Local Models for Forest Canopy Cover with Beta Regression

Lauri Korhonen, Kari T. Korhonen, Pauline Stenberg, Matti Maltamo and Miina Rautiainen

Korhonen, L., Korhonen, K.T., Stenberg, P., Maltamo, M. & Rautiainen, M. 2007. Local models for forest canopy cover with beta regression. Silva Fennica 41(4): 671–685.

Accurate field measurement of the forest canopy cover is too laborious to be used in extensive forest inventories. A possible alternative to the separate canopy cover measurements is to utilize the correlations between the percent canopy cover and easier-to-measure forest variables, especially the basal area. A fairly new analysis technique, the beta regression, is specially designed for modelling percentages. As an extension to the generalized linear models, the beta regression takes into account the distribution of the model residuals, and uses a logistic link function to ensure logical predictions. In this study, the beta regression method was found to perform well in conifer dominated study area located in central Finland. The same model shape, with basal area, tree height and an additional predictor (Scots pine: site fertility, Norway spruce: percentage of hardwoods) as independent variables, produced good results for both pine and spruce dominated sites. The models had reasonably high pseudo R-squared values (pine: 0.91, spruce: 0.87) and low standard errors (pine: 6.3%, spruce: 5.9%) for the fitting data, and also performed well in a cross validation test. The models were also tested on separate test plots located in a different geographical area, where the prediction errors were slightly larger (pine: 8.8%, spruce: 7.4%). In pine plots, the model fit was further improved by introducing additional predictors such as stand age and density. This improved also the performance of the models in the cross validation test, but weakened the results for the external data set. Our results indicated that the beta regression method offers a noteworthy alternative to separate canopy cover measurements, especially if time is limited and the models can be applied in the same region where the modelling data were collected.

Keywords beta regression, canopy cover, forest canopy

Addresses *L. Korhonen* & *Maltamo*: Univ. of Joensuu, Faculty of Forest Sciences, P.O. Box 111, FI-80101 Joensuu, Finland; *K.T. Korhonen*: Finnish Forest Research Institute, Joensuu Research Unit, P.O. Box 68, FI-80101 Joensuu, Finland; *Stenberg* & *Rautiainen*: Univ. of Helsinki, Dept of Forest Resource Management, P.O. BOX 27, FI-00014 University of Helsinki, Finland
E-mail lauri.korhonen@joensuu.fi, matti.maltamo@joensuu.fi, kari.t.korhonen@metla.fi, pauline.stenberg@helsinki.fi, miina.rautiainen@helsinki.fi

Received 15 April 2007 **Revised** 13 September 2007 **Accepted** 4 October 2007
Available at http://www.metla.fi/silvafennica/full/sf41/sf414671.pdf

1 Introduction

Interest in different methods and techniques for estimating forest canopy cover has recently increased significantly. The reasons for this trend include, for example, the need to incorporate ecological measures into traditional forestry, fast development of remote sensing techniques, and comparability of international forestry statistics. However, there still exists no fast and reliable method for estimating the canopy cover in boreal forests, even though several studies concerning the issue have been published since the 1940s (e.g. Robinson 1947, Sarvas 1953, Bonnor 1967, Johansson 1985, Bunnell and Vales 1990, Jennings et al. 1999, Rautiainen et al. 2005, Korhonen et al. 2006). Such estimation method would be useful, for instance, in national forest inventories (NFI's), as well as in numerous ecological and remote sensing applications.

Three alternative approaches to the problem of finding a new canopy cover estimation method have been presented: field measurements, statistical modelling and remote sensing. According to our recent study concerning canopy cover estimation with ground-based measurements (Korhonen et al. 2006), there seem to be few satisfactory options if the field measurement is required to yield quickly precise and unbiased results. For example, in the Finnish NFI, the estimation is done with a highly subjective ocular method (Valtakunnan metsien… 2005). Because no significant improvement in current ground measurement techniques can be expected, attention must be turned to the other approaches. This study is a sequel to our previously mentioned article (Korhonen et al. 2006) and focuses on the second option, the use of statistical modelling in estimation of canopy cover. The third alternative, estimation of canopy cover using different remote sensing materials such as satellite, aerial or laser scanned images, is another possible answer to this problem but falls beyond the scope of this study.

The biggest advantage of the statistical modelling approach in canopy cover estimation is that no separate measurements or materials are needed since the outcome is predicted from standard forest characteristics. Stand parameters such as basal area, mean stem diameter at breast height (DBH), and others, are usually measured in forest inventories to obtain an estimate of the growing stock. There are two options for applying this information in the statistical modelling of canopy cover. The first approach is to utilize the strong correlation between stem DBH and crown diameters (e.g. Ilvessalo 1950, Muinonen 1995, Gill et al. 2000, Bechtold 2003) and build models that predict the area covered by each crown. The task is simplified if the locations of the trees in the plot are mapped, in which case the crown overlap can be calculated (if the crowns are assumed to be regularly shaped) and reduced from the total area of individual crowns to give an estimate of canopy cover (Muinonen 1995, Gill et al. 2000, Williams et al. 2003). In other cases particular overlap correction functions that have been created for this purpose must be used (Crookston and Stage 1999, Gill et al. 2000, Shaw 2005).

The second option in the statistical modelling of canopy cover is to ignore the individual tree crown diameters and directly make models with canopy cover, or alternatively canopy closure*, as the dependent variable, and measured stand parameters as independent variables (Kuusipalo 1985, Mitchell and Popovich 1997, Knowles et al. 1999, Korhonen 2006). Of the possible predictor variables, basal area has been shown to correlate best with canopy closure (Kuusipalo 1985, Mitchell and Popovich 1997, Buckley et al. 1999, Knowles et al. 1999) as well as with canopy cover (Korhonen 2006). Kuusipalo (1985) used stand age and stand density as additional predictors, since they had, after basal area, the highest correlation with canopy closure. In Kuusipalo's study, mean height and mean DBH of growing stock did not correlate well enough with canopy closure to be taken into the model by the stepwise regression procedure. In the study by Mitchell and Popovich (1997), stand density also entered the stepwise regression but was rejected due to a high variance inflation factor, so that only basal area remained in the final model. Knowles et al. (1999)

* The difference between canopy cover and canopy closure percentages is that the cover is always measured in vertical direction, whereas the closure is measured with an instrument having an angle of view. For more information, see Jennings et al. (1999) or Korhonen et al. (2006).

used a nonlinear Chapman–Richards model with basal area and crown ratio as predictor variables to estimate the canopy closure of silvopastoral land in New Zealand. Korhonen (2006) presented several models for canopy cover; of the three alternative model shapes that were tested fairly simple models with basal area and mean DBH as predictors were found to yield better results than the more complicated models with stand density and the proportion of deciduous trees as additional variables. This was mainly due to the fact that the overly complicated model was in some way more sensitive to abnormal stand parameters. Thus, in the structurally atypical plots, the predictions were weaker than expected.

