An equation for yield prediction for *Pinus taeda* L. as a function of soil properties

Itamar A. Bognola[A]; Osmir J. Lavoranti[A], Antonio Rioyei Higa[B], João Bosco V. Gomes[A], Shizuo Maeda[A] and Carla M. Camargo Corrêa[B]

[A]Embrapa Forestry, Colombo, PR, Brazil, Email iabog@cnpf.embrapa.br, osmir@cnpf.embrapa.br, jbvgomes@cnpf.embrapa.br, maeda@cnpf.embrapa.br
[B]Forestry Department, University of Federal Paraná, Curitiba, PR, Brazil, Email higa@ufpr.br; camargocorrea@ufpr.br

Abstract
This study was developed in order to obtain an equation for yield prediction as a function of soil properties, for plantations established with *Pinus taeda* L. in the plateau located in the northern part of the State of Santa Catarina, Brazil. As a case study, one of the farms of Battistella Florestal Ltd. was chosen. Data were obtained from permanent sample plots from a Continuous Forest Inventory as well as from a detailed soil map of a 2,252 ha forest estate. Soil properties and yield of *P. taeda* were analysed with stepwise multiple linear regression. By these means an equation was developed for site index prediction with 93.97 % precision. It was also found that soils with well developed surface drainage and high levels of organic matter are more suitably for growth of *P. taeda* under local environmental conditions.

Key Words
Multivariate statistical methods, soil survey, forest planning.

Introduction
Forest owners need low cost and precise information about stand yield before harvesting operations are implemented in order to improve planning of the supply to log yards as well as for planning purposes and forest inventory. Growth and yield of planted forests depend on physiological responses to the interactions between biotic and abiotic environmental factors. Climate, physiographic and soils factors are the most important environmental elements influencing productive capacity in a given location, these represent site quality (Carmo and Resende 1990). Soil class incorporates important information such as depth, texture, nutrient and organic matter, chemical activity of the coloyd fraction and compacted layers that could restrict root growth and water movement (Rigatto *et al.* 2005). On the other hand, there are few studies about the joint analysis of such factors and their correlation with different ecosystems, site quality and pine species. Evaluation of potential and limiting factors affecting environmental quality for productive capacity of forest site evaluation cannot be based on isolated attributes, but on a syntesis of qualities and limitations of the ecosystem from an integrated perspective. Hence, it can be observed that when physical, chemical and physical-hydric soil characteristics, geology, terrain and climate are jointly examined and correlated to the different scenarios, the overall ranking of the influences is easier (Van Den Berg 1995; Rigatto *et al.* 2005). On the other hand, understanding the behavior of ecosystems is difficult due to the complex interactions and quite frequently demands predictive models. For this reason, when a detailed analysis is needed in order to understand the relations between quantitative attributes of the trees and physical environment the choice of a statistical method that optimizes resources without reducing precision of the estimation process is of fundamental importance (Mello *et al.* 2005). In this context, and according to Bognola (2007), this research was conceived in order to study physical environmental factors affecting growth of *Pinus taeda* L. in a commercial plantation as well as to develop an equation for yield prediction using multivariate techniques and regression analysis.

Material and methods
Working area
The study was developed in one of the forest farms belonging to Battistella Florestal, and located on the plateau of the northern part of the state of Santa Catarina, Brazil, in Rio Negrinho e Doutor Pedrinho counties (Figure 1). Climate type is *Cfa* according to Köppen's classification (tropical climate, with warm Summer, without any dry season, average temperatures of the coldest month under 18°C and above -3°C). The rainfall of the region is high (1,700 mm/y) and well distributed over the year.

1 – 6 August 2010, Brisbane, Australia. Published on DVD.

Data collection

Data were collected in 500 m^2 permanent sample plots of a Continuous Forest Inventory. In addition to DBH and height of individual trees, the following information was also obtained: a) physical, chemical and physical-hydric soil characteristics at 0–20 cm and 30–50 cm depth for all sample plots considered; b) aspect, geology and physiographic descriptive information obtained from a detailed soil survey (scale 1:10,000) of the area. Statistical analysis was performed using SAS® - Statistical Analysis System (SAS Institute Inc. 1993), licensed for Embrapa Florestas.

