Journal of Tropical Forest Science 35(2): 130–140 (2023)
https://doi.org/10.26525/jtfs2023.35.2.130
ISSN: 0128-1283, eISSN: 2521-9847

QUANTILE REGRESSION AS A COMPLEMENTARY TOOL FOR MODELLING BIOLOGICAL DATA WITH HIGH VARIABILITY

Antúnez P[1, *], Wehenkel C[2], Hernández-Díaz JC[2] & Garza-López M[3]

[1]*División de Estudios de Posgrado, Instituto de Estudios Ambientales, Universidad de la Sierra Juárez, Av. Universidad S/N, 68725, Oaxaca, México*
[2]*Instituto de Silvicultura e Industria de la Madera, Universidad Juárez del Estado de Durango, Km 5.5 Carretera Mazatlán, 34120 Durango, México*
[3]*Área de investigación. Sylvatica, S.C., Av. Nichupté-20, Supermanzana 19, 77500 Cancún Quintana Roo, México*

**pantune@uni-goettingen.de*

Submitted January 2022; accepted August 2022

Biological data are usually heterogeneous, fluctuating and have outliers. When these data are examined using least squares regression models, difficulties often arise in identifying the impact of regressors on specific segments of the point cloud. Traditional models cannot be applied when regression assumptions are not met. The objective of this study was to examine the robustness of quantile regression (QR) in modelling data with high presence of extreme values. Using QR and least squares methods (ordinary and non-linear), we evaluated the change in biomass contents in different organs of mahogany (*Swietenia macrophylla*). The results suggest that QR significantly reduces the mean absolute error and the leverage effect. It also identifies the unit impact of the regressor on a specific quantile of the distribution. One of the main novelties of this approach was that greater interpretative capacity was possible for the different sectors of the conditional distribution, especially for those points far from the mean and the median, revealing more detailed behavioural patterns of the response variable. With this information, the rate of change of one variable due to the unit change of the other is more clearly understood.

Keywords: Biomass prediction, correction of heteroscedasticity, leverage effect correction, forest modelling, reducing the effect of outliers

INTRODUCTION

Regression analysis using least squares method in its linear or non-linear form is widely used in forestry to model variables of interest as functions of several predictors. Examples of such variables include biomass or carbon accumulated in plant biomass (Chave et al. 2005), timber volume available for industrial harvesting (Zianis et al. 2005), and many other variables of interest to foresters. The input data in model building in forestry or similar areas are often heterogeneous, fluctuating and have outliers. Moreover, acquiring these data can be costly for forest owners or managers (Robinson & Hamann 2010).

It is highly desirable to have analytical tools that allow users to optimise resources in forestry. Ordinary least squares (OLS) is one of the most widely applied methods in this field. However, when using OLS, some assumptions must be met, including homogeneity of the error variance and independence of errors. Also, although it is not essential, it is desirable that the independent variables have a normal distribution (Osborne & Overbay 2004, Osborne 2013). If these assumptions are not met, the use of OLS is not appropriate (Cade & Noon 2003). Faced with this challenge, the data analyst can choose to either transform the data in order to meet the OLS requirements or apply another method. Forcing the application of OLS can create problems such as the construction of weak models due to the leverage caused by extreme values (Everitt & Skrondal 2010).

The use of quantile regression (QR), has only been scarcely applied in forestry to make predictions, despite giving successful results (Mäkinen et al. 2008, Bohora & Cao 2014, Cao & Dean 2015, Gao et al. 2017, Zhang et al. 2020). In fact, QR could be a complementary technique to the use of OLS. In other fields, QR has been used to examine the relationship between environmental factors and live organisms

(animals or plants) (Schröder et al. 2005). It is also used to study density changes in live organisms as a function of ecological patterns (Cade et al. 1999, Cade & Noon 2003); diversity and invasiveness of plant communities (Brown & Peet 2003); prediction of the maximum growth rates of marine phytoplankton (Bissinger et al. 2008); and prediction of the probability of biological impairment based on habitat assessments (Doll 2011). In recent years, proposals have emerged to strengthen the analysis capabilities of QR, including the Bayesian approach (Yu & Moyeed 2001) and the use of machine learning algorithms in random forests (quantile regression forests) (Meinshausen 2006). The use and application of QR has now been extended to the construction of prediction intervals for finite samples, without making distributional assumptions (Romano et al. 2019).

QR estimates a regression line for each quantile of interest in the distribution of a response variable in the model (Koenker 2000). It, therefore, provides a more accurate picture of the possible strong relationship between variables, allowing more precise calculations of growth curves and other reference values for estimating functional associations between dependent and independent variables (Cade & Noon 2003). QR has advantages compared to conventional regression techniques, for example, in-depth insight into the effects of the covariates that are often missed with conventional linear regression. QR allows construction of prediction intervals, fulfilling at all times the bases of inferential statistics, including regression, robustness and extreme value theory (Yu et al. 2003, Benoit & Van den Poel 2009, Das et al. 2019).

In this study we intended to help answer some questions that a forest analyst often faces such as: (1) how to build a robust model when the data show several extreme values, (2) how to build a robust model when the data are widely fluctuating or if the residuals do not follow a normal distribution, and (3) which is the impact of a regressor at a specific point or segment in the variable of interest. By answering these questions, the information generated by the analyst is important to improve the decision-making capabilities of the forest manager. As a contribution to achieve this objective, in this work, we examined the efficacy of QR to predict the number of leaves and the biomass of mahogany (*Swietenia macrophylla*).

