Annals of Forest Science (2017) 74: 29
DOI 10.1007/s13595-017-0629-y

ORIGINAL PAPER

Forest modelling: the gamma shape mixture model and simulation of tree diameter distributions

Rafał Podlaski[1]

Received: 27 December 2016 / Accepted: 10 March 2017 / Published online: 3 April 2017

Abstract

• ***Key message*** **New types of distribution functions are needed to model the dynamics of stands where important age classes are represented by few trees. In this study the gamma shape mixture model and two simulation methods were used for generating tree diameter data**.

• ***Context*** To analyse forest dynamics, it is necessary to know distribution of the characteristics (mainly tree diameters) of trees forming particular developmental phases. In many forest inventories, the measurement of large diameter at breast height (DBH) samples is practically impossible. In this case, DBH distributions can be generated using theoretical models.

• ***Aims*** The aim of this study was to assess the precision of the approximation of empirical DBH data using the gamma shape mixture (GSM) model and kernel density estimation. The strengths and weaknesses of the two simulation methods were presented and discussed.

• ***Methods*** The GSM model was adopted to approximate empirical DBH data collected in 20 near-natural stands. Two simulation methods were used: (a) the procedure based on a multimodal distribution and gamma random numbers (MDGR procedure) and (b) MCMC techniques with Metropolis–Hastings sampling (MH method).

• ***Results*** The GSM model precisely fitted the investigated DBH distributions. The MDGR procedure was slightly more precise than the MH method, especially in the case of the samples of 250 DBHs. The level of homogeneity within the drawn DBH sets was similar for all samples.

• ***Conclusion*** The GSM model is very flexible. The DBH random variates, generated with the use of analysed procedures, represented all tree generations being significant from a biological point of view.

Keywords Gamma shape mixture (GSM) model · Bayesian analysis · Diameter distribution · Simulation of diameter data · Near-natural forest

Handling Editor: Aaron R. Weiskittel

✉ Rafał Podlaski
r_podlaski@pro.onet.pl

[1] Department of Nature Protection and Plant Physiology, Institute of Biology, Jan Kochanowski University, ul. Świętokrzyska 15, 25-406 Kielce, Poland

1 Introduction

Disturbances occurring in forest ecosystems are one of the most important determinants of spatio-temporal development in stands (Gratzer et al. 2004). Due to disturbances of different spatial scales, gaps of varying sizes are formed. These processes have a significant effect on the structure of forests. The specific vertical structure is closely related to the shape of the diameter at breast height (DBH) distribution (Lawton and Putz 1988; Denslow et al. 1998). Many tree stands in various geographic regions contain cohorts of old trees, which are represented by only few individuals, but play a great role in stand structure and in ecosystem functioning. It is difficult to find distribution functions to represent these few large trees. During the approximation of these highly skewed and heavy-tailed DBH distributions, there is often the smoothing problem, which in turn requires the use of methods that are able to fit a tail probability well.

Different models have arisen naturally across a range of problems when modelling DBH in forestry (e.g. Pretzsch 2010). Single flexible theoretical distributions (e.g. Weibull, gamma) have often been used to fit empirical DBH data more or less asymmetrically with a positive skewness (Merganič

and Sterba 2006; Gove et al. 2008). Mixture distributions with a few components are an appropriate tool for modelling bi- and multimodal empirical DBH distributions (Zhang et al. 2001; Zasada and Cieszewski 2005; Podlaski 2011a, b; Zasada 2013).

In order to model the dynamics of forest stands with cohorts of old trees, new types of distribution functions are needed. A new approach for density estimation of highly skewed and heavy-tailed distributions, the gamma shape mixture (GSM) model, employs a mixture of gamma density functions with unknown weights (Venturini et al. 2008). A general Bayesian approach allows the creation of a flexible model characterised by a single parameter for all the gamma components and the ordinary set of mixture weights (Jasra et al. 2005; Venturini et al. 2008). This method significantly improves predictive performance in estimating tail probabilities compared to standard approaches employing e.g. single flexible theoretical distributions and mixture distributions with a few components (Venturini et al. 2008). A particularly important advantage of the GSM model is the possibility to use a great number of mixture components. In the case of two-generation stands where the two generations significantly differ in the number of trees, the model makes possible, among other things, to generate random DBH data, taking into account the existence of small local DBH maxima. These maxima, representing the older generation and creating longer-than-normal right tails, cannot be treated as atypical observations. The data are indispensable to correctly present DBH distributions in the case of two-generation stands, in which the older generation is formed by single old trees. Thereby, proposals that overcome the problem of atypical observations in distributions (e.g. by their identification and next, elimination) cannot be used.

