ISSN 0104-7760
v.25 n.4 2019

Luan Demarco Fiorentin[1a+], Wagner Hugo Bonat[1b], Allan Libanio Pelissari[1c], Sebastião do Amaral Machado[1d], Saulo Jorge Téo[2a], Gabriel Orso[1e]

GENERALIZED LINEAR MODELS FOR TREE SURVIVAL IN LOBLOLLY PINE PLANTATIONS

FIORENTIN, L. D.; BONAT, W. H.; PELISSARI, A. L.; MACHADO, S. A.; TEO, S. J.; ORSO, G. Generalized linear models for tree survival in loblolly pine plantations. **CERNE**, v. 25, n. 4, p.347-356, 2019.

HIGHLIGHTS

Regularization methods were tested for selecting covariates in loblolly pine survival models.

Stepwise method was the most suitable approach for selecting the covariates.

Link functions can influence the covariates selection procedure.

The survival models presented great ability to predict alive trees.

Keywords:
Elastic net
Link function
Logistic regression
Ridge regression
Stepwise method

ABSTRACT

To quantify the surviving trees in a forest stand and estimate the probability of an individual tree to survival are a fundamental task in forest management planning. Therefore, the main goal of this paper was to estimate the tree survival probability in loblolly pine (*Pinus taeda* L.) plantations based on generalized linear models (GLM). The data set was obtained from forest inventories carried out in the Midwest of Santa Catarina State, Brazil. The data analysis combined strategies for selecting covariates and different specifications of link functions in a Bernoulli GLM. We performed strategies for covariate selection at plot-level along with the standard stepwise procedure, where we considered the elastic net approach, as well as its special cases the lasso and ridge penalization. Our analyses showed that the stepwise procedure combined with the complementary log-log link function provide the best fit. The variables that most contributed to assess tree survival were basal area, number of individuals, maximum diameter, diameter of the average cross-sectional area and the diameter coefficient of variation per plots. This model presents 81.5% of accuracy given by ROC curve. Finally, we evaluated the fitted model by means of the half-Normal plots and randomized quantile residuals, whose results showed evidence of a suitable fit. We suggest the stepwise procedure for selecting covariates for a tree survival probability model, besides a complementary log-log link function.

Historic:
Received 25/06/2019
Accepted 13/12/2019

[+]Correspondence:
luandfiorentin@gmail.com

DOI:
10.1590/01047760201925042649

[1] Federal University of Paraná, Curitiba, Paraná, Brazil- ORCID: 0000-0003-1884-9849[a], 0000-0002-0349-7054[b], 0000-0002-0915-0238[c], 0000-0003-1010-4623[d], 0000-0001-5002-0649[e]
[2] University of Western Santa Catarina, Xanxerê, Santa Cararina, Brazil - ORCID: 0000-0003-4279-2635[a]

INTRODUCTION

Species of *Pinus* genus are cultivated in large-scale in the Southern region of Brazil, especially in Paraná and Santa Catarina States, mainly due to great adaptation to climatic conditions and their high timber economic potential. According to IBÁ - Indústria Brasileira de Árvores (2017), *Pinus taeda* L. and *Pinus elliottii* Engelm planted area covered more than 1.6 million of hectares in the base year of 2016, which represent 20.4% of the total planted area in the country.

The extensive *Pinus* planted area in Brazil implies that the trees are submitted to a wide range of environmental conditions and forest management systems, which results in a large range of timber productions. Therefore, statistical models able to express the forest developing in different conditions has become an important tool on the growth and yield planning. Despite important in individual tree growth simulators, tree survival is still few explored, probably because it is a rare phenomenon of high variability (Avila and Burkhart, 1992).

The tree survival or mortality in both planted and natural forest stands is a phenomenon associated to many factors (Adame et al., 2010) which include the competition among individuals; forest management practices, such as thinnings; climatic conditions (Diéguez-Aranda et al., 2005; Das and Stephenson, 2015; Thapa and Burkhart, 2015, Miranda et al. 2017; Téo, 2017); as well as the species genetic diversity. Thus, it is not completely clear how the tree mortality or survival occurs in a forest, once that individuals with similar features may present different outcomes.

To quantify the number of surviving trees over time is important in forest plantations. This information indicates the number of trees expected in the silvicultural rotation; and the potential timber assortments for being explored in the industry. Based on this, the tree survival probability at different site conditions and management systems can be obtained by statistical tools. Furthermore, regression models are essential on forest planning because can assist to identify factor associated to high or low survival probabilities.

One of these tools is logistic regression, a statistical approach widely used for estimating the tree survival probability in forest plantations (Yao et al., 2001; Diéguez-Aranda et al., 2005; Thapa and Burkhart, 2015; Téo, 2017). This model allows to express the survival probability through a linear predictor, which is usually composed by a set of tree and plot-level covariates. The linear predictor is connected to the expectation of the survival probability by a link function, frequently specified by a logit or probit functions (Téo, 2017; Vanclay, 1991; Yang et al., 2003; Yao et al., 2001).

