# STATISTICAL TECHNIQUES APPLIED TO FOREST BIOMETRICS

LUAN DEMARCO FIORENTIN
Sebastião do Amaral Machado (Advisor)
Allan Libanio Pelissari (Co-advisor)
Saulo Jorge Téo (Co-advisor)
Wagner Hugo Bonat (Co-advisor)

Federal University of Paraná
Postgraduate Program in Forestry Engineering
FUPR/PPFE

Friday, January 24, 2020

# Sumário


1. CHAPTERS
2. CHAPTER I
   - Covariance generalized linear models: an approach for quantifying uncertainty in tree stem taper modeling
3. CHAPTER II
   - Joint marginal modeling of height and volume for *Araucaria angustifolia*
4. CHAPTER III
   - Generalized linear models for tree survival in forest stands
5. REFERENCES AND ACKNOWLEDGMENT

# CHAPTERS

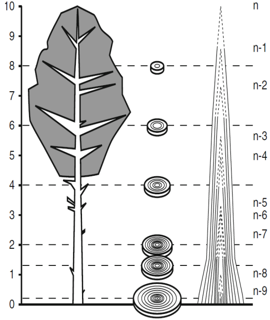
- *Chapter I*: **Covariance generalized linear models: an approach for quantifying uncertainty in tree stem taper modeling**
- *Chapter II*: **Joint marginal modeling of height and volume for *Araucaria angustifolia***
- *Chapter III*: **Generalized linear models for tree survival in forest stands**

# Sumário

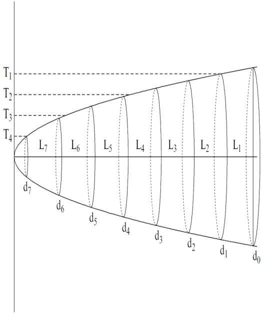

1. CHAPTERS
2. CHAPTER I
   - Covariance generalized linear models: an approach for quantifying uncertainty in tree stem taper modeling
3. CHAPTER II
   - Joint marginal modeling of height and volume for *Araucaria angustifolia*
4. CHAPTER III
   - Generalized linear models for tree survival in forest stands
5. REFERENCES AND ACKNOWLEDGMENT

# INTRODUCTION

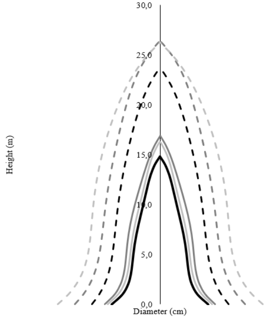
- Tree stem form modeling has special importance for the forest management;
- Requires multiples diameter measures within an individual tree:
    - Consequence is a correlation between observations.




(a) Colecting the data




(b) Measuring




(c) Modeling

# INTRODUCTION

- BONAT & JØRGENSEN (2016) developed the covariance generalized linear models (CGLM):
  - Quite flexible for modeling univariate and multivariate correlated data;
  - Considering response of mixed types;
  - And allow to define many covariance structures.
- CGLM is based on a marginal model specification and second-moment assumptions;
- Covariance are introduced by using a linear combination of known matrices.
- Thus, this approach presents a great potential in forest modeling.

# OBJECTIVES

- Hypothesis:

Covariance generalized linear models it will be suitable for modeling the behavior of *Pinus taeda* tree stem taper.

- Main goal:

To introduce the covariance generalized linear model framework in the context of forest biometrics.

# MATERIAL AND METHODS

- Taper data set was obtained by measuring 427 samples trees;
- Repeated measures of response were taken at: 0%, 0.5%, 1%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% of total height;
- Dataset was split in four age classes:
  - C1: 4 to 7 years old (without thinning);
  - C2: 8 to 11 years old (1 thinning);
  - C3: 12 to 19 years old (2 thinning);
  - C4: 20 to 30 years old (3 thinning).

# MATERIAL AND METHODS

Marginal specification of the covariance generalized linear model is given as

$$E[\boldsymbol{Y}] = \boldsymbol{\mu} = g^{-1}(\boldsymbol{X}\beta)$$

$$Var[\boldsymbol{Y}] = \boldsymbol{\Sigma} = \Omega(\boldsymbol{\tau}) = \tau_0 Z_0 + \ldots + \tau_D Z_D,$$

$\boldsymbol{Y}$ is an $N \times 1$ response vector;

$\boldsymbol{X}$ is an $N \times k$ design matrix;

$\beta$ is an $k \times 1$ regression parameters vector;

$g$ is the link function;

$Z_D$ with $d = 0, \ldots, D$ are known matrices reflecting the covariance structure;

$\boldsymbol{\tau} = (\tau_0, \ldots, \tau_D)$ is a $(D + 1) \times 1$ dispersion parameters vector.

