USE OF THE CORRELATION COEFFICIENT BETWEEN PLOTS IN ORDER TO IMPROVE THE ACCURACY OF FOREST INVENTORIES

Daniela Cunha da Sé[1]*, José Márcio de Mello[1], João Domingos Scalon[1], Joel Augusto Muniz[1],
Marcelo Silva de Oliveira[1], José Roberto Soares Scolforo[1]

*Corresponding author: danicunhase@gmail.com

ABSTRACT: Forest inventories are usually compiled without taking into account the existing correlations between sampling units, which is debatable particularly where the calculations involve environmental variables. When the potential correlations between sampling units are overlooked, the accuracy of such inventories becomes distorted in terms of the confidence interval range for the variable of interest, which is volume in cubic meters. The magnitude and form of such distortion will vary according to the correlation intensity between sampling units. This study aimed to present an analysis of the addition of the correlation coefficient to the calculation of the variance of the mean in a systematic sampling procedure of a native forest population or area, as well as its impact on the accuracy of the resulting estimates, with the assumption of independence between sampling units and with the addition of a correlation between sampling units as suggested by Cochran. Results revealed that, where the correlation coefficient was added to the variance of the mean formula, it increased inventory accuracy by about 14.3%, leading to the conclusion that such an effect will occur in any forest inventory being compiled for any forest population or area of interest.

Key words: Sampling, variance of the mean, native forest.

O USO DO COEFICIENTE DE CORRELAÇÃO ENTRE PARCELAS VISANDO AUMENTO DE PRECISÃO EM INVENTÁRIOS FLORESTAIS

RESUMO: *Os inventários florestais têm sido realizados sem levar em consideração a relação existente entre as unidades amostrais. Esse fato se torna altamente questionável, principalmente quando nos cálculos estão envolvidas variáveis de natureza ambiental. Ao serem desprezadas as possíveis relações entre as unidades amostrais, a precisão advinda desses inventários florestais se torna distorcida em termos de amplitude do intervalo de confiança da variável estudada, sendo esta o volume em metros cúbicos. A magnitude e a forma dessa distorção variam conforme a intensidade da correlação existente entre unidades amostrais. Dessa forma, no presente trabalho, objetivou-se apresentar uma análise da incorporação do coeficiente de correlação no cálculo da variância da média no procedimento de amostragem sistemática, em uma população ou área de floresta nativa e, seu impacto na precisão das estimativas geradas, com a suposição de independência entre as unidades amostrais e, com a adição da correlação entre as unidades amostrais, sendo esta incorporação sugerida pelo autor Cochran. Os resultados obtidos mostram que o coeficiente de correlação, quando incorporado na fórmula da variância da média, aumentou em média 14,3% a precisão dos inventários florestais avaliados neste trabalho, fato este que conclui que tal efeito ocorrerá em qualquer inventário florestal realizado para qualquer população ou área florestal a ser avaliada.*

Palavras-chave: Amostragem, variância da média, floresta nativa.

1 INTRODUCTION

Preparing a forest management plan with its relevant procedures is only possible by knowing or by estimating the parameters of the forest population in question.

A forest inventory can be defined as an activity that seeks to obtain quantitative and qualitative information on existing forest resources in a preestablished area (population), therefore a forest inventory consists in partially measuring a population, that is, measuring sampling units or plots to then subsequently generate estimates for the total area (LEITE; ANDRADE, 2002). More specifically, it estimates the biophysical characteristics of a given forest from direct measurement of individual trees in sampling plots that are representative of the tree population constituting such forest (RODRIGUEZ et al., 2010). Its purpose is to apply and evaluate sampling systems capable of generating accurate estimates of the population being sampled.

Much of the costs incurred in the forestry sector are concentrated in obtaining information necessary to carry out planning activities. Countless forestry-related studies were conducted looking to optimize the cost-accuracy relationship, including: Druszcz et al. (2010), Nakajima et al. (1998), Soares et al. (2004) and Vasquez (1988). It is thus extremely critical to obtain this necessary information not only at the lowest possible cost but also with the highest possible accuracy (SOARES et al., 2004).

[1] Universidade Federal de Lavras – Lavras, Minas Gerais, Brazil

This fact justifies seeking more specific methodologies, in terms of sampling, for the various forestry sectors. The idea of improving accuracy by using sampling procedures that have new estimators added without incurring extra costs while being easy to apply is thus very attractive.