In ecology, for analysis of forest dynamics, based on simulation studies, one should use data sets (mainly DBH) characterising the investigated stand in particular developmental phases. Tree lists, minimally a set of DBHs with an indicator of tree species, obtained from measurements made in selected plots are used to define the initial condition. The measurement of large DBH samples is practically impossible in many forest inventories due to economic limitations (e.g. Roesch et al. 2015). In this case, the DBH distributions can be generated using theoretical functions (e.g. Thompson 2000, Gehringer and Turnblom 2014).

Forest growth models based on progressing distributions are characterised by the inclusion of stand heterogeneity in the simulation approach, providing information on tree dimensions (e.g. Porté and Bartelink 2002). The accuracy of such models is primarily determined by the flexibility of the underlying type of theoretical function (e.g. Pretzsch 2010). Stand development is presented as a periodic progression of the frequency distributions. Each developmental phase is represented by a theoretical function of specified parameters. By changing these parameters, the BDH distribution can be shifted along the time axis. The DBH data generation makes it possible to increase the number of DBHs for small samples and then allows comparison of the model outputs with independent data.

Procedures based on Markov chain Monte Carlo (MCMC) techniques are frequently used methods for generating random numbers from probability distributions (Liu 2001). The Monte Carlo methods have become one of the most important tools to sample from complex distributions (e.g. Liu 2001; Robert and Casella 2004). There have been several classes of Monte Carlo techniques, e.g. MCMC techniques with Metropolis–Hastings sampling, sequential Monte Carlo techniques that include for example sequential importance resampling or particle filtering (Kong et al. 1994) and recent development of methods with equi-energy sampling (Kou et al. 2006).

The aims of this study are (1) to compare the precision of the approximation of empirical DBH data employing the GSM model and kernel density estimation (parametric and non-parametric methods) and (2) to assess the suitability of two methods for generating random DBH data from the GSM model: (a) the procedure using a multimodal distribution and gamma random numbers and (b) MCMC techniques with Metropolis–Hastings sampling. The GSM model has not been previously used for the analysis of forest data.

2 The gamma shape mixture model

The GSM model is defined as follows (Lehmann and Casella 1998; Venturini et al. 2008):

$$f(x|\pi_1, ..., \pi_J, \theta) = \sum_{j=1}^{J} \pi_j f_j(x|\theta) \tag{1}$$

where J is the number of mixture components (known and fixed), $\pi_1, ..., \pi_J$ are mixture weights (proportions) (unknown) and $1/\theta$ is the scale parameter for the whole GSM model (unknown). The gamma distribution $f_j(x|\theta)$ has a probability density function (PDF) given by

$$f_j(x|\theta) = \frac{\theta^j}{\Gamma(j)} x^{j-1} e^{-\theta x} \tag{2}$$

Each gamma distribution in the GSM model is indexed by a component-specific shape parameter (j) and has a single scale parameter ($1/\theta$).

The GSM model could also be defined as follows (Venturini et al. 2008):

$$p(x_1, ..., x_n|z_1, ..., z_n, \theta) = \frac{\theta^{\sum_{i=1}^{n} z_i}}{\prod_{i=1}^{n} \Gamma(z_i)} \left(\prod_{i=1}^{n} x_i^{z_i - 1} \right) e^{-\theta \sum_{i=1}^{n} x_i} \tag{3}$$

where $z_1, ..., z_n$ are the missing elements of the sample (Dempster et al. 1977; Diebolt and Robert 1994). Given $x_1, ..., x_n$, an integer z_i between 1 and J could be associated to each x_i that identifies the component of the mixture generating observation x_i; this auxiliary variable z_i identifies to which component the observation x_i belongs.