# MATERIAL AND METHODS

Mean structure is given as

$$E[\boldsymbol{Y}] = g^{-1}(\boldsymbol{X}\beta) = \beta_1(\boldsymbol{X} - 1) + \beta_2(\boldsymbol{X}^2 - 1) + \beta_3(\alpha_1 - \boldsymbol{X})^2 I_1 + \beta_4(\alpha_2 - \boldsymbol{X})^2 I_2,$$

$\boldsymbol{Y}$ is a response vector of relative diameter;

$\boldsymbol{X}$ is a predictor variable vector of relative height;

$\alpha_s$ are the inflexion points to be estimated $(s = 1, 2)$;

$\beta_t$ are the parameters to be estimated $(t = 1,\ 2,\ 3,\ 4)$;

$I_q = 1$ if $\boldsymbol{X} \leq \alpha_q$ and 0 otherwise, which are a dummy indicator variables vector;

$g(\cdot)$ is an identity link function.

# MATERIAL AND METHODS

Specification of a matrix linear predictor

- *Strategy 1 – VarStr* : components were specified without incorporating the repeated measures structure.
  - Variance structure was directly modeled based on $\boldsymbol{I}$; $\boldsymbol{H_r}$; $\boldsymbol{A}$; $\boldsymbol{H_r^2}$; $\boldsymbol{A^2}$; $\boldsymbol{H_r}:\boldsymbol{A}$; and $\boldsymbol{H_r^2}:\boldsymbol{A}^2$.

$$\boldsymbol{I} = \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix}; \quad \boldsymbol{H_r} = \begin{bmatrix} H_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & H_{ij} \end{bmatrix}; \quad \boldsymbol{A} = \begin{bmatrix} A_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & A_i \end{bmatrix}; \dots; \quad \boldsymbol{A^2} = \begin{bmatrix} A_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & A_i^2 \end{bmatrix}.$$

# MATERIAL AND METHODS

Specification of a matrix linear predictor

- *Strategy 2 – CovStr* : components were built considering the correlation structure among response variable.
  - Covariance structure was directly modeled based on $\boldsymbol{I}$; $\boldsymbol{MA}(p)$; $\boldsymbol{ED}_o$; and $\boldsymbol{ED}_h$.

$$\boldsymbol{I} = \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix}; \quad \boldsymbol{MA}(1) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}; \dots; \quad \boldsymbol{ED}_h = \begin{bmatrix} 0 & \dots & dh_{1j}^{-1} \\ \vdots & \ddots & \vdots \\ dh_{i1}^{-1} & \dots & 0 \end{bmatrix}.$$

# MATERIAL AND METHODS

Specification of a matrix linear predictor

- *Strategy 3 – RwStr* : we proposed to model the matrix linear predictor as a random walk model. The model is specified by the inverse of the dispersion matrix as

$$\mathbf{\Omega}\left(\delta, \rho\right)^{-1} = \delta\left(\mathbf{D} - \rho\mathbf{W}\right),$$

where $\boldsymbol{W}$ is a neighborhood matrix; $\boldsymbol{D}$ is a diagonal matrix with the number of neighborhoods in the main diagonal; $\delta$ is a precision parameter; and $\rho$ is a spatial autocorrelation parameter.

$$\mathbf{\Omega}(\boldsymbol{\tau})^{-1} = \tau_0\boldsymbol{D} + \tau_1\boldsymbol{W},$$

where $\tau_0 = \delta$; and $\tau_1 = -\delta\rho$.

# MATERIAL AND METHODS

Components of the matrix linear predictor for the first three observations taken within-tree were given as

$$\boldsymbol{W} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

# MATERIAL AND METHODS

Specification of a matrix linear predictor

- *Strategy 4 – MmStr* : we presented the marginal specification of a mixed-effect model for taking into account the repeated measures effects within-tree for the covariates

$$Z_1 = (\mathbf{X} - 1) \quad \text{and} \quad Z_2 = \left(\mathbf{X}^2 - 1\right).$$

$\boldsymbol{X}$ is a predictor variable vector;

# MATERIAL AND METHODS

- Our focus was to quantify the uncertainty associated to the response;
- Confidence intervals for the response were calculated by

$$CI(\boldsymbol{\mu}, \gamma) = \hat{\boldsymbol{\mu}} \pm Z_{\gamma/2}\sqrt{\hat{\boldsymbol{C}}},$$

where $\hat{\boldsymbol{\mu}}$ is an $N \times 1$ vector of expected value; $Z_{\gamma/2}$ is a quantile of normal distribution for $\gamma$ confidence level; and $\hat{\boldsymbol{C}}$ is a main diagonal of covariance matrix of the fitted model.

# MATERIAL AND METHODS

Model selection:

- Score Information Criterion (SIC) was proposed by Bonat et al. (2017) for selecting components of the matrix linear predictor:
  - Function used: *mc_sic_covariance* from the mcglm package (BONAT, 2018) of the *R statistical software* (R CORE TEAM, 2018).
- Goodness-of-fit statistics:
  - Log-likelihood (LogLik);
  - Akaike information criterion (AIC);
  - Bayesian information criterion (BIC);
  - Mean Squared Error (MSE).

