USE OF QUANTILE REGRESSION AND RANSAC ALGORITHM IN FITTING VOLUME EQUATIONS UNDER THE INFLUENCE OF DISCREPANT DATA

Jadson Coelho de Abreu[1*], Carlos Pedro Boechat Soares[2], Helio Garcia Leite[3], Daniel Henrique Breda Binoti[4], Gilson Fernandes da Silva[5]

[1*]Universidade do Estado do Amapá, Colegiado de Engenharia Florestal, Macapá, Amapá, Brasil - e-mail: jadson.abreu@ueap.edu.br
[2]Universidade Federal de Viçosa, Programa de Pós-Graduação em Ciência Florestal, Viçosa, Minas Gerais, Brasil - e-mail: csoares@ufv.br
[3]Universidade Federal de Viçosa, Programa de Pós-Graduação em Ciência Florestal, Viçosa, Minas Gerais, Brasil - e-mail: hgleite@gmail.com
[4]DAP Florestal, Jerônimo Monteiro, Espírito Santo, Brasil - e-mail: danielhbbinoti@gmail.com
[3]Universidade Federal do espírito Santo, Departamento de Engenharia Florestal, Jerônimo Monteiro, Espírito Santo, Brasil - e-mail: fernandes5012@gmail.com

Received for publication: 29/01/2020 – Accepted for publication: 18/09/2020

Resumo

Uso da regressão quantílica e algoritmo RANSAC no ajuste de equações de volume sob influência de dados discrepantes. O objetivo desse estudo foi avaliar três métodos de estimação no ajuste de equação de volume na presença de dados influentes ou dados de alavanca (ou de alavancagem). Para isso, foram utilizados dados do inventário florestal realizado pela Fundação Centro Tecnológico de Minas Gerais para a formação Cerradão. O modelo de Schumacher e Hall (1933), em sua forma não linear, foi ajustado considerando o uso da regressão quantílica (RQ), pelo algoritmo RANSAC e o método dos Mínimos Quadrados Ordinário não linear (MQO). Como critérios de avaliação das performances dos métodos avaliados, foram utilizados o coeficiente de correlação ($r_{y\hat{y}}$) entre os volumes observados e estimados, raiz quadrada do erro médio (RQEM), bem como análise gráfica da dispersão e distribuição dos resíduos. Após as análises, observou-se que o método dos mínimos quadrados não linear (MQO) apresentou resultado levemente superior em termos das estatísticas avaliadas, no entanto alterou a tendência esperada da curva ajustada devido à presença de dado de influência, o que não ocorreu com a RQ e o algoritmo RANSAC, sendo esses mais robustos na presença de dados discrepantes.
Palavras-chaves: Cerradão, dados de influência, dados de alavancagem, métodos de estimação.

Abstract

The objective of this study was to evaluate three estimation methods to fit volume equations in the presence of influential or leverage data. To do so, data from the forest inventory carried out by the Centro Tecnológico de Minas Gerais Foundation were used to fit the Schumacher and Hall (1933) model in its nonlinear form for Cerradão forest, considering the quantile regression (QR), the RANSAC algorithm and the nonlinear Ordinary Least Squares (OLS) method. The correlation coefficient ($r_{y\hat{y}}$) between the observed and estimated volumes, root-mean-square error (RMSE), as well as graphical analysis of the dispersion and distribution of the residuals were used as criteria to evaluate the performance of the methods. After the analysis, the nonlinear least squares method presented a slightly better result in terms of the goodness-of-fit statistics, however it altered the expected trend of the fitted curve due to the presence of influential data, which did not happen with the QR and the RANSAC algorithm, as these were more robust in the presence of discrepant data.
Keywords: Cerradão, influential data, leverage data, estimation methods.

INTRODUCTION

Estimating tree volumes is one of the main purposes of forest surveys. Estimates of standing tree volumes can be obtained by means of form factors, form quotient, volume equations, multiple volume or taper equations, among others.

The volume equation is the most common way of estimating the volume, in which the wood volume is a function of other tree magnitudes or variables (usually the diameter at breast height and height) which can be measured and estimated by non-destructive means.

