation bands for the different techniques in the Scenario A.



**Figure 3:** The solid line shows the theoretical curve and the dashed line shows the fit for the quantiles  $\tau \sim 0.3, 0.5, 0.7$  using the respective techniques in the Scenario A. Results are shown for the 100<sup>th</sup> simulation.

The results of this study are shown in Table 1 and similar results can be appreciated for all techniques. As can be seen with the MSE criterion: under Scenario A, the LP-based technique presents the lowest mean values and the BOOSTING technique shows the second lowest; for the Scenarios B and C, the BOOSTING technique presents the lowest mean values and the LP-based technique shows the highest mean values in Scenario B. Following the MADE criterion: in Scenarios A and B, the LP-based and BOOSTING techniques present the highest mean values; although under Scenario C the BOOSTING technique shows the lowest mean values in 5<sup>th</sup> and 7<sup>th</sup> percentile and the LP-based technique presents the second lowest, being the opposite in the 3<sup>rd</sup> percentile. When comparing standard deviation for the MSE and the MADE values, can be appreciated that the LMS and GAMLSS technique shows the lowest values in Scenarios A and B although the LP-based and BOOSTING techniques present the lowest values in Scenario C.

As mentioned above, we have not seen a clear winner in Table 1. But when graphing this, a clear change has been noticed and we can see the improvement of working with the boosting methodology. In the graphical presentations, the 95% simulation bands for the median and the quantile curves corresponding to different values of  $\tau$  are shown in Figures 2 and 3.

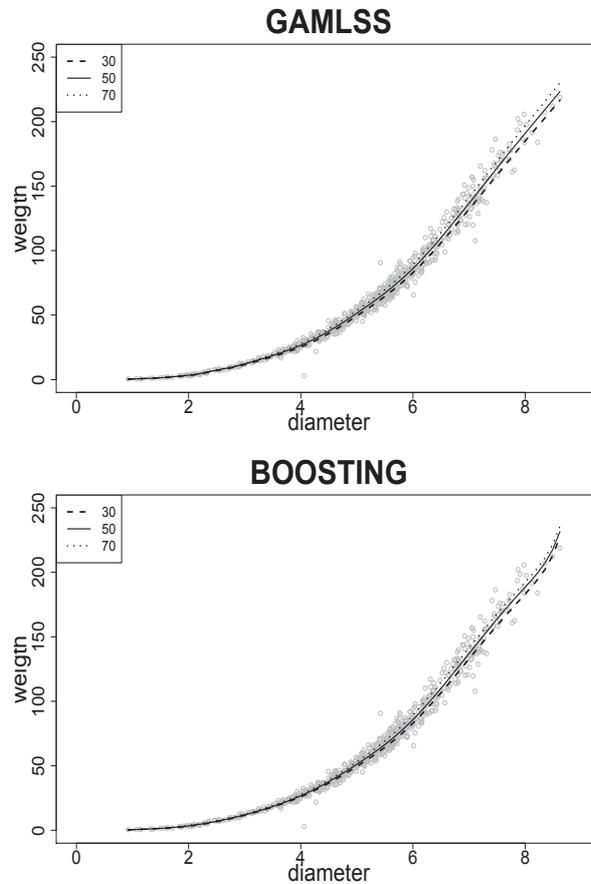
As can be seen from Figure 2, the inability of the LP-based, LMS and GAMLSS techniques to capture the variability of the data completely gave rise to problems in the simulation bands, and in the initial values of the covariate in particular.

When boosting algorithms were used, however, an improvement in the fit was observed across the entire scenario, with this being especially evident in the initial values referred to above. These characteristics can likewise be discerned in the calculation of the percentiles corresponding to  $\tau \sim 0.3, 0.5, 0.7$  shown in Figure 3.

#### 4. Application to the exploitation of marine resources

The study was undertaken at the following two sites along Galicia's Atlantic seaboard (NW Spain): Punta Area das Vacas (42°06'54" N; 008°54'30" W) (intertidal 1) situated on the Vigo estuary (*Ría de Vigo*); and Lago (42°19'25" N; 008°49'37" W) (intertidal 2) located on Aldán Bay (*Ensenada de Aldán*), at the southern edge of the Pontevedra estuary (*Ría de Pontevedra*). Both sites are representative of populations with a great abundance of *P. lividus* on the Galician coast.

