Eur J Forest Res (2013) 132:983–997
DOI 10.1007/s10342-013-0715-0

ORIGINAL PAPER

A flexible stem taper and volume prediction method based on mixed-effects B-spline regression

Edgar Kublin · Johannes Breidenbach · Gerald Kändler

Received: 25 March 2013 / Revised: 24 June 2013 / Accepted: 26 July 2013 / Published online: 31 August 2013

Abstract Modelling stem taper and volume is crucial in many forest management and planning systems. Taper models are used for diameter prediction at any location along the stem of a sample tree. Furthermore, taper models are flexible means to provide information on the stem volume and assortment structure of a forest stand or other management units. Usually, taper functions are mean functions of multiple linear or nonlinear regression models with diameter at breast height and tree height as predictor variables. In large-scale inventories, an upper diameter is often considered as an additional predictor variable to improve the reliability of taper and volume predictions. Most studies on stem taper focus on accurately modelling the mean function; the error structure of the regression model is neglected or treated as secondary. We present a semi-parametric linear mixed model where the population mean diameter at an arbitrary stem location is a smooth function of relative height. Observed tree-individual diameter deviations from the population mean are assumed to be realizations of a smooth Gaussian process with the covariance depending on the sampled diameter locations. In addition to the smooth random deviation from the population average, we consider independent zero mean residual errors in order to describe the deviations of the observed diameter measurements from the tree-individual smooth stem taper. The smooth model components are approximated by cubic spline functions with a B-spline basis and a small number of knots. The B-spline coefficients of the population mean function are treated as fixed effects, whereas coefficients of the smooth tree-individual deviation are modelled as random effects with zero mean and a symmetric positive definite covariance matrix. The taper of a tree is predicted using an arbitrary number of diameter and corresponding height measurements at arbitrary positions along the stem to calibrate the tree-individual random deviation from the population mean estimated by the fixed effects. This allows a flexible application of the method in practice. Volume predictions are calculated as the integral over cross-sectional areas estimated from the calibrated taper curve. Approximate estimators for the mean squared errors of volume estimates are provided. If the tree height is estimated or measured with error, we use the "law of total expectation and variance" to derive approximate diameter and volume predictions with associated confidence and prediction intervals. All methods presented in this study are implemented in the R-package *TapeR*.

Keywords Stem taper · Timber volume · Forest inventory · Mixed-effects models · Splines

Communicated by Aaron Weiskittel.

E. Kublin · G. Kändler
Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg, Freiburg, Germany

J. Breidenbach (✉)
Norwegian Forest and Landscape Institute, Ås, Norway
e-mail: johannes.breidenbach@skogoglandskap.no

Introduction

Taper functions are useful tools in modern forest inventory and management systems giving information on diameter at any point of a stem. Because of their flexibility, taper functions are often used in forest inventories to estimate timber volume and assortment structures. Attempts of a mathematical description of stem form can be traced back over more than 100 years. The first published taper

function is attributed to Höjer (1903), who used a logarithmic function for diameter estimation. Subsequently, more or less complicated models, ranging from simple polynomials, parametric linear and nonlinear and multivariate regression models, were published. For a detailed list with literature references on this subject and a comparison of more than 15 different taper equations we refer to Rojo et al. (2005). The geometrical shape of a stem can in general be described by a neiloid in the lower, a paraboloid in the middle, and a cone in the upper part of the bole. Simple parametric models are often too stiff to deal with heavily varying curvature along the trunk. In order to circumvent these difficulties, the stem taper is often approximated by splines. Splines are segmented polynomials which are continuously linked at joining points, the so-called knots. Taper models based on spline functions were introduced by Max and Burkhart (1976) and were used for example by Demaerschalk and Kozak (1974), Hradetzky (1980), and Fang and Bailey (2001), among others. Later, variable-exponent taper functions were introduced (e.g., Kozak 1988, 2004; Newnham 1988, 1992). Variable-exponent taper functions consist of an exponential mean function whose exponent is continuously changing along the trunk to describe the neiloid, paraboloid, and several intermediate geometrical shapes from stump to the top.

Most models use diameter at breast height (dbh) and total height of a tree to predict stem diameter at an arbitrary location. In scientific studies or for large-scale inventories with a wide range of stem forms to be modelled, often a second diameter from the upper part of the trunk is included as a predictor variable in the regression model. For example, in the Austrian National Forest Inventory (NFI), an additional diameter at 30 % of the tree height was measured in order to reduce the between-tree form variability in a given dbh–height class (Pollanschütz 1965). In the German NFI, upper diameters are recorded at varying measurement heights supplementary to the dbh and height at a subsample of trees. The taper model is based on a two-dimensional tensor product spline which can be interpreted as a varying coefficient model with diameters at 5 %, and 30 % of height and tree height as predictors. The coefficients are spline functions of the relative stem height with common knots at 0, 30, 50, 70, and 90 % of total height (e.g., Kublin 2003; Kublin et al. 2008).

Data used for fitting taper models frequently are diameters obtained by stem analysis, i.e., diameters are recorded in regular or irregular distances between measuring positions. As a result, multiple measurements from individual sample trees are correlated, meaning that the individual stem profiles systematically lie above or below the mean stem profile averaged over trees of a given dimension like a dbh–height class. This is especially true when only dbh and height are used as predictors in the taper model. These correlations have been ignored in taper modelling for a long time, and independent, identically distributed residuals were assumed instead. As a result, statistical tests and inferences based on independent residuals are not reliable, because estimates of the residual covariance matrix are biased. Nonetheless, irrespective of disregarded error correlations, the parameter estimates remain unbiased.

There are mainly two ways how the problem of correlated within-tree residuals was addressed in other studies. In one approach, mixed-effect modelling techniques are used. Fixed-effect parameters are used to model the population mean common to all units of the sampled population. Random effects implicitly define the within-tree error correlation structure and are used to localize the tree-specific deviation from the population mean. In addition, the correlation following from the random-effect structure can be supplemented by heteroskedastic or autocorrelated residual error components. However, there is a trade-off between the two correlation components, and a too complex correlation structure may give rise to numerical problems during the model fit. Zhang et al. (2002) investigated fertilization and treatment effects on stem form and volume with a nonlinear mixed model where within- and between-tree correlations are implicitly modelled by random effects. Trincado and Burkhart (2006) used the segmented taper equation of Max and Burkhart (1976) in combination with nonlinear mixed effects in order to account for within- and between-tree variations. Within-tree variation was accounted for by a first-order autoregressive error structure and a variance function. Between-tree variation is modelled by adding random effects on some model parameters. The successful application of nonlinear mixed-effects techniques in combination with autocorrelated errors structure has also been demonstrated in studies by Fang and Bailey (2001) and Garber and Maguire (2003) for instance. In most studies using mixed model techniques, the focus is on diameter and volume prediction for a sample tree. The error structure and random-effect parameters are used to calibrate individual stem profile curves when diameter measurements are available for a sample tree. Prediction errors for diameter and volume, even though theoretically available, are usually not provided.

Following another approach, Lappi (2006) introduced a multivariate nonparametric regression model for the diameter prediction on a discrete grid of stem locations. Diameter predictions for locations not included in the discrete grid are calculated through interpolation. For all possible pairs of grid locations, a correlation and variance function is derived from sectional data with nonparametric smoothing methods. As in the case of diameter prediction,

correlation and variance estimates for locations not included in the discrete grid are interpolated. Assuming normally distributed random errors and the variance covariance structure to be known, confidence intervals (CIs) for diameter and volume predictions are calculated. Moreover, standard linear prediction theory is used to calibrate diameter and volume predictions when additional upper stem diameter measurements are available for a sample tree. The model of Lappi (2006) provides diameter predictions also when only dbh is measured on the sample tree. With respect to the mean function, this approach is comparable to the multivariate linear regression model presented by Sloboda et al. (1998) but the approach of Lappi (2006) is more flexible in the underlying error covariance structure. Kublin et al. (2008) proposed a functional regression model for diameter prediction on a continuous rather than on a discrete measurement grid which was used by Lappi (2006). The population mean function is a 2-dimensional tensor product spline with dbh and tree height as predictor variables. The within-tree covariance is assumed to be a smooth function depending on diameter and tree locations. Within-tree correlations were estimated by double smoothing the cross-products of standardized taper curve residuals with a nonparametric smoothing method. Kublin et al. (2008) applied principal component techniques in order to guarantee the necessary positive definiteness of the estimated covariance functions. Calibrated diameter predictions together with pointwise and simultaneous CIs are calculated within the methodological framework of linear prediction theory and functional principal component analysis. The numerical effort for these calculations is demanding.

In extension to the approach proposed by Kublin et al. (2008), we present a semiparametric mixed-effects taper model, which is a compromise between the usual parametric nonlinear mixed model and a pure nonparametric model. We assume that the population mean taper curve as well as the tree-specific random deviation thereof is a smooth function of relative height. Furthermore, we assume that the stem taper may be reasonably well approximated by piecewise cubic polynomials with a low number of internal knots, known as splines. By choosing a given number of knots for the population mean taper curve and the smooth random deviation, the regression parameters as well as the random-effect parameters for the calibration of the tree-individual taper function can be calculated within the approved methodological framework of linear mixed models. The same applies to prediction errors and CIs for diameter and volume estimates.

Most taper models require one or two diameter measurements from predefined measuring locations and total height of a sample tree as predictors. In our model, there are no such restrictions. Neither the number of the recorded diameter measurements per sample tree is fixed nor their positions. Moreover, recorded diameters and their positions may vary from sample tree to sample tree. To transform the diameter positions to a relative scale, which is used in our model, total tree height or a suitable estimate thereof is needed for every sample tree. When height is estimated or measured with error, we use the "law of total expectation and variance" (Weiss 2005) to calculate diameter and volume estimates with prediction errors that take the additional uncertainty introduced by the height estimation into account.