MATERIALS AND METHODS

Variables and sample

A random sample of 1000 nine-month-old seedlings of *S. macrophylla* were used for this study. We measured four variables, namely, diameter at the stem base (DB, mm), diameter of the stem at the height of the first live leaf or twig (DFL, mm), total height (TH, cm) and length of the leafless stem (LSL, cm). These data were obtained from a plant nursery located at the Higher Technological Institute of Venustiano Carranza, north-west of the State of Puebla, Mexico (20° 30' N, 97° 40' W), at an average elevation of 113 m above sea level. This nursery is located within the Neotropical ecoregion, on the north-eastern coastal plain of the Gulf of Mexico. The climate is tropical, with rains during seven months of the year (mainly in summer) and mild temperature oscillations. The diameters of the seedlings were measured to the nearest 0.5 mm, while the heights were approximated to the nearest millimeter. A total of 42 independent specimens were selected at random and destructively sampled to measure the biomass components. For this purpose, these specimens were separated into three components, namely, (1) stems, (2) roots and (3) leaves and/or twigs. Each component was weighed fresh and after oven drying at 60 ºC (since seedlings are softer than wood) to constant weight, using an analytical balance with approximation to 0.01 g, following the standard method (Brown & Lugo 1984, Acosta et al. 2002).

Data analysis

To examine the relationship between the dependent and independent variables listed in Table 1, we followed the steps below:

(1) Independent variables that had the strongest association with the dependent variables were identified. For this, we used the non-parametric covariation coefficient C (equation 1), proposed by Gregorius et al. (2007):

$$C = \frac{\sum_{i<k}(X_i - X_j).\,(Y_i - Y_j)}{\sum_{i<j}|(X_i - X_j).\,(Y_i - Y_j)|} \quad (1)$$

Table 1 Descriptive statistics of *Swietenia macrophylla* variables used in the analyses

Variable	Unit	Variable type	Min	Max	Mean	SD	Sk	SW p-value
			Data used for modelling the number of leaves					
Number of leaves	-	DV	1.00	22.00	9.01	3.51	0.60	< 0.0001*
Diameter at the stem base (DB)	mm	IV	1.90	6.90	4.14	0.88	0.14	0.0166*
Total height (TH)	cm	IV	9.00	56.00	37.03	8.56	-0.75	< 0.0001*
			Data used for biomass modelling					
Diameter at the stem base (DB)	mm	IV	2.32	5.40	4.39	0.93	-1.03	0.0185*
Total height (TH)	cm	IV	6.500	11.600	7.479	1.608	0.004	0.0205 *
Diameter of the stem at the height of the first live leaf or twig (DFL)	mm	IV	2.04	4.88	3.56	0.91	-0.45	0.1285
Length of the leafless stem (LSL)	cm	IV	11.00	30.50	20.44	5.82	-0.20	0.0790
Leaf and twig biomass	g	DV	0.080	1.630	0.912	0.408	-0.135	0.694
Stem biomass	g	DV	0.110	1.470	0.729	0.303	-0.070	0.760
Root biomass	g	DV	0.150	2.390	1.210	0.568	0.050	0.725
Total biomass	g	DV	0.710	5.130	2.851	1.193	-0.111	0.425

SD = Standard deviation, Sk = skewness value, SW = Shapiro–Wilk's test, * = data does not follow a normal distribution ($p < 0.05$), DV = dependent variable in the regression, IV = independent variable in the regression

The C values range from -1 to +1, with C equals to 1 for strictly positive covariation and C equals to -1 for strictly negative covariation (Gregorius et al. 2007). It should be noted that, prior to the correlative study, the raw data showed high variability and marginal distribution of both variables was not uniform (as the dimensions of the plants increased, the variability also increased). Table 1 provides an overview of such data.

(2) Each dependent variable was modelled as a function of its selected predictor (in the previous step) using QR and the results were compared to those of least squares methods in both its OLS and non-linear forms (NLS). The *lm* and *nls* functions of R were used to adjust OLS and NLS respectively. The fitted non-linear equation was taken from the general allometry with two parameters of the form $y = ax^b$, proposed by Huxley (1950).

The QR, on the other hand, was computed with the *rq* function of the *quantreg* package of R, generating regression lines in the quantiles 0.10. 0.25, 0.50, 0.75 and 0.95, where the syntax was rq(y ~ x, data = dataset, tau = τ^{th} quantile). For quantile 0.10, the R code was rq(y ~ x, data = dataset, tau = 0.10). Examples for both, fitting and obtaining the graphs in R, can be found in Koenker (2013).

Equations (2) and (3) reproduce the initial mathematical reasoning of the quantile regression. Let Y be a random variable, characterised by its original distribution function,

$$F(y) = \text{Prob}\ (Y \leq y) \qquad (2)$$

Let τ be a particular Y value, then for any τ the following expression is called the τ^{th} quantile of X,

$$Q(\tau) = \inf \{y:F(y) \geq \tau\} \qquad (3)$$

Similar to the original distribution function, the quantile function also provides a complete characterisation of the random variable (Y) (Koenker 2000). Finally, the quantile function may be formulated as the solution to a simple optimisation problem. Full description of the QR equation, the graphic illustration of its verification function, and other expressions derived from equations (2) and (3), are presented and explained in Koenker and Bassett (1978), Koenker (2000) and Koenker and Hallock (2001).

(3) To find the most suitable quantile model for the different sectors of the conditional distribution, we identified which τ^{th} model minimised the prediction error for each of the observed values. The label "a" was assigned to the i^{th} value where the error was minimal with $\tau = 0.10$. The label "b" was used for the i^{th} value where the error was minimal with $\tau = 0.25$, and so on for the other three modelled quantiles and so, for $\tau = 0.95$, the i^{th} value with the minimum error was labelled with the letter "e".
(4) Parametric estimators were recalculated for each sector of the conditional distribution (pooled in the previous step) and are reported in the results section as QR coefficients.
(5) To assess the robustness of the models, we analysed graphics and numerical results for evidence of the significant contribution of each independent variable in each model. As goodness of fit, we calculated the root mean square error (RMSE), mean absolute error (MAE) (in the traditional way using residuals), the quantile–quantile plots (Q–Q plots) and the Akaike information criterion (AIC) (Akaike 1974). In addition, we generated confidence and prediction bands (95%) for each regression line.
(6) Finally, we compared the averages of the MAEs by the Kruskal–Wallis test, to demonstrate whether or not there was any significant difference between the MAEs given by QR and the least squares models.