A general Bayesian approach for estimating the unknown parameters of the GSM model is often used. The $\pi_1, ..., \pi_J$ and θ are independent a priori and the following conjugate prior distributions are specified (Venturini et al. 2008):

$$\pi_1, ..., \pi_J \sim D_J\left(\frac{1}{J}, ..., \frac{1}{J}\right) \tag{4}$$

and

$$\theta \sim G(\alpha, \beta) \tag{5}$$

where $D_J(\bullet)$ is a Dirichlet distribution and $G(\bullet)$ is a gamma distribution, J, α and β are the hyperparameters. The posterior distribution is (Venturini et al. 2008)

$$p(\pi_1, ..., \pi_J, \theta | x_1, ..., x_n, z_1, ..., z_n) \propto \left(\prod_{j=1}^{J} \pi_j^{(1/J)+n_j-1}\right) \theta^{\alpha+\left(\sum_{i=1}^{n} z_i\right)-1} e^{-\left(\beta+\sum_{i=1}^{n} x_i\right)\theta} \tag{6}$$

where

$$n_j = \sum_{i=1}^{n} \mathrm{I}(z_i = j)$$

as well as $j = 1, ..., J$ and I(•) is the indicator function.

The posterior distribution is estimated using a Gibbs sampler, the parameter θ is derived analytically through integration. After having integrated out θ the posterior distribution is (Venturini et al. 2008)

$$p(\pi_1, ..., \pi_J, \theta | x_1, ..., x_n, z_1, ..., z_n) \propto \prod_{j=1}^{J} \pi_j^{(1/J)+n_j-1} \tag{7}$$

The primary advantage of this strategy is that the Markov chain runs in a smaller space (Robert 1996; MacEachern et al. 1999; Venturini et al. 2008).

3 Materials and methods

3.1 Field measurements

The plots were sampled in two-generation stands with fir *Abies alba* Mill. and beech *Fagus sylvatica* L., in protected, near-natural forests in the Świętokrzyskie Mountains (Świętokrzyski National Park, 50° 50′–50° 53′ N, 21° 01′–21° 05′ E). The study area lies at an elevation between 320 and 590 m above sea level. The most common plant associations are *Dentario glandulosae-Fagetum* and *Abietetum polonicum* (nomenclature after Matuszkiewicz 2008). In these stands, 30 circular plots from 0.2 to 0.4 ha were randomly selected. The radius of each plot was chosen so that the whole plot was situated within the boundaries of a homogenous patch of similar vertical stand structure. The age of trees, determined on the basis of increment core analysis, carried out during the present study and earlier dendrochronological research, shows that in the investigated area fir and beech trees of the older generation were usually characterised by DBHs >70 cm (Podlaski 2008, 2011a, b; Podlaski and Żelezik 2012). In each plot, the DBH was measured for all living trees >6.9 cm in diameter.

3.2 Forest data

To identify similar DBH structures in the investigated plots, 21 were used variables: fractions of the tree number (10 variables) and fractions of the basal area (10 variables) at 10-cm intervals from 7 to 107 cm, and the number of main extremes for DBH distributions (1 variable). The hierarchical cluster analysis (HCA) was employed with the Jaccard measure and the Ward's minimum variance agglomeration method. The 20 plots were clustered in three main groups (Fig. 1):

1. Group RS includes DBH distributions showing the rotated-sigmoid (RS) shape (10 plots) (Fig. 2).
2. Group BMS includes DBH distributions showing the typical bimodal M-shape (5 plots) (Fig. 3).
3. Group UID includes the unimodal irregularly descending distributions (5 plots) (Fig. 4).

The remaining 10 plots, in which the share of fir and beech assessed on the basis of a tree number was smaller than 80% as well as DBH distributions forming transitional structures, were not used in further studies.

In the investigated plots basal area for all species together was from 10.78 to 63.09 $m^2\ ha^{-1}$. The number of trees ranged from 86 to 234 stems per plot. Fir and beech definitely dominated and the appropriate values of the basal area varied from 6.62 to 53.5 $m^2\ ha^{-1}$ for fir and from 0.10 to 25.58 $m^2\ ha^{-1}$ for beech.

