Age-varying bivariate distribution models for growth prediction

Edmundas Petrauskas, Petras Rupšys

Abstract—Height-diameter models are classically analyzed by fixed or mixed linear and non-linear regression models. In order to possess the among-plot variability, we propose stochastic differential equations that are deduced from the standard deterministic dynamic ordinary differential equations by adding the process variability to the growth dynamic. The advantage of the stochastic differential equation framework is that it analyzes a residual variability, corresponding to measurements error, and an individual variability to represent heterogeneity between subjects. An analysis of 1575 Scots pine (Pinus sylvestris) trees provided the data for this study. The results are implemented in the symbolic computational language MAPLE.

Keywords—Age-varying bivariate density, diameter, height, normal bivariate copula, stochastic differential equation.

I. INTRODUCTION

Diameter at breast height outside bark and total height are the variables most often measured in forest inventories. In most sample plot systems diameter at breast height is conventionally measured for all trees sampled, but height is measured for only a sub-sample. Height-diameter relationship is used to estimate the heights of trees measured only for diameter at breast height [1], and to determine the dominant height for measuring the site productivity [2]. Complex tree characteristic such as height, volume always appear to be multivariate event that are characterized by a few correlated variables. A complete understanding of these events needs to investigate joint probabilistic behaviour of these correlated variables such as diameter at breast height, total height, density, etc. When it comes to modelling dependent random variables, in practice as a first resort is used the multivariate normal (or lognormal distribution), mainly because it is easy to parameterize and to deal with. In recent years, particular in the quantitative analysis of tree variables, we have also witnessed a trend to tree variable marginal distributions and dependence structure, using the copula concept. The normal (or Gaussian) copula has even come to the attention of the foresters due to its use in the valuation of merchantable forest products [3], [4]. A lognormal distribution is one of frequently selected candidates for characteristics frequency analysis. In this paper, the multivariate lognormal distribution will serve as an important tool for analyzing a multivariate tree episode.

Although many refinements and extensions are possible, the basic dynamic model for diameter \( D(t), \ t \geq t_0 \) and height \( H(t), \ t \geq t_0 \) processes can be described by univariate or bivariate parametric stochastic differential equation [5]

\[
dX(t) = \mu(X(t), \Theta)dt + \sigma(X(t), \Theta)dW(t)
\]

where \( W(t), \ t \geq t_0 \) is a standard Brownian motion. Parametric approach assume that the drift \( \mu(X(t), \Theta) \) and diffusion \( \sigma(X(t), \Theta) \) are known functions except for an unknown parameter vector \( \Theta \). Examples include [6]-[16].

Following the recent trend in stochastic continuous-time models, we develop a multivariate stochastic height model using the Gompertz shape tree growth. Multivariate models can deal, for instance, with multiple explanatory factors (diameter, stand density, basal area) in asset tree height. Their use permits the analysis of the movement and correlation of these tree state variables over-age. In many cases, however, the margins are relatively easy to describe, while an explicit expression of the joint distribution may be difficult to obtain. In this paper making an attempt on broaden the class of non-linear multivariate diffusion processes that could lead to closed form likelihood functions was applied so-called copula function method [17]-[18]. These functions provide an interesting tool to model a multivariate distribution when only marginal distributions are known. Such an approach is, thus, particularly useful in situations where multivariate normality does not hold. By copula approach we first construct univariate non-linear stochastic processes of tree diameter and height leading to closed form transition probability density functions based on reducible stochastic differential equations. The derived univariate marginal transition probability density functions are then welded together by an appropriate copula function to obtain a joint transition probability density function of diameter and height which accounts for the correlation across diameter and height. Evidently, the stochastic Gompertz bivariate model can be used to estimate height like copula approach too.