Correct data sampling for fitting equations of any nature, including volume equations, constitutes a fundamental step so that the estimated parameters of the equations are not biased, and consequently the estimates are obtained without bias such as the case of volumetric stock measurements of forests.

However, even in observing all the precautions in data collection for fitting volume equations, discrepant data, also called leverage or influential points or outliers may occur due to the nature of the vegetation being studied. It is worth noting that the extreme value or an atypical outlier value is an observation with "high" residual. In this case, we have an observation whose dependent variable is unexpected given the value of the explanatory

variable. On the other hand, we say that an observation is a leverage point (or high leverage) if the explanatory variable is extreme. An influential point is an observation which can influence any part of the regression analysis, such as the prediction of a response, the slope or the result of hypothesis tests.

For example, in obtaining data in native tropical forests in which a high diversity of species and tree shapes is found, there can be different relationships between the trunk height, diameter and volume, defining observations which seem to be discrepant from the remaining data set (BARNETT and LEWIS 1995), implying asymmetry in the data distributions and loss of efficiency in using traditional methods of estimating the parameters of the regression models, since these can be adjusted using the least squares methods, maximum likelihood, and its estimates refer to average (expected) values, so the average is influenced by extreme values which influence the regression curve.

In this situation, it is possible to use robust estimators such as Quantile Regression (QR), which, unlike traditional estimation methods, does not use the mean (central value) in its estimators. The QR generalizes this explanation to any quantile of interest using the median (BARROSO *et al*., 2015). In addition, they can be characterized as semi-parametric regression models, as they do not require any probability distribution for the response variable (dependent) and enables fitting models for conditional quantile functions through a weighting in minimizing errors (KOENKER and BASSETT, 1978).

The standard regression model for the mean response is:

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip}.. \text{ i=1.....n}$$

And β_js are estimated by solving the problem of ordinary least squares.

$$min_{\beta 0 \ldots . \beta p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_i \beta_j \right)^2$$

In contrast, the regression model for the quantile level of the response is:

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau) x_{i1} + \cdots \beta_p(\tau) x_{ip}, .. \text{ i=1.....n}$$

and $\beta_j(\tau)$s are estimated, solving the minimization problem

$$min_{\beta 0 \ldots . \beta p(\tau)} \sum_{i=1}^{n} \left(y_i - \beta_0(\tau) - \sum_{j=1}^{p} x_i \beta_j(\tau) \right)$$

In which: $\rho_\tau(r) = \tau \max(r, 0) + (1 - \tau) \max(-r, 0)$. The $\rho_\tau(r)$ function is referred to as the loss of verification, because its shape looks like a check mark.

The solution to the minimization problem produces a different set of regression coefficients for each level of quantile τ. Note that τ = 0.5 corresponds to the median regression, and 2ρ0.5(r) is the absolute value function.

Although Quantile Regression has great potential for use in several areas, it is still little used in forest measurement and especially in Brazil. One of the works is that published by Araújo Junior et al. (2016), who used it to classify forest sites. More work is found in forest measurement outside of Brazil, such as forecasting growth in diameter (BOHARA and CAO, 2014), growth in height (MEYER et al., 2005), tapering functions (CAO and WANG, 2015) and forest production dynamics (WOODALL *et al*., 2008). Still regarding forestry research in other countries, Quantile Regression has been used in studies on tolerance to invasive plants (THIFFAULT *et al*., 2012) and species response to environmental factors (SCHRÖDER *et al*., 2005).

Another option with robust estimators is the Random sample consensus algorithm (RANSAC) and its variations, which was proposed to obtain estimates of model parameters by removing outliers and keeping the model estimates stable against noise (LE *et al*., 2017). It is an algorithm which generates candidate solutions using the minimum amount of observations necessary to estimate the parameters of a given predefined model, unlike conventional methods which use the largest amount of data possible to obtain the initial solution.

