SHORT COMMUNICATION

Comparing Johnson's S_{BB}, Weibull and Logit-Logistic bivariate distributions for modeling tree diameters and heights using copulas

José J. Gorgoso-Varela[1]; Juan D. García-Villabrille[2]; Alberto Rojo-Alboreca[2]; Klaus von Gadow[2]; Juan G. Álvarez-González[2]*

[1] Departamento de Biología de Organismos y Sistemas. Universidad de Oviedo. Escuela Politécnica de Mieres. Mieres, Spain. [2] Unidad de Gestión Forestal Sostenible (UXFS). Departamento de Ingeniería Agroforestal. Escuela Politécnica Superior. Universidad de Santiago de Compostela. Lugo, Spain.

Abstract

Aim of study: In this study we compare the accuracy of three bivariate distributions: Johnson's S_{BB}, Weibull-2^P and LL-2^P functions for characterizing the joint distribution of tree diameters and heights.

Area of study: North-West of Spain.

Material and methods: Diameter and height measurements of 128 plots of pure and even-aged Tasmanian blue gum (*Eucalyptus globulus* Labill.) stands located in the North-west of Spain were considered in the present study. The S_{BB} bivariate distribution was obtained from S_B marginal distributions using a Normal Copula based on a four-parameter logistic transformation. The Plackett Copula was used to obtain the bivariate models from the Weibull and Logit-logistic univariate marginal distributions. The negative logarithm of the maximum likelihood function was used to compare the results and the Wilcoxon signed-rank test was used to compare the related samples of these logarithms calculated for each sample plot and each distribution.

Main results: The best results were obtained by using the Plackett copula and the best marginal distribution was the Logit-logistic.

Research highlights: The copulas used in this study have shown a good performance for modeling the joint distribution of tree diameters and heights. They could be easily extended for modelling multivariate distributions involving other tree variables, such as tree volume or biomass.

Keywords: Plackett copula; normal copula; *Eucalyptus globulus*.

Introduction

Stand volume, one of the most important variables in forest management, is usually estimated based on sampled tree diameters and heights (Wang & Rennolls, 2007). The common practice is to obtain the height data from a subsample of trees for which diameters are available, and to fit an empirical height-diameter relationship to estimate the average height per diameter class. Tree volume is then estimated using an individual-tree volume equation. Although this approach may appear satisfactory, it is often not appropriate because one tends to ignore the fact that height may vary considerably for a given diameter due to genetic, environmental or silvi-cultural factors. Therefore, the height residuals are seldom homoscedastic and normally distributed and in many forests the variance about the diameter-height regression is heterogeneous (Zucchini *et al.*, 2001).

An alternative approach for improving stand volume estimation, which takes into account those variations, involves the use of a bivariate distribution (Zucchini *et al.*, 2001; Wang *et al.*, 2008; Mønness, 2015). The joint bivariate distribution of tree diameter and height provides a detailed impression of the relationship between the two variables, which is not given by the two marginal distributions (Rupsys & Petrauskas, 2010). Moreover, bivariate distributions of diameter and height are also useful for assessing timber value based on

price sizes (Schreuder & Hafley, 1977) and stand structural diversity (Staudhammer & LeMay, 2001).

Hence there has been considerable interest in identifying suitable bivariate distributions to describe diameter-height frequency data. For many years, the bivariate extension of the S_B distribution, the S_{BB} (Johnson, 1949), has been the only bivariate distribution used for modeling bivariate tree diameter-height frequency data (e.g. Hafley & Schreuder, 1977; Knoebel & Burkhart, 1991; Tewari & Gadow, 1999; Castedo Dorado *et al*., 2001; Zucchini *et al*., 2001). Johnson's S_{BB} is developed by applying a four-parameter logistic transformation to each of the component variables of a standard bivariate normal distribution (Johnson, 1949; Rennolls & Wang, 2005). The construction of any other analytic bivariate distribution without resorting to a transformation of a bivariate normal distribution is complicated (Wang & Rennolls, 2007). However, the use of a copula function has provided a general way of constructing multivariate distributions. During recent years, several authors made use of the approach described by Sklar (1973) joining a multivariate distribution based on their one dimensional marginal distributions (e.g. Li *et al*., 2002; Wang & Rennolls, 2007; Wang *et al*., 2008).