Although their popularity, both logit and probit link functions share a limitation for the reason they are symmetric (McCullagh and Nelder, 1989). In practice, this feature can be a limitation depending of the data sets. Thus, Cauchit and complementary log-log asymmetric link functions are available in the statistical literature as alternative approaches. However, the suitability of these link functions is not well-known in the context of forest management research, doing this subject quite relevant.

The specification of a statistical model for modeling tree survival has at least two crucial choices: a suitable link function and which covariates will compose the linear predictor. In general, forest researchers have been used forward, backward, or stepwise selection procedures, where satisfactory results have been reported (Téo, 2017; Zhang et al., 2017). In this paper, we introduce an alternative approach for selecting covariates at plot-level based on regularization methods. The main idea of this methods is to fit a regression model whose parameter estimates are penalized or shrunken toward to zero. In this approach, the goal is to obtain estimates with lower variance at the cost of introducing some bias in the parameter estimates. This feature of the regularization methods can be used for selecting covariates measured in forest plantations, once that they present high value of correlation among them, which implies in large standard errors.

Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge regression are a frequently applied regularization technique (Tibishirani, 1996). An extension of these strategies is the Elastic Net approach which is a combination of Lasso and Ridge penalizations (Zou and Hastie, 2005). These techniques penalize the covariates by shrinking the parameter estimate and enabling the removal of the covariates whose estimated effects approach zero. Thus, our research hypothesis is that the regularization methods are appropriated for selecting correlated covariates, once this approach can reduce the variance of the parameter estimates.

The aim of this paper was to estimate the probability of loblolly pine (*Pinus taeda*) survival in forest plantations; and to identify which factors are associated to the tree survival. Therefore, we obtained data from forest inventories carried out in the Midwest of Santa Catarina State, Brazil. Our data was composed by a set of covariates usually measured at plot-level. The response variable was a binary value that indicated whether the tree is alive or not. We investigated and compared Bernoulli's regression models fitted by the link

functions logit, probit, Cauchit and complementary log-log. Furthermore, we performed the covariates selection based on the standard stepwise procedure, as well as the methods based on regularization as the Lasso, ridge and elastic net approaches.

We present the data set, a brief description about the study area and the modeling strategies in the material and methods section. The results section describes an exploratory data analysis and the application of the models to the data. We also present a discussion section about the main results. Finally, concluding remarks are presented in the conclusion section.

MATERIAL AND METHODS

Study area

The study area corresponds to loblolly pine (*Pinus taeda*) plantations, located at Midwest region of Santa Catarina state, Brazil. The plantations are distributed on the municipalities of Caçador, Lebon Régis, Macieira, Rio das Antas, Santa Cecília, and Timbó Grande. According to IBGE - Instituto Brasileiro de Geografia e Estatística (2012), the region presents original vegetation belonging to Mixed Ombrophilous Forest (MOF), under the Montane Mixed Ombrophilous Forest. Based on the Köppen classification, the study region presents a Cfb climate type, that is, a wet subtropical zone, oceanic climate, without a dry season and with summers temperate. The average temperature of the warmest month is 19.7 °C and the coldest month is 11.5 °C, and the annual precipitation is 1,736 mm (ALVARES et al., 2014).

The forest plantations were planted with an average initial spacing of 2.5 x 2.5 m (1,600 trees per hectare). The ideal rotation age is 25 years, with three commercial thinnings usually performed at 10, 15, and 20 years old. In the first thinning, 50% of the trees per hectare were removed; while 40% of the remaining trees were removed in the second thinning; in the third and last thinning were removed 30% of the remaining trees.

Data set

The data set was obtained from forest inventory performed in two occasions carried out at 2009 and 2015. The age ranged from 5.5 to 35.2 years old. In addition, due to the difference of six years between both forest inventories and because we re-sampling a few sample units, we assume that there is no correlation between measures.

The plots had dimensions from 497.93 to 739.68 m^2, which were randomly allocated (simple random sampling) in the study area, by using a stratified sampling process. The stratum represented administrative divisions of the company (projects and stands). The diameter at breast height (DBH) was measured at 1.30 m of height in all trees inside each sub-sample. The total height of 20% of the trees in each plot was indirectly taken by using a hypsometer Vertex III. The trees dominant height was measured in individuals without bifurcation or defects over the stem and crown, and it was defined proportionally as the 100 trees with largest diameter at breast height per hectare.

The data set we used for modeling was composed by 13 random variables measured at plot-level. The number of trees selected was 40,556 trees. The description of each variable is given as follow:

survival: binary variable – takes value 1 if the tree is alive or 0 otherwise. The classification of alive tree was performed when the data was collected in the forest inventory. The tree was considered a dead individual when green branches were not observed on the field. In our approach, both regular and irregular mortality were combined. The regular mortality was due to the natural competition among trees and the senescence process. The irregular mortality was caused by irregular factors, as monkey attacks, which are quite common at the study area.

age: continuous variable – age of the tree (years);

gsample: continuous variable – sum of cross-section areas (m^2) of the trees inside plot.

nsample: discrete variable – number of trees inside plot;

daverage: continuous variable – average diameter (cm) of the trees inside plot;

dcv: continuous variable – coefficient of variation (%) of the diameters inside plot;

dg: continuous variable – quadratic average diameter (cm) of the trees inside plot;

dmax: continuous variable – maximum diameter (cm) of the trees inside plot;

ddom: continuous variable – dominant diameter (cm) of the trees inside plot. This variable was computed based on the average diameter of the one hundred largest trees per hectare, but proportionally to the size of each plot;

hdom: continuous variable – dominant height (m) of the plot. This variable was computed based on the height average of the one hundred largest trees by hectare, but proportionally to the size of each plot;

thinsample: binary variable – takes value 1 whether were performed thinnings on the plot or 0 otherwise;

gthin: continuous variable – sum of removed cross-sectional area on the plot during the thinnings;

nthin: discrete variable – number of trees removed during the thinnings on the plot.