Data and tree height estimation

To give an example of the proposed taper modelling approach, we used Norway spruce (*Picea abies* (L.) H. Karst.) trees measured in a study which was conducted to derive models for the estimation of tree biomass using the data from the third German NFI. The data were collected between 2008 and 2010 according to the regional distribution of Norway spruce in Germany. The data set was complemented with measurements from other field studies carried out in the north-western and south-eastern part of Baden-Württemberg. Overall, 387 sample trees with 6,787 diameter measurements were used for model fitting and validation. For each tree, a stump diameter and diameter measurements at 0.50 cm, 1 m, 2 m, and from 3 m up to the top in regular intervals of 2 m were available. In the original data set, dbh was not available but interpolated for each tree by spline smoothing. The smallest, largest, and mean dbh were 12.2, 84.2, and 35.2 cm, respectively. The tree height varied between 7.3 and 41.5 m with a mean tree height of 26.4 m.

Usually, tree height is only measured on a subset of trees at an NFI sample plot. The height of the other trees is estimated using a model. A simple height-diameter model was fitted in order to illustrate the influence of the uncertainty of the additional model. For the estimation of mean tree height and its variance given dbh, we used data from the current German NFI. The dbh–height scattergram with more than hundred thousand trees was smoothed with a low degree of freedom smoothing spline. The empirical variance within 2 cm dbh-classes was smoothed in order to get estimates of the height variability given a dbh-class. In a second step, we adapted the regression functions to the data of the biomass study in order to get unbiased estimates for the sample trees in the study. We do not describe all details here because height-diameter models are not the focus of this paper and the estimated tree heights are only used to demonstrate the method of taper curve prediction when height is estimated or measured with error.

Methods

Taper curve model: mixed-effects B-spline regression

Stem tapering is clearly nonlinear in nature. Moreover, data on tree shape are generally hierarchically organized since repeated measurements, i.e., the diameter measurements along the stem, are taken from sampled trees. The interest is not only in the population mean diameter–height relation but also on effects of stand and management conditions, for instance, on the tree-individual taper curve. Therefore, the most obvious method for the analysis is nonlinear mixed models. However, nonlinear models are often too rigid and predetermine the estimates up to a certain degree in advance. This is especially unfavourable if factors influencing the tree shape are of interest but underlying mechanisms and interactions are largely unknown. Moreover, nonlinear mixed models are computationally expensive and numerically instable. Semiparametric mixed models, on the other hand, are model alternatives which are flexible enough to let "data speak" and are numerically tractable and stable.

We propose a semiparametric linear mixed approach, where we assume that the population mean taper curve as well as the tree-specific random deviation thereof is a smooth function of relative height which may be reasonably well approximated by spline functions. The flexibility of a spline function is determined by the number and the location of the knots. The internal knots define the points where the polynomial pieces are connected with continuous second derivatives across the knots. Alternative spline representations exist which yield equivalent fits. We use a B-spline representation because it has good numerical properties and, at least to a certain extent, allows for a local interpretation—B-splines are different from zero only for a limited range depending on the order, the number, and the location of the knots of the B-spline basis. The local interpretation of the B-splines fits well to the different geometrical shapes within the tree.

For statistical inference, we further assume that the observed diameter deviations of the sampled trees are realizations of a Gaussian process with two components. The first component describes the smooth tree-specific random deviations from the population mean taper curve with zero mean and a covariance function depending on the position of the measurements. The second component is a vector of zero mean residual errors with a covariance matrix which can consider heteroskedastic and autocorrelated errors.

In the following, relative height is given by $h = H/H_t$, where H = height at a certain position along the stem and H_t = tree height. In our approach, we assume that the population mean diameter for a given relative height may be approximated by a cubic (polynomial of degree $p = 3$) B-spline function $f(h) = \sum_{l=0}^{d_1} \beta_l B_l^{(1)}(h)$, $h \in [0, 1]$, with a given partition of the unit interval defined by the knot sequence $0 = k_0^{(1)} < k_1^{(1)} \cdots < k_{K_1}^{(1)} = 1$, $d_1 \le K_1 + p$. The superscript (1) indicates the spline functions of the fixed-effects part of the model (population mean) where the number of internal knots is $K_1 - 1$. $K_1 + p$ is the maximal number of different B-spline basis functions given the knot sequence, from which we use only a subset of d_1 basis functions. This means, d_1 is the number of fixed-effect parameters in our regression model. The computation of the B-spline basis function $B_l^{(1)}(h) \equiv B_{l,p}^{(1)}(h)$ is based on the "*de Boor recurrence relation*" with

$$B_{l,0}^{(1)} = \begin{cases} 1 & \text{for } k_l^{(1)} \le h \le k_{l+1}^{(1)} \\ 0 & \text{otherwise} \end{cases},$$

as starting point followed by the recursion

$$B_{l,j+1}^{(1)}(h) = \alpha_{l,j+1} B_{l,j}^{(1)}(h) + \left[1 - \alpha_{l+1,j+1} B_{l+1,j}^{(1)}(h)\right],$$

where

$$\alpha_{l,j} = \begin{cases} \frac{h - k_l^{(1)}}{k_{l+j}^{(1)} - k_l^{(1)}} & \text{if } k_{l+j}^{(1)} \ne k_l^{(1)} \\ 0 & \text{otherwise} \end{cases}, \quad j = 0, 1, \ldots, p$$

with 0/0 defined as 0 for the above computations (de Boor 1978).

The tree-individual random deviation from the population mean taper curve is also modelled by a B-spline function $g(h) = \sum_{l=0}^{d_2} \gamma_l B_l^{(2)}(h)$, $h \in [0, 1]$ with a given knot sequence $0 = k_0^{(2)} < k_1^{(2)} \cdots < k_{K_2}^{(2)} = 1$, $d_2 \le K_2 + p$ to partition the unit interval and normally distributed random coefficients $\gamma = (\gamma_1, \ldots, \gamma_{d_2})^T$ with zero mean and covariance matrix G. The superscript (2) indicates the spline functions of the random-effects part of the model (tree-individual effects). The population mean function and the function describing the between-tree deviation are the basic elements of the semiparametric regression model which we propose for the jth diameter measurement $D(h_{ij})$ from the ith sample tree, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$:

$$\begin{aligned} D(h_{ij}) &= f(h_{ij}) + g_i(h_{ij}) + \epsilon_{ij}, \quad g_i(h) \sim \mathrm{GP}(0, \Gamma), \\ \epsilon_i &= (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T \sim N(0, R_i) \end{aligned} \tag{1}$$

where $f(h_{ij}) = \sum_{l=0}^{d_1} \beta_l B_l^{(1)}(h_{ij})$ is the smooth population mean and $g_i(h_{ij}) = \sum_{l=0}^{d_2} \gamma_{il} B_l^{(2)}(h_{ij})$, $i = 1, \ldots, n$ are the tree-specific deviations from $f(h_{ij})$, which are assumed to be realizations of a Gaussian process (GP) with mean 0 and covariance function $\Gamma(h, h') = \mathrm{cov}(g_i(h), g_i(h'))$ for any pair of within-tree locations $h, h' \in [0, 1]$. The vector of residual errors ϵ_i is assumed to be normally distributed with covariance matrix R_i. Note: the smooth part of the

within-tree correlation, i.e., the covariance function of the deviation processes $g_i(h)$, is directly related to the covariance matrix of the random effects γ_i and the B-spline base $B^{(2)}(h)$ through the "spatial" bilinear form $\Gamma(h, h') = B^{(2)}(h)^T G B^{(2)}(h')$ where $G = \mathrm{cov}(\gamma_i)$.

With the notation $x_{ij} := x(h_{ij}) := (B_0^{(1)}(h_{ij}), \ldots, B_{d_1}^{(1)}(h_{ij}))^T =: B^{(1)}(h_{ij})^T$ and $z_{ij} := z(h_{ij}) := (B_0^{(2)}(h_{ij}), \ldots, B_{d_2}^{(2)}(h_{ij}))^T =: B^{(2)}(h_{ij})^T$, $\beta = (\beta_0, \cdots, \beta_{d_1})^T$ and $\gamma_i = (\gamma_{i0}, \ldots, \gamma_{id_2})^T$, the semiparametric model (1) can be denoted as

$$D(h_{ij}) = x_{ij}^T\beta + z_{ij}^T\gamma_i + \epsilon_{ij},\ \gamma_i \sim N(0, G),\ \epsilon_i \sim N(0, R_i),\quad j = 1, \ldots, n_i,\ i = 1, \ldots, n. \tag{2}$$

Given the two sets of B-spline basis functions $B^{(1)}(h)$ and $B^{(2)}(h)$, model (2) is a standard linear mixed-effects (LMEs) model, which can be expressed in matrix notation as

$$D_i = X_i\beta + Z_i\gamma_i + \epsilon_i,\ \gamma_i \sim N(0, G), \epsilon_i \sim N(0, R_i),\quad i = 1, \ldots, n,\ \text{where} \tag{2'}$$

$$D_i = (D(h_{i1}), \ldots, D(h_{in_i}))^T,$$
$$X_i = (x_{i1}, \ldots, x_{in_i})^T = (B^{(1)}(h_{i1}), \ldots, B^{(1)}(h_{in_i}))^T$$

and $Z_i = (z_{i1}, \ldots, z_{in_i})^T = (B^{(2)}(h_{i1}), \ldots, B^{(2)}(h_{in_i}))^T$.

The LME model (2′) can easily be solved for $\hat\beta$, $\hat\gamma_i$ and the variance parameters $\hat\theta = (\hat\theta_G, \hat\theta_R)^T$ with $\hat G = G(\hat\theta_G)$ and $\hat R_i = R_i(\hat\theta_R)$ using standard statistical software such as the *lme* (Pinheiro et al. 2012) function in the *R*-package *nlme* or the SAS® procedure *PROC MIXED*. Usually, $\theta = (\theta_G, \theta_R)^T$ is a set of parameters that determine the covariance function of the random effects and the residual variance–covariance matrix. For a general covariance for the random effects, the components of the variance parameter θ_G are random-effect standard deviations and correlations, for which *nlme* uses log and logit transformations in order to guarantee positive standard deviation estimates and correlations ranging between -1 and $+1$. With homoscedastic and independent residual errors, the parameter θ_R is usually the logarithm of the residual standard deviation σ_ϵ (Pinheiro and Bates 2000, p. 93).

