






ARTICLE

Robust local quantile regression for reference curves

 Gianni M. S. Santos,^{*}¹  Carmen D.S. de André,² and  Julio M. Singer²

¹Applied Statistics Division, Federal University of São Paulo, São Paulo-SP, Brazil.

²Department of Statistics, University of São Paulo, São Paulo-SP, Brazil.

*Corresponding author. Email: giannisantos2014@gmail.com

(Received: January 3,2022; Revised: February 22,2022; Accepted: March 15,2022; Published: March 1,2023)

Abstract

We propose two non-parametric methods to construct locally fitted quantile reference curves that are robust with respect to outliers in the predictor variable. The first includes a weighting procedure and the second, the detection and subsequent elimination of outlying predictor variable values before the local fitting process. The reference curves fitted by the proposed methods generate quantile limits that are less affected in regions with a low frequency of the predictor variable values. The proposed procedures are used to fit reference curves to data extracted from a study conducted at the Heart Institute of the University of São Paulo Medical School.

Keywords: outliers, predictor variables, quantize regression.

1. Introduction

Reference curves are constructed to obtain quantiles of a response variable as a function of one or more predictor variables and are popular in many scientific areas. In Pediatrics, for example, reference curves for weight or height are built according to age and sex so that the growth of a child

can be compared to that of children with the same age and sex (Freeman *et al.*, 1995). In Economics, Fitzenberg *et al.* (2002) used reference curves to examine the relationship between development and economic growth of different countries. Martins *et al.* (2004) employed reference curves to study the distribution of wages as a function of schooling level. In Education, such curves are employed to evaluate the relationship between the distribution of public school student performance and certain characteristics such as parent income, classroom size and teacher qualification (Buchinsky, 1998). In the field of Biology and Ecology, reference curves may be used to estimate the effects of factors that may affect growth, survival and reproduction of certain systems (Cade *et al.*, 1999), as well as to evaluate the association between the size of prey and the size of predators (Scharf *et al.*, 1998).

Different methods for constructing reference curves are available. Harris and Boyd (1995) classified these methods as parametric, according to which the form of the relationship between the response variable Y and the predictor variable X is specified, or non-parametric, where this type of relationship is not completely specified. Parametric reference curves are considered in Royston (1991) and de Paula *et al.* (2005), for example. Non-parametric reference curves, mainly based on quantile regression models discussed in Koenker and Hallock (2001), are also considered for such purposes and, perhaps, constitute the current paradigm for such purposes. See, for example, Fan and Gijbels, (1996), Healy *et al.* (1988), Huang and N 'Guyen (2018), Waldmann (2018) and Muggeo *et al.* (2021), among others.

To motivate our investigation, we consider a dataset containing data from patients that sought the Heart Institute of University of São Paulo Medical School for check-ups. The complete data were analyzed by de Paula *et al.* (2005). For our purposes, we consider data from 349 females showing no symptoms of cardiac illnesses. The response variable is the mean heart rate (bpm), computed from measurements over a period of 24 hours and the predictor variables is age (years). As illustrated in Figure 1 (left panel), the reference curves obtained by standard local quantile regressions are strongly affected by points corresponding to the few individuals aged over 60 years.

In an attempt to reduce the influence of these points in regions with low frequency, we propose two robust methods for the construction of quantile reference curves. The first considers assigning additional weights to the observations as in Einbeck *et al.* (2004) and generates reference curves that are resistant with respect to regions with low frequency of the predictor variable. The second considers a procedure for the detection and subsequent elimination of values in such regions before locally fitting the reference curves.

In Section 2, we describe the standard local quantile regression model. In Section 3, we outline the two robust procedures and in Section 4, we show how they may be used to construct reference