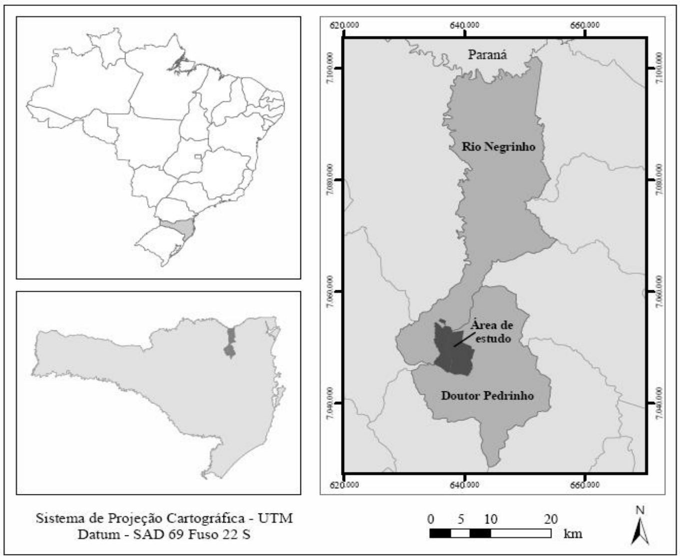

Figure 1. Working area location.

Regression Equation

A multiple regression analysis was performed using the following general linear model:

$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + ... + \beta_p Xp_i + \varepsilon_i$

where: Y_i is the observed value for the response variable (dependent) at the level i of the predictive (independent) variables X_i´s;

β_0 : regression constant (intercept of the regression equation with axis y);

β_1 : regression coefficient (variation of Y as a function of the variation of one unit of X1);

β_2 : regression coefficient (variation of Y as a function of the variation of one unit of X2);

β_p : regression coefficient (variation of Y as a function of the variation of one unit of Xp);

$X1_i$: value of variable X1, at the level i;

$X2_i$: value of variable X2, at the level i;

Xp_i : value of variable Xp, at the level i;

ε_i : error or deviation associated with the distance of the observed value Yi and the corresponding estimated value $\hat{Y}_i$ based on the regression equation.

As the coefficient of determination always increases with the inclusion of new a independent variable in the equation, even if it has no relation with the dependent variable, the corrected coefficient of determination was used for the degrees of freedom as defined by Ribeiro Júnior (2001):

$$R^2_{adjusted} = \overline{R}^2 = R^2 - \frac{p}{n-p-1}(1-R^2)$$

where *p* sets the number of regression coefficients (not including β_0).

In practice there can be a large number of variables influencing the response. Hence, a stepwise regression procedure was adopted for the selection of the best regression model. This procedure determines that the first variable entering the model is the variable X with the highest correlation with the response variable Y. Subsequently, through the largest values for the partial coefficient of determination of all other variables were not included in the model. The model is adjusted with the inclusion of the last variable and the values for the partial F test are determined. By those means any variable that produces a non-significant contribution (partial F) is removed from the model. This process continues until there are no other variables being added or removed (Lavoranti 2005).

Results

In order to identify the best regression equation for *P. taeda* plantations, for Site Index 15 (SI_15), based on 49 independent variables, related to physical environment characteristics, a principal component (PCA) multiple linear regression analysis was used. Regressive variables were selected in two stages: a) first, the most significant variables were identified by principal component and factor analysis; and b) the eleven most significant factors were identified and considered in a stepwise regression analysis. The reduction of variables through PCA has produced a 9 % loss of information on the total variance. In other words, the model developed with this procedure explained 91 % of the data variability. According to Royston (1992), an exploratory analysis has to be performed with the data in order to verify the normal distribution of the residuals for the structure of the regression model. In this context, it can be verified in Table 1 for the analysis of variance and, in Table 2, where the values for the coefficients of the regression model are defined. The variables selected are presented in equation 1.

Table 1. Analysis of variance for the multiple regression in the step 19 (stepwise method), for the set of independent variables of this study.