When G and R_i are known, the best linear unbiased estimate (BLUE) for the fixed-effects coefficients is

$$\hat\beta = \left(\sum_{i=1}^n X_i^T\Sigma_i^{-1}X_i\right)^{-1}\sum_{i=1}^n X_i^T\Sigma_i^{-1}D_i$$

with $\Sigma_i := \mathrm{cov}(D_i) = Z_iGZ_i^T + R_i$, $i = 1, \ldots, n$. When R and G are unknown, they are replaced by their estimates in which case $\hat\beta$ is called the empirically best linear unbiased estimate.

Taper curve calibration and prediction

Once the LME model is fitted, we use the fixed-effects estimate $\hat\beta$ and the estimates for the variance parameters $\hat\theta = (\hat\theta_G, \hat\theta_R)^T$ to predict tree-individual taper curves of trees that were not used to fit the taper model. For those trees that may have been measured in an inventory, at least one or a set of several diameter and corresponding height measurements are required to calibrate the tree-individual random effects. The diameter measurements for calibration are denoted $D_m = (D(H_1), \ldots, D(H_m))^T$ at heights $H_1, \ldots, H_m$ and tree height H_t. With these calibration data (D_m, H_t), we calculate an empirical best linear unbiased prediction (EBLUP) estimate for the random effects as

$$\hat\gamma_m = \hat E[\gamma|D_m, H_t] = \hat G Z_m^T\hat\Sigma_m^{-1}(D_m - X_m\hat\beta)$$

which is the basis for diameter prediction at an arbitrary location H on the stem

$$\begin{aligned}\widehat{E}[D(H)|D_m, H_t] &:= \widehat{E}[D(h)|D_m] = \hat f(h) + \hat g_m(h)\\ &= x(h)^T\hat\beta + z(h)^T\hat\gamma_m\end{aligned} \tag{3}$$

where $h = H/H_t$ is the relative height for which a diameter estimate is sought. The mean squared error (MSE) for the calibrated taper curve $\hat E[D(H)|D_m, H_t]$ is estimated by

$$\begin{aligned}\widehat{\mathrm{VAR}}\left[\widehat{E}[D(H)|D_m, H_t]\right] &= x(h)^T\hat\Sigma_{\hat\beta}\,x(h)^T + z(h)^T\hat V_m z(h)\\ &\quad + x(h)^T\hat C_m z(h) + z(h)^T\hat C_m^T x(h)\end{aligned}$$

with

$$\begin{aligned}V_m &:= \mathrm{VAR}[\hat\gamma_m - \gamma_m]\\ &= G - GZ_m^T\Sigma_m^{-1}Z_mG + GZ_m^T\Sigma_m^{-1}X_m\Sigma_{\hat\beta}X_m^T\Sigma_m^{-1}Z_mG,\end{aligned} \tag{4}$$

$C_m = \mathrm{COV}[(\hat\beta - \beta), (\hat\gamma_m - \gamma_m)] = -\Sigma_{\hat\beta}X_m^T\Sigma_m^{-1}Z_mG$ and $\Sigma_{\hat\beta} = \left(\sum_{i=1}^n X_i^T\Sigma_i^{-1}X_i\right)^{-1}$, see Vonesh and Chinchilli (1997). With the $(1-\alpha)$ percentile $\Phi^{-1}(1-\alpha)$ of the standard normal distribution, we get approximate estimates for the pointwise CIs of the expected mean diameter at position H

$$\widehat{E}[D(H)|D_m, H_t] \pm \Phi^{-1}(1 - \alpha/2)\sqrt{\widehat{\mathrm{VAR}}\left[\widehat{E}[D(H)|D_m, H_t]\right]} \tag{5}$$

and prediction intervals for a single diameter measurement

$$\widehat{E}[D(H)|D_m, H_t] \pm \Phi^{-1}(1 - \alpha/2)\sqrt{\widehat{\mathrm{VAR}}\left[\widehat{\mathrm{E}}[D(H)|D_m, H_t]\right] + \sigma_\epsilon^2(H)} \tag{6}$$

where $\sigma_\epsilon^2(H)$ is the residual error at height H.

Taper curve prediction when H_t is not measured

In most operational forest inventories, tree height is only measured for a subsample of trees. Therefore, one needs an estimate for tree height in order to use Eq. (3) for the prediction of the tree taper curve and approximate CIs. When tree height is not measured exactly but estimated using a model or measured with error, we seek an estimate of the BLUP $E[D(H)|D_m]$ of the mean diameter at position H. With some knowledge on the tree height distribution given D_m and the law of total expectation we get

$$E[D(H)|D_m] = E(E[D(H)|D_m, H_t]) = \int_{\mathbb{H}_t} E[D(H)|D_m, H_t] p(H_t|D_m)\, \mathrm{d}H_t \quad (7)$$

where $p(H_t|D_m)$ is the probability density function of the conditional tree height distribution $H_t|D_m$. It is assumed that $p(H_t|D_m)$ may be approximated by a normal probability density function $p_N(H_t|\mu_{H_t}, \sigma_{H_t})$. The parameters of the approximate tree height distribution are estimated as $\widehat{\mu}_{H_t} = T^H_\mu(D_m)$ and the deviation $\widehat{\sigma}_{H_t} = T^H_\sigma(D_m)$ with some suitable regression functions T. Using the EBLUP (Eq. 3), a natural estimate of the expected diameter at an arbitrary tree position H is

$$\hat{E}[D(H)|D_m] = \int_{\mathbb{H}_t} \hat{E}[D(H)|D_m, H_t]\, p_N(H_t|\hat{\mu}_{H_t}, \hat{\sigma}_{H_t})\, \mathrm{d}H_t. \quad (8)$$

Note that we define $D(H) \equiv 0$ for all $H \notin [0, H_t]$ throughout the paper. When calibration diameters D_m are available for a sample tree but tree height is not measured, pointwise CIs $\mathrm{CI}^D_{1-\alpha}(H|D_m)$ for the expected diameter at position H are obtained from

$$\mathrm{CI}^D_{1-\alpha}(H|D_m) : q^D_{\alpha/2}(H) \le E[D(H)|D_m] \le q^D_{1-\alpha/2}(H) \quad (9)$$

where the limits of the CI are the quantiles $q^D_{\alpha'}$ of the $E[D(H)D_m]$ distribution which are estimated by solving

$$\begin{aligned} \mathrm{P}\big(E[D(H)|D_m] \le q^D_{\alpha'}\big) &= E_{H_t}\big[P(E[D(H)|D_m, H_t] \le q^D_{\alpha'}|H_t)\big] \\ &= \int_{\mathbb{H}_t} P(E[D(H)|D_m, H_t] \le q^D_{\alpha'}) p(H_t|D_m)\mathrm{d}H_t \\ &\approx \int_{\mathbb{H}_t} P\big(N(\hat{\mu}_D(H|H_t), \hat{\sigma}_D(H|H_t)) \le q^D_{\alpha'}\big) \\ &\quad \times p_N(H_t|\hat{\mu}_{H_t}, \hat{\sigma}_{H_t})\, \mathrm{d}H_t = \alpha' \end{aligned} \quad (10)$$

for $\alpha' = \alpha/2$ and $1 - \alpha/2$.

The parameter estimates of the normal approximation $N(\hat{\mu}_D(H|H_t), \hat{\sigma}_D(H|H_t))$ are the EBLUPs $\hat{\mu}_D(H|H_t) = \widehat{E}[D(H)|D_m, H_t]$ and $\hat{\sigma}_D(H|H_t) = \sqrt{\widehat{\mathrm{VAR}}\left[\widehat{E}[D(H)|D_m, H_t]\right]}$ of the calibrated taper function (Eqs. 3, 4). In (10), $\widehat{\mu}_{H_t}$, $\widehat{\sigma}_{H_t}$ are the parameter estimates for the normal approximation of the $H_t|D_m$ distribution (Eq. 8).

Volume estimation when H_t is measured

Taper functions provide the possibility of volume estimation for the total stem and arbitrary stem intervals. Thus, assortment structures can be estimated using inventory data. Stem volume between height A and B is calculated by the integrated cross-sectional area of the taper function

$$\begin{aligned} E\big[\mathrm{VOL}^B_A\big] &= c_{[m]} E\left[\int_A^B D^2(H)\, \mathrm{d}H\right] \\ &= c_{[m]}\left\{\int_A^B E^2[D(H)]\mathrm{d}H + \int_A^B \mathrm{VAR}[D(H)]\mathrm{d}H\right\} \end{aligned} \quad (11)$$

with the scaling factor $c_{[m]}$. A scaling factor of $c_{[m]} = \left[\frac{\pi}{4}10^{-4}\right]$ converts diameters measured in centimetres and heights measured in metres to volume in cubic metres. In the above equation, we assume that the taper function is so regular that expectation and integration can be exchanged. The volume for the complete stem is given by setting $A = 0$ and $B = H_t$.