The RANSAC algorithm works as follows 1. Randomly select the smallest number of observations necessary to estimate the model parameters; 2. Obtain estimates of the model parameters for this data set; 3. Determine how many of the observations (z_i) in the initial set meet the established criteria to be included in the fitting; 4. If the number of z_i observations is greater than the consensus set, it appears that the current solution is better and should take its place. If the current iteration is not the last one, return to the first step; 5. When reaching the iteration limit, the model parameters must be estimated by least squares using all consensus observations (FISCHLER and BOLLES 1982).

RANSAC and its variants have been used for effective modelling approaches and tracking approaches under heavy noise and outliers (LI and GANS, 2017), for microbial metabolic data mining and phenotypic improvement (TEOH *et al*., 2016), for face recognition (VINAY *et al*., 2015), remote sensing (CHENG *et al*., 2012), among others. However, this algorithm has not yet been evaluated in the forest measurement area.

Based on the hypothesis that outliers influence the regression estimate, the objective of this study was to evaluate the Quantile Regression, the RANSAC algorithm and the nonlinear least squares method (OLS) in fitting the volume equation in the presence of discrepant data.

MATERIAL AND METHODS

Database

The data for this study come from the forest inventory carried out by the *Fundação Centro Tecnológico de Minas Gerais* for fitting application volume equations in the State and other regions of the country. Specifically, data from the forest formation called Cerradão were used, which contained 210 individuals divided into 46 species.

Information was collected from all trees in the rigorous cubing process to identify the species; diameters were measured at 1.30 m above the ground (d), in centimeters; and the total heights (ht), in meters. The bark volumes of individual tree stems (v), in cubic meters, were obtained by successive application of the Smalian formula (HUSCH *et al*., 1982), considering sections of 1m in length and minimum commercial diameter with bark of 4cm.

Estimation methods

The Schumacher and Hall (1933) model in its nonlinear form, given by:

$$v = \beta_0 . d^{\beta_1} . ht^{\beta_2} . \varepsilon \quad (1)$$

In which: v = total volume with bark in m³, d = diameter at breast height, ht = total height, β_0 to β_2 = model coefficients, ε = random error. It was adjusted considering the Quantile Regression (QR), the RANSAC algorithm and the nonlinear least squares method (OLS).

The estimates of the model parameters (expression 1) considering the nonlinear OLS were obtained using the R stats software package. The estimates for Quantile Regression were obtained according to Araújo Junior et al. (2016) using the Simplex algorithm in the quantreg library (Koenker, 2013) of the R program, considering a percentile (q) equal to 50%, which is equal to the median such that:

$$Min\ n^{-1} \sum_{i=X}^{n} \rho_\theta (Y_t - X'\beta) \quad (2)$$

In which: Y_t = vector of the dependent variable, X'= matrix of the independent variables, β = vector of the model coefficients, and ρ is a "check" function, defined by:

$$\rho_\theta (u) = \begin{cases} qu & u \geq 0 \\ (q-1)u & u < 0 \end{cases} \quad (3)$$

In which: q is a given percentile, and u is the error or residual.

The RANSAC software program (BINOTI and SILVA, 2016) was used in fitting using the RANSAC algorithm, and considering the following sequence of steps for applying the algorithm DERPANIS *et al.* (2010):

Determine:
n The lowest number of points needed
k The number of iterations needed
t The threshold used to identify a point that fits well;
d The number of nearby points needed to affirm a model fits well
Until k iterations have occurred
Draw a sample of n data points uniformly and randomly;
Fit model to this set of n points
For each out-of-sample data point
Test the distance from the point to the line in relation to *t*; if the distance from the point to the line is less than *t*, the point is an inlier, otherwise an outlier
End
If there are *d* or more points close to the line, there will be a good fit. Replace the line using all these points
End
Use the best fit in this collection, using the fit error as a criterion

In which: $k = \frac{\log(1-p)}{\log(1-w^n)}$

In which: *K* is maximum iterations, *p* is the probability that RANSAC selects only inliers of the input data when choosing *n* points from which the model parameters are estimated, and *w* is the probability of choosing an inlier. The generalized Cook distance test was used to verify the occurrence of influential points, in which h > 0.1 are considered influential data. The test was applied to the Schumacher and Hall model fit by nonlinear least squares, while it was not necessary to apply the Cook distance test for the methods using quantile regression and RANSAC, according to Souza (1998).