# MATERIAL AND METHODS

Our main goal was increasing the prediction ability.

- Conditional predictions:

$$\widetilde{\boldsymbol{\mu}}_{i+1|i} = \hat{\boldsymbol{\mu}}_{i+1} + \hat{\boldsymbol{C}}_{(i+1),i}\hat{\boldsymbol{C}}_{i,i}^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i),$$

where $\widetilde{\boldsymbol{\mu}}$ is a vector of conditional predictions of response; $\hat{\boldsymbol{\mu}}$ is a vector of marginal prediction of response; $\boldsymbol{y}$ is a vector of observed response; and $\hat{\boldsymbol{C}}$ is a covariance matrix of the fitted model.

- Prediction analysis was based on mean squared error (MSE) and the bias (B) over the stem.

# RESULTS

- Linear predictor of *CovStr:*

Table 1: Parameter estimates, standard errors (SE), Z-statistics and root mean square error (MSE) for *CovStr*

| Parameter | Estimates | SE | Z-statistics | MSE |
|---|---|---|---|---|
| $\beta_1$ | 3.3519 | 1.3324 | 2.5157 | 0.00622 |
| $\beta_2$ | -2.7030 | 0.7041 | -3.8392 | |
| $\beta_3$ | 23.8337 | 0.8543 | 27.8976 | |
| $\beta_4$ | 2.2442 | 0.6941 | 3.2331 | |
| $\alpha_1$ | 0.0900 | 0.0017 | 52.0427 | |
| $\alpha_2$ | 0.8726 | 0.0260 | 33.5535 | |

# RESULTS

- Components of the matrix linear predictor of *CovStr:*
  - *Identity matrix;*
  - *Euclidean distance between pairs of observations;*
  - *Moving average model of order 1, 2 and 3.*

$$Var[\boldsymbol{Y}] = \hat{\tau}_0 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + \hat{\tau}_1 \begin{bmatrix} 0 & 1.0 & 0.5 & 0.3 \\ 1.0 & 0 & 1.0 & 0.5 \\ 0.5 & 1.0 & 0 & 1.0 \\ 0.3 & 0.5 & 1.0 & 0 \end{bmatrix} + \tag{1}$$

$$\hat{\tau}_2 \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} + \hat{\tau}_3 \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} + \hat{\tau}_4 \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}. \tag{2}$$

# RESULTS

- Matrix linear predictor:

Table 2: Parameter estimates, standard errors (SE), z-statistics (Z-value), likelihood (LogLik), akaike (AIC) and bayesian (BIC) information criterion for the matrix linear predictor for *CovStr*

| Parameter | Estimates | SE | Z-statistics | LogLik | AIC | BIC |
|---|---|---|---|---|---|---|
| $\tau_0$ | 0.00629 | 0.00027 | 23.3944 | | | |
| $\tau_1$ | 0.01273 | 0.00063 | 20.2913 | | | |
| $\tau_2$ | -0.00728 | 0.00041 | -17.7324 | 12100.52 | -24179.04 | -24103.92 |
| $\tau_3$ | -0.00164 | 0.00012 | -13.9066 | | | |
| $\tau_4$ | -0.00031 | 0.00004 | -8.7549 | | | |

# RESULTS

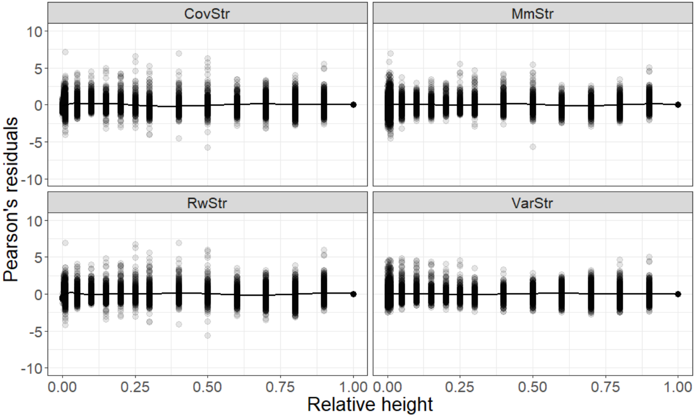



Figure 1: Pearson's residuals by relative height for different modeling strategies and fitted smooth curve in solid line