Given tree height H_t and calibration diameters D_m as well as the fitted taper model (Eqs. 3, 4), a volume estimate is obtained by

$$\begin{aligned} \hat{E}\big[\mathrm{VOL}^B_A|D_m, H_t\big] = c_{[m]}\Bigg\{&\int_A^B \hat{E}^2[D(H)|D_m, H_t]\mathrm{d}H \\ &+ \int_A^B \widehat{\mathrm{VAR}}\left[\hat{E}[D(H)|D_m, H_t]\right]\mathrm{d}H\Bigg\}. \end{aligned} \quad (11')$$

Using the equations for the multivariate moments of a bivariate normal distribution (see "Appendix"), the variance of the stem volume is calculated as

$$\begin{aligned}
&\mathrm{VAR}\left[\mathrm{VOL}_A^B\right] \\
&= c_{[m]}^2 \int_A^B \int_A^B \mathrm{COV}\left[\, D^2(H_1), D^2(H_2)\right] \mathrm{d}H_1 \mathrm{d}H_2 \\
&= c_{[m]}^2 \int_A^B \int_A^B \{\mathrm{VAR}[D(H_1)]\mathrm{VAR}[D(H_2)] \\
&\quad + 2\mathrm{COV}^2[D(H_1), D(H_2)] \\
&\quad + 4E[D(H_1)]E[D(H_2)]\mathrm{COV}[D(H_1), D(H_2)]\} \mathrm{d}H_1 \mathrm{d}H_2
\end{aligned} \tag{12}$$

When calibration data D_m and H_t are available, the variance of the stem volume may therefore be estimated by

$$\begin{aligned}
&\widehat{\mathrm{VAR}}\left[\mathrm{VOL}_A^B | D_m, H_t\right] \\
&= c_{[m]}^2 \int_A^B \int_A^B \Big\{ \widehat{\mathrm{VAR}}\left[\hat{E}[D(H_1)|D_m, H_t]\right] \widehat{\mathrm{VAR}}\left[\hat{E}[D(H_2)|D_m, H_t]\right] \\
&\quad + 2\,\widehat{\mathrm{COV}}^2\left[\widehat{E}[D(H_1)|D_m, H_t], \widehat{E}[D(H_2)|D_m, H_t]\right] \\
&\quad + 4\,\widehat{E}[D(H_1)|D_m, H_t]\,\widehat{E}[D(H_2)|D_m, H_t] \\
&\quad \times \widehat{\mathrm{COV}}\left[\widehat{E}[D(H_1)|D_m, H_t], \widehat{E}[D(H_2)|D_m, H_t]\right] \Big\} \mathrm{d}H_1 \mathrm{d}H_2
\end{aligned} \tag{12'}$$

where $\widehat{\mathrm{COV}}\left[\widehat{E}[D(H_1)|D_m, H_t], \widehat{E}[D(H_2)|D_m, H_t] = x(h_1)^T \hat{\Sigma}_{\hat{\beta}}\, x(h_2)^T + z(h_1)^T \hat{V}_m z(h_2) + x(h_1)^T \hat{C}_m z(h_2) + z(h_1)^T \hat{C}_m^T x(h_2)\right.$, with $h_1 = H_1/H_t$ and $h_2 = H_2/H_t$ (Eq. 4).

Volume estimation when H_t is not measured

When only some diameters D_m are available for calibration of a tree-individual taper curve, we seek for an estimate of $E[\mathrm{VOL}_A^B|D_m]$. With the same technique used above for diameter prediction (Eq. 7), we get

$$\begin{aligned}
E\left[\mathrm{VOL}_A^B | D_m\right] &= E_{H_t}\left[E\left[\mathrm{VOL}_A^B | D_m, H_t\right]\right] \\
&= \int_{\mathbb{H}_t} E[\mathrm{VOL}_A^B | D_m, H_t] p(H_t|D_m)\, \mathrm{d}H_t.
\end{aligned} \tag{13}$$

Moreover, applying the "law of the total variance" we also get

$$\begin{aligned}
\mathrm{VAR}\left[\mathrm{VOL}_A^B | D_m\right] &= \int_{\mathbb{H}_t} \mathrm{VAR}\left[\mathrm{VOL}_A^B | D_m, H_t\right] p(H_t|D_m)\, \mathrm{d}H_t \\
&\quad + \int_{\mathbb{H}_t} \left(E\left[\mathrm{VOL}_A^B | D_m, H_t\right]\right)^2 p(H_t|D_m) \mathrm{d}H_t \\
&\quad - \left(\int_{\mathbb{H}_t} E\left[\mathrm{VOL}_A^B | D_m, H_t\right]\, p(H_t|D_m) \mathrm{d}H_t\right)^2.
\end{aligned} \tag{14}$$

Replacing $p(H_t|D_m)$, $E\left[\mathrm{VOL}_A^B | D_m, H_t\right]$ and $\mathrm{VAR}\left[\mathrm{VOL}_A^B | D_m, H_t\right]$ by their estimates (Eqs. 8, 11′, 12′) provides the desired estimates $\widehat{E}\left[\mathrm{VOL}_A^B | D_m\right]$ and $\widehat{\mathrm{VAR}}\left[\mathrm{VOL}_A^B | D_m\right]$.

Applications

The spruce data from the biomass study were fitted with a mixed-effects B-spline regression model (1) with knots {0, 0.1, 0.75, 1.0} corresponding to the B-spline basis functions $B^{(1)}(h)$ for the population mean function and {0, 0.1, 1.0} for $B^{(2)}(h)$ to model the smooth deviation from the population average. We used the R-function *splineDesign* (Bates and Venables 2013) with modified knot sequences {0, 0, 0, 0, 0.1, 0.75, 1.0, 1.0, 1.0, 1.0} and {0, 0, 0, 0, 0.1, 1.0, 1.0, 1.0, 1.0} to implement the de Boor recurrence relation in order to calculate the B-spline basis functions $B^{(1)}(h)$ and $B^{(2)}(h)$. The results of these calculations are shown in Fig. 1. In order to guarantee diameter predictions to be zero at the top of the tree, we omit the rightmost basis function $B_6^{(1)}(h)$ and $B_5^{(1)}(h)$ in the model Eq. (1). This means, we use $d_1 = 5$ fixed-effect parameters for the population mean function and $d_2 = 4$ random-effect parameters for the tree-individual deviation. We fitted the mixed-effects B-spline regression Eq. (1) with the R-function *lme* (Pinheiro et al. 2012) with 6,787 diameter observations from 387 spruce trees.

The residuals corresponding to the population mean function, i.e., $r_{ij}^{(1)} = D(h_{ij}) - \hat{f}(h_{ij})$ are unbiased along the whole stem showing increasing variances towards the stump (Fig. 2a). We note that the high residual variability observed in the box plot, with residual deviations from the population mean of up to 40 cm, is mainly due to the fact that only relative height is used as predictor in our model. No tree diameter, such as dbh or the diameter at 1 m of height for instance, is incorporated as predictor variable in the population mean model. Using all diameter measurements available for a tree, the complete mixed-effects B-spline regression model, of course, has much more prediction power (Fig. 2b). Assuming all diameters measured, the empirical residual distribution of $r_{ij}^{(2)} = D(h_{ij}) - (\hat{f}(h_{ij}) + \hat{g}_i(h_{ij}))$ ranges approximately between -4 and $+4$ cm. Hence, the variability is nearly reduced by a factor of ten compared to that of $r_{ij}^{(1)}$. Moreover, the residual variance of $r_{ij}^{(2)}$ is more or less constant along the tree, but we still observe an increasing residual variance in the stump area. In addition, there is no evidence of systematic deviations from zero (Fig. 2b).

The EBLUP estimate of the individual taper curve $\widehat{E}[D(H)|D_m, H_t] = \hat{f}(h) + \hat{g}_m(h)$ calibrated according to Eq. (3) with height = 41 m and all diameter measurements available, $D_m = \{D(H_j), j = 1, \ldots n\}$ is illustrated in Fig. 3.

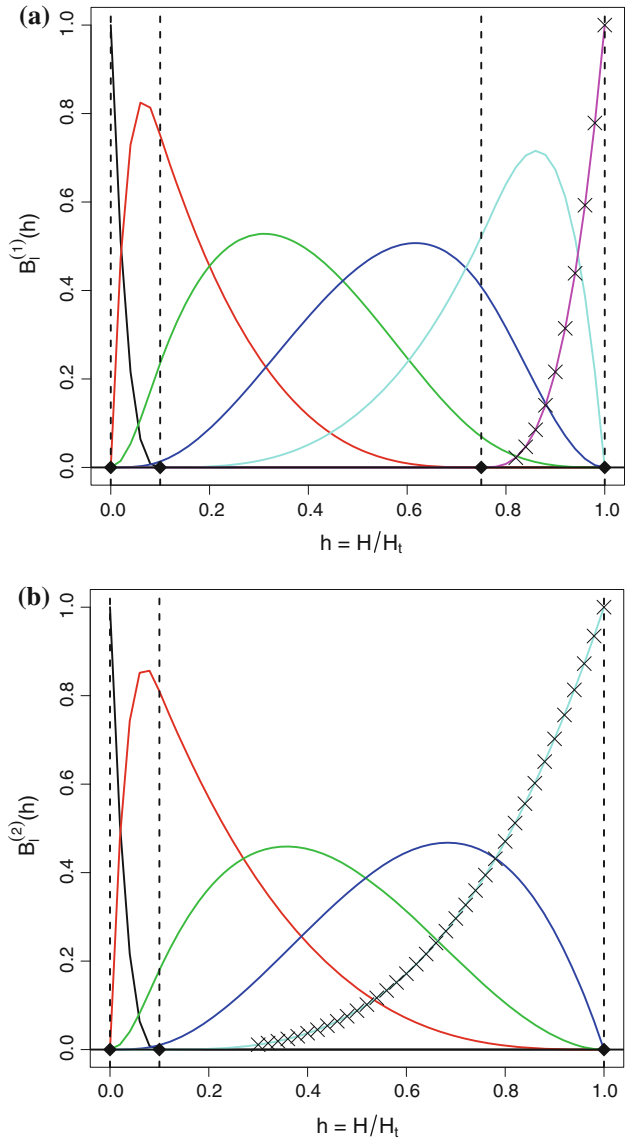

Fig. 1 **a** Population mean—fixed-effects B-spline basis $B_l^{(1)}(h)$, $l = 1,\ldots,6$. **b** Smooth deviation from population average—random-effects B-spline basis $B_l^{(2)}(h)$ $1 = 1,\ldots,5$. *Diamonds* and *dashed vertical lines* indicate knot positions. *Crosses* indicate omitted splines in order to obtain estimates of 0 at tree height