In order to reduce the probability of obtaining false positive results or Type I errors, in all cases, we used Bonferroni corrected significance levels (Huberty & Morris 1989, Koenker 2013). For this, we divided the original α value ($\alpha = 0.05$) by the number of hypotheses (m) in order to get the Bonferroni corrected values (Hochberg 1988). This significance level was only used as a reference value and not as a rigid dichotomous rule of rejection or non-rejection. To assess the statistical insignificance of a predictor, the p-value against the null hypothesis was observed. A value sensibly close to zero suggests very low evidence against the null hypothesis (Antúnez et al. 2021).

RESULTS

Examination of the size of the correlation coefficients (C) (Table 2) shows that DFL is the regressor most closely correlated with the response variables, followed by DB. By contrast, LSL was the least correlated regressor with any dependent variable, although it was significant in relation to the number of leaves (Table 2). Since there was a high correlation between DFL and DB, as well as between TH and LSL, we included only one of these pairs in the regression model (the one with the highest correlation coefficient with the dependent variable).

Tables 3 and 4 show estimators of the parameters of the regressors filtered by the size of significant coefficient C. These parameters indicated the impact that a marginal change in each explanatory variable had over plant biomass and the number of leaves, while maintaining the remaining explanatory variables constant.

Figures 1a and b show examples of relationship between one plant attribute and one variable of interest, according to the unitary effect of each covariate in each quantile. The horizontal axes of Figures 1a and b show the quantile values, and the vertical axes show the initial parameters estimated by QR. The contours of the scatter

Table 2 Values of correlation (C) and their respective probability values

Variable	Number of leaves		Leaf and twigs biomass		Root biomass		Stem biomass		Total biomass	
	C	p-value	C	p-value	C	p-value	C	p-value	C	p-value
DFL	0.579	0.02	0.680	< 0.0002*	0.724	< 0.0002*	0.600	0.002	0.737	< 0.0002*
DB	0.556	< 0.0002*	0.530	< 0.0002*	0.635	< 0.0002*	0.601	< 0.0002*	0.641	< 0.0002*
TH	0.467	< 0.0002*	0.631	< 0.0002*	0.582	< 0.0002*	0.284	1.3	0.570	< 0.0002*
LSL	-0.180	< 0.0002*	0.240	2.42	0.292	1.14	0.125	8.52	0.501	1.98

DFL = diameter of the stem at the height of the first leaf or live twig, DB = diameter at the base, TH = total height, LSL = leafless stem length, C = covariation coefficient with 10,000 permutations; * = significant coefficients (after Bonferroni correction)

Table 3 Parametric and goodness-of-fit indicators for each one of the adjusted models

Variable and adjustment indicator	τ = 0.10	τ = 0.25	τ = 0.50	τ = 0.75	τ = 0.95	OLS	NLS
Model to predict the number of leaves							
DFL	1.537*	2.079*	2.548*	3.171*	4.082*	2.614*	4.958*
AIC	435.087	233.724	429.326	553.946	496.156	4867.388	4959.604
RMSE	0.915	0.420	0.553	0.796	1.654	2.849	2.986
MAE	0.732	0.350	0.464	0.664	1.188	2.229	2.359
Model to predict leaf and twig biomass							
DFL	0.134*	0.227*	0.265*	0.307*	0.359*	0.262*	0.068*
AIC	-3.425	-27.143	-25.003	-20.081	-12.777	10.775	2.452
RMSE	0.160	0.050	0.039	0.063	0.045	0.263	0.237
MAE	0.122	0.043	0.033	0.054	0.034	0.204	0.178
Model to predict root biomass							
DFL	0.157*	0.275*	0.347*	0.419*	0.491*	0.352*	0.092*
AIC	-4.772	-13.283	-27.339	-19.087	-2.667	33.608	13.136
RMSE	0.129	0.082	0.072	0.067	0.105	0.347	0.270
MAE	0.086	0.069	0.065	0.056	0.086	0.280	0.222
Model to predict stem biomass							
DB	0.094*	0.136 *	0.166 *	0.200*	0.279*	0.170*	0.037*
AIC	-15.819	-12.756	-39.046	-35.744	2.316	-9.070	-24.356
RMSE	0.080	0.045	0.040	0.052	0.158	0.206	0.171
MAE	0.070	0.040	0.036	0.042	0.159	0.160	0.128
Model to predict total biomass							
DFL	0.456*	0.708*	0.857 *	0.943*	1.020*	0.823*	0.214*
AIC	-1.094	-0.277	-13.649	-12.807	3.782	85.007	66.588
RMSE	0.175	0.195	0.108	0.059	0.237	0.649	0.519
MAE	0.150	0.152	0.096	0.051	0.175	0.531	0.435

τ = model of the τ^{th} quantile adjusted by quantile regression (QR), OLS = model adjusted by ordinary least squares, NLS = model adjusted by non-linear least squares, DFL = diameter of the stem at the height of the first live leaf or twig, DB = diameter at the base, AIC = values of the Akaike information criterion, RMSE = root mean square error, MAE = mean absolute error; * = significant coefficients; for NLS, the variables are transformed (exp, log, square root, squared and cube); only results of the models that show the lowest mean absolute error are reported here and, in most cases, it happened when the predictor was raised to the squared