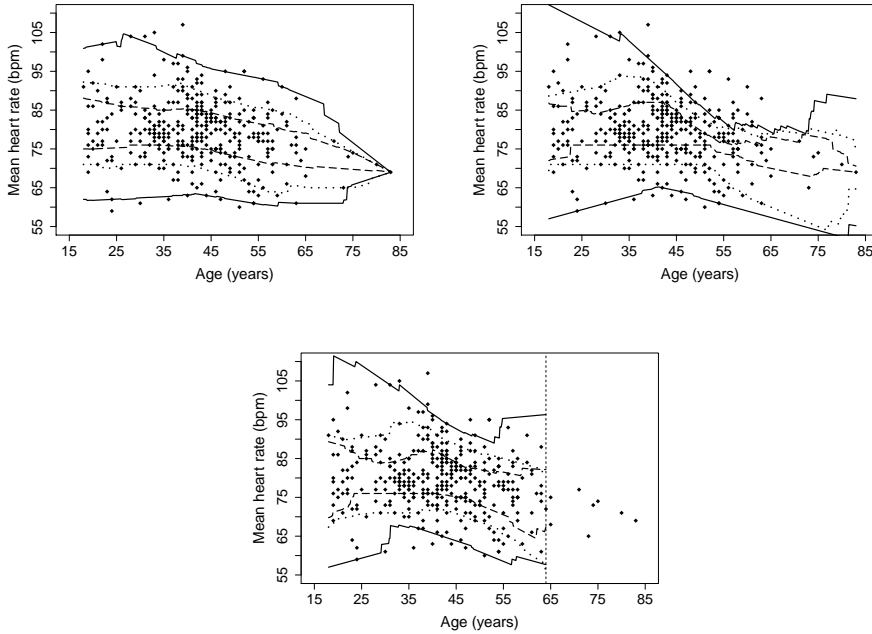


Figure 1. Standard (left panel), soft robustified (right panel) and hard robustified reference curves (bottom panel): 2.5% and 97.5%: solid lines, 10% and 90%: dotted lines, 25% and 75%: dashed lines.

curves for the motivating cardiology dataset. A brief discussion and proposal for future studies are outlined in Section 5.

2. Local quantile regression

Quantile regression, introduced by Koenker and Bassett (1978), characterizes the conditional quantile distribution of a response variable given the values of predictor variables, thus providing a broad description of the relationship between them in addition to being robust with respect to outlying values of the response variable.

Formally, the objective of an $(\alpha, \xi_\alpha(x))$ quantile regression is to estimate the $0 < \alpha < 1$ population conditional quantile of a response variable Y corresponding to the value x of a predictor variable X via an unspecified function $\xi_\alpha(x)$. It is expected that, for each value x , $(1 - \alpha)100\%$ of the population values of the response variable Y lie among the limits $\xi_{\alpha/2}(x)$ and $\xi_{1-\alpha/2}(x)$.

Fan and Gijbels (1996) showed that the reference value $\xi_\alpha(x_0)$ can be locally obtained in the vicinity of a target point $X = x_0$ by weighting the values of X in a pre-specified neighbourhood of x_0 via a kernel function. Specifically, these authors suggest that the local fit of $\xi_\alpha(x_0)$ involves the

minimization of

$$\sum_{i=1}^n \ell_{\alpha} \left[Y_i - \sum_{j=0}^p \beta_j (X_i - x_0)^j \right] K \left[\frac{X_i - x_0}{h(\alpha)} \right], \quad (1)$$

where $\ell_{\alpha}(t) = |t| + (2\alpha - 1)t$, $0 < \alpha < 1$, $\beta_j, j = 0, \dots, p$ are the coefficients of the polynomial adopted for the local fit, $K(\cdot)$ is a kernel function and $h(\alpha)$ is a bandwidth, depending only on α .

In their original work, Fan and Gijbels (1996) note that the choice of the bandwidth $h(\alpha)$ is an important aspect to be considered since it controls the amount of smoothing applied to the data. For small values of h , only observations lying in a small neighborhood of x_0 are taken into account and lead to a more complex model; otherwise, for large values of h , observations that are distant from x_0 will also affect the local fitting process. This problem has been addressed by many authors, without a clear definition of an optimal solution. In particular, Yu and Jones (1998) presented a simplified approach relating the parameter α used in the local fit of $\xi_{\alpha}(\cdot)$ to the bandwidth which minimizes the asymptotic mean squared error used in fitting the regression function. These authors obtained the following expression for the computation of an "optimal" bandwidth

$$h_{\text{opt}} \left[\frac{\alpha(1-\alpha)}{\phi(\Phi^{-1}(\alpha))^2} \right]^{1/5}, \quad (2)$$

where ϕ and Φ are, respectively, the density and the cumulative distribution functions of a standard normal distribution and h_{opt} is the plug-in optimal estimator proposed by Ruppert *et al.* (1995).