Several sampling procedures are available that could be used in forestry to address a population of interest, and the most commonly used are Simple Random Sampling, Stratified Random Sampling, Systematic Sampling, Cluster Sampling, among others (PÉLLICO-NETO; BRENA, 1997). It should be noted, however, particularly where native forests are concerned, that systematic sampling is usually the procedure obtaining best estimates for the parameters of interest, as it ensures better representativeness of the sampled area (AUBRY; DEBOUZIE, 2001) due to the systematic method of plot distribution across the field, consequently capturing best the variation in the area.

One drawback of using systematic sampling is that it does not allow deducing an estimator for the variance of the mean from data of a single sample. This is due to the fact that the selection of sampling units is not independent since only the first unit is random. Several methods have been proposed to best determine the approximation of the sampling error of a systematic sample (SOUZA, 2007). In populations with heterogeneity between the sampling units, or with a defined tendency, an alternative for estimating the variance and sampling error is to use the successive difference formula based on the premise that the sampling units are not completely independent (CAMPO; LEITE, 2006).

Cochran (1977) argues that systematic sampling is accurate when units within the same sample are heterogeneous and inaccurate when units are homogeneous, which is obviously intuitive, because if there is little variation within a systematic sample, successive sampling units will be repeatedly providing the same information.

Due to the great diversity present in native forests, even when they are stratified into sub-populations, obtaining accurate estimates can be difficult. With that in mind, a possible proposal for solving this problem is to add a variable to the estimators used in the systematic sampling procedure that will enable capturing the variation between the launched plots, in word words, besides the sample variance already existing in that estimator, to add a variable capable of better explaining the great diversity present. This variable is the correlation coefficient between plots as proposed by Cochran (1977) and used by Mello (2004).

The correlation coefficient follows classical statistics theory as it only considers the relationship between trial units rather than taking into account their spatial location, and thus acting as a measure of population homogeneity (COCHRAN, 1977). Overlooking these potential correlations between sampling units can distort estimates made for population variability. This means to say that if the correlation between sampling units is ignored, the resulting confidence intervals are overestimated or underestimated depending on the intensity of the correlation being disregarded (MINGOTI; FIDELIS, 2001).

The central idea is that the estimator used to obtain the variance of the mean in systematic sampling procedures of native forests fails to effectively provide the variance around the mean within a population, leading to confidence interval distortions. Even being the procedure that provides best spatial representativeness of the population, there is loss of information which is strongly captured when systematic sampling is used, for instance the relationship between sampling units. Therefore, the estimator proposed by Cochran (1977) can help capture better such relationships by adding the correlation coefficient to the estimator of the variance of the mean.

This study aimed to test the effectiveness of the correlation coefficient in improving accuracy of a forest inventory using the systematic sampling procedure, and also to compare performances of the traditionally used variance of the mean estimator and the estimator proposed by Cochran (1977).

2 MATERIAL AND METHODS

2.1 Study site

The study was conducted in a montane seasonal semideciduous forest 5.04 hectares in area and located in the municipality of Lavras (MG) at coordinates 21°13’40”S and 44°57’50”W, at an altitude of 925 meters (Figure 1). The local climate is Cwb type, according to Koppen classification, which means temperate with mild summers and dry winters. The local soil is predominantly a distrophic red latosol with a very clayey texture (CURI et al., 1990).

2.2 Data collection

Dendrometric data were collected from every individual in 126 contiguous plots 20 x 20 meters in size. In each plot, all individuals were marked with metal tags containing plot number and tree number. All individuals were measured to obtain merchantable height and circumference 1.30 meter above the ground (CBH).

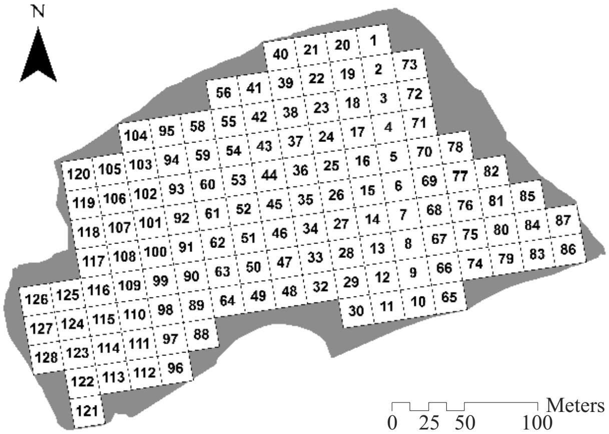

Figure 1 – Map of the study site with delimitated plots.