Good predictor models should always produce values that are within the application range. In this case, the predictions should be percentages, i.e. the model should be asymptotic in its both ends. In the case of a normal linear regression this is not the case; if one of the predictors has an extreme value, the model may produce illogical results beyond the application range. Using a linear model with quadratic shape (Korhonen 2006) helps to some extent, but the fitted quadratic curve may finally turn in the opposite direction instead of staying asymptotic. The problem can be avoided with a piecewise (Mitchell and Popovich 1997) or a nonlinear (Knowles et al. 1999) model, but these solutions become problematic if several predictors are required to describe the relationships affecting the canopy cover value, as is the case in boreal forests.

For the Finnish conditions, Kuusipalo (1985) has presented a model suitable for estimating canopy closure. However, the model was built for percent canopy closure, which was determined from hemispherical images, i.e. the model is not suitable for estimation of vertically projected canopy cover. Models by Korhonen (2006) were built with unbiased, but rather limited modelling data. In addition, both of these models lack the asymptotic nature required for sound percentage predictions. The objective of this study is to test whether statistical models based on stand parameters and implemented with asymptotic regression are a competitive alternative to the currently used field estimation techniques, especially the ocular method, in the estimation of canopy cover. In this study, the problem of asymptotes is solved with a fairly new statistical modelling technique, the beta regression (Ferrari and Cribari-Neto 2004), which is specially designed for modelling percentage variables, such as canopy cover. As the study area and partially also the data are the same as in the previous article (Korhonen et al. 2006), the performance of the new models was tested using the same control plots that were used in the previous study. This was done by including the comparison plots as a special case to the cross validation test. In addition, the prediction power of the models was tested in separate study plots located in a different geographical area to preliminarily assess whether the models are transferable to other regions.

2 Materials and Methods

2.1 Materials

The modelling data were collected during the summers of 2005 and 2006 at Suonenjoki, central Finland. There were two separate study sites located approximately 20 km from each other: the Hirsikangas site (62º38´N, 27º01´E) and the Saarinen site (62º40´N, 27º29´E). The plots were selected so that the data would include structurally very different forest stands. The study stands had to meet the following criteria: 1) the dominant tree species had to be either Scots pine (*Pinus sylvestris* L.) or Norway spruce (*Picea abies* (L.) Karst), and 2) tree height in the stand had to be at least two meters. Altogether the data consisted of 100 sample plots, 52 of which were pine and 48 of which were spruce dominated. In mixed stands, the species with the largest basal area was considered as the main species. Most of the pine stands were typical dry, poor or rather poor heaths (*Vaccinium*-type or worse; see Cajander 1949), but also a fair number (n = 10) of peatlands were included. Most of the peatlands were ditched and fully stocked, but some (n = 3) were natural, very poor mires with sparsely located, withered pines. The spruce stands were more fertile (*Myrtillus*-type or better), often mixed with deciduous trees, and included only a few slightly peat-covered sites.

More information on the stand structure for

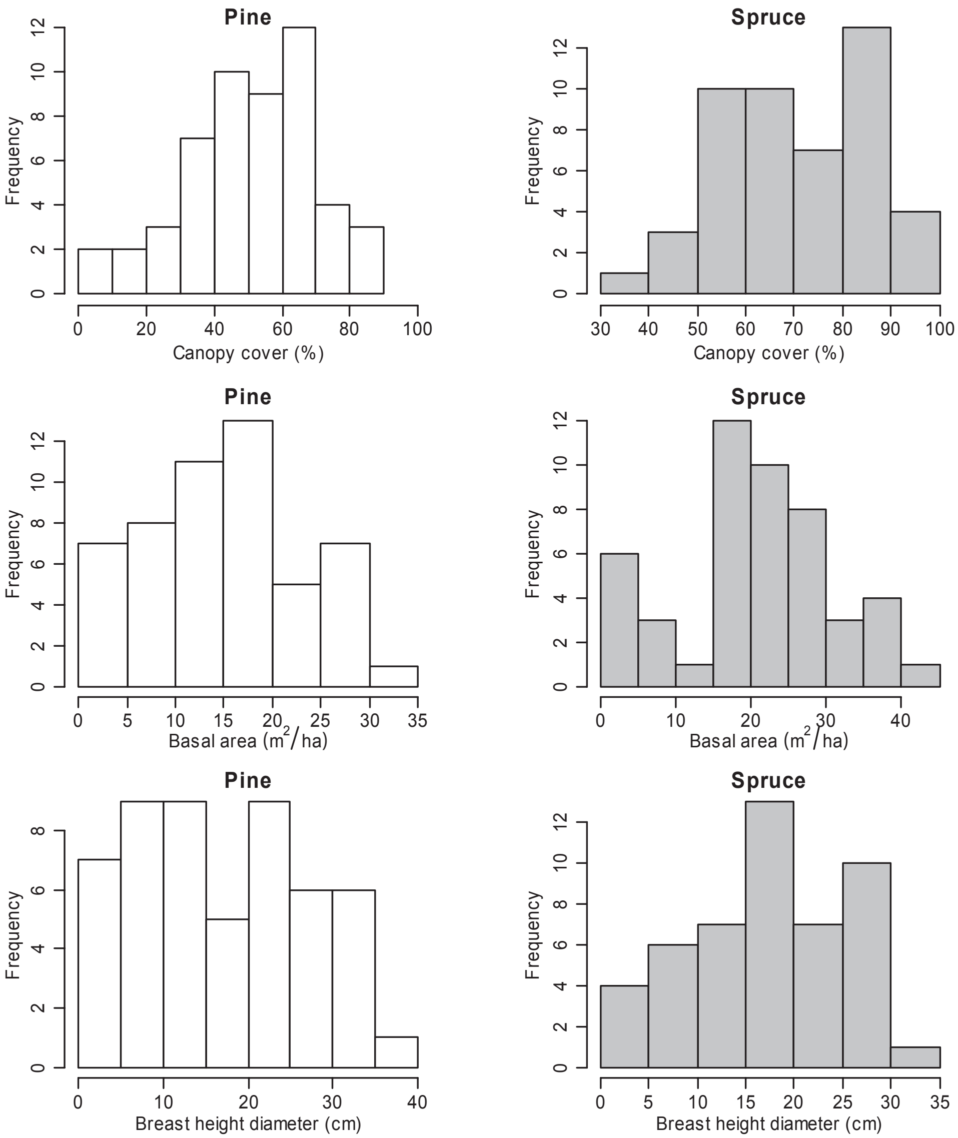

Fig. 1. Distributions of canopy cover, basal area and mean DBH in the modelling data.

both species in the modelling data is presented in Fig. 1 and Table 1. As can be seen, the Suonenjoki data covered a whole range of size classes from sapling to old-growth stands, including two pine seed tree stands. The pine stands had generally lower canopy cover percentages than the spruce stands, which is explained by the lush vegetation of the more fertile spruce stands. The number of young spruce stands in the data is rather small, mainly because such sites were scarce in the study area. Some of the stands had been recently thinned and were striped by forwarder tracks, which increased the heterogeneity inside the plots. Many of the sapling stands and younger spruce forests had a clearly grouped spatial structure, with dense groups of trees and large treeless gaps in between. The mature sites were generally more homogeneous.