# RESULTS

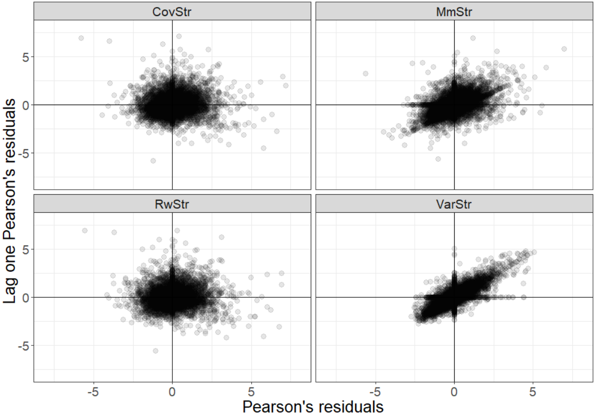



Figure 2: Correlation between lag one Pearson's residuals for response variable by different modeling strategies

# RESULTS

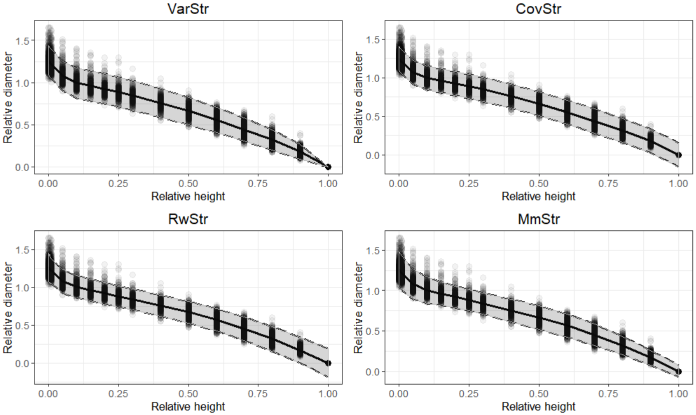



Figure 3: Uncertainty in the response. Observed values (full circles), fitted values (solid lines) and 95% confidence intervals (dashed lines) for response variable

# RESULTS

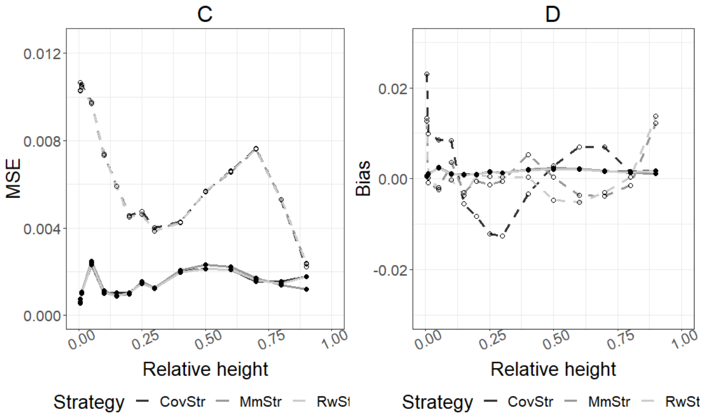



Figure 4: Mean squared error (C) and bias (D) for marginal (dashed lines) and conditional (solid lines) models

# CONCLUSION

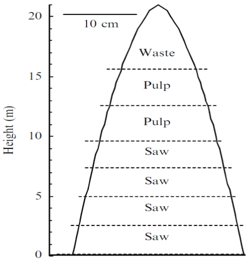
- The CGLM can be easily used for stem taper modeling;
- Advantage of our approach is obtaining a robust taper model.
- The *CovStr* approach was the best strategy.

# Sumário

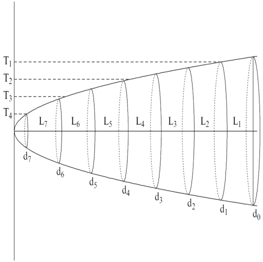

1. CHAPTERS
2. CHAPTER I
   - Covariance generalized linear models: an approach for quantifying uncertainty in tree stem taper modeling
3. CHAPTER II
   - Joint marginal modeling of height and volume for *Araucaria angustifolia*
4. CHAPTER III
   - Generalized linear models for tree survival in forest stands
5. REFERENCES AND ACKNOWLEDGMENT

# INTRODUCTION

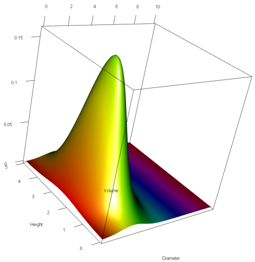
- Volume is an important information for evaluating the potential of a forest;
- Diameter and height are fundamental attributes at tree level;
- Usually, we fit models for describing variables such as:
    - Height and Volume.




(a) Height




(b) Volume




(c) Modeling

# INTRODUCTION

- BONAT & JØRGENSEN (2016) develop the so-called multivariate covariance generalized linear models (MCGLM);
- MCGLM allow to model response variables from distinct nature simultaneously;
- Besides, quantify the association between responses by using correlation parameters:
  - Also, we can include a covariance matrix;
  - And variance functions for different type of response variables.

# OBJECTIVES

- Hypothesis:

Correlation between response variables influence the fitting of the regression models.

- Main goal:

To analyze the fitting of univariate and multivariate regression models for describing the behavior of height and volume of *Araucaria angustifolia* in native forest.