This article is an attempt to design a model for predicting the height of a tree from its variables (diameter, height) age-varying bivariate distribution. For modelling the joint and marginal diameter and height processes, we use a class of the Gompertz shape stochastic differential equations that are reducible to an Ornstein-Uhlenbeck process.
II. MATERIAL AND METHODS

A. Height Models

In this paper we will focus on dynamics of tree diameter and height distribution subject to the age changes. We assume that the dynamics of tree diameter and height can be expressed in terms of the bivariate or univariate Gompertz shape stochastic differential equation with multiplicative noise (a linear volatility function). Two types of models (bivariate and univariate) were used in modelling the changes of a tree. Let us consider a bivariate or univariate diameter and height stochastic growth process facing stochastic fluctuations in the following Ito’s [5] system of stochastic differential equations

\[ dX(t) = A(X(t))dt + Q(X(t))dW(t), \]
\[ dD(t) = \alpha_1 D(t) - \beta_1 D(t) \ln(D(t))dt + \sigma_1 dW_1(t), \]
\[ dH(t) = \alpha_2 H(t) - \beta_2 H(t) \ln(H(t))dt + \sigma_2 dW_2(t), \]

where \( X(t) = (D(t), H(t))^T \), \( T \) denotes transposition, \( t \in [t_0, T] \), \( t_0 \geq 0 \), \( D(t) \) is a breast height diameter (in the sequel - diameter) at the age \( t \), \( H(t) \) is a height at the age \( t \), \( X(t_0) = (d_0, h_0)^T \) is a fixed vector \( d_0 \geq 0 \), \( h_0 \geq 0 \), and \( \{ W(t); t \in [t_0, T] \} \) is a bivariate standard Brownian motion.

The models parameters \( \alpha_i, \beta_i, \gamma_i, \sigma_i, \sigma_{ij}, 1 \leq i, j \leq 2 \), \( \rho \) \((|\rho| \leq 1)\) are unknown real numbers to be estimated. The transition probability densities for the proposed models (2)-(4) can be easily obtained by the standard transformation method of the distribution. The transition probability density function of the bivariate stochastic process \((D(t), H(t))\) defined by Eq. (2) has a bivariate lognormal distribution \( \Lambda_2(\mu(t), \Sigma(t)) \) with the mean vector

\[ \mu_1(t) = \ln d_0 e^{-\beta_i(t-t_0)} + \frac{1 - e^{-\beta_i(t-t_0)}}{\beta} \left( \alpha_1 - \frac{\sigma_{11}}{2} \right), \]
\[ \mu_2(t) = \ln h_0 e^{-\beta_i(t-t_0)} + \frac{1 - e^{-\beta_i(t-t_0)}}{\beta} \left( \alpha_2 - \frac{\sigma_{22}}{2} \right), \]

and the variance-covariance matrix

\[ \Sigma(t) = \frac{1 - e^{-2\beta_i(t-t_0)}}{2\beta} B. \]

The conditional mean height given diameter \( D = d \) can be derived as follows

\[ h = h(d, t) = E(H(t)|D(t) = d) = E(H(t)) + \rho(t) \sqrt{\text{Var}(H(t))} \frac{d - E(D(t))}{\text{Var}(D(t))}, \]
\[ \rho(t) = \frac{\text{Var}(H(t))}{\text{Var}(D(t))}, \]

where

\[ E(D(t)) = \exp \left( \mu_1(t) + \frac{\sigma_{11}}{4\beta} \left( 1 - e^{-2\beta(t-t_0)} \right) \right), \]
\[ V(D(t)) = \exp \left( 2\mu_1(t) + \frac{\sigma_{11}}{2\beta} \left( 1 - e^{-2\beta(t-t_0)} \right) \right) \times \left( \exp \frac{\sigma_{11}}{2\beta} \left[ 1 - e^{-2\beta(t-t_0)} \right] - 1 \right), \]
\[ V(H(t)) = \exp \left( 2\mu_2(t) + \frac{\sigma_{22}}{2\beta} \left( 1 - e^{-2\beta(t-t_0)} \right) \right) \times \left( \exp \frac{\sigma_{22}}{2\beta} \left[ 1 - e^{-2\beta(t-t_0)} \right] - 1 \right), \]
\[ \rho(t) = \left( \exp \frac{\sigma_{12}}{2\beta} \left[ 1 - e^{-2\beta(t-t_0)} \right] - 1 \right) \left( \exp \frac{\sigma_{11}}{2\beta} \left[ 1 - e^{-2\beta(t-t_0)} \right] - 1 \right)^{-\frac{1}{2}} \]
\[ \mu_D(t) = \ln d_0 e^{-\beta_i(t-t_0)} + \frac{1 - e^{-\beta_i(t-t_0)}}{\beta_1} \left( \alpha_1 - \frac{\sigma_{11}}{2} \right), \]
\[ \mu_H(t) = \ln h_0 e^{-\beta_i(t-t_0)} + \frac{1 - e^{-\beta_i(t-t_0)}}{\beta_2} \left( \alpha_2 - \frac{\sigma_{22}}{2} \right), \]
\[ \sigma_D(t) = \frac{1 - e^{-2\beta_i(t-t_0)}}{2\beta_1} \sigma_{11}. \]