Table 1. Summary of the Suonenjoki modelling data.

	Pine (n = 52)				Spruce (n = 48)			
	Mean	Min	Max	Sd	Mean	Min	Max	Sd
Canopy cover (%)	51.0	2.5	84	19.3	69.9	34.2	96.8	15.9
Basal area (m^2/ha)	15.0	0.3	34	8.4	20.9	1.0	44.8	10.3
Stand density (stems/ha)	2700	260	10500	2380	2880	450	15900	2740
Age (years)	65	7	160	40.2	51	11	112	26
Mean diameter (cm)	16.7	1.4	36.6	10.1	17.1	2.6	31.3	8.0
Mean height (m)	13.7	1.8	29.2	7.4	15.0	2.9	26.7	7.1
Deciduous trees (%)	1.8	0	25.9	4.6	8.9	0	47.3	10.6

Table 2. Summary of the Suonenjoki comparison plots.

	Pine (n = 10)				Spruce (n = 9)			
	Mean	Min	Max	Sd	Mean	Min	Max	Sd
Canopy cover (%)	57.6	35.6	83.3	15.9	65.7	34.2	88.4	17.1
Basal area (m^2/ha)	17.7	9.3	29	6.4	19.4	1.0	27.3	8.9
Stand density (stems/ha)	1700	260	4650	1500	3370	580	15900	4900
Age (years)	65	25	124	38	52	12	95	26
Mean diameter (cm)	17.9	6.3	34.6	9.2	17.2	3.3	27.3	8.5
Mean height (m)	15.1	5.8	26.4	6.4	15.7	3.4	25.6	7.7
Deciduous trees (%)	0.4	0	3.6	1.1	13.1	0	34.6	11.2

Table 3. Summary of the Koli test plots.

	Pine (n = 15)				Spruce (n = 6)			
	Mean	Min	Max	Sd	Mean	Min	Max	Sd
Canopy cover (%)	67.9	43.7	84.9	11.5	86.2	73.7	96.8	8.4
Basal area (m^2/ha)	24.9	16.6	34.3	5.5	37.2	32.3	45.4	4.8
Stand density (stems/ha)	1560	540	3500	960	1390	590	2480	640
Age (years)	98	40	161	35	77.0	43	138	36
Mean diameter (cm)	24.4	12.8	33.2	6.6	42.7	26.2	66.5	14.1
Mean height (m)	19.0	9.8	25.3	5.0	27.1	17.5	35.5	6.2
Deciduous trees (%)	11.7	0	41.9	13.1	34.5	13.5	47.2	12.3

In nineteen plots (Table 2), canopy cover was measured with multiple measurement techniques. These plots were used in the comparison of canopy cover estimation techniques in the preceding study (Korhonen et al. 2006). The same plots, described in more detail in the previous study, were used as test sites also in this study: the models found to perform well in the full data set were refitted into data from which the test plots had been omitted, and the test plot cover percentages were predicted with the refitted models for both species. The control plots did not include any mires, because they were all measured during the summer 2005, and the data was not expanded to cover peatland forests until summer 2006.

The transferability of the models was tested on a separate study site in the Koli national park (63º03´N, 29º52´E), located approximately 150 km north-east from the Suonenjoki sites. The Koli study site consisted of 15 pine dominated and 6 spruce dominated plots (Table 3). Many of the Koli plots were located in old growth stands, and no sapling stands were included. Consequently, the average tree size and growing stock were considerably larger than in Suonenjoki, especially in the spruce plots. The pine plots, however, were chosen to be fairly similar to commercially utilized forests outside the national park. There were larger differences in site fertility: several pine plots were considerably more fertile than the Suonenjoki pine plots, having a large percentage of deciduous trees, whereas the others were located in places where the growth potential was limited because of the soil rockiness. All spruce plots were fertile or very fertile, and thus more similar to their Suonenjoki counterparts. Nonetheless, there were differences in stand structure; the Koli spruce plots were unmanaged forests whereas the Suonenjoki spruce plots were more or less managed commercial forests. In mature stands, however, the differences in structure were fairly small.

In all the plots, canopy cover was estimated with the Cajanus tube using the dot count technique (Sarvas 1954, Johansson 1984, Jennings et al. 1999, Rautiainen et al. 2005, Korhonen et al. 2006). In Suonenjoki, the shape of the plot was a 24 m×25 m rectangle, and in Koli a 30 m×30 m square. The plots were covered with a 1 m (distance between measurement points on transect)×2.5 m (distance between parallell transects) dot grid, so that in the Suonenjoki plots 275 and in Koli plots 403 individual measurements were made. In unclear points, the decision whether the point was covered or not was made according to the rules described in Korhonen et al. (2006). In addition to the canopy cover, routine stand inventory parameters were determined for each plot. These included site type (Cajander 1949), basal area (m^2/ha), stand density (stems/ha), stand age, mean DBH (cm), mean height (m) and mean height of the base of the living crown (m). Stand age was determined either from the median tree or taken from stand registers. Stand density included all trees taller than 1.3 m. The mean parameters of the growing stock were measured from the basal area median tree of the dominant species. These variables were used as possible predictors in the canopy cover models.

2.2 Statistical Analysis

The models were built separately for pine and spruce using R statistical software and an additional betareg library, which allowed the construction of beta regression models. The beta regression technique (Ferrari and Cribari-Neto 2004) is an extension to the generalized linear models, described in detail by McCullagh and Nelder (1989). The generalized linear models differ from the standard linear regression in that the expected values μ_i of the random variable Y are replaced by a link function $g(\mu_i) = \eta$, where η is a linear combination of the predictor variables. The purpose of the link function is to stabilize the error variance and transform the fitted values to the desired application range. In addition, the error distribution of the model can be chosen independently, whereas in linear regression the error distribution is always assumed to be normal. In case of a continuous response variable restricted to the standard unit interval [0,1], such as the proportion of the ground covered by the canopy, the errors typically display asymmetry (Ferrari and Cribari-Neto 2004). The two-parameter beta distribution is very flexible and thus capable of describing the distribution of errors in such situations. Thus, typical features of the beta regression technique are the assumption that the model residuals are beta distributed and the use of a link function to transform the predicted values to the application range. The link function g can be chosen from several alternatives. In this study, the logistic link function (Eq. 1) (McCullagh and Nelder 1989, p. 108) was used:

$$\log(\frac{\mu}{1-\mu}) = \eta = \sum_{j=1}^{p} x_j \beta_j \qquad (1)$$

where μ = predicted canopy cover, η = linear combination of predictor variables, x_j = vector of predictor variables, and β_j = vector of model coefficients. The predicted values were obtained as the inverse of the logistic function (Eq. 2):

$$\mu = \frac{\exp(\eta)}{1+\exp(\eta)} \qquad (2)$$

The logistic link function is asymptotic in the range [0,1], i.e. the predicted values are automatically in the desired application range. Compared to alternative probit and cloglog link functions, logistic function reaches its asymptotes more slowly, which is useful in Finnish forests where canopy cover is seldom close to 100% or 0%. Thus, the beta regression allows the use of several predictors in model estimation and eliminates the risk of getting ineligible predictions. The estimation of model parameters is done with the maximum likelihood technique. However, the estimation procedure differs from the generalized linear models since the beta distribution does not belong to the exponential family (Ferrari and Cribari-Neto 2004).

The different models were evaluated utilizing the pseudo R-squared values (R_p^2), standard errors, residual plots and Akaike information criterion (AIC) (Sakamoto et al. 1986) produced by the software. The pseudo R-squared values were calculated as the square of the correlation between $g(y)$ and η (Ferrari and Cribari-Neto 2004). The residuals were calculated by subtracting the predicted cover from the true cover, i.e. negative residuals and statistics indicate overestimation, and the standard errors were calculated as the standard deviation of these residuals. The models were also tested with cross validation (Shao 1993). First, a sample of k plots was taken from the data. For pine, the sample size was ten and for spruce nine. The model was refitted with the remaining plots, and the cover estimates for the sampled plots were predicted with the refitted models. The process was repeated a hundred times, after which the averages and variances of the mean, median, standard deviation, interquartile range, minimum, and maximum of the residuals obtained in the prediction of the sampled plots were studied. The same set of samples was used for all compared models. As a special case, predictions for the plots used in the testing of different ground measurement techniques were obtained similarly*. The results obtained with the best predictor models were then compared to some of the techniques described in the previous article. In addition, the models fitted with the full Suonenjoki data were used to get estimates of canopy cover for the Koli plots. The predictive power of the models in Koli was evaluated using the same statistical coefficients as in the cross validation.

3 Results

3.1 The Relationships Between Canopy Cover and the Independent Variables

Of all the measured variables, basal area showed once again the best correlation with canopy cover in the modelling data (Fig. 2). The dependence was slightly nonlinear; for pine, which is a shade-intolerant species, the asymptote was at approximately 85% cover, and for the shade-tolerant spruce at 100% cover. Mean height also had a strong correlation with canopy cover, but the relationship was strongly nonlinear (Fig. 3). In the beginning, the percent cover increases with mean height, but as the stand density decreases (either because of thinnings or natural mortality), the growth of the individual tree crown diameter cannot compensate for the decrease. Accordingly, at the approximate height of 14 metres, canopy cover starts to decrease with increasing height. Similar relationships with canopy cover appeared also for the mean DBH and age, as they strongly depend on tree height as well as each other. Finally, as stand density remains fairly stable at the end of the forest succession, the growth of the branches will balance the effect of thinning and turn the tree height – canopy cover curve slightly upwards. According to Fig. 3, this seems to happen at the height of 20–30 metres.

The relationship between stand density and percent cover (Fig. 4) resembles the relationship shown in Fig. 3: at first, the cover increases rapidly with the growing stand density, but when the density approaches 4000 stems per hectare, canopy cover starts to decrease. However, an increase in canopy cover seems to take place again soon after 8000 stems per hectare. This seemingly paradoxical phenomenon occurs because the dense stands are young sapling stands, where the crown diameter is very small. However, if there are approximately 10 000 saplings per hectare, the number of crowns will eventually counterbalance the decrease in crown size.

* In the preceding article the plot shape was a circle with radius of 12.5 m, i.e. only the area inside this radius were taken into the account. Here the results were calculated for the full 24 × 25 meters rectangle.

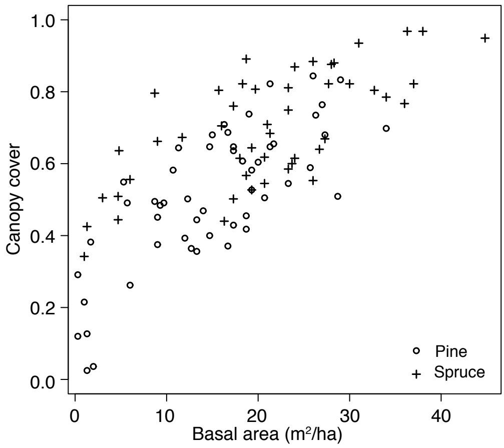

Fig. 2. The relationship between basal area and canopy cover.

3.2 Canopy Cover Models

Because of the nonlinear relationships between the independent variables and canopy cover described in the previous sections, the regression models were based on cubic forms of the independent variables, i.e. the cubic function (3):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \qquad (3)$$

The cubic form will produce a curve similar to the relationships observed in Figs. 3–4. For mean height, mean DBH and age, also quadratic curves might have been used, but since the shape of the quadratic curve is a parabola, extremely large values would have caused irrational results. This does not happen with a cubic model shape. However, also the cubic regression functions must be used with caution, because they have a tendency to produce extreme predictions when used to extrapolate. In this case, extreme values are eliminated by the logistic link function. However, if the second and third order coefficients were not statistically significant and increased the model AIC (low AIC indicates a good balance between model fit and the number of parameters), the linear relation was used instead.

The best of the tested models were based on the cubic form of basal area, which had the highest correlation with the percent cover for both pine ($R^2 = 0.63$) and spruce ($R^2 = 0.50$). The other variables were used as additional predictors. Tree height performed well as an additional predictor, whereas mean DBH was incapable of entering the models when height was present. Including stand density and age also improved the model fit for both species, and different dummy predictors and the percentage of deciduous also improved some of the model candidates. After comparing different model shapes based on their fit in the modelling data and their predictive power in the cross-validation test and in the separate test plots, two alternative model shapes were found

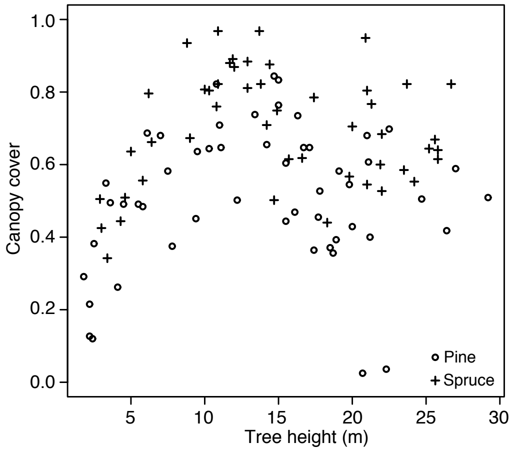

Fig. 3. The relationship between tree height and canopy cover. The outliers at the bottom right corner are the two pine seed tree stands.