The generalized linear model

Tree survival (*survival*) was the response variable, which takes a binary value, i.e., the response variable take value 1 whether the tree is alive and 0 otherwise. Therefore, we applied a Bernoulli's regression model due to the nature of response variable (McCullagh and Nelder, 1989). The systematic component was formulated by a linear combination of a set of predictor variables, besides a link function selected according to the behavior of response variable. The specification of the model is given as, where is the random variable, whose observed values are denoted by y_i, $i=1,2,...,n$; $x_{i1},...,x_{ip}$ are vectors of the predictor variables X_i; π_i is the probability of success, i.e., is the survival probability; g is a differentiable and monotone link function; η_i is the linear predictor; and $\beta_0, \beta_1, \ldots, \beta_p$ are parameters to be estimated.

$$Y_i|x_i \sim Bernoulli(\pi_i) \quad [1]$$

$$g(\pi_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad [2]$$

Linear predictor and link function selection

For composing the linear predictor, 12 covariates were available. We applied two strategies for selecting the covariates:

I) Stepwise: covariate selection was based on the minimization of the Bayesian Information Criterion (BIC), given by the following expression, where is the maximized log-likelihood value; is the number of observations; and is the number of parameters of the model. This algorithm is a combination of backward and forward procedure, where the covariates are added or removed in successive iterations until obtaining the smallest BIC. Thus, we assumed that this methodology is the standard approach in forest modeling due to its large applications.

$$BIC = -2\hat{l} + ln(n)p \quad [3]$$

II) Regularization: covariates selection was performed with regularization methods. This procedure is based on penalizations controlled by the parameter ; while the penalization intensity was quantified by parameter . The general formulation is given by equation 4. For the especial case where , we obtained a first order penalization, also called as lasso regularization method. A second order penalization was defined when , and the method is called as ridge regression. The Elastic Net is an intermediate penalization when , and we tested a large grid of penalization intensity. The optimum was determined by cross-validation, using the cv.glmnet function of the glmnet package (Friedman et al., 2010) on the R software (R Core Team, 2019). In this approach, our main goal was to identify the smallest loss for a sequence of . Still, we tested the loss function based on Mean Squared Error (MSE), Mean Absolute Error (MAE) and Deviance (DEV). Once that the , the penalization term has no effect when , and the parameter estimates are equal to the maximum likelihood estimates. However, when the penalization is strong and the parameter estimates tend to zero (Tibshirani, 1996). The covariates have different nature, what can influence on the selection procedure; thus, we standardized them for minimizing their scale effects.

$$\frac{1}{n}\sum_{i=1}^{n} \hat{l}\left(y_i, \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) + \lambda\left[\alpha \sum_{i=1}^{n} ||\beta_p|| + (1-\alpha)\sum_{i=1}^{n} ||\beta_p^2||\right] \quad [4]$$

After defining the best and parameters and the covariates selected on the regularized model, we specified four link functions. The Cauchit (01), complement log-log (02), logit (03) and probit (04) link functions were tested for verifying their influence on the selection of covariates for the stepwise approach. The most suitable link function was based on the smallest value of Bayesian Information Criterion (BIC), once that the models can present different number of covariates. The generalized linear model specification for each link function on linear predictor scale is given by equations 5, 6, 7 and 8, where *tan* is the tangent function; is the natural logarithm; and ϕ^{-1} is the inverse of probability density function of the standard normal distribution.

$$tan[\pi(\pi_i - 0.5)] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad [5]$$

$$ln[-ln(1-\pi_i)] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad [6]$$

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad [7]$$

$$\phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad [8]$$

Investigating the performance of non-normal models usually cannot be done by traditional residual analysis. Therefore, for evaluating the assumptions of the fitted models, we performed a diagnostic analysis based on half-Normal plots (HNP) with simulated envelope, built by hnp function of the hnp package (Moral et al., 2017) on the R software. The idea behind HNP is to verify whether the error distribution was specified in an appropriately way. Thus, for a well-fitted model, the simulated envelope is such that the model diagnostics measures are likely to fall within it. The main purpose of the envelope is to serve as an indicative of what we expect about the residuals under a well-fitted model

(Moral et al., 2017). Still, we computed the randomized quantile residuals (RQR) as a complementary analysis. In this case, if our model is correctly specified, we expect the residuals follow a normal distribution (Dunn and Smyth, 1996).