Next we propose a copula-based approach for the modelling of the bivariate diameter and height distribution. Flexible non-linear multivariate models can be constructed by joining univariate non-linear processes via appropriate copulas [19]. A copula-based approach accounts simultaneously for observed non-linearity and correlation across diameter and height. Welding the derived marginal transition probability densities of diameter and height via copula function \( C(u, v) \) we obtain an age-varying bivariate distribution function. Focusing on the univariate case, we denote \( F_{D_i}(d, t) \) and \( F_{H_i}(h, t) \) as the continuous marginal distribution functions of diameter and height, and \( F_c(d, h, t) \) the copula distribution function. Also let \( f_{D_i}(d, t) \) and \( f_{H_i}(h, t) \) be the marginal density functions, respectively, and \( f_c(d, h, t) \) the copula probability density function. For notational convenience, set \( u = F_{D_i}(d, t) \) and \( v = F_{H_i}(h, t) \). The Sklar’s Theorem states [20]

\[ C(u, v, \rho) = P(U \leq u, V \leq v), \]
\[ c(u, v; \rho) = \frac{\partial^2 C(u, v; \rho)}{\partial u \partial v}, \]

where \( U \) and \( V \) are uniform random variables on \([0,1]\), and \( c \) is the corresponding bivariate copula density function.

Many copula functions is derived, however, we focus on the normal copula. The density and distribution function of the bivariate standard normal distribution are defined by
\[ \phi_2(x, y; \rho) = \frac{1}{2\pi \sqrt{1 - \rho^2}} \exp \left( -\frac{x^2 - 2\rho xy + y^2}{2(1 - \rho^2)} \right), \]

\[ \Phi_2(x, y; \rho) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_2(u, v; \rho) \, du \, dv \]

with correlation parameter \( \rho \in (0,1) \). The bivariate normal (or Gaussian) copula with parameter \( \rho \) is then defined by application of Sklar’s Theorem [20]

\[ C(u, v; \rho) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \rho), \]

\[ c(u, v; \rho) = \frac{1}{\sqrt{1 - \rho^2}} \exp \left( -\frac{\rho(u^2 + v^2 - 2uzv)}{2(1 - \rho^2)} \right), \]

where \( z_x = \Phi^{-1}(u) \) and \( z_y = \Phi^{-1}(v) \). The bivariate normal copula log-likelihood function is given as

\[ \ln(c) = -0.5 \ln(1 - \rho^2) - \frac{\rho(u^2 + v^2 - 2uzv)}{2(1 - \rho^2)}. \]  (6)

Under the standard multivariate normal theory [20], mean and variance of transformed variable \( Z_H \) are defined by

\[ \mu_{Z_H} = E(Z_H | Z_D = z_D) = \rho x_D, \]

\[ \sigma_{Z_H}^2 = Var(Z_H | Z_D = z_D) = 1 - \rho^2, \]

where \( Z_D = \Phi^{-1}(F_D(D, t)) \) and \( Z_H = \Phi^{-1}(F_H(H, t)) \) are transformations for diameter, \( D \), and height, \( H \), respectively. Taking a backward transformation \( F^{-1}_H(\Phi^{-1}(\mu_{Z_H}), t) \), we get the median regression of height, given diameter and age, in the following form

\[ h = h_d(t) = F^{-1}_D(\Phi^{-1}(F_D(D, t)), t). \]  (7)