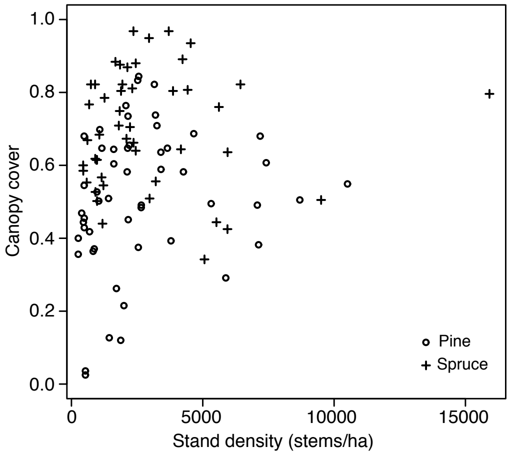

Fig. 4. The relationship between stand density and canopy cover.

for spruce and three for pine. The model coefficients are presented in Eqs. 4–8 and results of the performance tests in Tables 4–6.

Spruce, Model 1 (*standard model*, $R_p^2 = 0.871$, s.e. = 0.059, AIC = –127.0):

$$\eta = -0.48019 + 0.32488G - 0.0093056G^2 + 0.00011171G^3 - 0.15779H - 0.002459H^2 + 0.00015333H^3 + 1.5203HW \quad (4)$$

Spruce, Model 2 (*alternative model*, $R_p^2 = 0.901$, s.e. = 0.053, AIC = –137.1):

$$\eta = -2.4709 + 0.12843G - 0.003597G^2 + 6.4329\times10^{-5}G^3 + 0.13252T - 0.0027546T^2 + 1.5183\times10^{-5}T^3 + 0.00010652N \quad (5)$$

Pine, Model 1 (*standard model*, $R_p^2 = 0.914$, s.e. = 0.063, AIC = –131.9):

$$\eta = -1.1194 + 0.23663G - 0.0038168G^2 + 9.2475\times10^{-6}G^3 - 0.095561H + 0.16055F - 0.30635P \quad (6)$$

Pine, Model 2 (*alternative model*, $R_p^2 = 0.945$, s.e. = 0.052, AIC = –146.6):

$$\eta = -1.2616 + 0.21514G - 0.0057727G^2 + 7.4293\times10^{-5}G^3 - 0.0088222T - 1.8387\times10^{-4}T^2 + 1.2655\times10^{-6}T^3 + 4.3071\times10^{-5}N - 0.87605S \quad (7)$$

Pine, Model 3 (*comprehensive model*, $R_p^2 = 0.939$, s.e. = 0.043, AIC = –156.1):

$$\eta = -1.5392 + 0.19772G - 0.0039876G^2 + 4.3117\times10^{-5}G^3 - 0.032476T + 2.486\times10^{-4}T^2 - 5.3266\times10^{-7}T^3 + 0.16375H - 0.011887H^2 + 1.9104\times10^{-4}H^3 + 6.3301\times10^{-5}N \quad (8)$$

Abbreviations: G = basal area (m^2/ha), H = mean height (m), HW = percentage of hardwoods (in hundredths), T = age (years), N = number of stems (ha^{-1}), F = dummy for medium fertile (*Myrtillus*-type) or more fertile site type, P = dummy for poor (*Calluna*-type) or poorer site type, S = dummy for seed tree stands (stands having only some seed trees standing for natural regeneration). The model coefficients predict the parameter η in Eq. 1, i.e. the correct predicted values of canopy cover (μ) are obtained from the η values by inverting the logistic transformation (Eq. 2).

The "standard" model shape that included basal area, height, and percentage of the deciduous trees or site fertility as predictors performed fairly well for both spruce and pine (Fig. 5). This model shape never fell very far behind the other model shapes in any of the comparison tests (Tables 4–6). Moreover, the models are fairly simple, and predictor variables are easy to obtain in the field. The standard spruce model presented in Equation 4 yielded satisfactory results in all comparison tests. The standard error of 5.9% in the fitting data is fairly good; however, in the cross-validation tests the standard error increased to 6.8% and in the Koli data to 7.4%. The model predictions were reasonably reliable: both in the cross validation and in Koli data, the largest prediction errors were approximately 11%. The pine standard model (Eq. 6) differed from the spruce model in that the linear form of tree height was used instead of the cubic form, and dummy variables describing site fertility were used instead of the percentage of deciduous trees. In the modelling data, this model performed nearly as well as the standard spruce model, achieving standard error of 6.3%. In the cross validation the error rose to 7.0% and in Koli to 8.8%.

The "alternative" model shape differed from the standard shape in that height was replaced by cubic form of age and stand density was used as a third predictor. Moreover, a dummy variable representing seed tree stands was included in the alternative pine model (Eq. 7). In the standard model, a separate predictor for seed tree stands did not improve model performance, but in the alternative model the results improved significantly. Thus, the alternative pine model had slightly better standard errors than the standard model: 5.2%, 6.5%, and 8.1% in the modelling data, cross-validation and Koli data, respectively. However, it was seriously biased in the Koli plots, as the mean underestimation was as large as 6.8%, whereas with the standard model the mean underestimation was only 2.7%. The alternative spruce model (Eq. 5) had standard errors of 5.3%, 6.0% and 11.1%, respectively. For the spruce alternative model, the standard errors in fitting data and cross-validation were approximately one percent smaller than for the standard model, but in Koli

Table 4. Results of the cross-validation tests in the modelling data. The mean and sd values indicate the mean and standard deviation of each variable in one hundred simulations.

Model	Sd		Mean		Min		Max	
	mean	sd	mean	sd	mean	sd	mean	sd
Spruce, standard	0.068	0.019	–0.002	0.027	–0.111	0.019	0.107	0.052
Spruce, alternative	0.060	0.017	0.004	0.025	–0.098	0.041	0.090	0.052
Pine, standard	0.070	0.015	0.006	0.026	–0.105	0.041	0.117	0.032
Pine, alternative	0.065	0.012	–0.002	0.022	–0.101	0.027	0.101	0.032
Pine, comprehensive	0.055	0.013	0.001	0.022	–0.083	0.028	0.083	0.031

Table 5. Results of the models in the Suonenjoki comparison plots.