Many authors have considered modifications and improvements of the method proposed by Fan and Gijbels (1996). Among them, we mention Anas Knefati *et al.* (2016), who propose plugging a radial basis function neural network in the local linear quantile regression estimation and Liu *et al.* (2019), who consider an algorithm based on a normal scale-mixture representation of an asymmetric Laplace distribution that enjoys the same good design adaptation that ensures non-crossing quantile curves for any given sample. More recently, Muggeo *et al.* (2021) propose an iterative algorithm to select the smoothing parameters in additive quantile regression, wherein the functional forms of the predictor variables are unspecified and expressed via B-spline bases with different penalties on the spline coefficients.

The fitting methods proposed by these authors are robust with respect to response variable outlying values, but not with respect to predictor variable outlying values as in our motivating example. Based on the ideas of Einbeck *et al.* (2004), we consider two procedures for the estimation of quantile regression functions that are robust with respect to outlying values in the response and in the predictor variables simultaneously.

3. Predictor variable outlier-robust local quantile regression

3.1 Soft robustification method

In the context of local regression models, Einbeck *et al.* (2004) proposed to modify the standard fitting algorithm by adding weights associated to the values of the density function of the predictor variable in such a way that points lying in a sparse region are penalized. Specifically, for fitting the regression function at a point x_0 , the algorithm consists of the minimization of

$$\xi_{\alpha}(x_0) = \sum_{i=1}^n \left[Y_i - \beta_0 - \beta_1(X_i - x_0)^2 \right] \zeta(X_i) K\left(\frac{X_i - x_0}{h_n}\right), \quad (3)$$

where β_0 and β_1 are the parameters to be estimated, $\zeta(\cdot)$ is a monotonic increasing function of the density function $f(\cdot)$ of X , $K(\cdot)$ is a kernel function and h_n is the bandwidth.

The role of $\zeta(\cdot)$ in (3) is to reduce the effect of the outlying X values in the estimation of the regression function. The reduction can be more effective if we consider $\zeta(\cdot) = f(\cdot)^k$ for some $k > 1$ although k cannot increase arbitrarily, since estimation may become unstable.

Einbeck *et al.* (2004) use a consistent estimator for the density function of the variable X at the point x_0 , given by

$$\hat{f}(x_0) = \frac{1}{ns_n} \sum_{i=1}^n K\left(\frac{X_i - x_0}{s_n}\right),$$

where s_n is the bandwidth proposed by Silverman (1986), namely $s_n = 0.9An^{-1/5}$, with

$$A = \min\left(dp, \frac{IIQ}{1.34}\right), \quad (4)$$

where dp and IIQ are, respectively, the standard deviation and the interquartile sample range of the predictor variable X . Other non-parametric estimators of $f(\cdot)$ as those proposed by Fan and Gijbels (1996) may also be considered.

Along similar lines, we extend the results to address quantile regression models by considering the minimization of

$$\xi_{\alpha}(x_0) = \sum_{i=1}^n \ell_{\alpha} \left[Y_i - \beta_0 - \beta_1(X_i - x_0) \right] \zeta(X_i) K\left(\frac{X_i - x_0}{h_n}\right),$$

where $\ell_{\alpha}(t)$ and $K(\cdot)$ are defined in (1), h_n is the bandwidth defined in (2), $\zeta(\cdot)$ is defined in (3) and β_0 and β_1 are parameters to be estimated for each local linear fit.

3.2 Hard robustification

In the process of fitting $\xi_{\alpha}(\cdot)$ described in the previous section, the effect of the outlying X values is minimized, but not eliminated. In an attempt to further reduce this effect, Einbeck *et al.*

(2004) suggest an alternative procedure to detect and eliminate these values. For such a purpose, these authors consider the outlying values of X as those with estimated density below a certain threshold (cut-off value). To determine the amount of elimination, they assume that X follows a normal distribution with mean equal to the sample median, x_{md} , and standard deviation given by A in (4). Letting d be the proportion of expected outlying values of X , Einbeck *et al.* (2004) proposed a cut-off point given by

$$\delta = N_{\mu, A^2}(x_{md} + Az_{d/2}) = \frac{1}{A} \Phi(z_{d/2}), \quad (5)$$

where $z_{d/2}$ denotes the quantile of order $d/2$ of the standard normal distribution.