### B. Parameter Estimation

In this paper, we apply the theory of multi-stage maximum likelihood estimator for the normal copula bivariate density and a one-stage maximum likelihood estimator for the joint bivariate density. For each of the diameter and height data sets, the Gompertz shape stochastic models (2)-(4) are estimated. All three models have closed form likelihood functions. We estimate parameters of all three models using maximum likelihood estimator [14]. For the normal copula bivariate density, the correlation/dependency parameter, \( \rho \), can be estimated by maximum likelihood estimator (log-likelihood Eq. (6)). In this paper, the margins are estimated as the kernel distributions \( \hat{F}_D(d) \) and \( \hat{F}_H(h) \) which represent the smooth estimated stationary margins of diameter and height, respectively. The stationary kernel distribution functions \( \hat{F}_D(d) \) and \( \hat{F}_H(h) \) are defined by

\[ \hat{F}_D(d) = \frac{1}{m} \sum_{i=1}^{m} W \left( \frac{d - d_i}{h_d} \right), \quad \hat{F}_H(h) = \frac{1}{m} \sum_{i=1}^{m} W \left( \frac{h - h_i}{h_h} \right) \]

where: \( W(x) = \frac{1}{K(t)} dt, \) \( h_d \) and \( h_h \) denote bandwidths \( (h_d \to 0 \text{ and } h_h \to 0 \text{ as } n \to \infty \text{ and } m \to \infty), \) respectively, and \( K \) is a unimodal symmetric density function with support \([-1;1]\).

### III. Results and Discussion

#### A. Data

Measurements for 1575 Scots pine trees were used for height models analysis. All data were collected during 1979-2008 across the entire Lithuanian territory, except for Kuršių Nerija National Park. Diameter outside bark at breast height and total height of each tree in a plot was measured. Diameter was measured to an accuracy of 1 mm. Summary statistics for diameter outside bark at breast height (D), total height (H) and age (A) of all trees used for models comparison are presented in Table 1.

<table>
<thead>
<tr>
<th>Data</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>D(cm)</td>
<td>2.2</td>
<td>51.0</td>
<td>16.4</td>
<td>11.75</td>
</tr>
<tr>
<td>H(m)</td>
<td>2.3</td>
<td>33.5</td>
<td>13.6</td>
<td>8.74</td>
</tr>
<tr>
<td>A( yr)</td>
<td>12</td>
<td>102</td>
<td>37.9</td>
<td>23.72</td>
</tr>
</tbody>
</table>

#### B. Results and discussion

The bivariate and univariate Gompertz shape stochastic differential equations with multiplicative noise (2)-(4) enable us to write the maximum likelihood estimators in a closed form [14]. All the parameters \( \alpha_1, \alpha_2, \beta, \beta_1, \sigma_1, \sigma_{ij}, 1 \leq i, j \leq 2 \) of stochastic differential equations (2)-(4) were estimated simultaneously. The estimate of the dependency parameter, \( \rho \), of a normal copula was estimated via log-likelihood function (6). A MAPLE program was used to carry out calculations. In our paper, the pseudo-sample was constructed using standard methods of nonparametric statistics, for example, the kernel distributions \( \hat{F}_D(d) \) and \( \hat{F}_H(h) \) as margins were defined in the following form

\[ z_d = \Phi^{-1}(\hat{F}_D(d_i)), 1 \leq i \leq n, \quad z_h = \Phi^{-1}(\hat{F}_H(h_i)), 1 \leq i \leq n. \]

The Epanechnikov kernel was employed and the optimal bandwidth was approximated by Parzen [21]

\[ b_n = 0.9 \min(\hat{\sigma}, IQRn)^{\frac{1}{5}}, \]
where $\hat{\sigma}$ and $IQR$ are the sample standard deviation and sample inter-quartile range. The calculated optimal value of the bandwidth for diameter and height datasets was 2.42 and 1.80, respectively.