As criteria for evaluating the performance of the estimation methods, the correlation coefficient ($r_{y\hat{y}}$) between the observed and estimated volumes, the root-mean-square error (RMSE) was calculated in the original unit of the dependent variable volume (m^3), as well as a graphical analysis of the dispersion and distribution of residuals.

RESULTS

Considering the total sample trees used in the analyzes (210), the tree diameters (DBH) ranged from 5.10 to 34.40 cm and the heights from 3.40 to 15.40 m (Table 1). A total of 46 species were found in the Cerradão formation.

Table 1. Descriptive statistics of the dendrometric variables of the sample trees in the cerradão formation.
Tabela 1. Estatística descritiva das variáveis dendrométricas das árvores-amostras na formação cerradão

Variables	Minimum	Maximum	Mean ($\bar{X}$)	$Standard\ deviation$ (S)	No. of Observations	Species
DBH (cm)	5.10	34.40	12.86	5.73	210	46
Ht (m)	3.40	15.00	8.03	2.59		
No. of branches	0.00	35.00	4.97	5.24		
Volume (m^3)	0.0051	0.9983	0.1072	0.1354		

The estimation methods showed very close statistics with a slight superiority for the non-linear OLS method (Table 2). In addition, all parameters were statistically significant (p-value < 0.05).

Table 2. Estimates of the parameters of the Schumacher and Hall model (1933) obtained by nonlinear OLS, Quantile Regression (QR) and RANSAC algorithm and their respective statistics.
Tabela 2. Estimativas dos parâmetros referentes ao modelo de Schumacher e Hall (1933), obtidos por MQO não linear, Regressão Quantílica (RQ) e algoritmo RANSAC e suas respectivas estatísticas.

Methods	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$r_{\hat{y}y}$	**QRME**
OLS	0.00007	1.986000	0.90440	0.9068	0.0570
CI	0.00012; 0.000026	1.825263; 2.14674	0.61015;1.19865		
Quantile Regression (QR)	0.00007	2.050810	0.78890	0.9066	0.0588
CI	0.00005; 0.00009	1.87198; 2.22964	0.59545; 0.98235		
RANSAC	0.00011	2.133239	0.52925	0.9044	0.0579
CI	0.000083; 0.00013	1.947224; 2.31925	0.399473; 0.65903		

In Figure 1 it is possible to verify the influence of two points in the nonlinear least squares estimation for the generalized Cook distance. Observations 6 and 144 were identified as influential for the β estimates.