Using this criterion, we propose that the hard robustified quantile regression function $\widehat{\xi}_\alpha(\cdot)$ be obtained by minimizing

$$\sum_{i=1}^n \ell_\alpha \left[Y_i - \beta_0 - \beta_1(X_i - x_0) \right] \zeta(X_i) I[f(X_i) > \delta] K \left(\frac{X_i - x_0}{h_n} \right),$$

where $\ell_\alpha(t)$, $K(\cdot)$ are defined in (1), h_n is defined in (2), $\zeta(\cdot)$ is defined in (3), $I(\cdot)$ is the indicator function and β_0 and β_1 are parameters to be estimated for each local linear fit.

4. Analysis of the heart dataset

In this section we apply of the procedures for fitting the local quantile reference curves discussed in Section 3 to the dataset presented in Section 1.

In our proposal, we consider the bandwidth given in (2) and a Gaussian kernel function. Computation of h_{opt} may be carried out via the function `dpill()`, available in the package `KernSmooth` and the proposed quantile regression functions may be fitted via slight modifications of the function `lprq()`, available in the `quantreg` R package.

In Figure 1 (right panel) we present the 2.5%, 10%, 25%, 75%, 90% and 97.5% quantile reference curves fitted by the soft robustification method. Observe that the curves obtained by soft robustification unlike the ones fitted by the standard local quantile regression method are less affected by values of age over 60 years.

In the process of fitting quantile reference curves by hard robustification, we fixed the amount of points to exclude as $d = 5\%$; then, from (5) it follows that the outlying values of X were considered as those for which the estimated density is $\leq \delta = 0.05$. The estimated density along with the cut-off points are displayed in Figure 2.

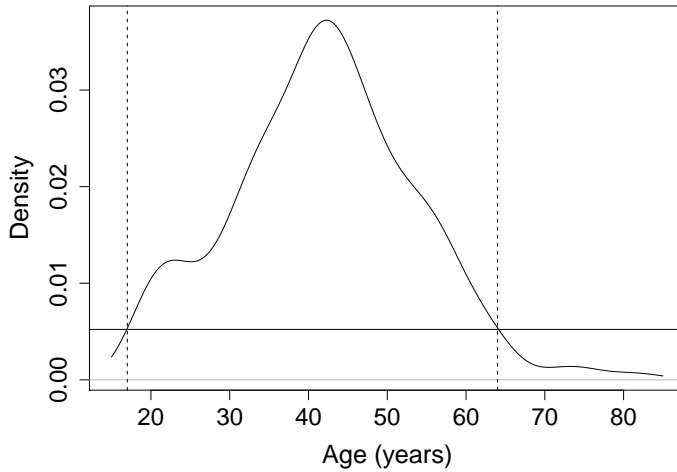


Figure 2. Estimated age density and cut-off limits.

The region where estimation of the reference curves is considered to be less reliable corresponds to ages less than 17 or greater than 64, and imply that 8 observations should be discarded. This corresponds to 2.3% of the sample size. The associated reference curves are displayed in the bottom panel of Figure 1, clearly indicating the region where the observations are sparse.

The local quantile reference curves displayed in Figure 1 are quite irregular; they may be smoothed by applying the `rqss()` function in the standard case and by fitting local polynomial models to the robust estimated response variable fitted values in both the soft and hard robust cases. This is easily accomplished via the `locpoly()` function. The smoothed curves are displayed in Figure 3.