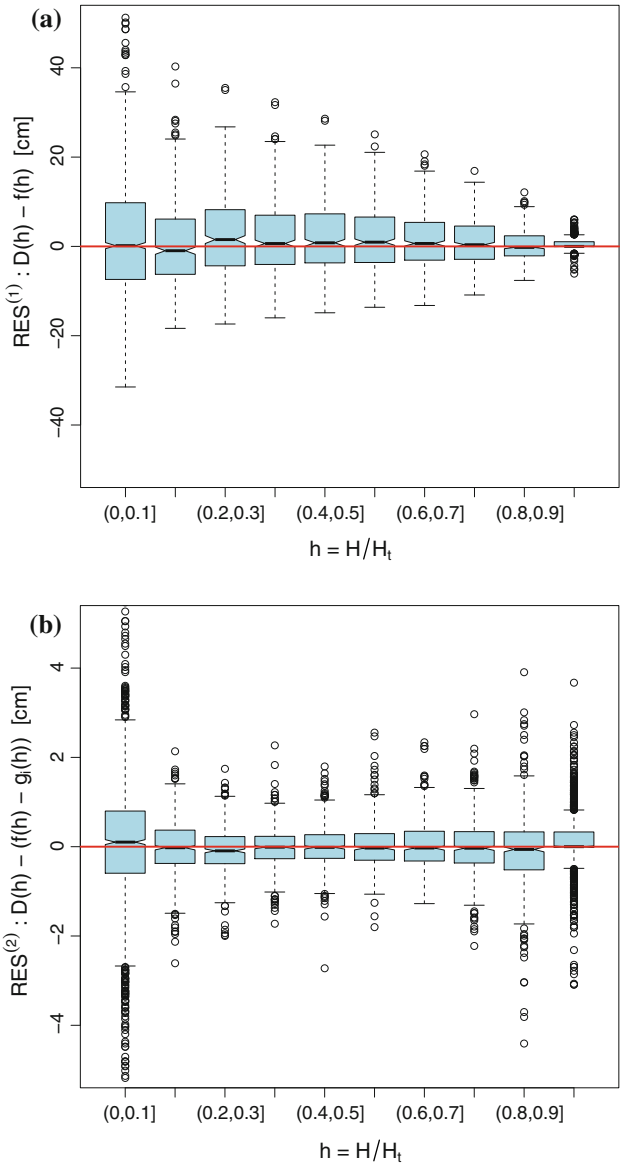

Fig. 2 *Box plots* of the empirical residual distribution **a** residuals of the population model $r_{ij}^{(1)} = D(h_{ij}) - \hat{f}(h_{ij})$ over relative height h_{ij} **b** residuals of the tree-individual models given all measured diameters $r_{ij}^{(2)} = D(h_{ij}) - (\hat{f}(h_{ij}) + \hat{g}_i(h_{ij}))$ over relative height h_{ij}

The estimated taper curve is smooth with a good fit for all diameter measurements. The pointwise CIs for the mean diameter (area shaded dark grey) are narrow. Moreover, all diameters lie within the corresponding prediction interval which is plotted as the area with light grey colouring. Although not so important from a practical point of view, this example demonstrates that mixed-effects B-spline regression is well suited to extract the smooth tree-individual taper curve from sectional data. But the main practical interest is in calibrating tree-specific taper curves when only one or two diameters and possibly tree height are recorded for sample trees in inventories.

The estimated taper curve of the same tree if only dbh = 61.2 cm and tree height $H_t = 41$ m is used for calibration is shown in Fig. 4a. The pointwise CIs for the mean (area of dark grey colour) remain nearly constant from 5 to 25 m with an MSE for the mean of about 2.4 cm and for the diameter prediction of 3.5 cm. For predictions in the upper part of the stem, the MSE is shrinking towards zero. The model fit meets the restriction of a zero mean diameter prediction at top height (Fig. 4a). However, the empirical taper curve is systematically underestimated between 5 and 25 m. Nevertheless, the sectional diameters are captured by the pointwise prediction intervals depicted

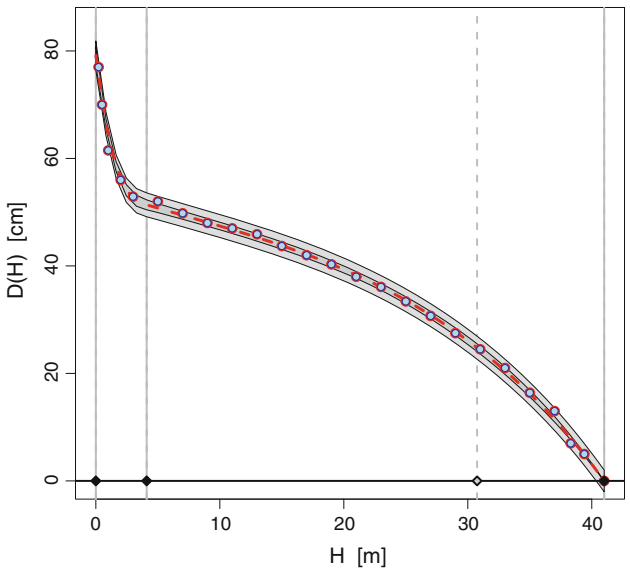

Fig. 3 Empirical taper curve with the smooth EBLUP estimate $\widehat{E}[D(H)|D_m, H_t]$ with $D_m = \{D(H_j), j = 1, \ldots, n\}$ as *dashed line*. *Circles* indicate diameter measurements (empirical taper curve). Pointwise CIs for the mean diameter are given in *dark grey* and prediction intervals for a diameter observation are given in *light grey*. The associated knots for the B-spline representation of the mean function are indicated on the abscissa as points and as *vertical lines*. The *dashed vertical line* indicates the knot that is only used by the population model

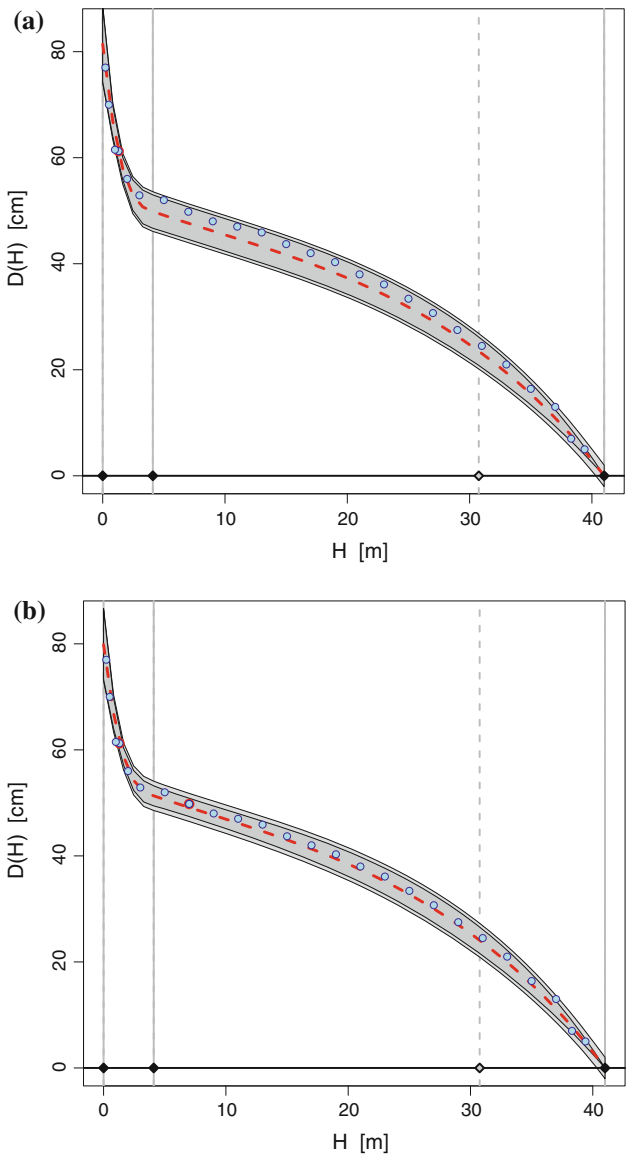

Fig. 4 Taper curve calibration **a** EBLUP estimate $\widehat{E}[D(H)|\text{dbh}, H_t]$ and **b** $\widehat{E}[D(H)|\{\text{dbh}, D_7\}, H_t]$. Pointwise CIs for the mean diameter are given in *dark grey* and prediction intervals for a diameter observation are given in *light grey*. *Circles* indicate diameter measurements (empirical taper curve). The associated knots for the B-spline representation of the mean function are indicated on the abscissa as points and as *vertical lines*. The *dashed vertical line* indicates the knot that is only used by the population model

by the area shaded in light grey colour. In the upper stem area, the model fit looks better. The corresponding result, when an additional upper diameter $D_7 = 49.8$ cm at 7 m is used for calibration, is illustrated in Fig. 4b. The overall fit is good with no observable systematic deviation from the empirical taper curve. The pointwise CIs for the mean diameter are smaller compared to those in Fig. 4a. Especially around the upper calibration diameter at 7 m, we see a clear reduction of the prediction variance. The RMSE for the mean is minimal at 7 m with 0.7 cm and increases up to 1.7 cm at 27 m before decreasing again continuously to zero at top height.

The calibration with one or two diameters results in practically unbiased estimates along the trunk of the trees (Fig. 5). The calibration based on dbh alone results in a symmetric and homogeneous variability band with an interquartile range of about 3 cm (±1.5 cm) (Fig. 5a). If D_7 is available for calibration in addition to dbh, the variability decreases with a minimum residual variance at relative position of $h = 0.25$. The relative height $h = 0.25$ coincides well with the relative height of the D_7 for the population mean tree height, which is 26.4 m. The positive effect of variance reduction caused by the incorporation of an upper diameter abates towards the stump and the top of the tree.

Similar observations as for tree taper were made for timber volume. For the sample tree used above, the volume prediction from Eq. (11′) with dbh alone is $\widehat{E}\left[\text{VOL}_0^{H_t}|\text{dbh}, H_t\right] = 4.55$ m^3 with an estimated prediction error $\widehat{\text{STD}}\left[\text{VOL}_0^{H_t}|\text{dbh}, H_t\right] = 0.5$ m^3 (Eq. 12′). The corresponding estimates when dbh, D_7 and H_t are used for calibration are 4.78 and 0.18 m^3. With the full information, i.e., $D_m = D_{\text{obs}} = \left\{D(H_j), j = 1, \ldots n\right\}$ and $H_t = 41$ m, we get a volume estimate 4.89 m^3 with standard deviation 0.05 m^3.