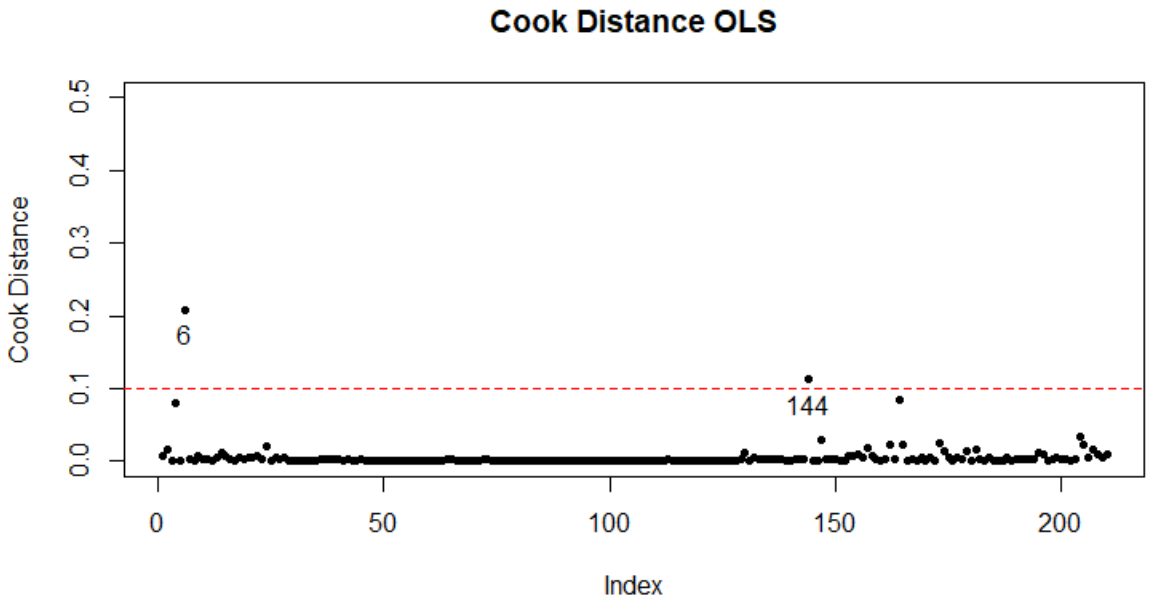

Figura 1. Análise de influência para os dados de volume.
Figure 1. Analysis of influence for volume data.

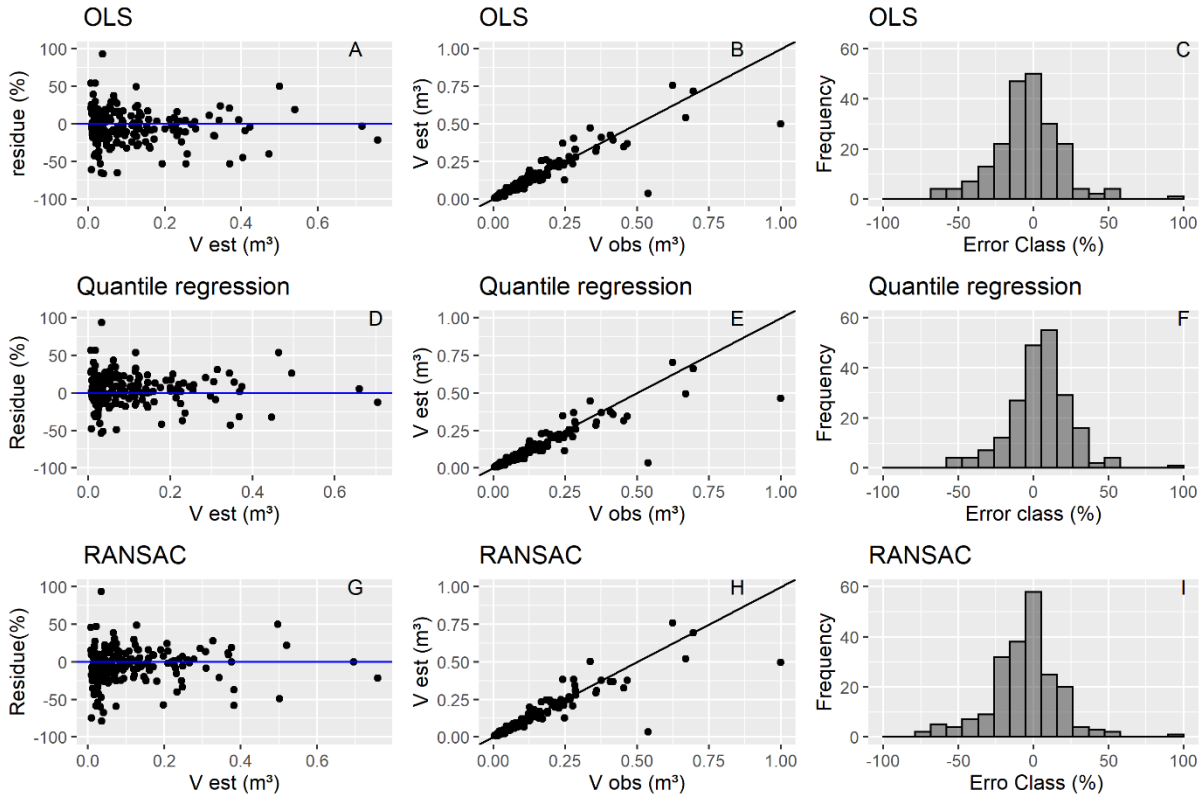

Figure 2. Residual distribution graphs, estimated versus observed value and histogram of error distribution for the three methods evaluated.
Figura 2. Gráficos de distribuição dos resíduos, valor estimado versus observado e histograma de distribuição dos erros para os três métodos avaliados.