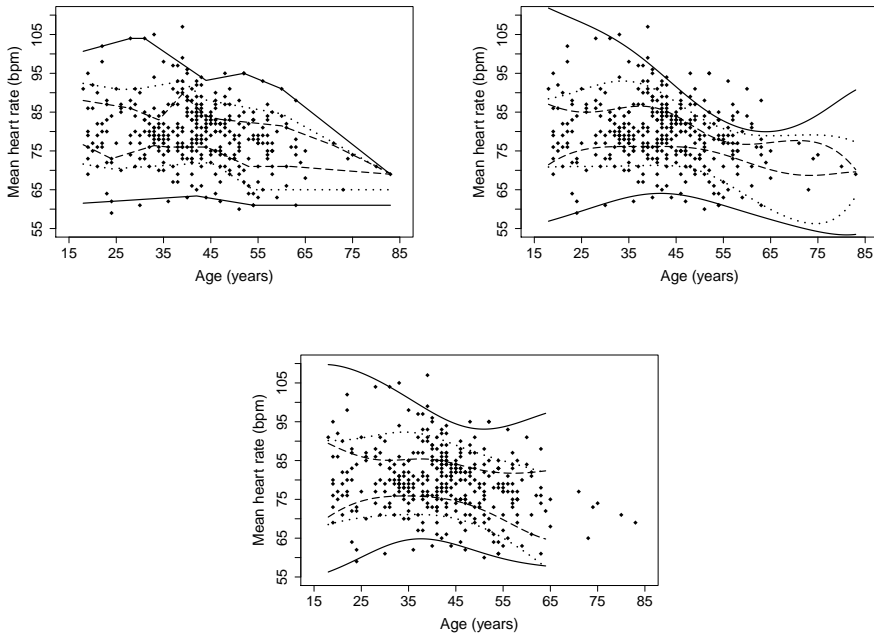


Figure 3. Standard (left panel), soft robustified (right panel) and hard robustified smoothed reference curves (bottom panel): 2.5% and 97.5%: solid lines, 10% and 90%: dotted lines, 25% and 75%: dashed lines.

5. Discussion

Although construction of quantile reference curves has been addressed by many authors, it is still a problem with no optimal solution, specially when data are irregular and/or sparse. Recent work by Liu *et al.* (2019) and Muggeo *et al.* (2021) are clear examples of the associated complexity. In particular, we mention the algorithm proposed by the latter, implemented via the function `gcrq()` available in the `quantregGrowth` package, where the generated reference curves accommodate different characteristics of the data and have interesting features, like avoiding crossing of the different quantile curves. Nevertheless, the choice of the parameters that govern the construction of the reference curves is not an easy task. An naive application of the function with the proposed optimal choice of parameters to the data described in the Introduction generates the reference curves presented in Figure 4.

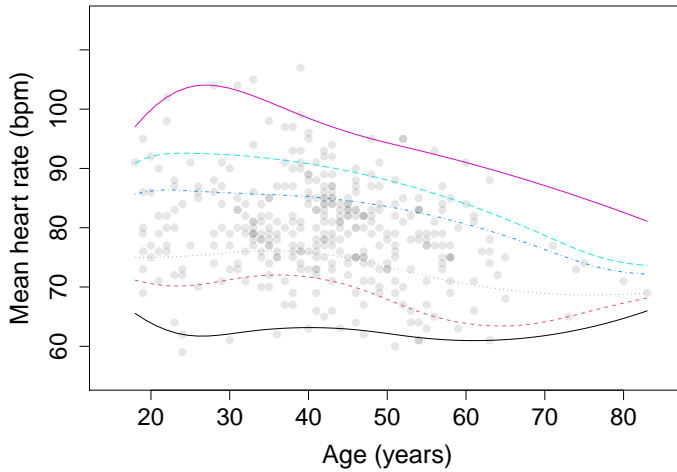


Figure 4. 2.5%, 10%, 25%, 75%, 90% and 97.5% reference curves generated by the `gcrq` function.

We consider two alternative approaches where regions with sparse predictor variable values are penalized. In the soft robustification approach, observations in these regions are downweighted, generating wider intervals; otherwise, in the hard robustification case, these observations are eliminated, defining regions where the reference curves construction is assumed unreliable. In this regard, we quote Einbeck *et al.* (2004), who mention that

Surely the question arises whether one can rely on estimation results in areas where the data were downweighted or even cut off. This, however, is a question inherent to any robust method. In particular, when applying soft robustification techniques, we must face the question of whether it is correct to downweight the data, on the one hand, i.e., to pretend not to trust the data, but to believe in the estimation results in the same region, on the other hand. Some decision has to be made and we suggest to base it on areas of confidence, which can be selected by means of density estimation. Within the areas of confidence, i.e., for all x with $\hat{f}(x) > \delta$, the estimation is considered to be reliable. Outside these areas, the reliability of the estimation procedures is questionable and interpretation of the estimated curve must be taken cautiously.

The choice between the reference curves must take the regions in which the explanatory variable values are sparse into account. These points can drastically affect the curves obtained by the standard method, as illustrated in the manuscript, and one of the two proposed methods may provide an alternative solution. The hard robustification method allows the elimination of the region with low frequency of explanatory variable values. This was not the case in our example. The few points in the region made up of patients over 64 can be attributed to the small population of women without

any heart disease who attend the Heart Institute. In this case, the points can be kept in the analysis and weighted to obtain robust reference curves via the soft robustification technique. Analytical methods to decide which alternative to use are not yet available and constitute an interesting topic for further research.