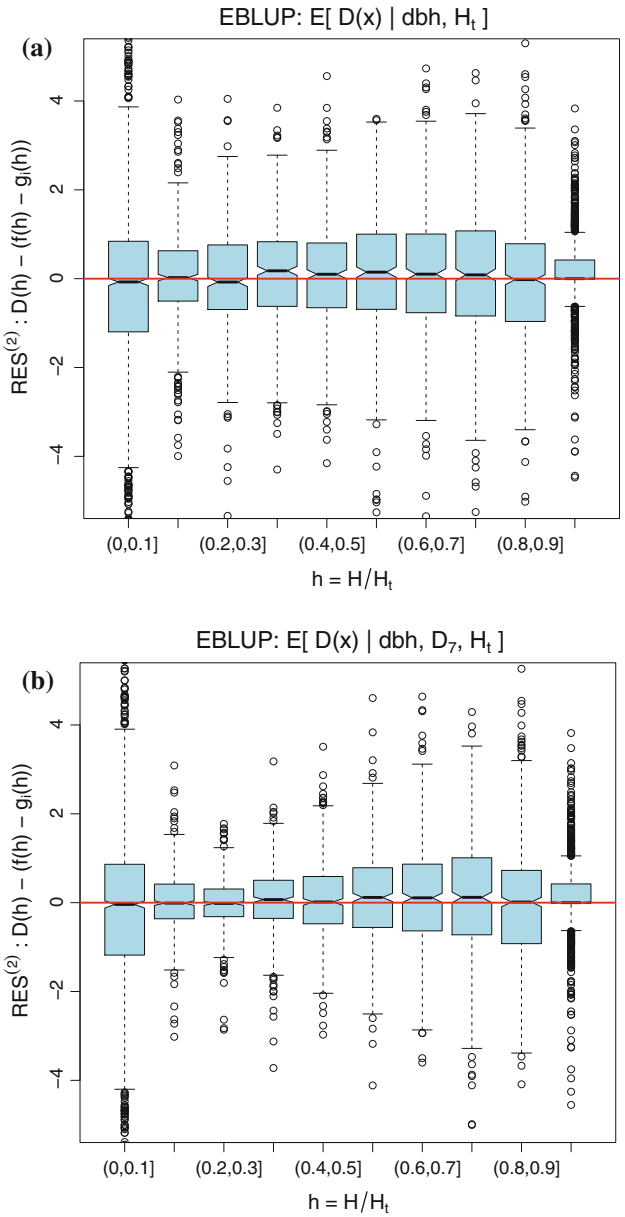

Fig. 5 *Box plots* of residual distributions. **a** Residuals (cm) based on calibration with dbh and H_t, EBLUP $\widehat{E}[D(H)|\text{dbh}, H_t]$ over relative height h_{ij}, **b** residuals (cm) based on calibration with dbh, D_7 and H_t, EBLUP $\widehat{E}[D(H)|\{\text{dbh}, D_7\}, H_t]$ over relative height h_{ij}

To compare the benefit from measuring D_7, we calculated the relative volume error $\text{CV} = (\widehat{\text{STD}}[\text{VOL}_0^{H_t}|D_m, H_t]/\widehat{E}[\text{VOL}_0^{H_t}|D_{\text{obs}}, H_t]$ for every sample tree with $D_m = \{\text{dbh}\}$ and $\{\text{dbh}, D_7\}$. The relative errors of the sample tree volume are displayed in Fig. 6. The reduction in relative error from using dbh together with D_7 instead of using dbh alone as calibration diameter is around 3 % for small trees and decreases to about 1.5 % for big trees.

When planning inventories, one has to decide whether or not the gain in precision is worth the price of recording an upper diameter. In the German NFI, upper diameters as

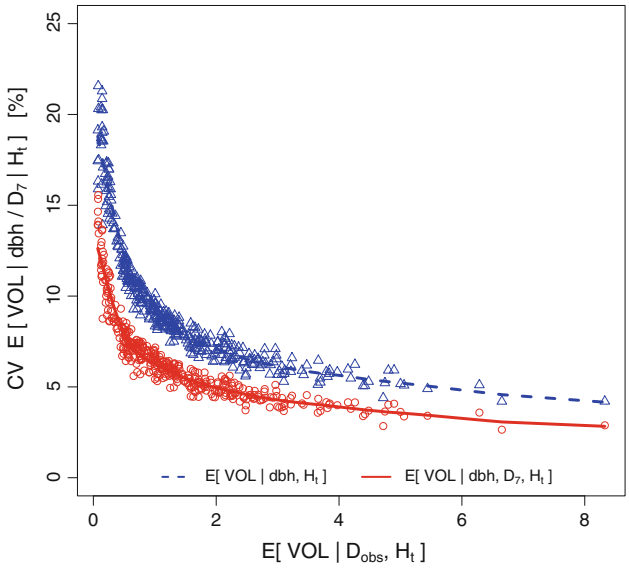

Fig. 6 Coefficient of variation (CV, in per cent) over volume predictions (m^3). **a** Calibration with dbh and H_t, EBLUP $\widehat{E}[D(H)|\{D_{1.3}\}, H_t]$ (*triangles* and *dashed line*). **b** Calibration with dbh, D_7 and H_t, EBLUP $\widehat{E}[D(H)|\{D_{1.3}, D_7\}, H_t]$ (*circles* and *solid line*)

well as tree height are only measured on a subsample of trees. For the other trees, height is estimated by a height-diameter model calibrated at the plot-level using heights of sample trees.

Usually, the additional variability caused by the estimation of height or erroneous height measurements is not considered in diameter and volume prediction. The effect of this additional source of uncertainty is demonstrated in Fig. 7a, b for the sample tree in Fig. 4 but with the measured height (41 m) replaced by an estimate of 35.6 m. Moreover, we use an estimated standard deviation for the height prediction of 4.2 m. The predicted taper curve where only dbh (61.2 cm) for calibration and a normal approximation $N(\mu_{H_t}, \sigma_{H_t})$ with $\hat{\mu}_{H_t} = 35.6\,\text{m}$ and $\hat{\sigma}_{H_t} = 4.2$ m for the conditional $H_t|\text{dbh}$ distribution is used (see Eq. 7) is shown in Fig. 7a.

The taper curve based on measured dbh and estimated height for the selected sample tree heavily underestimates the observed diameter measurements. This is especially true in the upper stem part. The reason for this is mainly the fact that the sample tree is much higher than expected from the mean regression in the NFI sample trees. The uncertainty of the tree height estimate is expressed in wide CIs for the mean diameter predictions. The dotted lines in Fig. 7a correspond to the EBLUP of the taper of a tree with dbh = 61.2 and a measured height of $\hat{\mu}_{H_t} - 2\hat{\sigma}_{H_t} = 27.2$ m and $\hat{\mu}_{H_t} + 2\hat{\sigma}_{H_t} = 44$ m. It is clear that for positions $H > \hat{\mu}_{H_t} + 2\hat{\sigma}_{H_t}$ only big trees with top height $H_t > \hat{\mu}_{H_t} + 2\hat{\sigma}_{H_t}$ are expected to have positive mean diameter

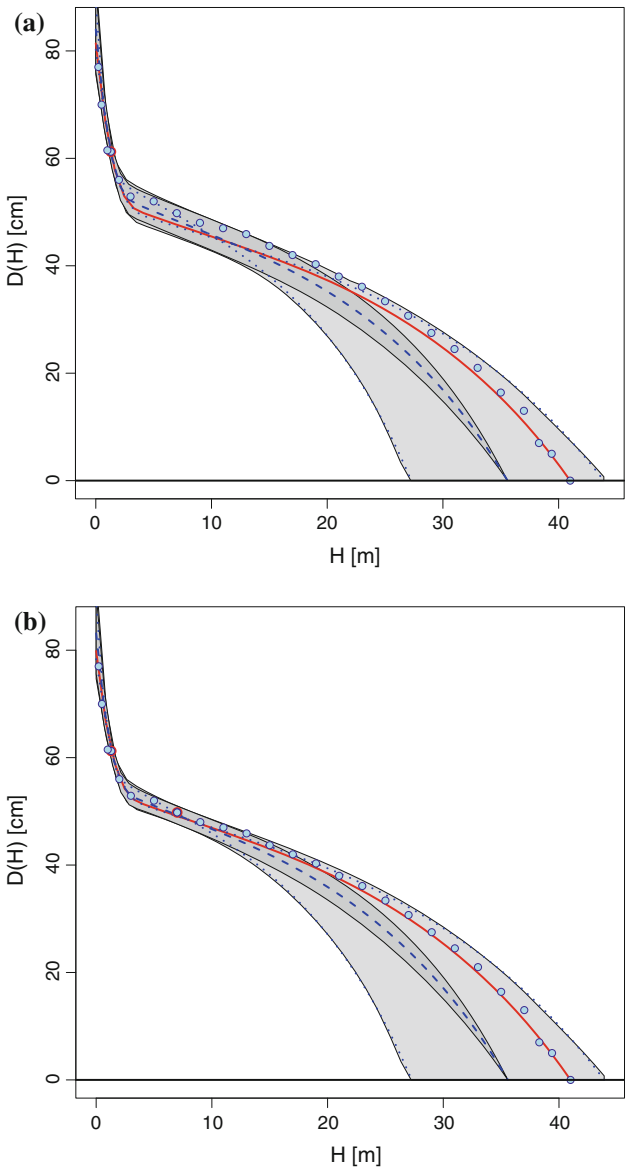

Fig. 7 Taper curve calibration—EBLUPs with confidence (*dark grey*) and prediction (*light grey*) intervals for **a** calibration with dbh = 61.2 cm and estimated $H_t = 35.6$ (*dashed line*), **b** calibration with dbh = 61.2 cm, $D_7 = 49.8$ cm and estimated $H_t = 35.6$ (*dashed line*). The *dotted lines* are EBLUPs based on using $\hat{\mu}_{H_t} \pm 2\hat{\sigma}_{H_t}$ as the tree height. The *solid line* indicates the taper curve estimated using dbh and tree height for calibration. *Circles* indicate diameter measurements (empirical taper curve)

predictions. From the normal approximation of the conditional $H_t|$dbh distribution, we know that this is true for about 2.5 % of the trees within the same dbh-class. In addition, the taper curves preserve some ordering (Kublin 2003). Similar arguments apply for diameter predictions at positions $H \sim \hat{\mu}_{H_t} - 2\hat{\sigma}_{H_t}$. Summarizing these heuristical arguments, we conclude that in the upper stem area, the EBLUP taper curves corresponding to calibration data $\{\text{dbh}, H_t = \hat{\mu}_{H_t} \pm 2\hat{\sigma}_{H_t}\}$ are good approximations for the upper and lower 2.5 % bounds of the CIs for the mean diameter prediction from Eq. (9). In the lower bole area (0–10 m), the diameter predictions are mainly influenced by the dbh as the calibration diameter. Tree height variation plays only a minor role in that area, i.e., we may expect diameter predictions to fall within the confidence limits of a tree with estimated mean height $H_t = \hat{\mu}_{H_t}$ in a given dbh-class, see the dark grey area in the plot. Figure 7a clearly shows that variability from tree height is the dominating cause of uncertainty in the upper stem area but the influence is decreasing towards the stump. This is underlined by Fig. 7b where dbh and D_7 is used for calibration and tree height is estimated. The variance reduction effect due to the incorporation of D_7 is strictly restricted to a small area around the upper diameter and much smaller than the variance reduction effects observed in Fig. 4b, where height is assumed to be measured without error.