The objective of the present study is to fit and compare the accuracy of three bivariate distributions: Johnson's S_{BB}, Weibull and Logit-Logistic fitted to diameter-height data from pure and even-aged stands of *Eucalyptus globulus* in Northwestern Spain. The Weibull and the Logit-Logistic (LL) bivariate distributions, denoted as Weibull-2^P and LL-2^P, were obtained from Weibull and LL marginal distributions by using the Plackett copula whereas the S_{BB} bivariate distribution was obtained from S_B marginal distributions using the Normal copula, i.e., a four-parameter logistic transformation to each of the component variables.

Material and methods

Data

All tree diameters and heights were measured in 128 field plots in Tasmanian blue gum (*Eucalyptus globulus* Labill.) stands in Galicia. The plots had been re-measured 1, 2 or 3 times resulting in a total of 308 inventories. The plots were established in pure and even-aged stands covering a wide variety of combinations of age, number of trees per hectare, site quality and method of regeneration. The sample plot size ranged from 375 to 900 m², depending on stand density. The objective was to assess a minimum of 30 trees per plot.

All trees in each plot were numbered; diameters at breast height were measured with a caliper, to the nearest 0.1 cm, and heights were measured with hypsometer to the nearest 0.1 m. The stand variables calculated in each inventory included the quadratic mean diameter, the number of trees per hectare, dominant height, basal area and mean height. A total of 17,588 trees were measured. The summary statistics of the main stand variables are presented in Table 1.

Copula functions

Wang *et al*. (2008) presented an exhaustive review of different one-parameter copulas which are useful for modeling bivariate tree diameter and height distributions. A copula is a function that joins a multivariate distribution function based on its one-dimensional marginal distributions. Suppose X and Y are two continuous random variables and $F(x) = \Pr(X \le x)$ and $G(y) = \Pr(Y \le y)$ are their marginal cumulative distri-

Table 1. Summary of the main descriptive statistics of stand variables.

		Variable	Mean	Maximum	Minimum	Standard deviation
Eucalyptus globulus (n=308)	Stand variables	dg	13.4	34.8	1.5	4.3
		N	1,174.4	2,386.8	435,7	339.1
		H_0	19.1	40.1	3.1	6.3
		G	16.5	63.5	0.2	8.3
		H_m	14.7	29.0	2.2	4.6
		Trees/plot	97.8	111	30	57.1
	Tree variables	D	12.2	69.8	0.1	6.1
		H	14.5	42.6	1.6	6.2
	Kurtosis	D	-0.1	6.7	-1.4	1.1
		H	0.7	14.8	-1.6	2.2
	Skewness	D	-0.1	1.4	-2.1	0.6
		H	-0.7	0.8	-3.3	0.7

dg: quadratic mean diameter (cm); N: number of trees·ha^{-1}; H_0: dominant height (m); G: basal area (m²·ha^{-1}), H_m: mean height (m), D: tree diameter (cm), H: tree height (m).

bution functions, respectively. The copula function C combines these two marginal to give the joint distribution function $H(x,y)$ as $H(x,y) = C\big(F(x),G(y)\big)$. If both marginal distribution functions and the copula are differentiable, the joint density function can be expressed as:

$$h(x,y) = f(x) \cdot g(y) \cdot c\big(F(x),G(y)\big) \qquad (1)$$

where $f(x)$ and $g(y)$ are the marginal density functions, and $c(F(x),G(y))$ is the used copula density.