Although the proposed quantile reference curves may be implemented via straightforward modifications of well established R package functions, we recognize that further research is needed to improve the solution. The choice of appropriate bandwidths and possible annoying crossing of the quantile curves are two topics under investigation.

Acknowledgements

Work of the third author was partially funded by grant # 304841/2019-6 from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil.

Conflicts of Interest

The authors declare no conflict of interest.

References

- [1] Anas Knefati, M., Cuauvet, P. E., Nguyen, S.; Bassam, D. Reference curves estimation using conditional quantile and radial basis function network with mass constraint. *Neural Process Letters*, **43**, 17-30 (2016).
- [2] Buchansky, M. Recent advances in quantile regression models: A practical guide for empirical research, *Journal of Human Resources*, **33**, 88-126 (1998).
- [3] Cade, B. S.; Terrel, J.W.; Schroeder, R.L. Estimating effects of limiting factors with regression quantiles, *Ecology*, **80**, 311-323 (1999).
- [4] de Paula, R.S.; Antelmi, I.; Vincenzi, M.A.; André, C.D.S.; Artes, R.; Grupi, C.G.; Mansur, A.J. Influence of age, gender and serum triglycerides on heart rate in a cohort of asymptomatic individuals without heart disease. *International Journal of Cardiology*, **105**, 152-158 (2005).
- [5] Einbeck, J.; André, C.D.S.; Singer, J.M. Local smoothing with robustness against outlying predictors. *Environmetrics*, **15**, 541-554 (2004).

- [6] Fan, J.; Gijbels, I. *Local Polynomial Modelling and Its Applications, Mono graphs on Statistics and Applied Probability*, London: Chapman & Hall, (1996).
- [7] Fitzenberg B.; Koenker, R.; Machad, J.A.F. *Economic Applications of Quantile Regression*, Berlin: Springer, (2002).
- [8] Freeman, J.V.; Cole, T.J.; Chinn, S.; Jones, P.R.M.; White, E.M.; Preece, M.A. Cross sectional stature and weight reference curves for the UK, 1990. *Archives of Disease in Childhood*, **73**, 17-24 (1995).
- [9] Harris, E.K.; Boyd, J.C. *Statistical Bases of Reference Values in Laboratory Medicine*, New York: Marcel Dekker, 1995.
- [10] Healy, M.J.R.; Rasbach, J.; Yang, M. Distribution-free estimation of age-related centiles. *Annals of Human Biology*, **15**, 17-22 (1988).
- [11] Huang, M.L.; Nguyen, C. A nonparametric approach for quantile regression, *Journal of Statistical Distributions and Applications*, **5**,(1), 1-3 (2018).
- [12] Koenker, R.; Bassett, G. Regression quantiles. *Econometrica*, **46**, 33-50 (1978).
- [13] Koenker, R.; Hallock, K.F. Quantile regression: An introduction. *Journal of Economic Perspectives*, **15**, 143-156 (2001).
- [14] Liu, X.; Yu, K.; Xu, Q.; Tang, X. Improved local quantile regression. *Statistical Modelling*, **19**, 501-523 (2019).
- [15] Martins, P.S.; Pereira, P.T. Does education reduce wage inequality? Quantile regression evidence from 16 countries. *Labour Economics*, **11**, 355-371 (2004).
- [16] Muggeo, V.M.R.; Torretta, F.; Eilers, P.H.C.; Sciandra, M.; Attanasio, M. Multiple smoothing parameters selection in additive regression quantiles, *Statistical Modelling*, **21**, 428-448 (2021).
- [17] Royston, P. Constructing time-specific reference ranges. *Statistics in Medicine*, **10**, 675-690, (1995).
- [18] Ruppert, D.; Sheather, S.J.; Wand, M.P. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257-1270 (1995).
- [19] Scharf, F.S.; Juanes, F.; Sutherland, M. Inferring ecological relationships from the edges of scatter diagrams: comparison of regression techniques. *Ecology*, **79**, 448-460, (1998).

- [20] Silverman, B.W. *Density Estimation for statistics and Data Analysis*, London: Chapman & Hall, (1986).
- [21] Waldmann, E. Quantile regression: a short story on how and why. *Statistical Modelling*, **18**, 203-218 (2018).
- [22] Yu, K.,; Jones, M.C. Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228-237 (1998).