# MATERIAL AND METHODS

- Data set was collected at Xanxerê municipality, Santa Catarina, Brazil;
- Trees were randomly selected on the forest fragment;
- Data set were composed by 169 independent sample trees and the variables measured were:
  - Diameter at breast height ($D$, cm);
  - Total height ($H$, m);
  - Individual volume with bark ($V$, $m^3$).

# MATERIAL AND METHODS

Generic formulation for the MCGLM is given as

$$E(\boldsymbol{Y}) = \boldsymbol{M} = \{g_1^{-1}(\boldsymbol{X}_1\beta_{\mathbf{1}}), ..., g_R^{-1}(\boldsymbol{X}_R\beta_R)\},$$

$$Var(\mathbf{Y}) = \boldsymbol{C} = \boldsymbol{\Sigma}_R \otimes \boldsymbol{\Sigma}_b,$$

The covariance matrix $\boldsymbol{\Sigma}_r$ for each response is given as

$$\boldsymbol{\Sigma}_r = V(\boldsymbol{\mu}_r; \boldsymbol{p}_r)^{1/2} h\{\Omega(\boldsymbol{\tau}_r)\} V(\boldsymbol{\mu}_r; \boldsymbol{p}_r)^{1/2}.$$

$\boldsymbol{\Sigma}_R$ is an $N \times N$ covariance matrix within response $r = 1, ...R$; $\boldsymbol{\Sigma}_b$ is a correlation matrix among response variables; $V(\boldsymbol{\mu}_r; \boldsymbol{p}_r)$ is a diagonal matrix, whose main entries denote the variance functions; $\boldsymbol{p}_r$ is a power parameter vector; $h\{\Omega(\tau_r)\} = \tau_0 Z_0 + ... + \tau_D Z_D$; and $h$ is a covariance link function.

# MATERIAL AND METHODS

Statistical analysis:

- Linear predictor:
    - Response variables: tree height ($H$) and tree volume ($V$);
    - Covariate: diameter at breast height ($D$).
- Matrix linear predictor:
    - Variance modeling was performed as a function of D;
    - Variance function was specified.

# MATERIAL AND METHODS

Statistical analysis:

- Linear predictor:
  - Response variables: tree height ($H$) and tree volume ($V$);
  - Covariate: diameter at breast height ($D$).
- Matrix linear predictor:
  - Variance modeling was performed as a function of D;
  - Variance function was specified.
- Modeling approach:
  - Univariate;
  - Multivariate.

# MATERIAL AND METHODS

Statistical analysis:

- Linear predictor:
    - Response variables: tree height ($H$) and tree volume ($V$);
    - Covariate: diameter at breast height ($D$).
- Matrix linear predictor:
    - Variance modeling was performed as a function of D;
    - Variance function was specified.
- Modeling approach:
    - Univariate;
    - Multivariate.
- Performance of the models:
    - Gaussian pseudo likelihood (PL);
    - Pseudo Bayesian's information criterion (PBIC).

# MATERIAL AND METHODS

Statistical analysis:

- Univariate and multivariate models were fitted on the R (R CORE TEAM, 2019);
- We use the *mcglm* package, version 0.5.0 (BONAT, 2018):
  - Package has an intuitive interface;
  - Many functions are available for building the components of the matrix linear predictor.

# RESULTS

- Exploratory data analysis

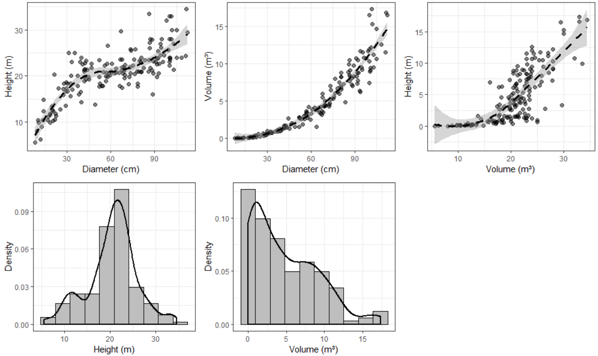



Figure 5: Histograms of response variables height (H) and volume (V); and scatter plot between response variables and covariate diameter (D)

# RESULTS

- Linear predictor:
  - $E(H_i) = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \beta_3 D_i^3$;
  - $E(V_i) = exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2)$.
- Matrix linear predictor:
  - $M1 = \boldsymbol{\Omega}(\boldsymbol{\tau}_r) = \tau_0 I$;
  - $M2 = \boldsymbol{\Omega}(\boldsymbol{\tau}_r) = \tau_0 I + \tau_1 Z$;
  - $M3 = \boldsymbol{\Omega}(\boldsymbol{\tau}_r) = \tau_0 I + \tau_1 Z + \tau_2 Z^2$;
  - $M4 = \boldsymbol{\Omega}(\boldsymbol{\tau}_r) = \tau_0 I + \tau_1 Z + \tau_2 Z^2 + \tau_3 Z^3$.
  - Variance function was included in all models.

where $I$ is an $N \times N$ identity matrix, being $N$ the number of observations; $Z$ is an $N \times N$ diagonal matrix whose main entries are constituted by tree diameters ($D$), where the effects varied until third degree.