Frequently used copulas are the Normal (Mardia, 1970) and the Plackett copula (Plackett, 1965). Their densities are given by (Wang *et al*., 2008):

$$\text{Normal copula } c(F(x),G(y)) = \frac{1}{\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{z_x^2 - 2\rho \cdot z_x z_y + z_y^2}{1-\rho^2}} \qquad (2)$$

Plackett copula

$$c(F(x),G(y)) = \frac{\omega\Big\{1+(\omega-1)\big[F(x)+G(y)-2F(x)\cdot G(y)\big]\Big\}}{\Big\{\big[1-(F(x)+G(y))(1-\omega)\big]^2 + 4\omega(1-\omega)F(x)\cdot G(y)\Big\}^{3/2}} \qquad (3)$$

where z_x and z_y are specific transformations of x and y, respectively and ω, defined as the cross-product ratio or odds-ratio, is a positive constant for all (x,y) for which neither F nor G assumes the value 0 or 1; ρ is a measure of the degree of association.

Fitting the SBB, Weibull-2[P] and Logit-logistic (LL-2[P]) bivariate distributions

The S_{BB} distribution was obtained from S_B marginal distributions using the normal copula. In this case, the variables x and y were defined as: $x = \big(D-\varepsilon_1\big)\big/\lambda_1$ and $y = \big(H-\varepsilon_2\big)\big/\lambda_2$ where ε_1 and ε_2 are the location parameters and λ_1 and λ_2 are the observed ranges of diameter (D) and height (H), respectively. The values of z_x and z_y were obtained from a four-parameter logistic transformation of x and y $z_x = \gamma_1 + \delta_1 \log\big(x/1-x\big)$ and $z_y = \gamma_2 + \delta_2 \log\big(y/1-y\big)$. These variables have a joint normal bivariate distribution with correlation ρ:

$$h(z_x, z_y, \rho) = \frac{f(x) \cdot g(y)}{\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{z_x^2 - 2\rho \cdot z_x z_y + z_y^2}{1-\rho^2}} \qquad (4)$$

The parameters ε were predetermined as d_{min}-0.5 and h_{min}-0.5 for diameter and height, respectively, whereas the parameter λ was set equal to the range of diameters

and the range of heights plus one, for the two marginal distributions, respectively.

The Weibull-2[P] and the LL-2[P] bivariate distributions were obtained using the Plackett copula and the marginal Weibull and Logit-Logistic density functions:

$$\text{Weibull density function } f(x) = \left(\frac{c}{b}\right)\left(\frac{x-\varepsilon}{b}\right)^{c-1} e^{-\left(\frac{x-\varepsilon}{b}\right)^c} \qquad (5)$$

Logit-logistic density function

$$f(x) = \frac{\lambda}{\sigma}\frac{1}{(x-\varepsilon)(\varepsilon+\lambda-x)}$$

$$\frac{1}{e^{-\mu/\sigma}\left(\frac{x-\varepsilon}{\varepsilon+\lambda-x}\right)^{1/\sigma} + e^{\mu/\sigma}\left(\frac{x-\varepsilon}{\varepsilon+\lambda-x}\right)^{-1/\sigma} + 2} \qquad (6)$$

where x is the diameter (D), y is the height (H), ε is the location parameter, b and c are the scale and shape parameters of the Weibull distribution, with $b, c > 0$; λ is the scale parameter and μ and σ are the shape parameters of the Logit-logistic distribution, with $\varepsilon < x < \varepsilon + \lambda$; $-\infty < \varepsilon < \infty$; $-\infty < \mu < \infty$; $\lambda > 0$; $\sigma > 0$.

The parameters were estimated by minimizing the negative log-likelihood function of equations (1) for Weibull-2[P] and LL-2[P] and (4) for S_{BB} using the R function optim (R Core Team, 2014). Assuming that the sample observations are independent with identical distributions, the negative log-likelihood function is the sum of single-tree terms (Wang & Rennolls, 2007). Both univariate distributions considered in this study to develop bivariate distributions using the Plackett copula (Weibull-2[P] and LL-2[P]) have a closed form of their cumulative distributions. If this is not the case, numerical methods should be used for evaluating the cumulative distribution in the model-fitting process.

Comparing the bivariate distributions and goodness-of-fit

Each bivariate model considered in this study has the same number of parameters, namely five: two specific parameters for each marginal distribution and one common parameter. Thus, the parameter values were used as goodness-of-fit criteria for comparison. The Wilcoxon signed-rank test was used to compare the related samples of the negative log-likelihood function calculated for each sample plot and each of the three distributions. This is a non-parametric paired difference test to assess whether the population mean ranks differ when the population cannot be assumed to be normally distributed.