Predictive performance

The predictive performance of the models was compared by standard methods. The data set was randomly split in two subsets. The fitting data was composed by approximately 90% of the observations and was used to fit the survival model. The validation data set was applied for evaluating the prediction performance of them models by Receiver Operating Characteristic curve (ROC) of the ROCR package (Sing et al., 2005) on the R software. The sensibility (*Sens*) and specificity (*Esp*) of each model was estimated for 0.75; 0.85; 0.90; 0.95 and 0.99 probability cut points. These measures indicate the performance of the models for classifying individuals in survival or non-survival, in which the more suitable cut point was obtained based on Youden and Closest Topleft rules (Unal, 2017), whose expression are given respectively as equation 9.

$$Youden = \max[Sens + Esp] \quad \text{and} \quad CT = \min[(1 - Sens)^2 + (1 - Esp)^2] \quad [9]$$

RESULTS

In this section, we presented an exploratory analysis of the variables and how they are related which others. We also showed the effect of the link functions in selecting covariates for composing the linear predictor of the generalized linear model, besides the main results obtained on the covariates selection procedure. Finally, we applied the best models in the validation data set for assessing their prediction performance.

Exploratory data analysis

Boxplots presented in Figure 1 suggested an asymmetric distribution of the covariates according to the response variable levels and a possible significant effect of the covariates based on *diameters* measures and *age*. Figure 2 presents a correlogram based on Spearman´s rank correlation coefficient, where the covariates were clustered by the centroid method (Mingoti, 2005). Three groups with high correlation values stand out, which suggest that multicollinearity can be a concerning problem for this data set and highlights the need of a covariate selection. The *nsample* covariate had a negative relationship with other covariates that directly express tree dimensions. This indicates that as the number of trees in the sample increase, the tree individual dimensions tend to decrease. Moreover, thinnings covariates showed high positive correlation among them, but negative correlation with almost all the other covariates. High positive correlation values were also observed for covariates directly computed based on tree-level measures, such as diameter and height.

Fitting the models

The stepwise procedure selected the covariates *gsample*, *nsample*, *dcv*, *dg* and *dmax* for composing the linear predictor. On the other hand, all covariates were selected by the regularization methods. The best value obtained by cross-validation was close to 0 for all sequences of that perform the Lasso, Ridge, and Elastic Net procedures, regardless of loss measure tested (MSE, MAE or DEV), indicating that the penalization term had no effect on the parameter estimates. However, even when we fitted the model with all covariates, only *gsample*, *nsample*, *dcv*, *dmax*, *gthin,* and *nthin* were significant (Table 2). Thus, we decided to continue the data analysis considering these covariates in their natural scale. So, we could easily interpret their effects on tree survival.

Bayesian information criterion (BIC) and residual deviance (RD) indicated that the complementary log-log link function provided the best fit in both modeling approaches (Table 1). However, when BIC values were compared between covariate selection approaches, the stepwise procedure provided the best fit for all link functions. This result is related to the largest penalty on the log-likelihood function of the model based on the regularization approach due to the largest number of parameters. Furthermore, complement log-log and probit link functions had the same covariates selected by the stepwise procedure. For logit link function, this method also selected the covariates related to thinnings, such as *gthin* and *nthin*, while Cauchit selected *ddom* and *daverage*.

We performed a graphical analysis for evaluating the assumptions of the fitted models. Our models were based on a Bernoulli specification of the binary response variable *survival*. Thus, the assumptions usually assumed for normal data are no longer demanded. The half-Normal plot presented in Figure 3 suggested that the models were properly specified, once the residuals do not exceed the simulated envelope. However, both models presented similar behavior, indicating a good fit and a suitable probability distribution of response variable. As a feature of the randomly quantile residuals, when the model is suitable to the data it should be expected a normal distribution of the residuals, regardless of the distribution of the response variable and selected