Model	Sd	Interq. range	Mean	Median	Min	Max
Spruce, standard	0.052	0.059	–0.028	–0.028	–0.130	0.040
Spruce, alternative	0.052	0.044	–0.020	–0.018	–0.090	0.069
Pine, standard	0.089	0.115	0.010	0.040	–0.129	0.137
Pine, alternative	0.070	0.083	0.021	0.017	–0.106	0.119
Pine, comprehensive	0.058	0.085	0.024	0.034	–0.071	0.085

Table 6. Results of the models in the Koli test plots.

Model	Sd	Interq. range	Mean	Median	Min	Max
Spruce, standard	0.074	0.071	–0.030	–0.048	–0.105	0.101
Spruce, alternative	0.111	0.092	–0.062	–0.047	–0.260	0.057
Pine, standard	0.088	0.098	0.027	0.008	–0.097	0.191
Pine, alternative	0.081	0.093	0.068	0.063	–0.039	0.231
Pine, comprehesive	0.099	0.094	0.052	0.028	–0.087	0.258

also the spruce model was seriously biased, as the mean overestimation was as high as 6.2%. Another disadvantage of the alternative models is that age and stand density are more difficult to estimate reliably than tree height and site fertility. However, in the modelling data, the alternative models produced slightly better estimates than the standard models for both species (Fig. 5).

The model shapes that achieved the lowest standard error in the modelling data were called "comprehensive" models. The comprehensive models included cubic forms of the basal area, height, and age, as well as stand density as a linear predictor. For spruce, the model shape was at best only equally good as the simpler models in the comparison tests, so it is not presented here. However, the pine comprehensive model had the lowest standard errors both in the modelling data (4.3%) and cross-validation test (5.5%). This model included eleven separate terms, but it also had very low AIC (–156.1), indicating that all the predictors contain useful information. In the Koli data, however, the standard error was larger than for the other two models (9.9%), and also the bias (5.2%) was considerably large. In addition, some of the residuals were unacceptably high (up to 25.8%). This result is not surprising, since complex models often describe the original modelling data very well, but predicting new, unknown observations with such models is very risky.

The results for the comparison plots (Table 5) were only a special case of the cross-validation test, so they were not emphasized in the model selection. The alternative spruce model and the comprehensive pine model achieved the best results in the comparison plots. The alternative spruce model had a slightly higher standard error than the standard model but was less biased, whereas the pine comprehensive model

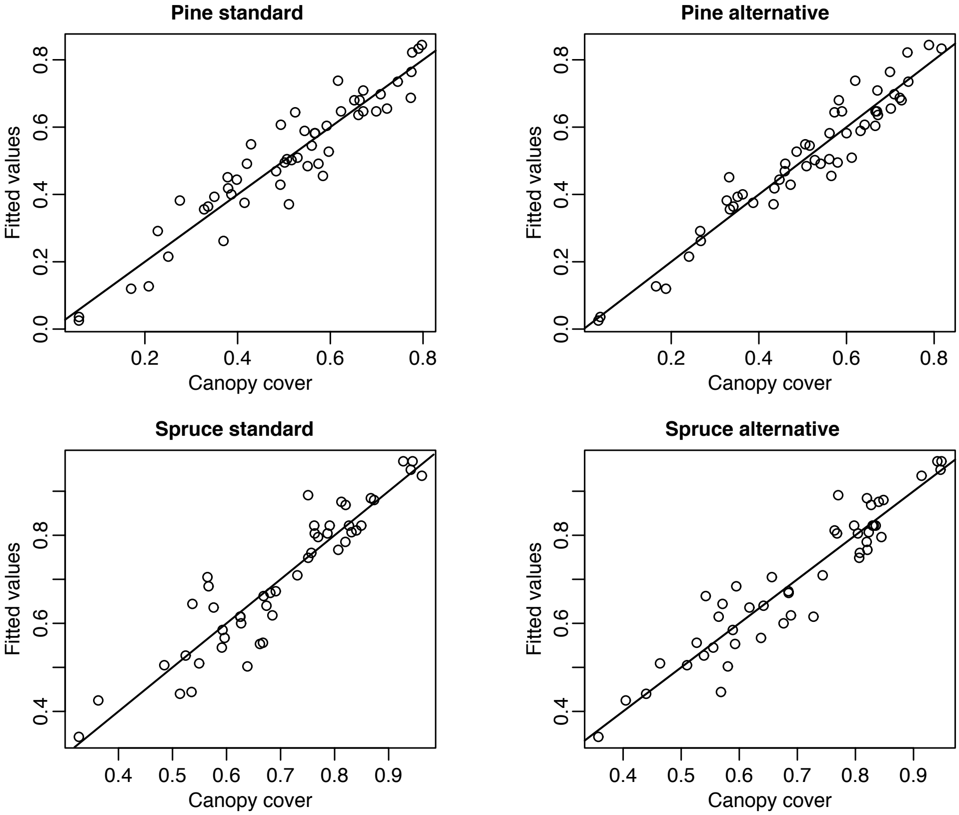

Fig. 5. Fitted values of the standard and alternative models plotted against the true canopy cover in pine and spruce data.

had clearly the lowest standard error and smallest extreme residuals. The predictions produced by these two models were combined and used as a single category called the "best model". The new "best model" technique had the following error statistics: mean = 0.28%, median = –1.0%, standard deviation = 5.8%, quartile range = 11.9%, minimum = –9.0%, and maximum 8.5%.

4 Discussion

4.1 Analysis of the Models

The most important feature of any good canopy cover estimation technique is reliability: the model predictions should never go badly wrong. In practice, finding a method that produces errors no larger than 10% and requires less than five minutes of field work time is set as a goal of our ongoing canopy cover research. As it appears that the separate canopy cover field measurements cannot achieve this aim (Korhonen et al. 2006), statistical modelling could be a solution in situations where reliable information on the standard forest characteristics is available. The results of the models tested in this study clearly indicate that canopy cover can be successfully described through regression structure, where easier-to-measure forest characteristics are used as predictors.

In all model construction it is important that the models are fairly simple and that they have a clear interpretation. Of the models introduced in this study, the standard spruce and pine models are the easiest to interpret: basal area describes the general amount of the growing stock in the plot, height represents the size of the trees in the stand and gives an idea of the expected mortality in such forests, and site fertility dummies or the proportion of the deciduous trees describe the relative length of the branches when compared to stands with otherwise similar composition. In the model equations, basal area and the percent of deciduous trees had positive coefficients, i.e. an increase in these variables results in an increase in the predicted cover. Tree height had a negative coefficient in the pine standard model, whereas in the spruce model the height curve pointed downward for most of the application range. This indicates that when trees grow in size the stand density will decrease so much that growth in tree size cannot compensate for the decrease. Thus, a seedling stand with small basal area and height will have high stand density, and thus, a larger canopy cover than a seed tree stand with equally small basal area, but full-grown trees. In the spruce model, the height curve turned slowly upwards again after 25 meters. This happened at the end of the succession when the relationship between mortality and tree size had come to an end. In the pine standard model, the cubic form of the height was not statistically significant, so in the pine model this did not occur. The "fertile site" dummy variable increased the cover slightly in places where site quality allowed the trees to develop wider crowns than usually. The "poor site" dummy had an opposite effect.