Results and discussion

The means, maxima, minima and standard deviations of the estimated parameters for the three bivariate distributions (bivariate Johnson's S_{BB}, Weibull-2^P and LL-2^P) are presented in table 2. The maximum likelihood estimation converged for all plots and for all three bivariate distributions. In a study in Chinese fir plantations (*Cunninghamia lanceolata* Lamb.) a number of sample plots did not converge for the LL-2^P bivariate and bivariate beta distributions, probably due to these plots having J-shaped marginal distributions (Wang & Rennolls, 2007). Our good results could be due to the fact that all sample plots were installed in even-aged forests.

Table 3 presents the between-model comparative performance of the three bivariate distributions in terms of their goodness-of-fit statistics and the Wilcoxon rank test. The best results were obtained with the Logit-lo-

gistic marginal distribution combined with the Plackett copula. Wang *et al*. (2008), in a study for Chinese fir plantations, comparing five different copulas including the Normal and the Plackett, found that the normal copula showed the best results. In this study, we cannot compare directly the copulas because we are using different marginal distributions with each copula. Moreover, as the authors pointed out, the age range of the Chinese fir plantations used in their study was very limited and older stands had different structures influencing the observed outcomes.

The S_{BB} distribution showed better results in terms of goodness-of-fit statistics than the Weibull distribution (Table 3), although the differences were not significant. The better performance of LL-2^P over S_{BB} and Weibull was expected since the logit-logistic univariate distribution is more flexible than the other two, covering a wide range of skewness-kurtosis combinations. However, the good results of the Weibull distribution were unexpected,

Table 2. Mean values, maximum, minimum and standard deviation of the parameters for the three bivariate distributions compared.

Marginal	Parameter		Mean	Max	Min	S.D.
S_{BB}	Diameter	ε	2.78	10.80	-0.42	1.79
		λ	22.90	70.30	3.10	8.25
		δ	0.81	1.15	0.55	0.13
		γ	0.01	1.00	-1.14	0.38
	Height	ε	4.75	13.80	0.10	2.47
		λ	21.50	43.10	3.80	6.91
		δ	0.86	1.40	0.54	0.16
		γ	-0.40	0.72	-1.59	0.38
	Common	ρ	0.88	0.98	0.59	0.06
	- Log-likelihood		-259.2	-43.93	-555.4	75.98
Weibull-2^P	Diameter	ε	2.78	10.80	-0.42	1.79
		b	10.53	21.61	1.50	3.45
		c	2.21	4.67	1.19	0.60
	Height	ε	4.75	13.80	0.10	2.47
		b	10.91	23.88	1.79	4.20
		c	3.09	6.50	1.23	0.92
	Common	ω	57.81	186.14	6.36	39.32
	- Log-likelihood		-258.7	-86.16	-531.0	72.11
LL-2^P	Diameter	ε	2.78	10.80	-0.42	1.79
		λ	22.90	70.30	3.10	8.25
		μ	-0.03	8.53	-4.22	0.89
		σ	0.81	9.88	0.43	0.86
	Height	ε	4.75	13.80	0.10	2.47
		λ	21.50	43.10	3.80	6.91
		μ	0.50	3.09	-2.36	0.63
		σ	0.71	5.73	0.33	0.42
	Common	ω	28.90	231.85	0.09	19.65
	- Log-likelihood		-263.0	-92.40	-589.2	79.00

Table 3. Between-model comparative performance of these three models in terms of their goodness-of-fit statistics and the Wilcoxon test. Ratio is the proportion of cases in which the row distribution model had a lower value of the negative log-likelihood function than the column distribution.