# RESULTS

Table 3: Parameter estimates and standard errors of the univariate and multivariate models for $H$

| Parameter | Estimates | Standard error | Estimates | Standard error |
|---|---|---|---|---|
| | Univariate | | Multivariate | |
| $\beta_0$ | 0.3776 | 1.7087 | 5.7281 | 1.5568 |
| $\beta_1$ | 0.8937 | 0.1114 | 0.5201 | 0.0994 |
| $\beta_2$ | -0.0131 | 0.0021 | -0.0061 | 0.0018 |
| $\beta_3$ | 0.00007 | 0.00001 | 0.00003 | 0.00001 |
| $\tau_0$ | 8.7814 | 1.1021 | 9.3747 | 1.1124 |

# RESULTS

Table 4: Parameter estimates and standard errors of the univariate and multivariate models for $V$

| Parameter | Estimates | Standard error | Estimates | Standard error |
|---|---|---|---|---|
| | Univariate | | Multivariate | |
| $\beta_0$ | -3.1707 | 0.1108 | -3.1528 | 0.1122 |
| $\beta_1$ | 0.0988 | 0.0036 | 0.0982 | 0.0036 |
| $\beta_2$ | -0.0004 | 0.00003 | -0.0004 | 0.00003 |
| $\tau_0$ | 0.0830 | 0.0102 | 0.0843 | 0.0095 |
| $p$ | 1.6350 | 0.0829 | 1.6152 | 0.0781 |

# RESULTS

Table 5: Estimated correlation ($\rho$) between response variables height ($H$) and volume ($V$) for the multivariate fitting

| Model | Estimates | Standard error |
|---|---|---|
| M1 | 0.5075 | 0.0595 |
| M2 | 0.4945 | 0.0604 |
| M3 | 0.4715 | 0.0619 |

# RESULTS

Table 6: Pseudo likelihood (PV) and pseudo bayesian's information criterion (PBIC) from univariate and multivariate models

| Model | PV | PBIC |
|---|---|---|
| Model for $H$ and $V$: univariate case | | |
| M1 | -570.57 | 1199.31 |
| M2 | -567.80 | 1205.41 |
| M3 | -564.70 | 1210.84 |
| Model for $H$ and $V$: multivariate case | | |
| M1 | -550.98 | 1165.95 |
| M2 | -549.95 | 1175.52 |
| M3 | -548.09 | 1183.44 |

# RESULTS

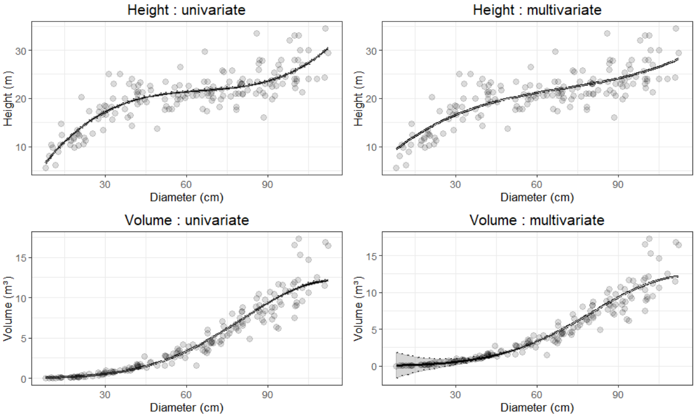



Figure 6: 95% Confidence intervals from univariate and multivariate models for response variables

# CONCLUSION

- Univariate and multivariate regression models were suitable;
- Correlation between response variables can influence the parameter estimates and standard errors;
- Variance function has potential to improve the performance of the models;
- Multivariate covariance generalized linear models have great potential to be applied to forest biometrics.

# Sumário


1. CHAPTERS
2. CHAPTER I
   - Covariance generalized linear models: an approach for quantifying uncertainty in tree stem taper modeling
3. CHAPTER II
   - Joint marginal modeling of height and volume for *Araucaria angustifolia*
4. CHAPTER III
   - Generalized linear models for tree survival in forest stands
5. REFERENCES AND ACKNOWLEDGMENT

# INTRODUCTION

- Tree survival is a phenomenon associated to many factors:
    - Competition; forest management practices; climatic conditions.
- Statistical tools able to predict the probability of a tree survive are essential;
- Logistic regression is widely used for estimating tree survival:
    - Require a linear predictor and a link function.
- Commom approach for selecting covariates for composing the linear predictor:
    - Forward, backward or stepwise.