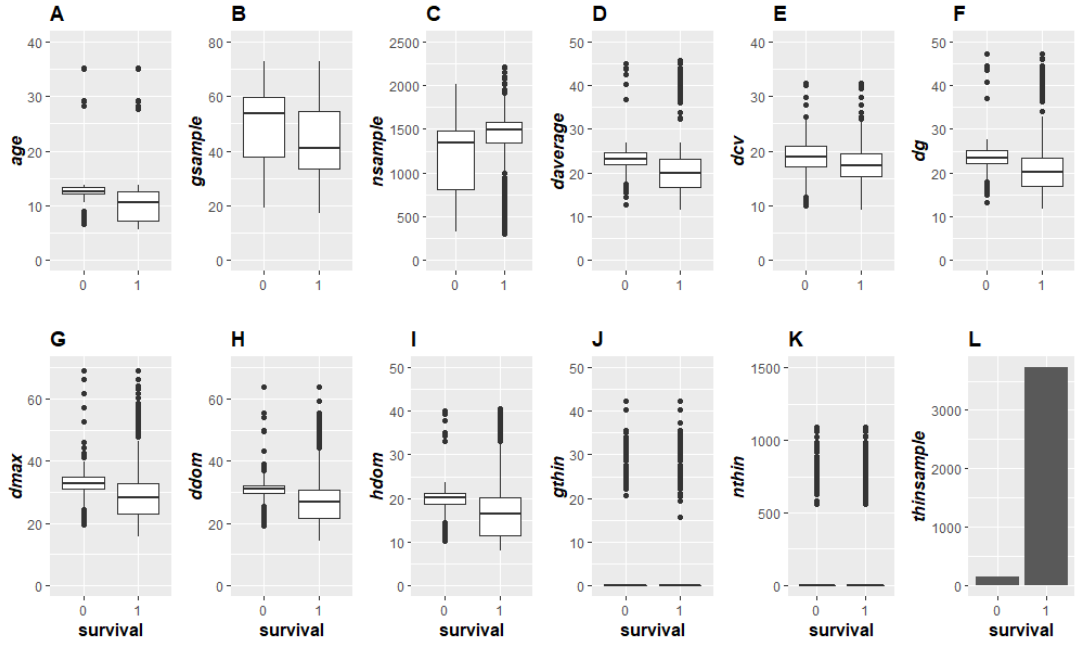

FIGURE 1 Boxplots (A:K) and barchart (L) of the covariates by survival index..

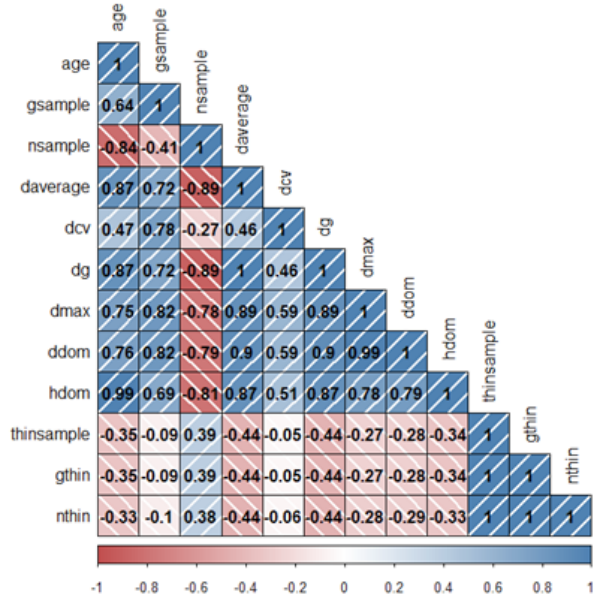

FIGURE 2 Correlogram between variables clustered by centroid method.

covariates. In our case, sample and theoretical residual quantile had a linear association (Figure 3), confirming a good performance of the fitted models and a normal distribution of the residuals.

Some preliminary analyses indicated that only the main effects of covariates were suitable for modeling the response variable *survival*, in which interaction terms are not required to be included in the linear predictor. Parameter estimates and standard errors for both covariate selection procedures are presented in Table 2. Point estimates of the fitted model based on stepwise selection suggested that the response variable has negative relation to *gsample* and *dcv*, since the associated parameters had a negative sign. In practice, larger cross-sectional area and higher diameter variability in the sample are associated with a lower individual survival probability. On the other side, *nsample*, *dg*, and *dmax* covariates are associated with higher values of survival probability.

TABLE 1 Bayesian information criterion (BIC) and residual deviance (RD) by link functions and covariate selection methods.

Link function	BIC (Number of variables)		Residual Deviance	
	Stepwise	Regularization	Stepwise	Regularization
Cauchit	7,068.31 (9)	7,094.78 (12)	6,963.30	6,958.20
C. log-log	6,847.46 (5)	6,904.20 (12)	6,784.40	6,768.30
Logit	6,874.73 (7)	6,910.75 (12)	6,790.70	6,774.20
Probit	6,851.50 (5)	6,906.84 (12)	6,788.50	6,769.30

TABLE 2 Parameter estimates, standard errors (SE) and p-value of the fitted models with complementary log-log link function on the linear predictor scale.

Parameter	Estimate	SE	*p*-value	Estimate	SE	p-value
	Regularization			Stepwise		
intercept	-0.2940	0.3917	$p > 0.05$	-0.3973	0.2305	$p \leq 0.10$
age	-0.0097	0.0128	$p > 0.05$	-	-	-
gsample	-0.0404	0.0034	$p \leq 0.05$	-0.0413	0.0024	$p \leq 0.05$
nsample	0.0017	0.0001	$p \leq 0.05$	0.0017	0.0001	$p \leq 0.05$
daverage	-0.4005	0.3486	$p > 0.05$	-	-	-
dcv	-0.0484	0.0146	$p \leq 0.05$	-0.0411	0.0047	$p \leq 0.05$
dg	0.4891	0.3469	$p > 0.05$	0.0575	0.0118	$p \leq 0.05$
dmax	0.0451	0.0093	$p \leq 0.05$	0.0312	0.0069	$p \leq 0.05$
ddom	-0.0426	0.0224	$p > 0.05$	-	-	-
hdom	0.0028	0.0102	$p > 0.05$	-	-	-
thinsample	-0.0238	0.1884	$p > 0.05$	-	-	-
gthin	0.0230	0.0098	$p \leq 0.05$	-	-	-
nthin	-0.0008	0.0003	$p \leq 0.05$	-	-	-