In addition to taper curve calibration with dbh and estimated height H_t, we calculated estimates for tree volume. For the previously used sample tree with dbh = 61.2 cm and estimated tree height $\hat{H}_t = 35.6$ m and $\sigma_t = 4.2$ m (Fig. 7a), we get a tree volume estimate $\widehat{E}\left[\text{VOL}_0^{\hat{H}_t}|\text{dbh}\right] = 4.2\,\text{m}^3$ with prediction error $\widehat{\text{STD}}\left[\text{VOL}_0^{\hat{H}_t}|\text{dbh}\right] = 0.36\,\text{m}^3$ from Eqs. (13) and (14). This means, the relative volume prediction error is 8.5 %. If we ignore height estimation and assume that the mean height $H_t = 35.6$ (0) m is measured, the prediction error is $\widehat{\text{STD}}\left[\text{VOL}_0^{H_t}|\text{dbh}\right] = 0.22\ \text{m}^3$ which is a relative prediction error of 5.2 %. If we calculate the "true" volume from the empirical taper curve with $D_m = D_{\text{obs}}$ and $H_t = 41$ m, we get a volume estimate of 4.89 (0.05) m^3.

Discussion

Model building and fitting

When constructing taper functions for volume and assortment estimation, we face two problems. First, taper functions used in large-scale inventories have to provide unbiased estimates for a wide range of stem dimensions. Therefore, the taper functions on the one hand have to be flexible enough to describe the nonlinear nature of tapering with different degrees of curvature from the stump to the top of a tree. On the other hand, taper functions have to be rigid enough to guarantee reliable estimates when extrapolating. Second, taper curve fitting is usually based on multiple observations from sample trees (i.e., longitudinal) data. Therefore, it is reasonable to assume that the within-tree observations are spatially correlated.

We propose a mixed B-splines regression model for diameter and volume prediction with relative heights as

predictor variables. Local flexibility of the B-spline regression model enables the reproduction of different degrees of curvature observed along the stem with a population mean function, and mixed model framework is used to accommodate nonindependence and/or heterogeneity. In the example application, we used a B-spline representation with fixed internal knots at 10 and 75 % of total height for the population mean function and a single internal knot at 10 % of height for the B-spline representation of the smooth tree-individual random deviation from the population mean. The knot positions are based on extensive experience. It is also in accordance with the very early attempts of taper curve modelling with a frustum of a neiloid at the stump, a frustum of paraboloid, and a cone at the top of the tree. The placement of the lower knot in the stump area corresponds to findings in other studies where the knot positions were estimated from the data (e.g., Max and Burkhart 1976; Sharma and Burkhart 2003 and Trincado and Burkhart 2006). With knot positions not fixed in advance, the B-spline regression is nonlinear. With prefixed knots, we can benefit from the advantages of the linear mixed regression methodology. In contrast to most studies, we do not use dbh as an additional predictor variable. As a consequence, the deviations of the diameter measurements from the population mean have a high and inhomogeneous variability. The variability is drastically reduced and homogenized when a tree-individual smooth random B-spline component is added to the population mean function. We used a fully unspecified positive definite matrix to model the random coefficient covariance together with a diagonal covariance matrix for homogenous and independent residual errors.

In our approach, the random effects do not only model the between-tree variation but also the major part of the within-tree correlation. The correlation between two diameter deviations from the mean function at different heights is a function of the diameter positions and the random-effects covariance. The local character of the B-spline representation, where $B_1^{(2)}(h)$ and $B_2^{(2)}(h)$ are used to model the random diameter deviations in lower, i.e., $0<h<0.2$, $B_l^{(2)}(h)$, $l = 2,3$ in the middle and $B_4^{(2)}(h)$ in the upper stem part, is responsible for considerable intercorrelation between the corresponding random effects. Estimates for the fixed-effects parameters and the variance parameters were calculated with the R-function *lme* (Pinheiro et al. 2012) using REML. The estimated correlation matrix of the random effects shows a clear spatial ordering with high within-tree correlations $\widehat{\mathrm{COR}}\begin{bmatrix}\gamma_1\\ \gamma_2\\ \gamma_3\\ \gamma_4\end{bmatrix} = \begin{bmatrix}1 & & & \\ 0.944 & 1 & & \\ 0.928 & 0.937 & 1 & \\ 0.894 & 0.933 & 0.854 & 1\end{bmatrix}$. This means that if we observe a large diameter deviation from the population mean at position h, it is very likely that we will observe a large deviation at neighbouring within-tree positions. The correlation decreases with increasing distances between two diameters. The standard deviation $\widehat{\mathrm{STD}}[\gamma] = (17.9, 10.6, 8.5, 7.9)^T$ is decreasing from the bottom towards the top of the tree, which is plausible, because at the top ($h = 1$) with an expected zero mean diameter we do not have any variation.

Trincado and Burkhart (2006) used spline regression with a truncated power basis representation (Max and Burkhart 1976) with the square of the relative diameter $(D(h)/\mathrm{dbh})^2$ as response and relative height (h) as predictors. The knot positions of the spline basis were treated as nonlinear fixed-effect parameters and estimated from the data. In order to account for the between-tree variability, the fixed-effects parameters of the mean function were partially overlaid by random effects. The truncated power basis representation does not have a local interpretation comparable to that of our B-spline representation. Therefore, it is not surprising that the random-effects correlations in the model by Trincado and Burkhart (2006) did not provide dependencies with a clear spatial structure. The estimated variances of the random effects are extremely inhomogeneous and difficult to explain. Moreover, the lack of a clear spatial structure in the random effects is probably the reason why the consideration of a autocorrelated residual error term (CAR(1)) with continuously decreasing exponential autocorrelation function $\phi^{|h-h'|}$ provided a significant improvement of the model fit over the model with independent residual errors. The estimated autocorrelation coefficient was $\hat{\phi} = 0.54$ (Trincado and Burkhart 2006). We were not able to identify residual autocorrelation in our model when we used a fully unspecified covariance function since the algorithms failed to converge.

As stated by Pinheiro and Bates (2000, p. 204), there are two components in a linear mixed effect model that can be used to model heteroscedasticity and correlation: the random-effects component and the within-tree component. Both components can be used to model autocorrelation. They compete in the sense that in practice one can use a complex random-effects component added to a simple within-tree component or a simple between-tree random-effects component in combination with a more elaborated within-tree component to model the overall dependency structure in the data. Pinheiro and Bates (2000) also note that there will generally be a trade-off between the complexity of the two components and some care must be taken to prevent nonidentifiability of the model parameters and numerical problems during parameter estimation.

With a diagonal random-effects covariance matrix, we were able to detect residual autocorrelation with an estimated $\hat{\phi} = 0.26$, i.e., with the assumption of independent random effects we fully shift the problem of modelling spatial dependence to the residual error component. The

estimated BIC value was 26,813 which, compared to a BIC value of 24,714 for the mixed-effects B-spline regression with fully unspecified covariance and independent random errors, shows that modelling spatial correlation with auto-correlated residual errors is clearly outperformed by the model where spatial correlation is modelled by the B-spline random effects (Pinheiro and Bates 2000, p. 10). Moreover, in practice, when only some diameter measurements D_m are available for a sample tree, we are mainly interested in the prediction of the smooth tree-individual taper curve $f(h) + g_m(h)$. For the calculation of the predictions and the corresponding prediction errors, we make use of the spatial covariance $\Gamma(h, h') = \mathrm{cov}(f(h) + g_m(h), f(h') + g_m(h'))$, see Eqs. (1) and (3). This and the findings from other studies, which have shown that modelling random effects with an unstructured variance–covariance matrix in many cases accounts for the majority of the correlation among the observations, see Yang et al. (2009) and references therein, were arguments in favour of the mixed B-spline regression model with an unstructured variance–covariance matrix and independent residual errors.

Tree-individual stem profile curves and volume prediction

Once having fitted the mixed B-spline regression, one borrows strength from the model information derived from sectional data of intensively callipered sample trees to localize tree-individual stem profile curves when only height and a few diameters are available from a sample tree. In contrast to most taper models, where one diameter, e.g., dbh, and total height are used as predictor variable in the mean function and tree-individual localizing is based on an additional upper diameter, we used only relative heights as predictor variables in the population mean function. In our model, calibration diameters may be recorded at arbitrary stem positions with no restriction on the number of measurements. Since our model is a linear mixed regression, the numerical calibration can be performed using any standard statistical package. Linear mixed model theory also provides confidence and prediction intervals for the calibrated taper curve, see Eqs. (5) and (6). Although standard and with some matrix algebra easily implemented, confidence and prediction intervals are rarely provided in the numerous subject-specific publications. We note that the MSE estimate of the calibrated taper curve (Eq. 4) ignores the variability induced by the estimation of the variance parameters. An approximately unbiased second-order MSE estimate is provided in Rao (2003, Chap. 6). All information needed to calculate the correction of the bias in the MSE estimate is provided by the R-function *lme* (Pinheiro et al. 2012).