Table 4 Parametric and goodness-of-fit indicators for allometric model to predict the number of leaves as a function of the total height of 1000 nine-month-old seedlings of *S. macrophylla*

Model constant and adjustment indicator	τ = 0.10	τ = 0.25	τ = 0.50	τ = 0.75	τ = 0.95	NLS
a	0.085*	0.382*	0.646*	0.748*	0.913*	0.585*
b	2.249*	1.607*	1.441*	1.489*	1.535*	1.516*
AIC	421.872	284.981	481.510	579.121	506.718	5027.397
RMSE	0.859	0.484	0.606	0.836	1.636	3.088
MAE	0.710	0.423	0.530	0.704	1.258	2.435

τ = model of the τ^{th} quantile adjusted by non-linear quantile regression (QR), NLS = model adjusted by non-linear least squares, letters a and b = constants of the general allometric equation, AIC = value of the Akaike information criterion, RMSE = root mean square error, MAE = mean absolute error; * = significant constants after Bonferroni correction

plot, outlined in grey, indicate the lower and upper limits of the 95% confidence bands for each estimator derived from QR. The middle horizontal line represents the mean, estimated by the OLS method, and the two extreme dashed lines correspond to the upper and lower limits of the 95% confidence intervals of this estimator. Due to heteroscedasticity, the values of the parameters estimated for the QR models are different for each quantile (Tables 3 and 4). However, the QR parameter estimated at the median ($\tau = 0.5$) was generally close to the line of the estimator of the OLS mean (Figures 2a–d).

Most parameters estimated by QR fell within the confidence interval of the OLS estimator of the conditional mean, like the diameter at the base against the stem biomass (Figure 1a), but not so in diameter of the stem at the height of the first live leaf against the number of leaves (Figure 1b).

Observing the adjustment indicators of the models, AIC, RMSE and MAE, it can be seen that the QR indicators of any line corresponding to any τ^{th} quantile are much better than the values obtained for the corresponding general model adjusted by OLS or NLS (Tables 3 and 4). Likewise, the Kruskal–Wallis test showed significant reduction in the MAE given by QR models compared with the OLS and NLS. Differences of the means are significant in both cases considering a level of significance of 0.05 (Figure 3).

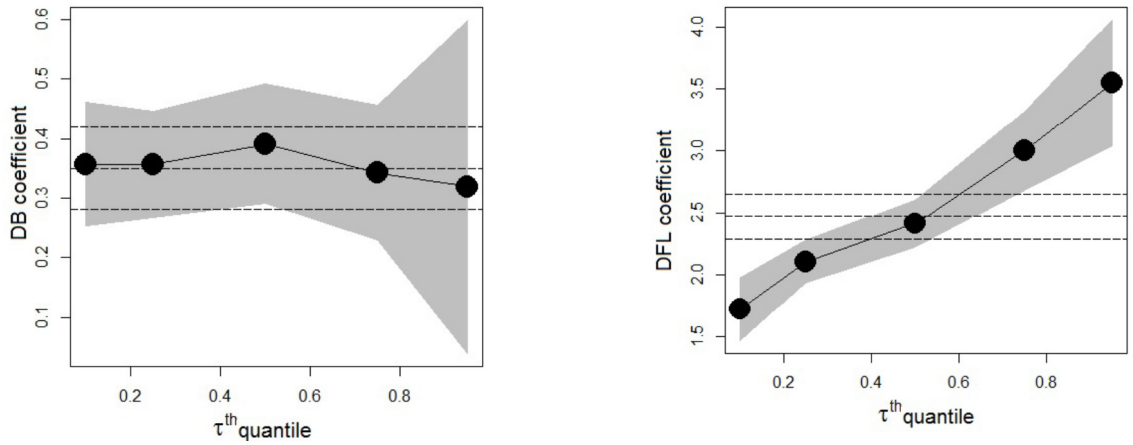
The values of AIC indicate that the joint effect of a given set of covariates differs greatly between different quantiles (Table 3). For instance, in the model of number of leaves the AIC values were 233.724 and 429.326 respectively for the quantiles 0.25 and 0.50, while much larger AIC values were obtained for the 0.75 and 0.95 quantiles (553.946 and 496.156 respectively). The AIC values were also used to determine the quantile which had a better fit. Based on their AIC values, the models of the 0.25, 0.50 and 0.75 quantiles were more robust than the models of the 0.10 and 0.95 quantiles (Table 3).

In addition to the numerical adjustment indicators (Tables 3 and 4), evaluation of the predictive power of each QR model by a Q–Q plot showed that most of the residual values of the models were within the 95% confidence range. Yet, high variability of the data increased the prediction error, as is the case of Figure 4a, where the difference between the QR modelled distribution (dotted line) and the ideal distribution (straight central line) was minimal. Conversely, in Figure 4b, the difference was greater, and many values were outside the confidence band.

Finally, when comparing the residuals in a single plot, those of QR had homogeneous distribution while those obtained by OLS and NLS showed heteroscedasticity. To illustrate this, using stem biomass data (continuous data), it was observed that extreme quantile models (0.10, 0.95) reduced the effects of outliers (Figure 5).