The dispersion and graphic distribution of the residuals (Figure 2A; Figure 2D; Figure 2G) shows a greater dispersion in the smallest volumes. The greatest amplitude in this range of values was that obtained by the RANSAC algorithm and the smallest amplitude by the Quantile Regression QR. The presence of outliers was observed in all three methods when the estimated data were plotted in function of those observed (Figure 2B; Figure 2E; Figure 2H). The method that presented the least amplitude of error in the residual histogram was Quantile Regression, underestimating the volumes in class 10 (Figure 2F). Despite presenting the greatest amplitude of error, the model fit by the RANSAC algorithm was the methodology which had the largest number of observations in the error class 0 (zero) (Figure 2I).

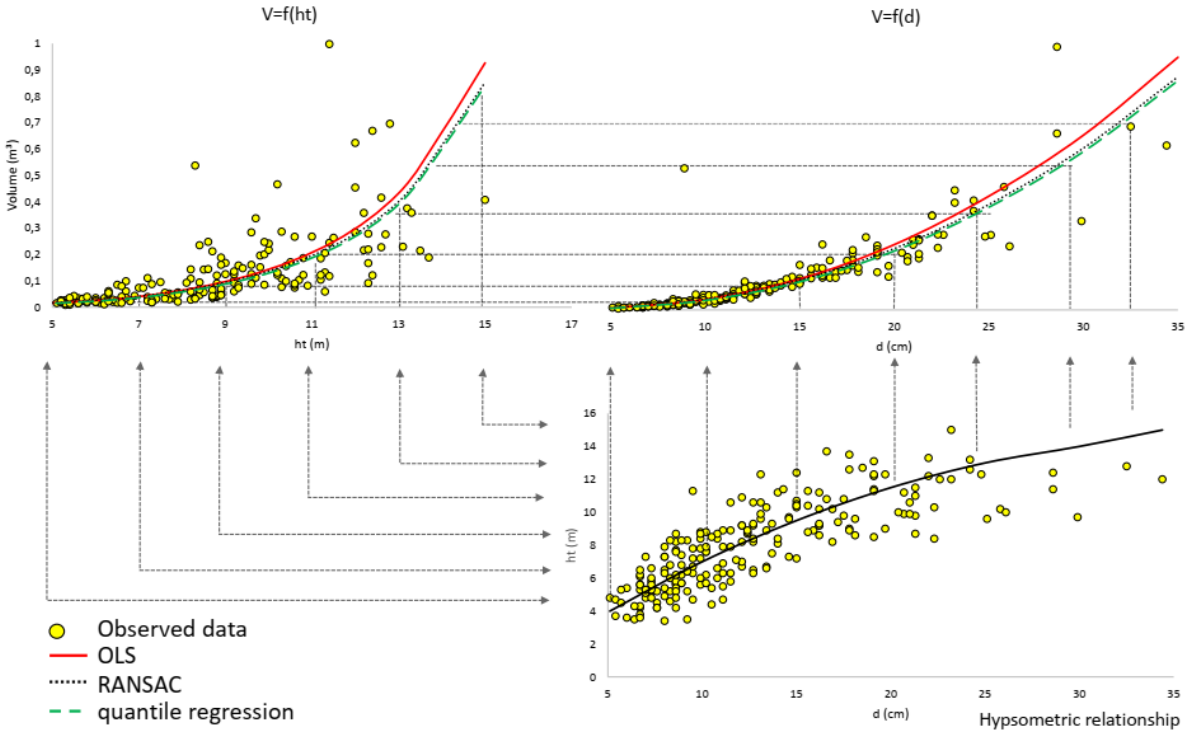

Figure 3. Trend of the curves obtained by estimation methods: nonlinear ordinary least squares (OLS), quantile regression and RANSAC algorithm (dotted line).
Figura 3. Tendência das curvas obtidas pelos métodos de estimação: mínimos quadrados ordinários não lineares (MQO), Regressão Quantílica e algoritmo RANSAC (linha pontilhada).