Another source of variability which is often ignored in MSE calculation is the uncertainty caused by using tree height estimates instead of measurements. We used the "law of total expectation and variance" together with a normal approximation of the tree height distribution given calibration diameter D_m to predict the expected mean diameter $E[D(H)|D_m]$ at position H, see Eqs. (7) and (8). The numerical integration in Eq. (8) was calculated by Gauss–Legendre quadrature with the help of the R-function *gaussLegendre* (Borchers 2013). The 2.5 and 97.5 % quantiles of the EBLUP distribution as lower and upper boundaries were calculated according to Eq. (10) with an iterative root-solver which is implemented in the R-function *uniroot* (Soetaert 2013). These calculations were time consuming and hence not applicable in practical situations. We gave heuristical arguments and demonstrated in Fig. 7a, b that height estimation dominates expected mean diameter variability in the upper part of a stem. We further demonstrated that in the upper area, the confidence limits for the mean stem profile are bounded by taper curves with tree heights corresponding to the 2.5 % and the 97.5 % quantiles of the $H_t|D_m$ distribution. In the stump region, the variability of height estimation only plays a minor role. All taper curves are more or less fixed by the calibration diameters, from which we concluded that in the stump area, the taper of the tree with mean height $H_t = \hat{\mu}_{H_t}$ provides reasonable approximations for the CIs. Thus, a suitable overall approximation of the upper/lower limit of the 95 % CI of the expected mean diameter is given by

$$\mathrm{CI}_{u/l}(H) \approx \mathrm{MAX/MIN}\{\hat{\mu}_D(H|H_t = \hat{\mu}_{H_t}) \pm 2\hat{\sigma}_D (H|H_t = \hat{\mu}_{H_t}),\ \hat{\mu}_D(H|H_t = \hat{\mu}_{H_t} \pm 2\hat{\sigma}_{H_t})\}$$

where we use the notations from Eqs. (8) and (10). We emphasize at this point that an elaborated model for tree height estimation is needed in order to keep the diameter prediction error as small as possible. However, as already mentioned, height estimation was not the focus of this study.

When tree height is measured, the volume estimate is the integral over the cross-sectional area derived from the calibrated taper curve (Eq. 11′). For the numerical integration, we used Gauss–Legendre quadrature. The MSE estimation of the volume prediction requires the numerical calculation of a two-dimensional integral, see (12′). For this, we have adapted an algorithm by Press et al. (1986, Chap. 4). When tree height is estimated or measured with error, we used the same techniques as with diameter prediction, i.e., the "law of total expectation and variance" and a normal approximation of the tree height distribution to calculate volume predictions and prediction errors, see Eqs. (11)–(14). As an alternative to the normal

approximation in Eqs. (13) and (14), Lappi (2006) proposed a two-point distribution $\tilde{p}(H_t|D_m)$ putting equal probability weight = 0.5 on $H_t = \hat{\mu}_{H_t} \pm \hat{\sigma}_{H_t}$. This approximation speeds up the calculations and provides similar estimates.

All methods presented in this study including the fitting of the mixed-effects B-spline regression are implemented in the R-package *TapeR* (Kublin and Breidenbach 2013).

Acknowledgments We would like to thank Dr Juha Lappi with The Finnish Forest Research Institute and two anonymous reviewers for their valuable comments that helped improving an early version of the manuscript.

Appendix: Derivation of Eq. (12)

For the bivariate normal distribution with density function

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right),$$

the bivariate moments are

$$\begin{aligned}
\mu_{22} &= \mathrm{COV}\left[(X-\mu_X)^2,(Y-\mu_y)^2\right] \\
&= (1+2\rho_{XY}^2)\sigma_X^2\sigma_Y^2 = \sigma_X^2\sigma_Y^2 + 2\mathrm{COV}^2[X,Y] \\
\mu_{12} &= \mathrm{COV}\left[(X-\mu_X),(Y-\mu_y)^2\right] \\
&= \mathrm{COV}\left[X,(Y-\mu_y)^2\right] = 0 \\
\mu_{21} &= \mathrm{COV}\left[(X-\mu_X)^2,(Y-\mu_y)\right] = \mathrm{COV}\left[(X-\mu_X)^2,Y\right] \\
&= 0
\end{aligned} \tag{15}$$

(see Kendall and Stuart 1977, p. 85).

With the identity $X^2 = (X-\mu)^2 - \mu^2 + 2X\mu$, we get

$$\begin{aligned}
\mathrm{COV}(X^2,Y^2) &= \mathrm{COV}\Big[(X-\mu_X)^2-\mu_X^2+2X\mu_X, \\
&\qquad (Y-\mu_Y)^2-\mu_Y^2+2Y\mu_Y\Big] \\
&= \mathrm{COV}\left[(X-\mu_X)^2,(Y-\mu_Y)^2\right] \\
&\quad - \mathrm{COV}\left[\mu_X^2,(Y-\mu_Y)^2\right] + 2\mu_X\mathrm{COV}\left[X,(Y-\mu_Y)^2\right] \\
&\quad - \mathrm{COV}\left[(X-\mu_X)^2,\mu_Y^2\right] + \mathrm{COV}\left[\mu_X^2,\mu_Y^2\right] \\
&\quad - 2\mu_X\mathrm{COV}\left[X,\mu_Y^2\right] + 2\mu_Y\mathrm{COV}\left[(X-\mu_X)^2,Y\right] \\
&\quad - 2\mu_Y\mathrm{COV}\left[\mu_X^2,Y\right] + 2\mu_X2\mu_Y\mathrm{COV}[X,Y] \\
&= \mathrm{COV}\left[(X-\mu_X)^2,(Y-\mu_Y)^2\right] + 2\mu_X2\mu_Y\mathrm{COV}[X,Y] \\
&= \sigma_X^2\sigma_Y^2 + 2\mathrm{COV}^2[X,Y] + 4\mu_X\mu_Y\mathrm{COV}[X,Y]
\end{aligned}$$

This follows since COV(X, Y) is bilinear in X and Y. With the relations in (15), we finally have

Note: This equation is also used in Lappi (2006) but the term $\sigma_X^2\sigma_Y^2$ is missing there. In the practical application, the value of $\sigma_X^2\sigma_Y^2$ is so small that the published results in Lappi (2006) remain practically unchanged when the correct formula is used (Lappi, personal communication).

References

Bates DM, Venables WN (2013) Regression spline functions and classes. R-package version 2.15.1

Borchers HW (2013) pracma: practical numerical math functions. R-package version 1.4.0

de Boor C (1978) A practical guide to splines. Springer, New York

Demaerschalk JP, Kozak A (1974) Suggestions and criteria for more effective regression sampling. Can J For Res 4:341–348

Fang Z, Bailey RL (2001) Nonlinear mixed effects modeling for slash pine dominant height growth following intensive silvicultural treatments. Forest Sci 47:287–300

Garber SM, Maguire DA (2003) Modeling stem taper of three central Oregon species using nonlinear mixed effects models and autoregressive error structures. Forest Ecol Manage 179(1/3):507–522

Höjer AG (1903) Tallens och granens tillväxt. In: Lovén FA (ed) Om vara barrskogar. Uddeholms Aktiebolag, Stockolm, pp 87–120

Hradetzky J (1980) Spline Funktionen und ihre Anwendung in der forstlichen Forschung. Forstwissenschaftliches Centralblatt 100:45–59

Kendall M, Stuart A (1977) The advanced theory of statistics, vol 1. Charles Griffin, London

Kozak A (1988) A variable-exponent taper equation. Can J Forest Res, NRC Res Press 18:1363–1368

Kozak A (2004) My last words on taper equations. Forest Chron 80:507–515

Kublin E (2003) Einheitliche Beschreibung der Schaftform–Methoden und Programme–BDATPro. Forstwissenschaftliches Centralblatt 122:183–200

Kublin E, Breidenbach (2013) TapeR—Flexible tree taper curves based on semiparametric mixed models. R-package version 0.3.0. Available on CRAN: http://cran.r-project.org/web/packages/TapeR/

Kublin E, Augustin NH, Lappi J (2008) A flexible regression model for diameter prediction. Eur J Forest Res 127:415–428

Lappi J (2006) A multivariate, nonparametric stem-curve prediction method. Can J Forest Res 36:1017–1027

Max TA, Burkhart HE (1976) Segmented polynomial regression applied to taper equations. Forest Sci 22:283–289

Newnham R (1988) A variable-form taper function. Information Report PI-X-083. Petawawa National Forestry Institute, Chalk River, p 33

Newnham R (1992) Variable-form taper functions for four Alberta tree species. Can J Forest Res 22:210–223

Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-Plus. Springer, Berlin

Pinheiro JC, Bates D, DebRoy S, Sarkar D, The R Development Core Team (2012) nlme: linear and nonlinear mixed effects models. R-package version 3.1-106

Pollanschütz J (1965) Eine neue Methode der Formzahl- und Massenbestimmung stehender Stämme. Hochschule für Bodenkultur Wien

Press W, Flannery B, Teukolsky S, Vetterling W (1986) Numerical recipes. Cambridge University Press, Cambridge

Rao J (2003) Small area estimation. Wiley Interscience, London

Rojo A, Perales X, Sanchez-Rodriguez F, Alvarez-Gonzalez JG, von Gadow K (2005) Stem taper functions for maritime pine (*Pinus pinaster* Ait.) in Galicia (Northwestern Spain). Eur J Forest Res 124:177–186

Sharma M, Burkhart H (2003) Selecting a level of conditioning for the segmented polynomial taper equation. Forest Sci 49:324–330

Sloboda B, Gaffrey D, Matsumura N (1998) Presentation of tree individual taper curves and their dynamics by spline functions and generalization by linear taper curve models. Allgemeine Forst und Jagdzeitung 169:29–39

Soetaert K (2013) rootSolve: nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations. R-package version 1.6

Trincado G, Burkhart HE (2006) A generalized approach for modelling and localizing stem profile curves. Forest Sci 52:670–682

Vonesh E, Chinchilli V (1997) Linear and nonlinear models for the analysis of repeated measurements. Marcel Dekker, New York

Weiss NA (2005) A course in probability. Addison-Wesley, Reading, MA

Yang Y, Huang S, Trincado G, Meng S (2009) Nonlinear mixed-effects modeling of variable-exponent taper equations for lodgepole pine in Alberta, Canada. Eur J Forest Res 128:415–429

Zhang YJ, Borders BE, Bailey RL (2002) Derivation, fitting, and implication of a compatible stem taper-volume-weight system for intensively managed, fast growing loblolly pine. Forest Sci 48:595–607