Figura 1 *– Mapa da área experimental com as parcelas delimitadas.*

Individual tree volume was estimated using an equation selected by Scolforo et al. (1994), and individual volumes were subsequently added together to find the aggregate plot volume.

2.3 Data processing

Data were processed to generate and assess the percent error of the inventory and the confidence interval, in two situations: (1) calculation based on the variance of the mean by the estimator of simple random sampling, (2) adding the correlation coefficient between sampling units to the estimator of simple random sampling.

Because all individuals in the area were enumerated (census), the parameters became known. That enabled simulations of eleven systematic samplings in the area, in which what changes is k (sampling interval) as a function of the allowable error (E%). Calculations derived two possible samples with an allowable error of 7.5%, k=2 plots and n=63 plots; three possible samples with an allowable error of 10.6%, with k=3 plots and n=42 plots; and six possible samples with an allowable error of 17%, with k=6 plots and n=21 plots. Allowable errors were selected for practical reasons. An allowable error of 7.5% is generally used in commercial forest inventories, while errors of 10.6% and 17% were accepted so as to obtain a larger number of simulations of possible systematic samples in the area.

Data were organized in such a way as to produce eleven different databases of the same trial site, that is, eleven possible systematic samples for the area. The eleven databases are divided into three major groups: Group A, whose sampling interval was 2 plots, thus providing two databases; Group B, whose sampling interval was 3 plots, thus providing three databases; and Group C, whose sampling interval was 6 plots, thus providing 6 databases.

Based on the eleven databases, part one of inventory processing was performed using estimators of simple random sampling, with assumption of independence between samples. These estimators are cited in several books directed at forest inventory sampling, including Cochran (1977), Scolforo and Mello (2006) and Thompson (1992). Next is the usual estimator of the variance of the mean (1).

$$S_{\bar{y}}^2 = \frac{S^2}{n} * \left(\frac{N-n}{N}\right) \tag{1}$$

Part two of inventory processing was performed similarly to the usual procedure (part one), the difference lies in the estimate of the variance of the mean, to which the correlation coefficient is added according to formula (2) and as proposed by Cochran (1977) and used by Mello (2004).

$$V_{(\bar{y})} = \frac{S^2}{n}\left(\frac{N-n}{N}\right)\left[1+(n-1)\rho_w\right] \tag{2}$$

where N: number of sampling units applicable to the area, n: sampling intensity used in the area, S^2: sample variance of data obtained in the survey, and ρ_w: correlation coefficient between paired units from the same systematic sample, as defined by formula (3):

$$\rho_w = \frac{2}{(n-1)(N-1)S^2}\sum_{i=1}^{k}\sum_{j<u}(y_{ij}-\bar{Y})(y_{iu}-\bar{Y}) \tag{3}$$

where y_{ij}: member of order j of systematic sample of order i, so that j=1,2,...,n, i=1,2,...,k, $\bar{Y}$: sample mean of individuals, n: sampling intensity in the area, N: number of plots applicable to the area, and S^2: sample variance of the data.

All analyses, charts and calculation routine of the correlation coefficient were performed using software R Development Core Team (2010). Requests for use should be submitted to the author.

3 RESULTS AND DISCUSSION

Table 1 provides the main descriptive statistics of forest inventory processing, plus the correlation coefficient between sampling units. This information should be submitted to exploratory analysis. Data refer to the three groups of systematized plots launched in the area. The value of the estimated mean was found to be similar among all three groups, regardless of the sampling intensity. This is due to the estimator of the sample mean not being biased, according to the statistical properties of this estimator (MAGINA et al., 2010).

Table 1 – Estimates of mean, coefficient of variation (CV%) and correlation coefficient (ρ_W), divided into groups according to the sampling intensity, for the variable volume (m^3).