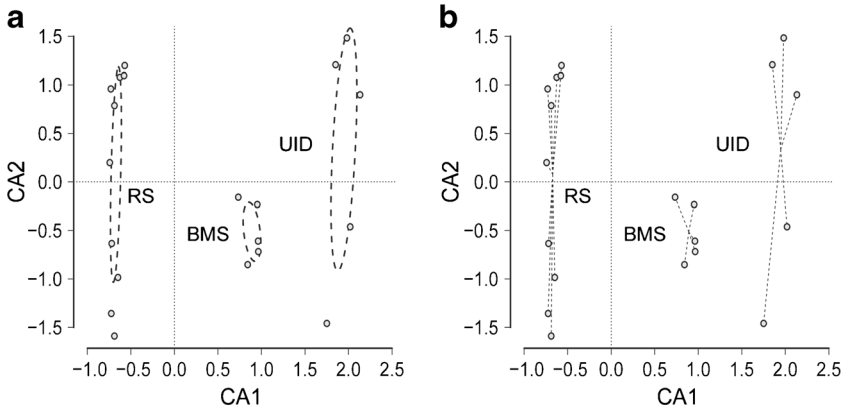

Fig. 1 Correspondence analysis (CA) ordination diagrams (CA1 and CA2 are ordination axes); 21 variables were used in the analysis to describe empirical tree DBH distributions. **a** '*Ellipse*' *diagram*—the weighted correlation defines the direction of the principal axis of the ellipse. **b** '*Spider*' *diagram*—each point is connected to the group centroid (*large black circles*). Cluster RS—rotated-sigmoid DBH distributions, cluster BMS— typical bimodal M-shape DBH distributions, cluster UID—unimodal, irregularly descending DBH distributions

3.3 Data analysis

Fitting with the GSM model requires three hyperparameters: the number of components J, and the α and β from the conjugate prior on θ. During the approximation of the empirical DBH data using the GSM model, it was assumed that the value of $J = 250$ and the weight of the prior information $\omega = 0.35$ (ω values between 0.2 and 0.5 are usually choices; for detailed information, see Venturini et al. 2008). With these assumptions for each plot, the α and β values were calculated as follows (Venturini et al. 2008):

$$\beta = \frac{\omega \sum_{i=1}^{n} x_i}{1-\omega} \tag{8}$$

$$\alpha = \frac{J}{\max(x_1, ..., x_n)}\beta \tag{9}$$

Kernel-type estimators are commonly used as non-parametric estimators for density functions. Let $x_1, ..., x_n$ be sample DBHs from an unknown density f. Then, its kernel estimate f is

$$f(x|h) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x-x_i}{h}\right) \tag{10}$$

where $K(\bullet)$ is a kernel function and h is a bandwidth. In this study, a Gaussian density as the kernel and a bandwidth $h = 2$ cm were used; the width for the DBH classes was chosen to be 2 cm (see also Lopez-de-Ullibarri 2015).

Two statistics were proposed for comparing the precision of the approximation of empirical DBH data using the GSM model and the kernel density estimation:

$$B_{\mathrm{DIF}} = |B_{\mathrm{GSM}}| - |B_{\mathrm{ker}}| \tag{11}$$

$$A_{\mathrm{DIF}} = A_{\mathrm{GSM}} - A_{\mathrm{ker}} \tag{12}$$

with

$$B_{\bullet} = \frac{1}{l}\sum_{q=1}^{l}\left(n_q - n_q\right) \tag{13}$$

$$A_{\bullet} = \frac{1}{l}\sum_{q=1}^{l}\left|n_q - n_q\right| \tag{14}$$

where n_q and n_q are the observed and predicted numbers of trees for the GSM model ($B_{\bullet} \equiv B_{\mathrm{GSM}}$ and $A_{\bullet} \equiv A_{\mathrm{GSM}}$) or for the

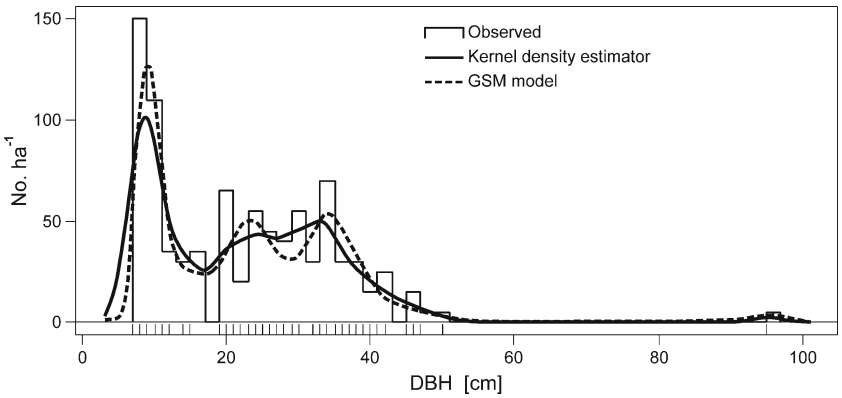

Fig. 2 Approximation of the empirical DBH distribution of an example stand from the group RS using the kernel density estimator and the GSM model (plot No. RS07)