Source	D.F.	SQ	MQ	F_0	Pr > F_0
Regression	5	4.609,03071	921,80614	162,71	< 0,0001
Error	11	62,31929	5,66539		
Total	16	4.671,35			

$$SI\_15_{est} = b_1 KPA1\_10 + b_2 KPA2\_10 + b_3 HAL1 + b_4 DENSID1 + b_5 MO1 + \varepsilon \qquad \text{(eq. 1)}$$

Where:

SI_15est: Site Index estimated for *P. taeda* at age 15 years (m);

KPA1_10: Water removed from the soil (cm^3/cm^3), tension 10 kPa, for the surface layer (5 – 10 cm) (moisture content, at the field capacity – CC);

KPA2_10: Water removed from the soil (cm^3/cm^3), tension 10 kPa, for the sub-superficial layer (30 – 50 cm) (moisture content, at the field capacity – CC);

HAL1: exchangeable hydrogen + aluminum contents (cmolc dm^{-3}), in the superficial layer (5 – 10 cm);

DENSID1: Soil density (kg/dm^3) for the supeficial layer (5 – 10 cm);

MO1: Organic matter (g/dm^3), in the superficial layer (5 – 10 cm);

ε: Residual;

$b_1,..., b_5$: Coefficients of the model.

Table 2. Coefficients of the regression model.*

Coefficient (β_i)	Estimated Parameter	Error (ε)	t	p > t
b_1	-520.392	210.213	-247.555	0.03082
b_2	215.549	53.809	400.418	0.00207
b_3	0.01799	0.00768	234.109	0.03910
b_4	904.679	0.70485	1.283.509	< 0.0001
b_5	0.39932	0.06772	589.679	0.00010

* Coefficient of Determination: R^2 = 0.9867.

It was observed that among all variables selected for the development of the predictive model, those related to the physical-hydric soil properties were the most important in that they presented the highest values for the coefficient of correlation. The analysis of residual distribution as a function of the estimated values allowed the observation of homoscedasticy, as well as the absence of outliers. Fitting the model was adequate for all the extentions examined which justifies the used of the model for the estimation of Site Index as a function of the variables chosen.

Conclusion

The equation defined through multivariate regression analysis of the case study data allows site index estimates with a 93.97 % precision.

It was also found that soils with well developed surface drainage and high levels of organic matter are most suited for the growth of *P. taeda* under local environmental conditions.

References

Bognola IA (2007). Unidades de manejo para P. taeda L. no planalto norte catarinense, com base em características do meio físico. Curitiba, 160 f. thesis (DSc. Silviculture), Centro de Ciências Florestais, Universidade Federal do Paraná.

Carmo DN, Resende M, Silva TCA (1990) Avaliação da aptidão das terras para eucalipto. (NF Barros, NF, RF Novais) pp.187-235. (Relação Solo-Eucalipto: Viçosa, Folha de Viçosa).

Lavoranti OJ (2005). Análise de Regressão Linear Múltipla. Curitiba, 41f. Apostila da Discplina CE-216 do Curso de Pós-graduação - Setor de Ciências Exatas – Departamento de Estatística, Universidade Federal do Paraná (UFPR).

Mello JM, Batista JLF, Ribeiro Jr PJ, Oliveira MS (2005) Ajuste e seleção de modelos espaciais de semivariograma visando à estimativa volumétrica de Eucalyptus grandis. *Scientia Forestalis, Piracicaba, SP,* **69**, 25-37.

Ribeiro Júnior JI (2001) 'Análises estatísticas no SAEG'. (Viçosa: UFV).

Rigatto PA, Dedecek RA, Mattos JLM de (2005) Influência dos atributos do solo sobre a produtividade de P. taeda. *Revista Árvore, Viçosa* **29**, 701-709.

SAS - Statistical Analysis System (1993) 'Programa de computador, ambiente VM. Cary,versão 6.08'.

Van den Berg E (1995) Estudo florístico e fitossociológico de uma floresta ripária em Itutinga, MG, e análise das correlações entre variáveis ambientais e a distribuição das espécies de porte arbóreo-arbustivo. Lavras, MSc Dissertation, Federal University of Lavras.