Gaussian copula belongs to the elliptical-copula family. As would be expected, the estimate for the dependency parameter ($\rho = 0.9891$) is very close to linear diameter-height correlation 0.9414. Fig. 1 shows the estimated bivariate lognormal transition and bivariate normal copula probability densities of tree diameter and height. These surfaces indicate that the density of tree diameter and height is steeper for normal copula bivariate distribution than the joint bivariate distribution.

Gaussian copula belongs to the elliptical-copula family. As would be expected, the estimate for the dependency parameter ($\rho = 0.9891$) is very close to linear diameter-height correlation 0.9414. Fig. 1 shows the estimated bivariate lognormal transition and bivariate normal copula probability densities of tree diameter and height. These surfaces indicate that the density of tree diameter and height is steeper for normal copula bivariate distribution than the joint bivariate distribution.

In the present study, the height-diameter models (Eqs. (5), (7)) were based on the Gompertz shape stochastic differential equation. Earlier modeling efforts focused on describing the mean parameter values of the height-diameter relationships. According Petrauskas and Rupšys [1], the q-exponential trend was consistently the best among the tested height-diameter regression models. For comparing the predictive ability of our presented models was additionally evaluated the q-exponential height-diameter regression model

$$H = \left[ \frac{\beta_1 - \beta_2}{\beta_2 (1 - \exp(1 - \beta_3 \beta_4 D))} \right]^{\frac{1}{1 - \beta_4}}$$

where $\beta_1 - \beta_4$ are parameters estimated from the data. The estimators were specified as: $\beta_1 = 0.9757$, $\beta_2 = 0.4835$, $\beta_3 = -0.2284$, $\beta_4 = 0.6696$ (q-exponential model). Results of the adjusted coefficient of determination showed that the all used height models (Eqs. (5), (7), (8)) account for at least 92.9% of the variation in height. The results of goodness of fit statistics for all models applied to the tree analysis data set are summarised in Table 2. The joint bivariate height model (5) gave better results than the other models used here.

<table>
<thead>
<tr>
<th>Model</th>
<th>B</th>
<th>RE%</th>
<th>AIC</th>
<th>$\bar{R}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint density</td>
<td>-0.033</td>
<td>14.82</td>
<td>13817</td>
<td>0.946</td>
</tr>
<tr>
<td>Copula density</td>
<td>0.045</td>
<td>14.91</td>
<td>13829</td>
<td>0.946</td>
</tr>
<tr>
<td>q-exp</td>
<td>-0.045</td>
<td>17.09</td>
<td>14252</td>
<td>0.929</td>
</tr>
</tbody>
</table>

Fig. 2 illustrates the designated models (Eqs. (5), (7)) of a tree height. When fixing the age and applying the joint bivariate height versus diameter to the observed data set ($\{d_i, h_i, t_i\}$, $1 \leq i \leq n$) for the joint bivariate model, predictions of height showed a more linear consistent pattern over diameter at breast height than the normal copula bivariate model. The height curves are quite similar until ages of 30-35 years.
Graphical diagnostics of residuals (see Fig 3) for the height predictions showed that the residuals of the joint bivariate height model (5) had more homogeneous variance than the other height models. The relationship between the observed and predicted height is shown in Fig. 4 together with non-parametric kernel regression curve. The solid line in Fig. 4 generated by the Nadaraya-Watson kernel regression [22] indicates the bias between the observed and predicted heights. Fig. 4 showed better performance between the observed and predicted height of the joint bivariate model (Eq. (5)) in comparison with other models (Eqs. (7), (8)).

IV. CONCLUSIONS

We examined two age-varying height-diameter models under the stochastic differential equations framework. These models are an alternative to classical nonlinear mixed models whose deterministic regression function is too limiting to simulate the stochastic component of growth variability. This methodology seems applicable to other data-bank growth modeling situations. It would also be useful to expand or modify the presented relationships with more growth predictive variables (for example, trivariate case, four-variate case).

REFERENCES