Samples were collected from January 2002 to February 2003 along the lower intertidal zone of both sites (intertidal 1 and intertidal 2), and in the sublittoral area of Lago (site 2—sublittoral). The samples were randomly collected, with each comprising a total of 25 specimens of *P. lividus*. A total of 725 specimens were finally studied. The specimens were weighed and measured while fresh. The parameters considered for study purposes were the following two continuous variables: fresh weight, which is a good indicator of the commercial potential of sea urchins and was taken into account by

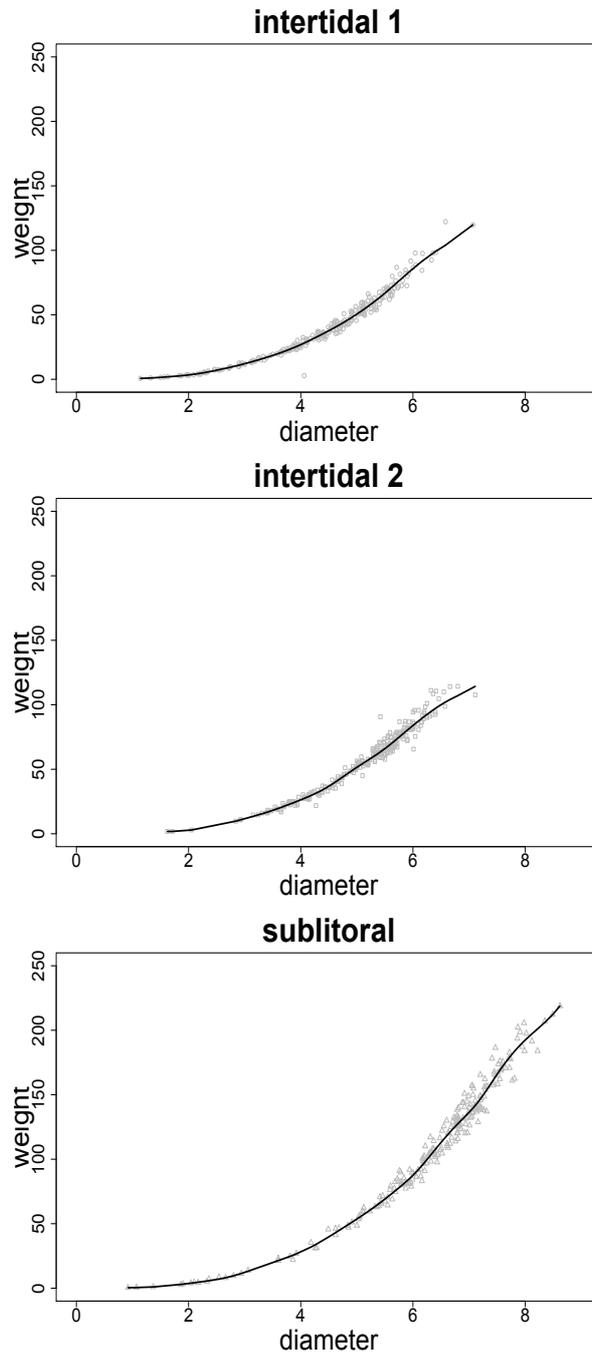


**Figure 4:** Global population: depiction of the fits for the  $\tau^{\text{th}}$  percentiles ( $\tau \in \{30, 50, 70\}$ ) with the GAMLSS and BOOSTING techniques.

**Table 2:** This table shows the values of the estimates obtained at the global sample and at the various sites for different diameter (Diam.) values for the median ( $\tau = 0.5$ ). These estimates were computed using the GAMLSS(G-T) and BOOSTING(B-T) techniques.