Tabela 1 – *Estimativas da média, coeficiente de variação (CV%) e coeficiente de correlação (ρ_W), sendo estes divididos em grupos de acordo com a intensidade amostral, para a variável volume (m^3).*

Group	Sample	Mean	CV(%)	n	ρ_w
A	1	4.5151	32.39	63	-0.0082
A	2	4.5445	31.31	63	-0.0081
B	3	4.2956	33.61	42	-0.0080
B	4	4.5402	27.62	42	-0.0080
B	5	4.7536	33.48	42	-0.0081
C	6	4.5215	35.01	21	-0.0083
C	7	4.6864	27.33	21	-0.0081
C	8	4.6299	34.53	21	-0.0082
C	9	4.0697	31.65	21	-0.0081
C	10	4.3939	28.23	21	-0.0081
C	11	4.8773	33.08	21	-0.0084

Variability had a random pattern within the three groups formed. This random variation denotes the heterogeneity of volume values between plots, which was corroborated by the low correlation coefficient (ρ_w), as described by Cochran (1977).

Formula (2) reveals that a positive correlation between the units of a single sample inflates the variance of the mean sample value. Even a small positive correlation can have a strong effect, due to the multiplier (n-1).

In referring to natural populations, Cochran (1977) argues that there is reason to expect that two observations, y_i and y_j, be approximately similar when i and j are neighbors in a sample than when they are farther apart. The author maintains that this happens whenever natural forces produce slow changes as one progresses with the sample. Forming a mathematical conception of this effect, one can assume that y_i and y_j are positively correlated and that this function depends solely on the distance that separates them, thus decreasing to the extent that the distance increases. Although this conception is an oversimplified notion, it may represent an important aspect in many native forest populations.

All sampling procedures used in forest inventories are grounded in the assumption of independence between sampling units, which is debatable particularly where calculations involve environmental data. By overlooking the potential correlations between sampling units, one could be distorting the estimates of variability for a given population (MINGOTI; FIDELIS, 2001).

The correlation results were found to be small and negative. According to Mundstock (2006), when the correlation coefficient is high and positive, the units of a systematic sample will be homogeneous, whereas when the correlation coefficient is low, whether positive or negative, the units of the systematic sample will be heterogeneous. This is an indication that correlation coefficient is a measure of homogeneity of a systematic sample.

Table 2 provides error results for the simulated inventories of the forest population in question, along with the percent differences, depending on whether the inventory is following the assumption of independence between sampling units or whether a measure of correlation is added.

Table 2 – Estimates of inventory error (%) for different systematic samples, simulated in the study site, in which Error^1 follows the assumption of independence between sampling units, Error^2 considers the correlations between sampling units and dif (%) is the percent difference between the two estimated errors.

Tabela 2 – *Estimativas obtidas do erro (%) do inventário para as diferentes amostras sistemáticas simuladas na área de estudo, onde Erro^1 obtido são os erros seguem a suposição de independência entre as unidades amostrais, Erro^2 aqueles que consideram as correlações entre as unidades amostrais e dif (%) é a diferença percentual entre os dois erros estimados.*

Sample	Allowable error (%)	Error^1 (%)	Error^2 (%)	dif (%)
1	7.5	5.77	4.09	29.116
2	7.5	5.58	3.96	29.032
3	10.6	8.55	7.01	18.012
4	10.6	7.03	5.76	18.065
5	10.6	8.52	6.98	18.075
6	17.0	14.51	13.30	8.339
7	17.0	11.32	10.38	8.304
8	17.0	14.31	13.11	8.386
9	17.0	13.11	12.02	8.314
10	17.0	11.07	10.72	3.162
11	17.0	13.71	12.56	8.388

An analysis of Tables 1 and 2 reveals, first of all, that within each group the sample having the greatest variance was also the sample having the greatest percent difference between the two estimated errors.

That statement proves that systematic sampling in native forest stands is more accurate when the correlation coefficient is used for estimating the variance of the mean. It proved effective in the simulations run here in improving the accuracy of the forest inventory. And consequently, it can be said that the usual formula of variance of the mean fails to efficiently capture the variability present in the area when systematic samplings are performed.

Inventory errors were invariably smaller with addition of a correlation measure between the sampling units than the errors obtained when the inventory was based on the assumption of independence between the sampling units, for the area in question (Figure 2). On average, a reduction of 14.3% was noted when the correlation coefficient was added to data processing, all three groups considered.