DISCUSSION

In this study, we modelled the response of variables against changes in predictors, even



(a) (b)

Figure 1 Parametric coefficients estimated by QR for (a) diameter at the base (DB), to predict stem biomass, and (b) diameter of the stem at the height of the first live leaf or twig (DFL), to predict number of leaves of *Swietenia macrophylla*; confidence bands (shaded area) of the QR in comparison with the 95% confidence intervals of the mean (dashed lines) estimated by ordinary least square

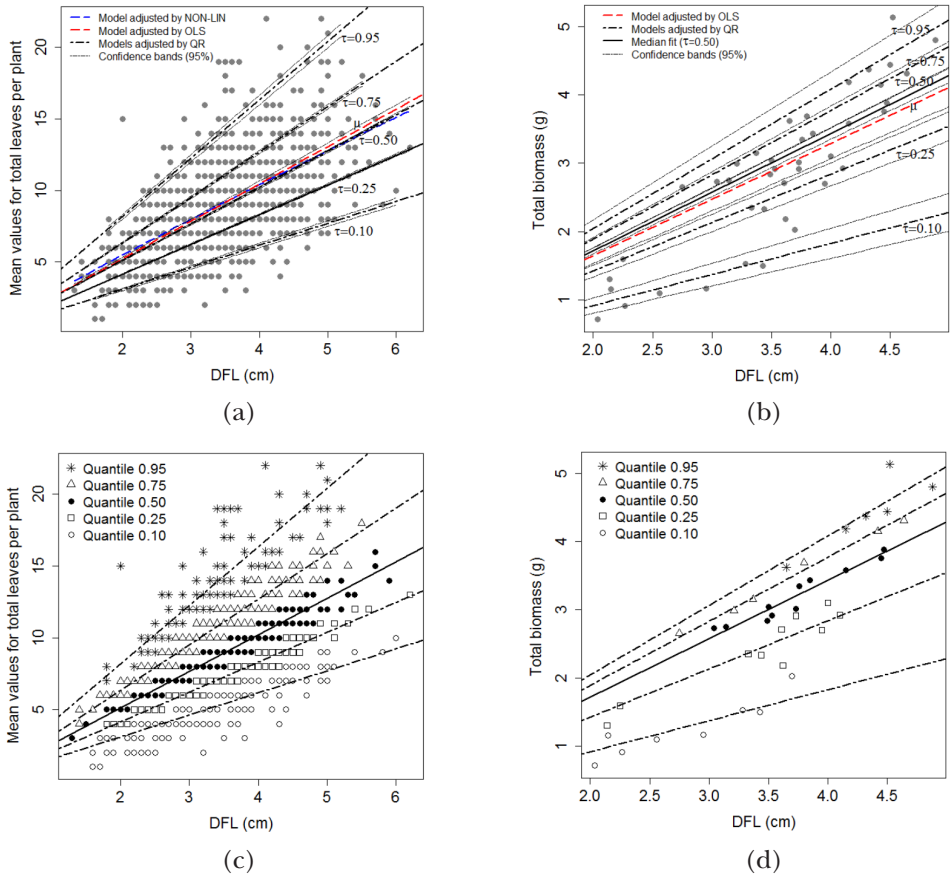


Figure 2 Scatter plots and regression lines fitted to different quantiles (τ) for (a) number of leaves as a function of the diameter of the stem at the height of the first live leaf or twig (DFL), (b) total biomass as a function of the diameter of the stem at the height of the first live leaf or twig (DFL), (c) number of leaves as a function of the diameter of the stem at the height of the first live leaf or twig (DFL), splitting the scatter plots according to the zone of influence of each QR; (d) total biomass as a function of the diameter of the stem at the height of the first live leaf or twig (DFL), splitting the scatter plots according to the zone of influence of each QR; each subset must be predicted using the parameters of the corresponding quantile line of that data segment

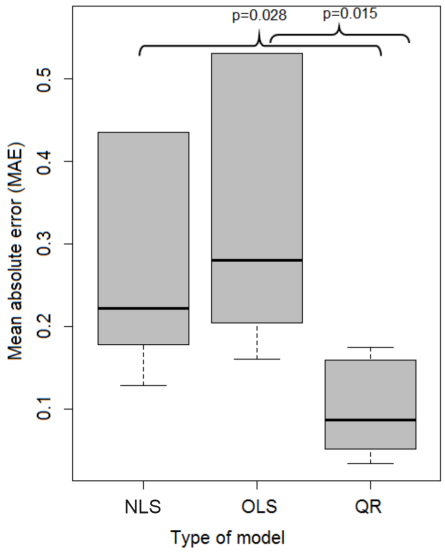
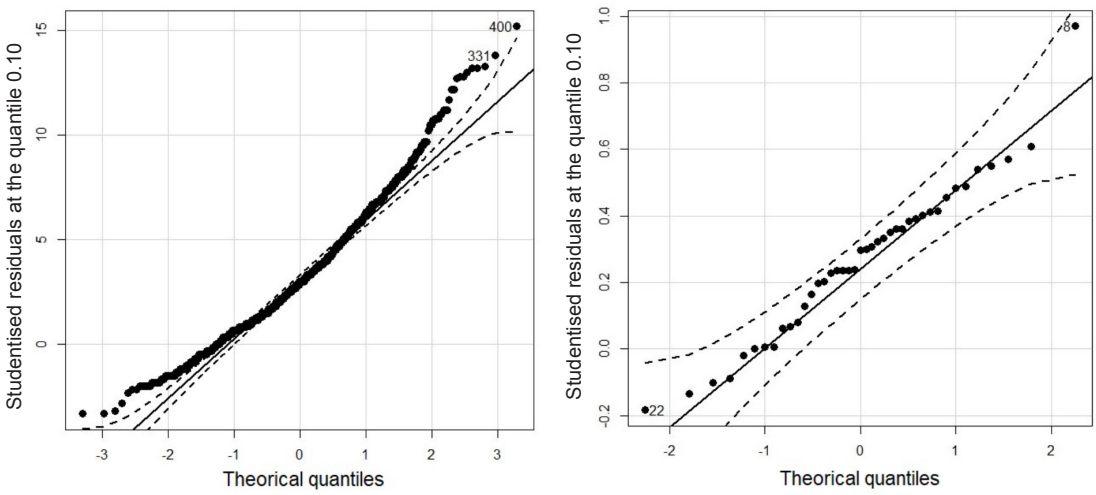