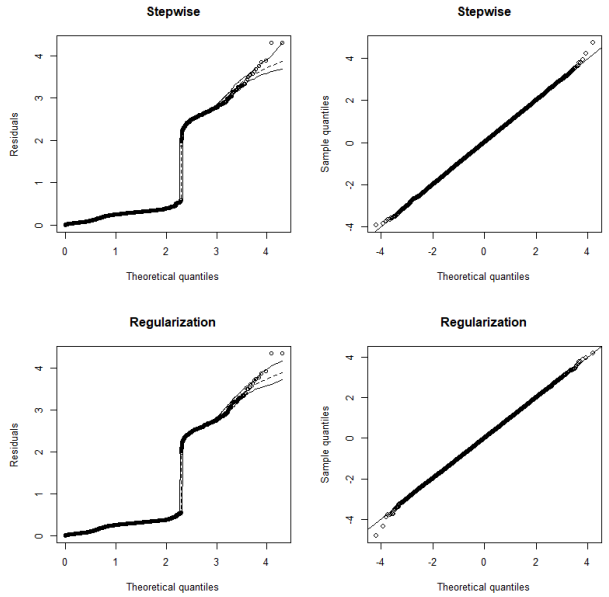

FIGURE 3 Half-Normal plot (left) and randomly quantile residuals (right) for assessing the fitted models.

Predictivity performance

A validation data set was used for comparing the performance of the fitted models in predicting the response variable, once the forest planning directly depends on the estimated number of alive trees in a forest stand. The ROC curves were similar for both models (Figure 4). However, the area under the curve was 0.805 for the model selected by stepwise procedure, and 0.814 for the model chosen by regularization method, indicating a slightly better predictions for the model with more parameters.

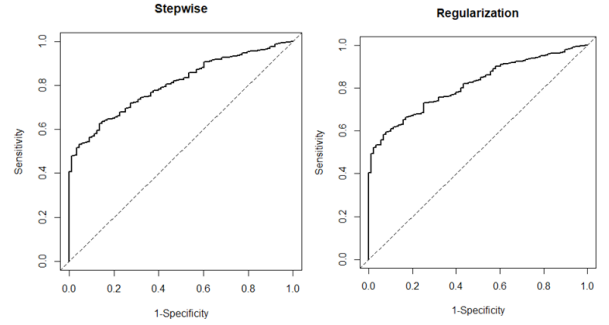

FIGURE 4 ROC curve of the models applied to the validation data set.

When we changed the cut point for defining a suitable probability value for classifying trees in survivors or non-survivors, the best result was obtained with a 0.99 probability cut point. This result was observed for both models, once that in this probability cut point was obtained the highest value in Younden´s rule and the lowest value in Closest Topleft´s rule (Table 3). We also noticed that the model based on the regularization procedure presented slightly higher values in the decision rules than the stepwise procedure, resulting in a better performance for classifying the individuals.

TABLE 3 Youden and Closest Topleft decision rules for different probability cut points of the models applied to the validation data set.

Model	Cut point	Youden	Closest Topleft
Stepwise	0.75	1.000	1.000
	0.85	1.010	0.977
	0.90	1.049	0.890
	0.95	1.300	0.372
	0.99	1.449	0.291
Regularization	0.75	1.000	1.000
	0.85	1.000	1.000
	0.90	1.049	0.890
	0.95	1.315	0.347
	0.99	1.454	0.286

The estimated sensitivity and specificity values for a 0.99 probability cut point are presented in Table 4. The results suggested that the models have great capacity to identify alive trees, due to the high sensitivity value. However, the low specificity value can be related to the rare non-surviving trees in the sample.

TABLE 4 Sensitivity and specificity of the selected models applied to the validation data set for a 0.99 probability cut point.

Model	Sensitivity	Specificity
Stepwise	0.989	0.460
Regularization	0.989	0.466

DISCUSSION

The main goal of this paper was to specify and fit a generalized linear model for estimating the tree survival probability in loblolly pine plantations. We tested four strategies of covariate selections based on stepwise and regularization procedures, such as ridge regression, lasso and elastic net method. We were also interested in analyzing the influence of link functions when selecting covariates for composing the linear predictor. Initially, we expected that the regularization method would be more appropriate for selecting correlated covariates, which are common in forest variables, because this approach can include some bias in parameter estimates in contrast to reduce their variance. Since the covariates are correlated and standard errors are larger, regularization procedures are quite promising in forest modeling. However, the penalization term had no effect in our model. As consequence, the stepwise procedure performed best due to the fewer selected covariates, making this a more parsimoniously procedure.

A different number of selected covariates for composing the linear predictor of the model can be obtained when we consider different link functions, what suggests that the link function must be appropriated for a specified data set. Despite preference by Logit link function on the tree survival probability modeling in forest plantations (Téo, 2017; Yang et al., 2003;

Yao et al., 2001), better results on BIC were obtained for complementary log-log and probit link functions, which provided models with a few parameters. The performance of the complementary log-log link function showed evidence that the behavior of tree survival probability is asymmetric when related to the linear predictor, once that the individual tree survival probability approaches to zero and one in different rate. Thus, considering a symmetric link function may not be a reasonable assumption in tree survival modeling (Jiang *et al.*, 2013). These results became relevant because the probability of success presented values quite near of one, where the link functions show more discrepancy.