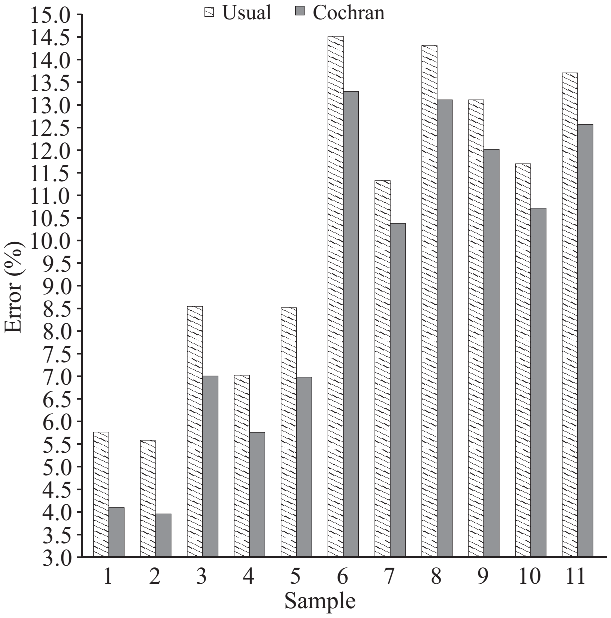

Figure 2 – Comparison of inventory errors when using the variance of the mean based on the assumption of independence between units (dashed line), as opposed to when using the Cochran formula (solid line).

Figura 2 *– Comparação do erro do inventário quando se usa a variância da média baseada na independência entre as unidades amostrais (tracejado), e quando se usa a formulação segundo Cochran (sólido).*

There being impact on error, changes will occur to the confidence interval range. This is illustrated in Figure 3, as when there is assumption of independence between sampling units, the confidence interval is overestimated in relation to reality. By adding a correlation measure between the sampling units, accuracy is increased and, consequently, the confidence interval is narrower than obtained previously.

From results, this study suggests that the processing of forest inventory data undergo an exploratory analysis. The exploratory analysis should place special emphasis on the issue of whether there is a correlation between the sampling units or not. There being a correlation, suitable estimators should be used that take such correlation into account.

Figure 3 illustrates that all generated confidence intervals were reliable in that they contained the population mean, noting that the intervals generated by adding the correlation coefficient proposed by Cochran (1977) not only contained the parameter mean (4.5298 m^3) but they also had a narrower range, confirming an improved inventory accuracy without loss in estimate veracity.

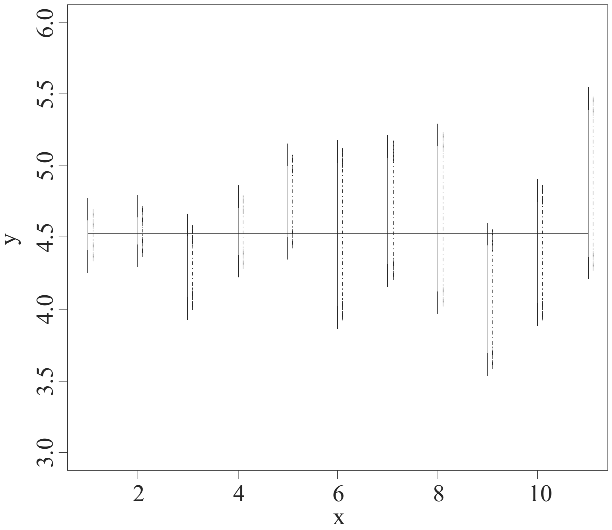

Figure 3 – Comparison between confidence interval coverage as generated by the usual estimator (solid line) and by the estimator that considers the correlation between plots (dashed line), and population mean of the stand which is 4.5298 m^3 (horizontal solid line).

Figura 3 *– Comparação entre as coberturas dos intervalos de confiança gerada pelo estimador usual (linha cheia) e pelo estimador considerando a correlação entre parcelas (linha tracejada) para as respectivas amostras simuladas, e média populacional do povoamento que corresponde 4,5298 m^3 (linha cheia na horizontal).*

4 CONCLUSIONS

The estimator of the variance of the mean proposed by Cochran (1977) proved effective in improving inventory accuracy and in consistently adding a measure of correlation between the sampling units of interest, thus causing a reduction in the inventory error which consequently led to a reduction in the confidence interval range.

5 REFERENCES

AUBRY, P.; DEBOUZIE, D. Estimation of the mean a two-dimensional sample the geoestatistical model-based approach. **Ecology**, Durham, v. 82, n. 5, p. 1484-1494, 2001.