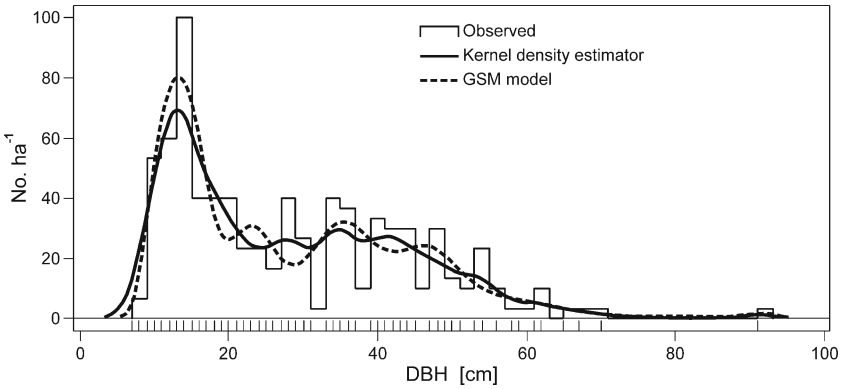

Fig. 3 Approximation of the empirical DBH distribution of an example stand from the group BMS using the kernel density estimator and the GSM model (plot No. BMS03)

kernel density estimation ($B_\bullet \equiv B_{\mathrm{ker}}$ and $A_\bullet \equiv A_{\mathrm{ker}}$), respectively, in the qth DBH class in the investigated plot; l is the number of DBH classes. The values of the $B_\bullet$ and $A_\bullet$ statistics indicate a measure of the bias and the flexibility of the analysed models, respectively.

3.4 Simulation studies

In order to generate random DBH data from the GSM model, the procedure using a multimodal distribution and gamma random numbers (hereinafter the MDGR procedure) and MCMC techniques with Metropolis–Hastings sampling (hereinafter the MH method) were employed. Gamma random numbers were generated in multinomial distribution cells using the acceptance-rejection principle with proper choice of the majorisation function (when the shape parameter was less than 1) or as the sum of two independent gamma variates (when the shape parameter was greater than or equal to 1) (Ahrens and Dieter methods; for detailed information, see Ahrens and Dieter 1974, 1982). The standard Metropolis–Hastings algorithm with jumping normal distribution was used (Robert and Casella 2004). For each plot, the following scheme was employed:

1. The empirical DBH distribution was fitted with the GSM model.
2. 50 samples of 100, 250 and 500 DBHs each were drawn using the GSM model and the MDGR procedure.
3. 50 samples of 100, 250 and 500 DBHs each were drawn using the GSM model and the MH method.

The k-sample Anderson-Darling tests (Scholz and Stephens 1987) were used to test the null hypotheses that (1) the samples come from the same but unspecified continuous distribution function and (2) the samples drawn using the MDGR procedure (block 1) and the MH method (block 2) come from the same but unspecified continuous distribution function (this function may change from block to block).

In the first case, the analyses were conducted for 50 samples containing 100, 250 and 500 DBHs for the MDGR procedure and the MH method; in total, six null hypotheses were tested for each plot. The Anderson-Darling k-sample test was employed; if AD is the Anderson-Darling criterion for k samples, its standardised test statistic is (Scholz and Zhu 2016)

$$T.AD = \frac{AD - \mu}{\sigma} \tag{15}$$

with μ and σ representing the mean and standard deviation of AD.