Figure 3 Comparison of averages of the mean absolute errors (MAEs) by the Kruskal–Wallis test; OLS—MAE of models adjusted by ordinary least square, NLS—MAE of models adjusted by non-linear least square and QR—MAE adjusted by quantile regression



(a) (b)

Figure 4 Examples of quantile–quantile plot for a variable of interest predicted as a function of a predictor; (a) number of leaves predicted as a function of the diameter of the stem at the height of the first live leaf or twig in the quantile 0.10, and (b) stem biomass predicted as a function of diameter at the base in the quantile 0.10

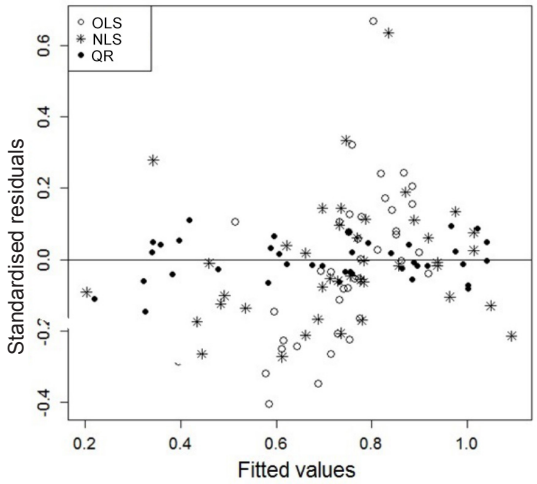


Figure 5 Example of residuals plot versus predicted values that resulted in predicting the stem biomass as a function of diameter at the base; OLS—residuals obtained by ordinary least squares, NLS—residuals obtained by non-linear least squares and QR—residuals obtained by quantile regression

within the extreme values of the distribution and with highly variable data (Table 1). For this purpose, we generated regression lines for each portion (quantile) of the distribution, including central and extreme quantiles (Figures 2a–d). This approach is more accurate than the use of a unique regression line to explain the effect of a predictor on a specific zone of the conditional distribution (Tables 3 and 4). This could be considered an advantage over the OLS method. If there is a lot of extreme points, it implies that more QR lines must be estimated to reduce the overall error (Koenker & Hallock 2001, Koenker 2013). In such cases, more effort, time and computer analyses are involved, which could be considered as disadvantages. However, once the parameters were estimated and corrected, the prediction error was significantly reduced, regardless if the relationship was linear (Table 3) or non-linear (Table 4).

Various authors have explored different indicators to ascertain the goodness of fit for QR. For example, Van Keilegom et al. (2008) suggested tests based on empirical distribution

of the residuals, while Fan et al. (2002) and Fan and Jiang (2007) proposed tests based on the verisimilitude function. In our study, we presented the MAE and the RMSE values which had intuitive interpretation. We also presented the AIC which measured the relative quality of the adjustment of each quantile regression. AIC allows measuring the difference between the projection of the model and the reality based on the theoretical criterion of minimum information (Akaike 1974). We are not only contrasting the results given by QR and least squares models, but also evaluating the behaviour of each QR model generated for each τ^{th} quantile (Tables 3 and 4). For instance, in reference to total biomass, AIC values fluctuated between -13.649 and 3.782 units, with a range of only 17.43 units between the highest and lowest values, suggesting a moderate and relatively similar goodness of fit of the models in each tested quantile. This finding is consistent with the other adjustment indicators (Table 3). Thus, the AIC values are useful and very practical to make a decision on generating more regression lines or not.

The lines modelled for different quantiles were not always parallel to each other (Figures 2a and b). This seems to respond to an irregular variance of the conditional distribution causing heteroscedasticity. The lines were equidistant if the distribution was uniform no matter if variability was high or low (see the quantile lines near the median of the Figures 2a–d). For a very irregular variance, a possible alternative would be to increase the number of samples to see if the data maintain the same distribution pattern (Ford 2015). Unfortunately, in this study, it was not possible to increase the sample size anymore because the owners of the forest nursery would not authorise further destructive analysis due to the difficulty of collecting seeds of the species studied. Therefore, the portion of the data in which the given quantile line is able to predict the variable of interest must be effectively delimited (Nava-Nava & Antúnez 2018).

Another novelty of this approach was that the estimated coefficients for each model showed changing trends between different quantiles. In some cases, there was a direct relationship, while in other cases, an inverse relationship (e.g. Figures 1a and b). This indicates that the variable of interest can respond positively (increasing its value) as the value of regressor increases, but another variable integrated in the model can have an opposite impact in some segments of the conditional distribution (Figures 2a–d). The latter is extremely important for an effective understanding of the relationship between the variables analysed.

That is probably the most valuable contribution of the QR method under the approach presented in this paper, since it attempts to describe the fluctuating nature of the biological modelled data. Also, the type of relationship (decreasing or increasing) and magnitude (reflected in the parameters) can help to identify the conditions under which an explanatory variable contributes significantly to the goodness of fit of a model (Koenker & Hallock 2001). For instance, for predicting the number of leaves, the DFL coefficients found were positively related (Figure 1b). This suggests a stronger relationship in higher than in lower quantiles (Cade & Noon 2003). In this way, the graphic result of QR shows the impact of a given regressor at a specific point or segment in the distribution of a variable of interest. In contrast, OLS and NLS did not clearly describe the impact of the predictor in those particular sectors of the distribution, since both drew a single line based on the mean (Figures 1a and b).