DISCUSSION

The least squares method is based on minimizing the sum of the squared errors, and therefore better statistics were expected than the alternative methods evaluated. It is worth noting that the QR and RANSAC methods do not discard the influential data in calculating the statistics, but only disregard it in obtaining the parameter estimates, causing the fit curve to pass through the observed data without their influence (CAO and WANG, 2015).

Quantile Regression was the method which most underestimated volumes; this fact occurred since the fit curve passes in the median of the data (KOENKER and BASSETT, 1978), and many of the observations of the Cerradão formation are above the median, making the model underestimate the volume. In addition, the more discrepant the point, the more distant the curves became, since the mean volume deviates from its median (ARAUJO JUNIOR *et al*., 2016).

The curve fit by the OLS was influenced by the discrepant data (Figure 3), given that the trend of the estimated curve was closer to them. The curve referring to the Quantile Regression and the RANSAC algorithm did not suffer influence of these data, maintaining the trend of the curves and very close to each other. In this situation in which the presence of influential data is observed, estimation methods which are not influenced by them should preferably be used (CHAMBERS and TZAVIDIS 2006).

It is important to note that the behavior of the curve fit by the RANSAC algorithm follows almost the same trend as the QR. However, it was fit according to an optimal configuration, which randomly selects the least amount of observations necessary to estimate the model parameters (DERPANIS *et al.*, 2010), removing influential data and keeping the model estimates stable against noise (LE *et al.*, 2017).

It is common to call discrepant data outliers and define some statistical criterion to remove them from the database. As a consequence, the removal of some data of this nature (outliers) in the example of this study would make the behavior of the curves obtained by the estimation methods very close, concluding with equality between them. However, a discrepant observation which may influence the tendency of the estimated curve must be carefully examined to determine the reasons for its peculiarity. Data of this nature sometimes provide information that other data could not provide given the combination of circumstances, which can be of vital importance or interest (DRAPPER and SMITH, 1998).

In the specific case of the Cerradão forest formation, trees with larger diameters may have smaller trunks due to bifurcation, so that the trunk volume is substantially smaller than trees of the same diameter, and are therefore characterized as discrepant data from the others, however without presenting any measurement error.

In addition, insufficient sampling to characterize the dispersion of tree sizes in natural forest vegetation in which there is a high diversity of species and environmental conditions (soil, climate, slopes, sun exposure, etc.) can induce a given observation to be discrepant from the others. This means it is not the observation itself which is discrepant, but the sampling was not correct to characterize the correct data trend.

CONCLUSION

- After the analysis, it can be concluded that Quantile Regression (QR) and the RANSAC algorithm are estimation methods indicated to be used in the presence of influential data, as in the case of the volume equations of the Cerradão formation, as they minimize the effect of these data on model parameter estimates.

REFERENCES

ARAÚJO JUNIOR, C. A.; SOARES, C. P. B.; LEITE, H. G. Curvas de índices de local em povoamentos de eucalipto obtidas por Regressão Quantílica. **Pesquisa agropecuária brasileira**, Brasília, v.51, n.6, p.720-727, jun. 2016.

BARROSO, L. M. A.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; SILVA, F. F.; CRUZ, C. D.; BHERING, L. L.; FERREIRA, R. P. Metodologia para análise de adaptabilidade e estabilidade por meio de Regressão Quantílica. **Pesquisa agropecuária brasileira.**, Brasília, v.50, n.4, p.290-297, abr. 2015.

BARNETT, V.; LEWIS, T. **Outliers in statistical data**. Chichester: John Wiley, 1995. 584 p.

BOHARA, S. B.; CAO, Q. V. Prediction of tree diameter growth using quantile regression and mixed-effects models**. Forest Ecology and Management**, Amsterdam, v.319, p.62–66, 2014.

BINOTI, D. H. B; SILVA, G. F. Aplicação da técnica Random Sample Consensus – RANSAC para o ajuste de modelos de regressão na mensuração florestal. Fundação de amparo a pesquisa do Espírito Santo. Vitória, 2016.