# INTRODUCTION

- Alternative approaches are regularization methods:
  - Lasso (Least Absolute Shrinkage and Selection Operator);
  - Ridge regression;
  - Elastic Net.
- Main idea: to fit a regression model which the parameter estimates are penalized or shrunken toward to zero.

# OBJECTIVES

- Hypothesis:

Regularization methods are appropriated for selecting correlated covariates, once this approach can reduce the variance of the parameters.

- Main goal:

To estimate the probability of *Pinus taeda* survival.

# MATERIAL AND METHODS

- Data set: forest inventory performed in two occasions (2009 and 2015);
- 13 variables were mesured at trees and sub-samples level:
  - Complete observations: 40,556 trees.
- Response variable: survival, which is a binary variable (1 if the tree is alive or 0 otherwise).
- Covariates: age; gsample; nsample; daverage; dcv; dg; dmax; ddom; hdom; thinsample; gthin; nthin.

# MATERIAL AND METHODS

The generelized linear model is given as

$$Y_i | x_i \sim Bernoulli\left(\pi_i\right)$$
$$g\left(\pi_i\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},$$

where $Y_i$ is the response variable; $x_{i1}, \ldots, x_{ip}$ are the predictor variables $X_i$; $\pi_i$ is the survival probability; $g$ is a link function; $\eta_i$ is the linear predictor; and $\beta_0$, $\beta_1$, ..., $\beta_p$ are parameters to be estimated.

- Covariates for composing the linear predictor were selected by:
  - Stepwise and Regularization.
- Four Link function were tested:
  - Cauchit, complement log-log, logit and probit.

# MATERIAL AND METHODS

- Stepwise was based on the minimization of the Bayesian´s Information Criterion (BIC):

$$BIC = -2\hat{l} + ln\,(n)\,p,$$

where $\hat{l}$ is the maximized log-likelihood value; $n$ is the number of observations; and $p$ is the number of parameters.

# MATERIAL AND METHODS

- Regularization was based on penalizations controlled by the parameters $\lambda$ and $\alpha$ that quantify the penalization intensity:

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

# MATERIAL AND METHODS

- Regularization was based on penalizations controlled by the parameters $\lambda$ and $\alpha$ that quantify the penalization intensity:

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

Lasso: first order penalization ($\alpha = 1$),

Ridge Regression: second order penalization ($\alpha = 0$).

Elastic Net: intermediate penalization ($0 < \alpha < 1$).

# MATERIAL AND METHODS

- Regularization was based on penalizations controlled by the parameters $\lambda$ and $\alpha$ that quantify the penalization intensity:

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

Lasso: first order penalization ($\alpha = 1$),

Ridge Regression: second order penalization ($\alpha = 0$).

Elastic Net: intermediate penalization ($0 < \alpha < 1$).

- Optimum $\lambda$: cross-validation based on *cv.glmnet* function of the glmnet package (FRIEDMAN et al., 2010);
- $\lambda = 0 \rightarrow \alpha$ not identifiable.

# MATERIAL AND METHODS

Influence of the link function on the covariate selection were tested:

- Cauchit:

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

- Complement log-log:

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

- Logit:

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

- Probit:

$$\phi^{-1}\left(\pi_i\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

# MATERIAL AND METHODS

Performance of the models:

- Half-Normal plots (HNP) by *hnp* function of the hnp package (MORAL et al., 2017);
- Randomized quantile residuals (RQR) (DUNN & SMYTH, 1996)

# MATERIAL AND METHODS

Performance of the models:

- Half-Normal plots (HNP) by *hnp* function of the hnp package (MORAL et al., 2017);
- Randomized quantile residuals (RQR) (DUNN & SMYTH, 1996)

Predictive performance: Data were split in fitting data (90%) and validation (10%);

- ROC (Receiver Operating Characteristic) curve of the ROCR package (SING et al., 2005);
- Sensibility (Sens) and specificity (Esp):
    - Estimated for 0.75; 0.85; 0.90; 0.95 and 0.99 probability values.

# RESULTS

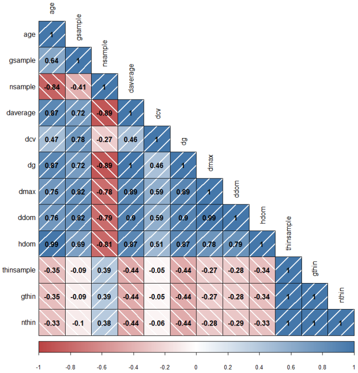



Figure 7: Correlogram between variables clustered by centroid method

# RESULTS

Covariates selected for composing the linear predictor:

- Stepwise: gsample, nsample, dcv, dg and dmax.
- Regularization: all covariates were selected:
    - Best $\lambda$ value was close to 0 for all sequences of $0 \leq \alpha \leq 1$,
    - regardless of loss measure we tested.
- We decided to continue the data analysis considering the natural scale of the covariates.