In the second case, the analyses were conducted for 50 samples containing 100, 250 and 500 DBHs; in total, three null hypotheses were tested for each plot; the combined Anderson-Darling k-sample test was employed. This multiple

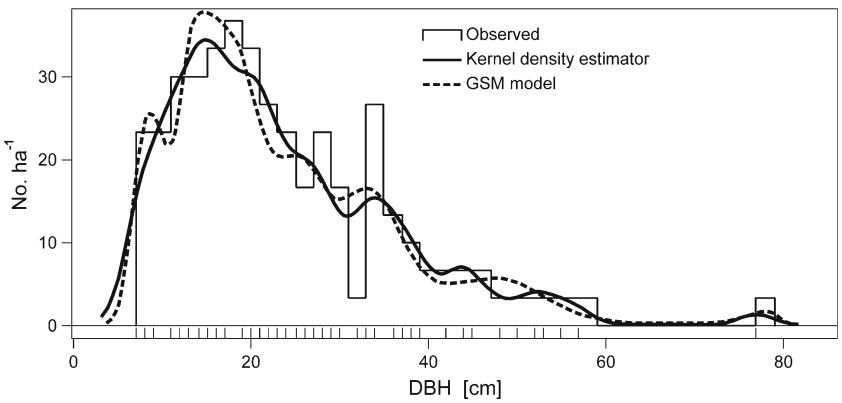

Fig. 4 Approximation of the empirical DBH distribution of an example stand from the group UID using the kernel density estimator and the GSM model (plot No. UID03)

procedure combines several independent k-sample Anderson-Darling tests into one overall test. If AD_i is the Anderson-Darling criterion for the ith block of k_i samples, its standardised test statistic is (Scholz and Zhu 2016)

$$T.AD_i = \frac{AD_i - \mu_i}{\sigma_i} \tag{16}$$

with μ_i and σ_i representing the mean and standard deviation of AD_i. The combined Anderson-Darling criterion is (Scholz and Zhu 2016)

$$AD_{\text{comb}} = \sum_{i=1}^{M} AD_i \tag{17}$$

and

$$T.AD_{\text{comb}} = \frac{AD_{\text{comb}} - \mu_c}{\sigma_c} \tag{18}$$

where

$$\mu_c = \sum_{i=1}^{M} \mu_i \tag{19}$$

$$\sigma_c = \sqrt{\sum_{i=1}^{M} \sigma_i^2} \tag{20}$$

and M is the number of blocks ($M = 2$).

These statistical analyses enabled the assessment of the level of homogeneity within drawn samples (Jamshidian and Jalal 2010). The k-sample Anderson-Darling tests do not require the user to assume that each analysed group belongs to a normal population and has the same variance. In all the cases, the first version of the Anderson-Darling test statistic was computed (for detailed information, see Scholz and Stephens 1987).

For each generated set of 50 samples, the fraction of samples with DBHs >70 cm was calculated. These fractions allowed assessment of the suitability of the two investigated methods in generating random DBH data from the GSM model; the main assessment criterion was the occurrence of trees of an older generation (characterised by DBH >70 cm).

Computational procedures were implemented using the statistical software R (R Core Team 2015); the *GSM* and the *kSamples* packages of R were also used (Venturini 2015; Scholz and Zhu 2016).

4 Results

In all plots, one to three trees representing the older generation (DBH exceeding 70 cm) were present. In three plots, the DBH of the thickest trees reached 100 cm. Trees of a DBH lower than 50 cm represented from 89 to 99% of all the trees in the investigated plots, whereas those with a DBH lower than 25 cm accounted for 46 to 74% of all the trees. The number of trees varied from 215 to 935 N ha^{-1}. The mean skewness for the plots was 1.3276. Generally, investigated DBH distributions are highly skewed and heavy-tailed (Figs. 2, 3 and 4).

The GSM model consists of 250 single gamma functions ($J = 250$). Each of these functions has a particular mixture weight ($\pi_1, ..., \pi_J$). The sum of all the mixture weights for a given model is equal to 1. The sums of 50-length intervals of mixture weights reflect the approximate distribution of these proportions (Table 1). In the plots, DBH distributions are asymmetrical and that is why mixture weight distributions also have longer-than-normal right tails. The mean sums of the 50-length intervals of mixture weights for the investigated plots varied from 0.478732 to 0.0111045 (from left to right; Table 1).

The B_{DIF} and A_{DIF} statistics compare the bias and the flexibility of the GSM model and the kernel density estimation (negative numbers show that the GSM model is 'better'). The B_{DIF} values were higher than zero in the case of all the 20 investigated plots (range 0.009–0.295), while the A_{DIF} values were lower than zero for 14 plots and higher than zero for 6 plots (from −0.619 to 0.218) (Table 1). The values of the calculated statistics indicate that the bias was lower for the kernel density estimation, while the GSM model was characterised by greater flexibility.