Our models performed well for fitting and predicting the survival probabilities. However, better results can be obtained whether more covariates are considered for composing the linear predictor, such as environmental variables, mainly whether the model is applied to large areas. Zhang et al. (2017) modeled the mortality of forest plantations located at China using climatic covariates, besides initial planting density and competitions indexes. The authors suggested the inclusion of climatic variables in mortality models can facilitate the projection of tree mortality under future climate change conditions. Thapa and Burkhart (2015) tested climatic and soil effects on tree mortality, and the predictions performed best when they included these covariates. However, climatic variables were significant just when the model was fitted for large areas, which suggests that only climatic effects play a minor role in small areas.

In forest research involving tree mortality or survival, tree competition indices are commonly used as predictor variables (Miranda et al., 2017; Téo, 2017, Zhang et al., 2017). However, these indexes are computed in function of covariates usually included in the linear predictor. As an example, basal area larger index (BAL) is obtained by summing the cross-sectional area of all trees with larger diameter than the object tree (Eid and Tuhus, 2001), then being a function of the diameter at breast height. This procedure can induce a correlation between both variables (Miranda, 2016; Schröder and Gadow, 1999). Consequence of correlated covariates is a larger standard error of the parameters estimates, that can compromise the hypothesis tests and inferences. In our preliminary analysis, changes in the parameters sign and standard error of the stepwise model were observed when we removed the covariates *dg* or *dmax*. This result is explained by the high correlation value (0.95) between them.

We tested *thinsample*, *gthin* and *nthin* covariates for accounting possible thinning effects on tree survival probability. However, similar to what was found by Avila and Burkhart (1992), no improvement was obtained in the predictions when those variables were added to the model. A possible reason is that the mortality is a quite rare phenomenon, and after thinning we also do not expect a relevant regular mortality. According to Bose et al. (2018), commercial thinning treatment replaced self-thinning of suppressed trees; thus, decreasing tree mortality in loblolly pine and Douglas-fir plantations in North America. The authors also highlighted that the thinning was effective for reducing long-term tree mortality in red spruce and balsam fir, confirming the significance of thinning intensity and basal area as relevant predictor covariates.

Our tree survival probability models presented a great ability to predict alive trees, as suggested by the sensitivity statistic table 4. Téo (2017) used logistic regression combined with logit link function for modeling *Pinus taeda* tree survival probability in Midwest of Santa Catarina. The sensitivity of his model was 98.9% and the specificity was 43.1% for irregular mortality, being similar that one obtained in this paper. When the author considered only regular mortality, sensitivity and specificity were 99.1% and 52.3%, respectively. These results suggest that the natural mortality is more regular than that one caused by external factors.

A possible reason for the discrepancy observed for sensitivity and specificity of the model was the different number of survived and non-survived trees. These imbalance between alive and dead trees classes have influence on the effectiveness of the model. In general, tree survival probability models usually do not present high values of specificity (Adame et al., 2010; Téo, 2017), what may be related to the few dead trees in a forest plantation when compared to the number of alive trees. As alternative, Kuhn and Johnson (2013) suggested to use more balanced prior probabilities or a balanced training set may help to deal with this class imbalance. However, this approach still requires detailed researches in forest modeling. Another possible reason for lower values of specificity is related to the lack of ability of the covariates usually measured at forest inventories in identifying the dead individuals. Thus, we recommend testing more covariates for increase the specificity of the survival models.

Finally, futures topics to be explored in survival modeling are related to the inclusion of forest inventories performed in several occasions, with several occasions of sample units measurements, defining a longitudinal study, due to the temporal dependency among observations. Applications of spatial statistics should be considered for improving the analysis of tree survival in forest stands, since the environmental gradient can influence the tree individual mortality.

CONCLUSION

In this study, we specify and fit a generalized linear model for estimating the probability of loblolly pine tree survival in forest plantations, considering covariates usually measured in forest inventories. The plot-level variables that most contributed to assess tree survival were basal area, number of trees, maximum diameter, diameter of the average cross-sectional area and the diameter coefficient of variation.

The stepwise procedure for selecting covariates was more parsimonious than the regularization procedures tested; and combined with complementary log-log link function was the procedure provided the most suitable model. The model presented a great prediction ability, mainly due to the high number of survival trees. Additional researches related to regularization techniques are recommended in forest modeling, mainly regarding survival and individual growth models.

ACKNOWLEDGMENTS

The authors thank CAPES and CNPq for their financial support for this research.

REFERENCES

ADAME, P.; RÍO, M. del; CAÑELLAS, I. Modeling individual-tree mortality in Pyrenean oak (*Quercus pyrenaica* Willd.) stands. **Annals of Forest Science**, v.67, n.8, p.810, 2010.

AVILA, O.B.; BURKHART, H.E. Modeling survival of loblolly pine trees in thinned and unthinned plantations. **Canadian Journal of Forest Research**, v.22, n.12, p.1878-1882, 1992.

BOSE, A.K.; WEISKITTEL, A.; Kuehne, C.; WAGNER, R.G.; TURNBLOM, R.; BURKHART H.E. Tree-level growth and survival following commercial thinning of four major softwood species in North America. **Forest Ecology and Management**, v.427, p.355-364, 2018.