	S_{BB}		Weibull-2^P		LL-2^P	
	Ratio	**Wilc. test**	**Ratio**	**Wilc. test**	**Ratio**	**Wilc. test**
S_{BB}	–	–	160/308	0.9167	104/308	0.0001
Weibull-2^P	148/308	0.9167	–	–	118/308	0.0196
LL-2^P	204/308	0.0001	190/308	0.0196	–	–

because the Weibull univariate turned out to be the least flexible of the three univariate distributions used. The reason for this may be the very regular shape of the marginal diameter and height distributions of our even-aged stands. Moreover, it also should be taken into account that the locations (ε) and the ranges (λ) of diameters (D) and heights (H) were fixed, affecting especially the performance of LL-2^P and S_{BB} bivariate distributions.

Both the normal and the Plackett copulas have shown a good performance for modeling the joint distribution of tree diameters and heights. They could be easily extended for modelling multivariate distributions involving other tree variables. However, it should be noted that the normal copula, in general, does not have a closed form for its joint density, except for the Normal or Johnson's marginal distributions. Another point to consider is the fact that the Plackett copula requires that the marginal has a closed form for its cumulative distribution ($F(x)$ and $G(y)$ in equation (3)), to avoid numerical methods for evaluating the cumulative distribution in the model-fitting process.

References

Castedo-Dorado F, Ruiz-González AD, Álvarez-González JG, 2001. Modelización de la relación altura-diámetro para Pinus pinaster Ait. en Galicia mediante la función de densidad bivariante S_{BB}. Invest. Agrar Sist Recur For 10(1): 111-125.

Hafley WL, Schreuder HT, 1977. Statistical distributions for fitting diameter and height data in even-aged stands. Can J For Res 7(3): 481-487. http://dx.doi.org/10.1139/x77-062

Johnson NL, 1949. Bivariate distributions based on simple translation systems. Biometrika 36: 297-304. http://dx.doi.org/10.1093/biomet/36.3-4.297

Knoebel BR, Burkhart HE, 1991. A bivariate distribution approach to modelling forest diameter distributions at two points of time. Biometrics 47: 241-253. http://dx.doi.org/10.2307/2532509

Li F, Zhang L, Davis CJ, 2002. Modeling the joint distribution of tree diameters and heights by bivariate generalized Beta distribution. For Sci 48(1): 47-58.

Mardia KV, 1970. Families of bivariate distributions. Griffin, London, UK. 231 pp.

Mønness E, 2015. The bivariate power-normal distribution and the bivariate Johnson system bounded distribution in forestry, including height curves. Can J For Res 45(3): 307-313. http://dx.doi.org/10.1139/cjfr-2014-0333

Plackett RL, 1965. A class of bivariate distributions. J Am Stat Assoc 60: 516-522. http://dx.doi.org/10.1080/01621459.1965.10480807

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rennolls K, Wang M, 2005. A new parameterization of Johnson's SB distribution with application to fitting forest tree diameter data. Can J For Res 35(3): 575-579. http://dx.doi.org/10.1139/x05-006

Rupsys P, Petrauskas E, 2010. The Bivariate Gompertz Diffusion Model for Tree Diameter and Height Distribution. For Sci 56(3): 271-280.

Schreuder HT, Hafley WL, 1977. A useful bivariate distribution for describing stand structure of tree heights and diameters. Biometrics 33: 471-478. http://dx.doi.org/10.2307/2529361

Sklar A, 1973. Random variables, joint distribution functions and copulas. Kybernetika 9: 449-460.

Staudhammer CL, LeMay VM, 2001. Introduction and evaluation of possible indices of stand structural diversity. Can J For Res 31: 1105-1115. http://dx.doi.org/10.1139/x01-033

Tewari VP, Gadow Kv, 1999. Modelling the relationship between tree diameters and heights using S_{BB} distribution. For Ecol Manage 119: 171-176.

Wang M, Rennolls K, 2007. Bivariate Distribution Modeling with Tree Diameter and Height Data. For Sci 53(1): 16-24.

Wang M, Rennolls K, Tang S, 2008. Bivariate Distribution Modeling of Tree Diameters and Heights: Dependency Modeling Using Copulas. For Sci 54(3): 284-293.

Zucchini W, Schmidt M., Gadow Kv, 2001. A model for the diameter-height distribution in an uneven-aged beech forest and a method to assess the fit of such models. Silva Fenn 35(2): 169-183. http://dx.doi.org/10.14214/sf.594