The alternative model shape differed from the standard model in that the cubic form of age replaced height, and stand density was included instead of fertility dummies or the proportion of deciduous trees. The age curve behaved similarly to the model height curve, turning clearly upwards after an age of 120 (pine) or 100 (spruce) years. After the age of 150 years, the effect of age on canopy cover increased very rapidly, suggesting that extrapolating with this model may be very risky. Instead of the dummy variables, stand density was used to give additional information on stand structure. As stand density included all trees reaching breast height, the density value used in the analysis was often significantly higher than the density of the dominant tree class. Thus, stand density was generally higher in fertile sites with a large number of understorey trees, and can be interpreted as an additional variable describing the multi-storey stand structure. The alternative model shape was very similar to the model proposed by Kuusipalo (1985), which indicates that this shape may be applicable also to other data sets. The model is also fairly simple with only three predictor variables and the seed tree dummy for pine stands.

The comprehensive pine model included the same predictors as the alternative model and, in addition, also the cubic form of the height. Stand age, however, had a high correlation with tree height (R = 0.84), which made including both in the model simultaneously problematic. Perhaps age can be seen as another factor affecting the stand structure through mortality and growth potential, whereas height represents tree size more accurately. Still, it is difficult to comprehensively interpret the relationships in such a complicated model. The model had four predictor variables and eleven separate terms, which can in some cases be considered too many for a good predictor model.

Further improvements of the models might be possible by including description of two tree layers and a grouping index, which would describe the degree of crown overlapping. Such a variable might be, for example, Fisher's index (Fisher et al. 1922), or a simple categorical variable which could be defined by eyesight and would range from very clumped to very regular structure. The fact that the models for spruce did not achieve as good a fit as the best pine models may be due to the more heterogeneous structure of the spruce stands. For example, thickets, windfalls and haulage tracks were more common in the spruce stands. Another factor, which obviously affects the results but was not too well present in these models, is the effect of thinning. The thinning effect is visible through decreasing basal area and increasing mean DBH, and after the thinning the branches of the remaining trees are shorter than expected, and take some time to recover. Also, some of the stands could be observed as outliers in the residual plots: especially the two pine seed tree stands and some of the sapling stands were problematic. Separate dummy variables can be used to eliminate these cases, as was done with the alternative pine model, but generally such fine-tuning should be avoided whenever possible.

However, in general and in spite of a fairly small base data, the models were capable of producing satisfactory predictions in structurally very different forest, ranging from seed tree stands and very sparsely stocked mires to dense young forests and old-growth forests. On the other hand, the modelling data were scarce for young spruce stands, mixed forests, and peatland forests, not to mention that deciduous stands were excluded altogether. If the modelling data sets were larger, it might be sensible to create separate models for the most difficult structural forest types, e.g. the seedling and sapling stands. In addition, the forests in Suonenjoki were managed; in naturally developed forests the effect of tree height or age on the canopy cover may be very different. As tree mortality rate is significantly slower in natural forests, it seems likely that the negative relationship between tree size and canopy cover will decrease or disappear.

4.2 Cross Validation and Koli Results

The errors of the models increased surprisingly little in the cross-validation tests: the average increase in the model standard error was only one percent. For some individual cases, the errors were certainly larger, but in average the models seem to yield reliable estimates in unknown plots in the same area. When the models were transferred to Koli, the standard error increased by 1.5–5.8 %-units from the fitted model. The increase was largest for the alternative spruce model and smallest for standard spruce model. In addition, the bias of the models increased significantly in the Koli plots, especially for pine. There are two main reasons for this phenomenon. First, some of the Koli pine plots were considerably more fertile than any of the Suonenjoki plots. They had a dense understorey and plenty of deciduous trees, which have wider crowns and thus create more cover than expected. If the Suonenjoki pine plots had contained more deciduous trees, it is likely that the percentage of deciduous trees would have been included in the predictors, as was the case in the spruce model. Secondly, all the tested models that included age were more biased in Koli than the standard model, which indicates that there may by systematic differences in the age estimates in the two areas. In Koli, the age estimates were taken from the existing stand registers, so it is reasonable to assume that in Suonenjoki the age values were more accurate, even though the Koli register ages were updated to the inventory date. This is probably the main reason in the Koli data for relatively poor performance of the models which include age. We can conclude that good, transferable prediction models should be fairly simple, and it should be easy to determine the predictor variables accurately. Both standard models fulfil this condition, but also the alternative model might be used if tree age and stand density can be determined reliably.

4.3 Comparison to Ground-based Measurement Techniques

Compared to the ground measurement techniques (Korhonen et al. 2006), the combined "best model" performed especially well. The model results were unbiased, and of the tested ground measurement techniques, only the time-consuming LIS-method (line intersect sampling) and a 102 points Cajanus tube grid had lower standard and minimum/maximum errors. The best of the three ocular observers and subjective ten points densiometer sample produced slightly less accurate results than the regression method. This means that the models cannot compete with time-intensive Cajanus tube measurements, which require a lot of field time (approximately an hour, depending on the plot), but are locally more reliable than any of the quick measurement techniques (densiometer, digital photographs) or ocular estimates. A carefully trained, experienced ocular observer might be able to achieve as good or even better results, but the skill is difficult to obtain. All in all, the regression models seem to be the only possible alternative for the ocular estimates, if the canopy cover estimation is included in a larger forest inventory, where the standard plot characteristics are measured carefully but there is little time for additional measurements. The predictions may be somewhat imprecise, but at least they should be unbiased and suitable for, at least, rough classification purposes.

5 Conclusion

According to the results of this study, the beta regression technique is well suited for canopy cover modelling based on standard forest characteristics. The models tested in this study were capable of producing satisfactory results locally, even though the modelling data included structurally very different forests. As the standard errors approached 5%, improving the models further may be difficult, since the random variation cannot be totally eliminated. Moreover, there will always be special situations were models fail to predict the cover with acceptable accuracy. In such situations, a system combining ocular estimates and regression estimated predictions might be safer. For instance, the final prediction could be calculated as a weighted average of the model prediction and field estimated value according to the estimated errors. Another solution would be that prediction of canopy cover would be done in the field immediately after the growing stock measurements, and this prediction would then be subjectively corrected. As an alternative to regression-based modelling, a nonparametric estimation technique, such as the most similar neighbour method (Moeur and Stage 1995), might also be tested. In addition, utilization of crown diameter based models with and without stem mapped data should be studied.