# RESULTS

Table 7: Bayesian information criterion (BIC) and residual deviance (RD) by link functions and covariate selection methods

| Link function | BIC (number of covariates) | | Residual deviance | |
|---|---|---|---|---|
| | Stepwise | Regularization | Stepwise | Regularization |
| Cauchit | 7068.31 (9) | 7094.98 (12) | 6963.30 | 6958.20 |
| C. log-log | 6847.46 (5) | 6904.20 (12) | 6784.40 | 6768.30 |
| Logit | 6874.73 (7) | 6910.75 (12) | 6790.70 | 6774.20 |
| Probit | 6851.50 (5) | 6906.84 (12) | 6788.50 | 6769.30 |

# RESULTS

Table 8: Parameter estimates, standard errors (SE) and p-value for the fitted models on the linear predictor scale, with complement log-log link function

| 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 | $p \leq 0.05$ | - | - | - |

# RESULTS

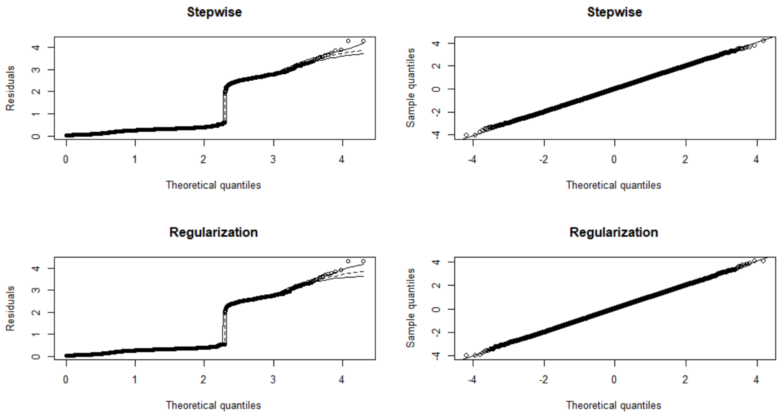



Figure 8: Half-normal plot (left) and randomly quantile residuals (right) for diagnosing the fitted models

# RESULTS

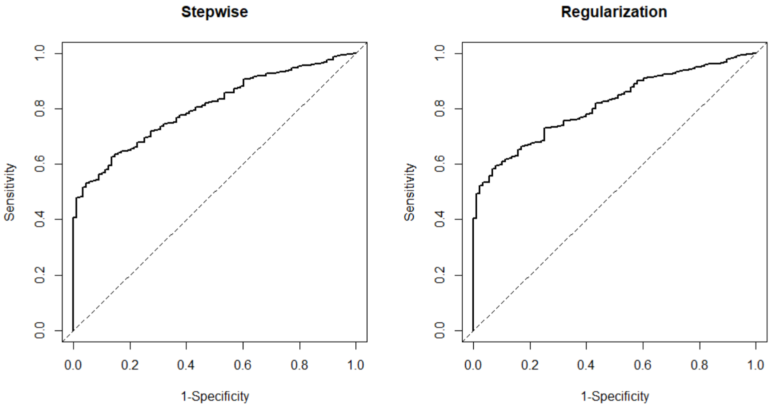



Figure 9: ROC curve of the models applied to the validation data set

# RESULTS

Table 9: Sensitivity and specificity by selected models applied to the validation data set for 0.99 probability cut point

| Model | Sensitivity | Specificity |
|---|---|---|
| Stepwise | 0.989 | 0.460 |
| Regularization | 0.989 | 0.466 |

# CONCLUSION

- Stepwise procedure for selecting covariates was more parsimoniously;
- Complementary log-log link function was the most suitable;
- Model presented a great prediction ability, mainly due to the high number of survival trees.

# Sumário


1. CHAPTERS
2. CHAPTER I
   - Covariance generalized linear models: an approach for quantifying uncertainty in tree stem taper modeling
3. CHAPTER II
   - Joint marginal modeling of height and volume for *Araucaria angustifolia*
4. CHAPTER III
   - Generalized linear models for tree survival in forest stands
5. REFERENCES AND ACKNOWLEDGMENT

# REFERENCES


- BONAT, W.H. Multiple response variables regression models in R: The mcglm Package. Journal of Statistical Software, v. 84, n. 4, 2018.
- BONAT, W.H.; JØRGENSEN, B. Multivariate covariance generalized linear models. Journal of the Royal Statistical Society. Series C: Applied Statistics, v. 65, n. 5, p. 649–675, 2016.
- DUNN, P.K.; SMYTH G.K. Randomized quantile residuals. Journal of Computational and Graphical Statistics, v. 5, n. 3, p. 236-244, 1996.
- 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.
- R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2019.

# ACKNOWLEDGMENT

Thank you all for everything!!!

(a) UFPR/PPGEF

(b) CNPq

(c) DEST/UFPR

(d) LEG