In general, regardless of the widely dispersed data, the differentiated QR prediction models constructed for different quantiles yielded substantial reduction in the error magnitude and the leverage effect of outliers, in comparison with the OLS method (Tables 3 and 4). Hence, QR could be one important tool to be considered in order to better explain the variables of interest in forestry, where the data are highly fluctuating and where it is often necessary to use the least possible number of predictors to reduce costs. QR can certainly be very useful as a complementary tool to the conventional OLS model. Although the effect of autocorrelation or multicollinearity is difficult to avoid when several attributes of living beings are simultaneously modelled, such effect may be reduced by including as few predictors as possible. Sometimes it becomes a challenge for the forest modeller to use information from the same individuals as regressors (diameter, height, crown height, diameter at the base or any other attribute of the same plant), which correlate with each other either in minor or greater scale, violating one of the criteria of standard regression models, which is the independence of regressors.

Finally, even though a significant difference was found between the average MAEs of the contrasted methods, results reported here are susceptible to improvement and could even be complemented with other analyses, such as the regression of the median (Ying et al. 1995, Koenker & Hallock 2001), non-parametric multiplicative regression (McCune 2006), and smoothing model, e.g. using penalised splines (Takeuchi et al. 2006). Likewise, a Bayesian approach to QR can improve modelling efficiency by using the asymmetric Laplace distribution to perform the likelihood function in a generalised linear model context (Yu & Moyeed 2001, Lancaster & Jun 2010, Feng et al. 2015).

The main advantage of the approach presented here is that it provides a greater interpretability of different sectors of the conditional distribution, especially for points away from the mean and median, revealing patterns of behaviour in a more detailed way. In addition, it also describes the influence of the independent variable on the range of the dependent variable and the shape of the conditional distribution (Wang & Jv 2021). It is also possible to capture the tail characteristics of the distribution.

The QR technique is gradually being used as an integral method of analysis in both linear and non-linear models (Yu & Moyeed 2001). Taking into consideration some drawbacks of the OLS technique, such as the greater investment of time (especially, if there are several predictors), we suggest using QR as a complementary tool to evaluate more accurately the relationship between variables and to predict the variables of high interest in forestry, plant ecology or related areas.

In particular, it is appropriate to use QR if the distribution of data is asymmetric, or if there is a high presence of outliers producing bias due to their leverage effect, or if the residuals do not follow a normal distribution. Using QR to generate constants for specific segments of the data, the reduction of bias is evident, regardless of whether there is a linear or curvilinear relationship. Therefore, QR gives more specific information about the properties of the relationship between each variable of interest and its respective predictors.

ACKNOWLEDGEMENT

Thanks to Salas Zúñiga A, Ortiz Coyular E and Jiménez Guzmán A for their support in data collection.

REFERENCES

Acosta MMA, Vargas HJ, Velázquez MA & Etchevers BJD. 2002. Estimación de la biomasa aérea mediante el uso de relaciones alométricas en seis especies arbóreas en Oaxaca, México. *Agrociencia* 36: 725–736.

Akaike H. 1974. A new look at the statistical model identication. *IEEE Transactions on Automatic Control* 19: 716–723.

Antúnez P, Rubio-Camacho EA & Kleinn C. 2021. Prueba de hipótesis en la investigación forestal, agropecuaria y en la ecología: retos y malentendidos sobre el uso de los niveles de significancia de 0.05 y 0.01. *Ecosistemas y Recursos Agropecuarios* 8: 591–600. http://dx.doi.org/10.19136/era.a8n1.2616

Benoit DF& Van Den Poel D. 2009. Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: an application in financial services. *Expert Systems with Applications* 36: 10475–10484. https://doi.org/10.1016/j.eswa.2009.01.031

Bissinger JE, Montagnes DJ Harples J & Atkinson D. 2008. Predicting marine phytoplankton maximum growth rates from temperature: improving on the Eppley curve using quantile regression. *Limnology and Oceanography* 53: 487–493. https://doi.org/10.4319/lo.2008.53.2.0487

Bohora SB & Cao QV. 2014. Prediction of tree diameter growth using quantile regression and mixed-effects models. *Forest Ecology and Management* 319: 62–66. https://doi.org/10.1016/j.foreco.2014.02.006

Brown RL & Peet RK. 2003. Diversity and invasibility of southern Appalachian plant communities. *Ecology* 84: 32–39. https://doi.org/10.1890/0012-9658(2003)084[0032:DAIOSA]2.0.CO;2

Brown S & Lugo AE. 1984. Biomass of tropical forests: a new estimate based on forest volumes. *Science* 223: 1290–1293. https://doi.org/10.1126/science.223.4642.1290

Cade BS & Noon BR. 2003. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* 1: 412–420. https://doi.org/10.1890/1540-9295(2003)001[0412:AGITQR]2.0.CO;2

Cade BS, Terrell JW & Schroeder RL. 1999. Estimating effects of limiting factors with regression quantiles. *Ecology* 80: 311–323. https://doi.org/10.2307/176999

Cao QV & Dean TJ. 2015. Using nonlinear quantile regression to estimate the self-thinning boundary curve. In Holley A et al. (eds) *Proceedings of the 17th Biennial Southern Silvicultural Research Conferenc*e. e-Gen. Tech. Rep. SRS-203. Forest Service, Southern Research Station, Asheville.

Chave J, Andalo C, Brown S et al. 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia* 14: 87–99. http://dx.doi.org/10.1007/s00442-005-0100-x

Das K, Krzywinski M & Altman N. 2019. Quantile regression. *Nature Methods* 16: 451–452. https://doi.org/10.1038/s41592-019-0406-y

Doll JC. 2011. Predicting biological impairment from habitat assessments. *Environmental Monitoring and Assessment* 182: 259–277. https://doi.org/10.1007/s10661-011-1874-4

Everitt BS & Skrondal A. 2010. *The Cambridge Dictionary of Statistics. Fourth edition*. Cambridge University Press, Cambridge.