Diam.	Global		intertidal 2		intertidal 2		sublittoral	
	G-T	B-T	G-T	B-T	G-T	B-T	G-T	B-T
2.0	3.42	3.46	3.49	3.40	2.90	2.66	3.83	3.68
2.5	7.48	7.05	7.24	7.03	7.03	6.14	7.60	6.74
4.0	26.55	26.75	27.17	27.00	27.38	26.32	32.94	28.10
5.5	67.10	66.47	65.93	66.84	66.74	66.04	71.76	69.11
6.0	86.13	87.75	81.52	85.79	83.22	83.94	91.36	87.55
8.0	191.96	189.43	—	—	—	—	182.15	192.37

being treated as a variable of interest; and diameter which, according to (Lustres-Pérez (2006)), is an indicator of size and strongly correlated with age, and was deemed to be a covariate in the model fitted.



**Figure 5:** Suitable zone by reference to resource exploitation: depiction of the fits for the  $\tau^l$ h percentiles ( $\tau \in \{30, 50, 70\}$ ) with the *BOOSTING* technique.

In order to show the growth of the urchin population at different percentiles, two of the techniques applied to the global population studied are considered. Figure 4 shows

the weight change versus diameter for the  $\tau^{th}$  percentile, ( $\tau \sim 0.3, 0.5, 0.7$ ), to GAMLSS and BOOSTING techniques. The results are similar in both cases but present slight differences as can be seen in Table 2.

Since the sample was collected in three separate locations, the behaviour of previous percentiles in each of the zones has also been studied. In this case only one of the techniques studied, the boosting technique, has been used.

As can be seen in Figure 5 and in Table 3, our results showed that specimens of the sublittoral population displayed important differences with respect to those collected from the two intertidal populations. For any given size, sublittoral sea urchins were thus observed to register higher weights than those that inhabited the intertidal strip, across all the population quantile curves. These divergences increased from the point at which *P. lividus* attained the stipulated commercial size (diameter 5.5 cm). Furthermore the existence of a greater number of larger-sized specimens in the sublittoral population was also in evidence.

**Table 3:** This table shows the values of the estimates obtained at the various sites for different diameter (Diam.) values and for three different percentiles ( $\tau$ ). These estimates were computed using the BOOSTING technique.

Diam.	intertidal 1			intertidal 2			sublittoral		
	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$
2.0	3.15	3.40	3.67	2.67	2.66	2.68	3.30	3.68	3.80
2.5	6.68	7.03	7.43	6.09	6.14	6.08	6.48	6.74	8.15
4.0	26.37	27.00	27.80	25.18	26.32	27.61	26.16	28.10	29.45
5.5	64.45	66.84	68.88	64.08	66.04	68.28	67.48	69.11	73.01
6.0	82.12	85.79	87.06	81.46	83.94	86.60	84.79	87.55	90.86
8.0	–	–	–	–	–	–	186.64	192.37	196.73

## 5. Discussion

The results yielded by the simulation process suggest that the methods are competitive for fitting quantile regression models. Estimation of parameters and selection of variables cannot be made at a single stage of the estimation, nor can the degree of smoothing be selected automatically with the LP-based technique. The LMS method is likewise unable to select the smoothing parameter automatically, is very sensitive to data-dispersion and displays problems when it comes to working with negative-value responses. This last-mentioned aspect makes it necessary for translations to be made before and after fitting the model, to ensure that the results obtained can be properly assessed. As with the two previously described techniques, the GAMLSS methodology requires selection of the degree of smoothing. The boosting-based method is the one which (1) estimates the parameters, (2) selects the variables at a single stage of the estimation, and (3) implements automatic selection of the degree of smoothing. Furthermore, in the light of

the results shown in Figures 2 and 3, among the four methods discussed, the boosting-based method is the one for which the data best fits both small and large values of the covariate. The drawback of this last-mentioned methodology arises due to the fact that the percentile curves are calculated separately, and this leads to problems with the cross-tabulation of quantiles. With respect to application to real data, as Figure 5 and Table 3 show, there is a clear difference between the populations considered. The study confirms that sublittoral populations display conditions better suited to exploitation of *P. lividus*, due to:

- the existence of a greater number of commercial specimens; data corroborated in earlier studies undertaken on the Galician coast, such as those by Fernández-Pulpeiro (1999) and Lustres-Pérez (2006). In the latter case, a study of 206 intertidal and 63 sublittoral sites showed that the percentage of commercial sea urchins was 7% at the intertidal site and exceeded 50% at the sublittoral site; and,
- the greater development of sublittoral versus intertidal sea urchins, i.e., higher weights for any given diameter. This in turn means that during the harvesting periods on the Galician coast (from October to April), the quantity of gonads extracted from each sea urchin (the substance that is marketed) is appreciably higher.