CAMPO, J. C. C.; LEITE, H. G. **Mensuração florestal:** perguntas e respostas. 2. ed. Viçosa, MG: UFV, 2006. 470 p.

COCHRAN, W. G. **Sampling techniques**. 3rd ed. New York: Wiley, 1977. 555 p.

CURI, N.; LIMA, J. M.; ANDRADE, H.; GUALBERTO, V. Geomorfologia, física, química e mineralogia dos principais solos da região de Lavras, MG. **Ciência e Prática**, Lavras, v. 14, p. 297-307, 1990.

DRUSZCZ, J. P.; NAKAJIMA, N. Y.; PÉLLICO-NETTO, S.; YOSHITANI-JÚNIOR, M. Comparação entre os métodos de amostragem de Bitterlich de área fixa com parcela circular em plantações de *Pinus taeda*. **Floresta**, Curitiba, v. 40, n. 4, p. 739-754, out./dez. 2010.

LEITE, H. G.; ANDRADE, V. C. L. Um método para condução de inventários florestais sem o uso de equações volumétricas. **Revista Árvore**, Viçosa, v. 26, n. 3, p. 321-328, maio/jun. 2002.

MAGINA, S.; CAZORLA, I.; GITIRANA, V.; GUIMARÃES, G. Concepções e concepções alternativas de média: um estudo comparativo entre professores e alunos do ensino fundamental. **Educar em Revista**, Curitiba, p. 59-72, 2010. Número especial.

MELLO, J. M. **Geoestatística aplicada ao inventário florestal**. 2004. 110 p. Tese (Doutorado em Recursos Florestais) - Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, 2004.

MINGOTI, S. A.; FIDELIS, M. T. Aplicando a geoestatística no controle estatístico de processos. **Produto & Produção**, Porto Alegre, v. 5, n. 2, p. 55-70, 2001.

MUNDSTOCK, E. C. **Amostragem II**. Porto Alegre: Artmed, 2006.

NAKAJIMA, N. Y.; KIRCHNER, F. F.; SANQUETTA, C. R.; POSONSKI, M. **Elaboração de um sistema de amostragem para estimativa de valores correntes e mudança/crescimento em reflorestamento de Pinus**. Curitiba: UFPR, 1998. 33 p.

PELLICO-NETO, S.; BRENA, D. A. **Inventário florestal**. Curitiba: [s.n.], 1997. 316 p.

R DEVELOPMENT CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2010. Disponível em: <http://www.R-project.org>. Acesso em: 10 mar. 2011.

RODRIGUEZ, L. C. E.; POLIZEL, J. L.; FERRAZ, S. F. B.; ZONETE, M. F.; FERREIRA, M. Z. Inventário florestal com tecnologia *laser* aerotransportada de plantios de *Eucalyptus spp* no Brasil. **Ambiência**, Guarapuava, v. 6, p. 67-75, 2010. Edição especial.

SCOLFORO, J. R. S.; MELLO, J. M. **Inventário florestal**. Lavras: UFLA, 2006. 561 p.

SCOLFORO, J. R. S.; MELLO, J. M.; LIMA, C. S. A. Obtenção de relações quantitativas para estimativa de volume do fuste em floresta estacional semidecídua montana. **Cerne**, Lavras, v. 1, p. 123-134, 1994.

SOARES, T. S.; SCOLFORO, J. R. S.; FERREIRA, S. O.; MELLO, J. M. Uso de diferentes alternativas para viabilizar a relação hipsométrica no povoamento florestal. **Revista Árvore**, Viçosa, v. 28, n. 6, p. 845-854, nov./dez. 2004.

SOUZA, F. N. **Desempenho de estimadores da precisão na amostragem sistemática em inventários florestais**. 2007. 49 p. Monografia (Graduação em Engenharia Florestal) - Universidade Federal de Lavras, Lavras, 2007.

THOMPSON, S. K. **Sampling**. New York: Wiley, 1992. 343 p.

VASQUEZ, A. G. **Método de amostragem em linhas:** desenvolvimento e aplicação em uma floresta implantada com *Pinus taeda*. 1988. 129 f. Dissertação (Mestrado em Engenharia Florestal) - Universidade Federal do Paraná, Curitiba, 1988.

Received: August 22, 2011; accepted: May 24, 2013.