Fan J & Jiang J. 2007. Nonparametric inference with generalized likelihood ratio tests. *Test* 16: 409–444. https://doi.org/10.1007/s11749-007-0080-8

Fan J, Zhang C & Zhang J. 2002. Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of statistics* 30: 153–193. http://dx.doi.org/10.1214/aos/1043351258

Feng Y, Chen Y & He X. 2015. Bayesian quantile regression with approximate likelihood. *Bernoulli* 21: 832–850. doi: 10.3150/13-BEJ589

Ford C. 2015. Getting started with quantile regression. http://data.library.virginia.edu/getting-started-with-quantile-regression/.

Gao H, Bi H & Li F. 2017. Modelling conifer crown profiles as nonlinear conditional quantiles: an example with planted Korean pine in northeast China. *Forest Ecology and Management* 398: 101–115. https://doi.org/10.1016/j.foreco.2017.04.044

Gregorius HR, Degen B & König A. 2007. Problems in the analysis of genetic differentiation among populations—a case study in *Quercus robur*. *Silvae Genetica* 56: 190–199. https://doi.org/10.1515/sg-2007-0029

Hochberg Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75: 800–802. https://doi.org/10.1093/biomet/75.4.800

Huberty CJ & Morris JD. 1989. Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin* 105: 302–308. https://psycnet.apa.org/doi/10.1037/0033-2909.105.2.302

Huxley JS. 1950. Relative growth and form transformation. *Proceedings of the Royal Society of London. Series B—Biological Sciences* 137: 465–469. https://doi.org/10.1098/rspb.1950.0055

Koenker R. 2000. Quantile regression. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons Ltd. https://doi.org/10.1002/9781118445112.stat07557

Koenker R. 2013. *quantreg: Quantile Regression. R package version 5.05*. R Foundation for Statistical Computing, Vienna. http://CRAN. R-project. org/package=quantreg.

Koenker R & Bassett G. 1978. Regression quantiles. *Econometrica* 46: 33–50.

Koenker R & Hallock K. 2001. Quantile regression. *Journal of Economic Perspectives* 15: 43–56. doi: 10.1257/jep.15.4.143

Lancaster T & Jun SJ. 2010. Bayesian quantile regression methods. *Journal of Applied Econometrics* 25: 287–307. https://doi.org/10.1002/jae.1069

Mäkinen A, Kangas AS, Kalliovirta J, Rasinmäki J & Välimäki E. 2008. Comparison of treewise and standwise forest simulators by means of quantile regression. *Forest Ecology and Management* 255: 2709–2717. http://dx.doi.org/10.1016/j.foreco.2008.01.048

McCune B. 2006. Nonparametric multiplicative regression for habitat modelling. *Journal of Vegetation Science* 17: 819–830.

Meinshausen N. 2006. Quantile regression forests. *Journal of Machine Learning Research* 7: 984–999.

Nava-Nava A & Antúnez P. 2018. Application of quantile regression to predict stem volume: a case study. *Ecosistemas y recursos agropecuarios* 5: 591–600. https://doi.org/10.19136/era.a5n15.1498

Osborne JW. 2013. Normality of residuals is a continuous variable, and does seem to influence the trustworthiness of confidence intervals: a response to, and appreciation of, Williams, Grajales, and Kurkiewicz (2013). *Practical Assessment, Research, and Evaluation* 18: Article 12. https://doi.org/10.7275/6k0p-s133

Osborne JW & Overbay A. 2004. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation* 9: Article 6. https://doi.org/10.7275/qf69-7k43

Robinson AP & Hamann JD. 2010. *Forest Analytics with R. An Introduction*. Springer, New York.

Romano Y, Patterson E & Candes E. 2019. Conformalized quantile regression. Advances in neural information processing systems. Pp 3549–3558 in Wallach HM et al. (eds) in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*. 8–14 December 2019, Vancouver.

Schröder HK, Andersen HE & Kiehl K. 2005. Rejecting the mean: estimating the response of fen plant species to environmental factors by non-linear quantile regression. *Journal of Vegetation Science* 16: 373–382. https://doi.org/10.1111/j.1654-1103.2005.tb02376.x

Takeuchi I, Le QV, Sears TD & Smola AJ. 2006. Nonparametric quantile estimation. *Journal of Machine Learning Research* 7: 1231–1264.

Van Keilegom KI, Manteiga WG & Sellero CS. 2008. Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *Test* 17: 401–415. https://doi.org/10.1007/s11749-007-0044-z

Wang ZX, & Jv YQ. 2021. A novel grey prediction model based on quantile regression. *Communications in Nonlinear Science and Numerical Simulation* 95: 1–17. https://doi.org/10.1016/j.cnsns.2020.105617

Ying Z, Jung SH & Wei LJ. 1995. Survival analysis with median regression models. *Journal of the American Statistical Association* 90: 178–184. https://doi.org/10.2307/2291141

Yu K & Moyeed RA. 2001. Bayesian quantile regression. *Statistics & Probability Letters* 54: 437–447. https://doi.org/10.1016/S0167-7152(01)00124-9

Yu K, Lu Z & Stander J. 2003. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)* 52: 331–350. https://doi.org/10.1111/1467-9884.00363

Zhang B, Sajjad S, Chen K et al. 2020. Predicting tree height–diameter relationship from relative competition levels using quantile regression models for Chinese fir (*Cunninghamia lanceolata*) in Fujian Province, China. *Forests* 11: 183. https://doi.org/10.3390/f11020183

Zianis D, Muukkonen P, Mäkipää R & Mencuccini M. 2005. Biomass and stem volume equations for tree species in Europe. *Silva Fennica Monographs* 4. https://doi.org/10.14214/sf.sfm4