In conclusion, statistical modelling may be the most cost-effective and feasible solution for obtaining canopy cover estimates for large areas (given that standard forest characteristics are available in a data base), especially if some method for eliminating the most inaccurate predictions can be developed. However, before statistical modelling can be applied at a large scale, it is necessary to obtain geographically representative, accurately measured canopy cover data sets, which cover the whole range of different forests. The data used in this study were rather limited, and the models presented here should not be applied elsewhere as such. These models are also not suited for certain special situations, such as naturally developed or very sparse forests. The expansion of the regression method to cover large areas and special situations will require a substantial amount of field work and further testing of different modelling techniques, but it is still the most likely solution to the near-future needs of canopy cover information. In the long run, remote sensing of canopy cover may also become a frequently used method along with the improved availability and reduced cost of accurate high resolution remote sensing materials and the fast development of physically based forest reflectance models. Nevertheless, for the development and calibration of remote sensing techniques, reliable ground truth measurements and models of canopy cover must be available.

Acknowledgements

The authors wish to thank following persons: Pekka Voipio for his valuable help during the Suonenjoki field campaigns; Jussi Peuhkurinen, Maria Villikka, Mikko Vehmas, and Raita Säynäjoki for providing the Koli field data; Hilkka Stotesbury for improving the English; and the referees for feedback concerning the content. Different phases of the study have been supported by the Graduate School in Forest Sciences, University of Helsinki research funds, Academy of Finland, Metsämiesten säätiö, and Marjatta ja Eino Kollin säätiö.

References

Bechtold, W.A. 2003. Crown-diameter prediction models for 87 species of stand-grown trees in the Eastern United States. Southern Journal of Applied Forestry 27(4): 269–278.

Bonnor, G.M. 1967. Estimation of ground canopy density from ground measurements. Journal of Forestry 65(8): 544–547.

Buckley, D.S., Isebrands, J.G. & Sharik, T.L. 1999. Practical field methods of estimating canopy cover, PAR, and LAI in Michigan oak and pine stands. Northern Journal of Applied Forestry 16(1): 25–32.

Bunnell, F.L. & Vales, D.J. 1990. Comparison of methods for estimating forest overstory cover: differences among techniques. Canadian Journal of Forest Research 20: 101–107.

Cajander, A.K. 1949. Forest types and their significance. Acta Forestalia Fennica 56. 71 p.

Crookston, N.L. & Stage, A.R. 1999. Percent canopy cover and stand structure statistics from the Forest Vegetation Simulator. Gen. Tech. Rep. RMRS-GTR-24. Ogden, UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 11 p. Available at: http://www.treesearch.fs.fed.us/pubs/6261.

Ferrari, S. & Cribari-Neto, F. 2004. Beta regression for modelling rates and proportions. Journal of Applied Statistics 31(7): 799–815.

Fisher, R.A., Thornton, H.G. & MacKenzie, W.A. 1922. The accuracy of the planting method of estimating the density of bacterial populations, with particular reference to the use of Thornton's agar medium with soil samples. Annals of Applied Botany 9: 325–359.

Gill, S.J., Biging, G.S. & Murphy, E.C. 2000. Modeling conifer tree crown radius and estimating canopy cover. Forest Ecology and Management 126: 405–416.

Ilvessalo, Y. 1950. On the correlation between the crown diameter and the stem of the trees. Communicationes Instituti Forestalis Fenniae 38(2). 32 p.

Jennings, S.B., Brown, N.D., & Sheil, D. 1999. Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. Forestry 72(1): 59–74.

Johansson, T. 1985. Estimating canopy density by the vertical tube method. Forest Ecology and Management 11: 139–144.

Knowles, R.L., Horvath, G.C., Carter, M.A. & Hawke, M.F. 1999. Developing a canopy closure model to predict overstorey/understorey relationships in Pinus radiata silvopastoral systems. Agroforestry systems 43: 109–119.

Korhonen, L. 2006. Havumetsän latvuspeiton mittaaminen ja ennustaminen puustotunnuksista. M. Sc. thesis. University of Joensuu, Faculty of Forest Sciences. 71 p. Available at: http://cc.joensuu.fi/~lakorhon/latvuspeitto.pdf. (In Finnish).

— , Korhonen, K.T., Rautiainen, M & Stenberg, P. 2006. Estimation of forest canopy cover: a comparison of field measurement techniques. Silva Fennica 40(4): 577–588.

Kuusipalo, J. 1985. On the use of tree stand parameters in estimating light conditions below the canopy. Silva Fennica 19(2): 185–196.

McCullagh, P. & Nelder, J.A. 1989. Generalized linear models, 2nd edn. Monographs on statistics and probability 37. Chapman and Hall, London. 511 p.

Moeur, M. & Stage, A.R. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. Forest Science 41(2): 337–359.

Mitchell, J.E. & Popovich, S.J. 1997. Effectiveness of basal area for estimating canopy cover of ponderosa pine. Forest Ecology and Management 95(1): 45–51.

Muinonen, E. 1995. Metsikön heijastussuhteen ennustaminen geometrisella latvustomallilla. Licenciate of Science thesis. University of Joensuu, Faculty of Forest Sciences. 48 p. (In Finnish).

Robinson, M.W. 1947. An instrument to measure forest crown cover. The Forestry Chronicle 23: 222–225.

Rautiainen, M., Stenberg, P. & Nilson, T. 2005. Estimating canopy cover in Scots pine stands. Silva Fennica 39(1): 137–142.

Sakamoto, Y., Ishiguro, M. & Kitagawa, G. 1986. Akaike information criterion statistics. KTK Scientific Publishers, Tokyo. 290 p.

Sarvas, R. 1953. Measurement of the crown closure of the stand. Communicationes Instituti Forestales Fenniae 41(6). 13 p.

Shao, J. 1993. Linear model selection by cross-validation. Journal of the American Statistical Association 88: 486–494.

Shaw, J.D. 2005. Models for estimation and simulation of crown and canopy cover. In: Proceedings of the fifth annual forest inventory and analysis symposium; 2003 November 18–20; New Orleans, LA. Gen. Tech. Rep. WO-69. Washington, DC: U.S. Department of Agriculture Forest Service. 222 p. Available at: http://www.treesearch.fs.fed.us/pubs/14293.

Valtakunnan metsien 10. inventointi (VMI10) 2005. Maastotyön ohjeet 2005. Koko Suomi. [10th National Forest Inventory (NFI10) 2005. Field Instructions. Whole Finland.] The Finnish Forest Research Institute. 181 p. (In Finnish).

Williams, M.S., Patterson, P.L. & Mowrer, H.T. 2003. Comparison of ground sampling methods for estimating canopy cover. Forest Science 49(2): 235–246.

Total of 27 references