CAO, Q.V.; WANG, J. Evaluation of methods for calibrating a tree taper equation. **Forest Science**, Washington, v.61, p.213-219, 2015.

CHENG, L.; LI, M.; LIU, Y.; CAI, W.; CHEN, Y.; YANG, K. Remote sensing image matching by integrating affine invariant feature extraction and RANSAC. **Computers and Electrical Engineering**, Amsterdam, v.38, p.1023-1032, 2012.

CHAMBERS, R.; TZAVIDIS, N. M-quantile models for small area estimation. **Biometrika**, Oxford, v.93, p.255-268, 2006.

DERPANIS, K. G. Overview of the RANSAC algorithm. 2010. http://www.cse.yorku.ca/~kosta/CompVis_Notes/RANSAC.pdf. 16 abr. 2019.

DRAPPER, N.R.; SMITH, N. **Applied regression analysis**. 3 ed. New York: Wiley-Interscience, 736p. 1998.

FISCHLER, M. A; BOLLES, R. C. Ramdom sample consensus: a paradigma for model fitting with aplications to image analysis and automated cartography. **Comm. Of the ACM**, New York, v 24, p.381-395,1981.

HUSCH, B.; MILLER, C. I.; BEERS, T. W. **Forest mensuration**. New York: Wiley & Sons, 1982.

LE, V. H.; VU, H.; NGUYEN, T. T.; LE, T. L.; TRAN, T. H. Acquiring qualified samples for RANSAC using geometrical constraints. **Pattern Recognition Letters.** Amsterdam, v.102, p.58-66, 2017.

LI, Y.; GANS, N. R. Predictive RANSAC: Effective Model Fitting and Tracking Approach Under Heavy Noise and Outliers. **Computer Vision and Image Understanding**. Amsterdam, v.161, p.99-113, 2017.

KOENKER, R.; BASSETT, G. Regression quantiles. **Econometrica**, Ohio, v.46, n.1, p.33-50, 1978.

KOENKER, R. quantreg: Quantile regression. R package version 5.05. 2013. https://cran.r-project.org/web/packages/quantreg/index.html. 17 Abr. 2019.

MEYER, K.M.; WARD, D.; MOUSTAKAS, A.; WIEGAND, K. Big is not better: small *Acacia mellifera* shrubs are more vital after fire. **African Journal of Ecology**, Oxford, v.43, p.131-136, 2005.

SCHRÖDER, H. K.; ANDERSEN, H. E.; KIEHL, K. Rejecting the mean: Estimating the response of fen plant species to environmental factors by non-linear quantile regression. **Journal of Vegetation Science**, New York, v.16, p.373-382, 2005.

SCHUMACHER, F. X.; HALL, F. D. S. Logarithmic expression of timber-tree volume. **Journal of Agriculture Research**., Washington, v. 47, n. 9, p. 719-734, 1933.

SOUZA, G. S. **Introdução aos modelos de regressão linear e não linear**. Embrapa, 1998, 505p.

TEOH, S. T.; KITAMURA, M.; NAKAYAMA, Y. PUTRI, S.; MUKAI, Y.; FUKUSAKI, E. Random Sample Consensus combined with Partial Least Squares regression (RANSAC-PLS) for microbial metabolomics data mining and phenotype improvement. **Journal of Bioscience and Bioengineering**, Amsterdam, v.122, p.168-175, 2016.

THIFFAULT, N.; PICHER, G.; AUGER, I. Initial distance to Kalmia angustifolia as a predictor of planted conifer growth. **New Forests**, New York, v.43, p.849–868, 2012.

VINAY A.; RAO, A. S.; SHEKHAR, V. S.; KUMAR, A. MURTHY, K. N. B.; NATARAJAN, S. Feature Extraction using ORB-RANSAC for Face Recognition. **Procedia Computer Science**. Amsterdam, v. 70, p.174-184, 2015.

WOODALL, C. W.; RUSSELL, M. B.; WALTERS; B. F.; D'AMATO, A. W.; ZHU, K.; SAATCHI, S. S. Forest production dynamics along a wood density spectrum in eastern US forests. **Trees**, New York, v.29, p:299–310, 2015.