Accordingly, we feel that it would be advisable for exploitation of *P. lividus* to be basically undertaken in the sublittoral area and always in a controlled manner. This would prevent the harvesting of a sizeable quantity of specimens with low commercial yields. Inappropriate extraction leads to a greater depletion of specimens, which limits the regeneration of populations of this echinoderm and, in turn, brings about a greater alteration in coastal ecosystems, bearing in mind the fundamental role that this species plays in the equilibrium of the habitats in which it lives (e.g. Benedetti-Cecchi, 1995; Kitching, 1961 and Ruitton, 2000).

## Acknowledgements

The authors would like to express their gratitude for support received in the form of National Research Projects MTM2008-01603 and MTM2010-09213-E from the Spanish Ministry of Science and Innovation and the Galician Regional Authority Research Project INCITE08PXIB208113PR.

## References

- Benedetti-Cecchi, L. and Cinelli, F. (1995). Habitat heterogeneity, sea urchin grazing and the distribution of algae in littoral rock pools on the west coast of Italy (western Mediterranean). *Marine Ecology Progress Series*, 126, 203–212.

- Breiman, L. (1998). Arcing classifiers (with discussion). *Annals of Statistics*, 26, 801–849.
- Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, 11, 1493–1517.
- Brian, S., Cade, B. and Noon, R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1, 412–420.
- Buehlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, 22, 477–505.
- Cole, T. J. (1988). Using the lms method to measure skewness in the nchs and dutch national height standards. *Annals of Human Biology*, 16, 407–419.
- de Boor, C. (1978). *A Practical Guide to Splines*. Springer.
- Fenske, N., Kneib, T. and Hothorn, T. (2009). *Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression*. Technical report, Department of Statistics University of Munich.
- Fenske, N., Kneib, T. and Hothorn, T. (2011). Childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association*, 106, 494–510.
- Fernández Pulpeiro, E., César Aldariz, J., Lustres-Pérez, V. and Ojea Bouzo, C. (1999). *Ordenación integral del espacio marítimo-terrestre de Galicia. Fauna asociada a sustratos rocosos*. Informe final, 1996-1999. Technical report, Consellería de Pesca, Marisqueo e Acuicultura-Xunta de Galicia.
- Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119–139.
- Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.
- Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics*, 28, 379–407.
- González Barcala, F. J., Cadarso-Suárez, C., Valdés Cuadrado, L., Leis Trabazo, R., Cabanas, R. and Tojo Sierra, R. (2008). Lung function reference values in children and adolescents aged 6 to 18 years in Galicia. *Arch Bronconeumology*, 44, 295–302.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.
- Kitching, J. and Ebling, F. (1961). The ecology of Lough Ine. xi. the control of algae by *Paracentrotus lividus* (Echinoidea). *Journal of Animal Ecology*, 30, 373–383.
- Klein, B. and Korsholm, L., editor (2001). *The GAMLSS Project: a Flexible Approach to Statistical Modelling*.
- Koenker, R. W. and D’Orey, V. (1994). Computing regression quantiles. *Applied Statistics*, 43, 410–414.
- Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15, 143–156.
- Koenker, R. W. and Bassett, G. W. (1978). Regression quantiles. *Econometrica*, 46, 33–50.
- Lustres-Pérez, V. (2006). *El erizo de mar: Paracentrotus lividus (Lamarck, 1816) en las costas de Galicia*. PhD thesis. Universidad de Santiago de Compostela.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics*, 54, 507–554.
- Ruitton, S., Francour, P. and Boudouresque, C. (2000). Relationships between algae, benthic herbivorous invertebrates and fishes in rocky sublittoral communities of a temperate sea (Mediterranean). *Estuarine, coastal and Shelf Science*, 50, 217–230.
- Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.
- Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *Journal of Royal Statistical Society-Series B*, 58(3), 481–493.