A desirable method of random variates generation must include various criteria, especially precision. For precise criterion p value parameters based on the Anderson-Darling k-sample test were calculated (Table 2).

1. With the MDGR procedure—from 0.0109 to 0.9253 for samples of 100 DBHs, from 0.0719 to 0.9798 for samples of 250 DBHs and from 0.0172 to 0.8839 for samples of 500 DBHs
2. With the MH method—from 0.0001 to 0.8791 for samples of 100 DBHs, from 0.0001 to 0.7821 for samples of 250 DBHs and from 0.0001 to 0.8897 for samples of 500 DBHs

In terms of precision, the MDGR procedure provides higher p values than the MH method, but the differences are small (Table 2). Therefore, the MDGR procedure is slightly more precise than the MH method. This is especially so in the case of the samples of 250 DBHs. The presented results are confirmed by the combined Anderson-Darling k-sample test (Table 3). The p values were from 0.0001 to 0.9036 for samples of 100 DBHs, from 0.0022 to 0.9964 for samples of 250 DBHs and from 0.0001 to 0.8838 for samples of 500 DBHs (Table 3). The high p values show that the level of the homogeneity within drawn DBH sets was similar for all generated

Table 1 Sums of mixture weights ($\pi_1, ..., \pi_J$; $J = 250$), scale parameter ($1/\theta$) and goodness-of-fit statistics (B_{DIF}, A_{DIF}) for the gamma shape mixture (GSM) model

Plot	Sum of mixture weights $\sum_j \pi_j$					$1/\theta$	B_{DIF}	A_{DIF}
	From j = 1 to j = 50	From j = 51 to j = 100	From j = 101 to j = 150	From j = 151 to j = 200	From j = 201 to j = 250			
RS01	0.67689	0.19323	0.10282	0.01428	0.01278	0.400	0.207	–0.301
RS02	0.64994	0.20898	0.10591	0.01632	0.01885	0.399	0.199	–0.280
RS03	0.67065	0.21081	0.08248	0.02813	0.00793	0.394	0.211	–0.533
RS04	0.64571	0.16443	0.13710	0.03988	0.01288	0.351	0.202	–0.425
RS05	0.68832	0.16493	0.10556	0.01542	0.02576	0.364	0.288	–0.398
RS06	0.44696	0.48711	0.04808	0.00634	0.01152	0.365	0.255	–0.397
RS07	0.43762	0.51367	0.04080	0.00116	0.00675	0.376	0.119	–0.275
RS08	0.39801	0.39664	0.19245	0.00317	0.00972	0.309	0.223	–0.367
RS09	0.56332	0.31142	0.11604	0.00139	0.00783	0.363	0.247	–0.619
RS10	0.48905	0.36658	0.13131	0.00619	0.00688	0.322	0.295	–0.485
BMS01	0.47679	0.20201	0.27207	0.04256	0.00657	0.356	0.040	–0.048
BMS02	0.35270	0.21065	0.29226	0.13513	0.00927	0.324	0.010	–0.041
BMS03	0.40667	0.30650	0.24149	0.03985	0.00548	0.365	0.010	0.183
BMS04	0.33827	0.26670	0.26833	0.10116	0.02554	0.308	0.032	–0.010
BMS05	0.39048	0.33637	0.19457	0.07229	0.00628	0.354	0.009	0.218
UID01	0.48104	0.33537	0.15545	0.01402	0.01412	0.402	0.016	0.047
UID02	0.38603	0.43741	0.15317	0.01523	0.00815	0.361	0.009	–0.055
UID03	0.37308	0.36297	0.19275	0.06103	0.01017	0.309	0.058	0.064
UID04	0.34382	0.50031	0.14395	0.00434	0.00759	0.397	0.037	0.053
UID05	0.35929	0.46679	0.10878	0.05712	0.00802	0.324	0.020	0.015

samples and for all the three groups of DBH distributions (RS, BMS and UID; Table 3).

The greatest fractions for generated samples containing DBHs >70 cm were achieved for the MDGR procedure in the case of simulations of 500 DBHs in a sample (maximal fraction was equal 1.00 for ten plots; Table 4). The smallest fractions were obtained for the MH method in the case of simulations of 100 DBHs in a sample (maximal fraction was equal 0.96 for one plot; Table 4).