DAS, A.J.; STEPHENSON, N.L. Improving estimates of tree mortality probability using potential growth rate. **Canadian Journal of Forest Research**, v.45, n.7, p.920-928, 2015.

DÍEGUEZ-ARANDA, U.; CASTEDO-DORADO F.; ÁLVAREZ-GONZÁLEZ J.G.; RODRÍGUEZ-SOALLEIRO. Modeling mortality of Scot Pine (*Pinus sylvestris* L.) plantations in the northwest of Spain. **European Journal of Forest Research**, v.124, p.143-153, 2005.

ALVARES, C.A.; STAPE, J.L.; SENTELHAS, P.C.; GONÇALVES, J.L.M.; SPAROVEK, G. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v.22, n.6, p.711-728, 2014.

DUNN, P.K.; SMYTH G.K. Randomized Quantile Residuals. **Journal of Computational and Graphical Statistics**, v.5, n.3, p.236-244, 1996.

EID, T.; TUHUS, E. Model for individual tree mortality in Norway. **Forest Ecology and Management**, v.154, p.69-84. 2001.

FRIEDMAN J.; HASTIE, T.; TIBSHIRANI R. Regularization Paths for Generalized Linear Models via Coordinate Descent. **Journal of Statistical Software**, v.33, n.1, p.1-22, 2010.

IBÁ. **Industria brasileira de árvores**. Available in http://iba.org/images/shared/Biblioteca/IBA_RelatorioAnual2017.pdf. 2017.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Manuais técnicos em geociências: Manual técnico da vegetação brasileira**, n.1, 2°ed. Rio de Janeiro: IBGE, 2012. 275p.

JIANG, B.X.; DEY, D.K.; PRUNIER, R.; WILSON, A.M.; HOLSINGER, K.E. A new class of flexible link functions with applications to species co-occurrence in Cape florist region. **The Annals of Applied Statistics**, v.7, n.4, p.2180-2204, 2013.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. Springer, 2013.

MCCULLAGH, P.; NELDER, J.A. **Generalized Linear Models**. Chapman & Hall, 2°ed., 1989.

MIRANDA, R.O.V. **MODELAGEM DE ÁRVORES INDIVIDUAIS PARA POVOAMENTOS NÃO DESBASTADOS DE *Pinus taeda* L.** 169 f. Dissertação (Mestrado em Engenharia Florestal) – Universidade Federal do Paraná, Curitiba, 2016.

TÉO, S.J. **Modelagem do crescimento e produção de árvore individual independente da distância, para *Pinus taeda* L., na região meio oeste do estado de Santa Catarina**. 2017. 283 p. PhD thesis Universidade Federal do Paraná, Curitiba.

MIRANDA R.O.V.; FIGUEIREDO FILHO, A.; MACHADO, S.A.; CASTRO, R.V.O.; FIORENTIN, L.D.; BERNETT, L.G. Modelagem da mortalidade em povoamentos de *Pinus taeda* L. **Scientia Forestalis**, v.15, n.115, p.435-444, 2017.

MINGOTI, S.L. **Análise de Dados Através de Métodos de Estatística Multivariada**. Belo Horizonte: Editora UFMG, 2005.

MORAL, R.A.; HINDLE, J.; DEMÉTRIO, C.G.B. Half-Normal Plots and Overdispersed Models in R: The hnp Package. **Journal of Statistical Software**, v.81, 2017.

R Core Team. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria, 2019.

SCHÖDER, J.; GADOW, K. von. Testing a new competition index for Maritime pine in northwestern Spain. **Canadian Journal of Forest Research**, v.29, n.2, p.280-283, 1999.

SING T.; SANDER O.; BEERENWINKEL N.; LENGAUER T. ROCR: visualizing classifier performance in R. **Bioinformatics**, v.21, n.20, p.7881, 2005.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society**, v.58, n.1, p.267-288, 1996.

UNAL, I. Defining an optimal cut-point value in ROC: analysis: an alternative approach. **Computational and Mathematics Methods in Medicine**, 2017.

THAPA, R.; BURKHART, H.E. Modeling stand-level mortality of loblolly pine (*Pinus taeda* L.) using stand, climate, and soil variables. **Forest Science**, v.61, n.5, p.834-846, 2015.

VANCLAY, J.K. Mortality functions for North Queensland rain forests. **Journal of Tropical Forest Science**, v.4, n1, p.15-36, 1991.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal Royal of Statistical Society – Series B**, v.67, n.2, p.301-320, 2005.

YANG, Y.; TITUS, S.J.; HUANG, S. Modeling individual tree mortality for white spruce in Alberta. **Ecological Modelling**, v.163, n.3, p.209-222, 2003.

YAO, X.; TITUS, S.J.; MACDONALD, S.E. A generalized logistic model of individual tree mortality of aspen, white spruce, and lodgepole pine in Alberta mixedwood forests. **Canadian Journal of Forest Research**, v.31, n.2, p.283-291, 2001.

ZHANG, X.; CAO, Q.V.; DUAN, A.; ZHANG J. Modeling tree mortality in relation to climate, initial planting density, and competition in Chinese fir plantations using a Bayesian logistic multilevel method. **Canadian Journal of Forest Research**, v.47, p.1278-1285, 2017.