The simulations that were carried out showed that both of the investigated methods are capable of simulating the DBH data from the GSM model, but the MDGR procedure was slightly more effective than the MH method.

5 Discussion

For generating the DBH data sets from the GSM model, one can use the MDGR procedure and, to a lesser degree, the MH method, preferably to simulate large sets containing e.g. 500 DBHs. In the case of smaller sets, it is always necessary to check if within the generated data there are DBHs representing trees from an older generation. A similar procedure can be used in all stands, in which one of the tree generations is represented but by few trees.

This paper has compared two methods for generating random DBH data from the GSM model fed with real data from forests with fir and beech in one geographical region. Future research can be concerned with forests consisting of different species and growing in different regions.

The GSM model is very flexible and thus it allows precise approximation of irregular data sets with local extremes. Increasing the value of J, we can increase the precision of the approximation but this may cause numerical problems. If we want to include empirical irregularity in the GSM models, then we should increase the value of J but if multimodality is random, then we should decrease the value of J. In the case of existence of specific subpopulations, it is desirable to use mixture models, in which component densities represent these subpopulations (Podlaski and Roesch 2014). However, it is necessary to remember that mixture models are not very useful where there is a significant difference in the number of elements constituting the subpopulations, as exemplified by highly skewed and heavy-tailed distributions in which one of the subpopulations forms the distribution tail. The very small number of elements of this subpopulation usually makes

Table 2 The p values for the Anderson-Darling k-sample test comparing DBH distributions within drawn DBH samples

Plot	Samples of 100 DBHs		Samples of 250 DBHs		Samples of 500 DBHs	
	MDGR procedure[a]	MH method[b]	MDGR procedure	MH method	MDGR procedure	MH method
RS01	0.9253	0.0297	0.8078	0.1976	0.3301	0.0009
RS02	0.0582	0.0038	0.9798	0.0037	0.6010	0.0291
RS03	0.6648	0.0001	0.7463	0.0001	0.0924	0.0001
RS04	0.8603	0.0305	0.1516	0.0577	0.8856	0.0047
RS05	0.4721	0.0031	0.8919	0.0001	0.0263	0.0006
RS06	0.4708	0.5851	0.8513	0.4444	0.0935	0.2412
RS07	0.4452	0.1675	0.6376	0.0951	0.5768	0.8390
RS08	0.3950	0.0810	0.2444	0.2489	0.7798	0.8061
RS09	0.1200	0.0070	0.0719	0.0118	0.0330	0.2469
RS10	0.4200	0.2767	0.6434	0.0179	0.1769	0.8099
BMS01	0.6946	0.8743	0.9177	0.1956	0.1090	0.4082
BMS02	0.5359	0.0172	0.7526	0.1888	0.7488	0.5131
BMS03	0.5717	0.6526	0.4623	0.2471	0.5513	0.5977
BMS04	0.5778	0.8791	0.3161	0.5320	0.7599	0.2769
BMS05	0.2923	0.6314	0.9241	0.7821	0.8939	0.0861
UID01	0.2080	0.7176	0.1280	0.5187	0.2531	0.0930
UID02	0.1098	0.7715	0.5172	0.0355	0.7183	0.1302
UID03	0.2167	0.0469	0.8524	0.4696	0.2260	0.8897
UID04	0.3961	0.8530	0.2067	0.3511	0.3385	0.4795
UID05	0.0109	0.1558	0.9657	0.1492	0.0172	0.5158

[a] Gamma random numbers were generated in multinomial distribution cells using the acceptance-rejection principle with proper choice of the majorisation function (when the shape parameter was less than 1) or as the sum of two independent gamma variates (when the shape parameter was greater than or equal to 1) (Ahrens and Dieter methods; for detailed information, see Ahrens and Dieter 1974, 1982)

[b] The standard Metropolis–Hastings algorithm with jumping normal distribution was used (Robert and Casella 2004)

impossible to associate the component of the mixture model with the subpopulation.

Fir and beech trees from the older generation usually create only small local DBH maxima within a lower threshold of over 70 cm. This kind of highly skewed and heavy-tailed distribution is correctly approximated by the GSM model. The precision of the GSM model was comparable to the approximation precision obtained with the use of the kernel density estimation. This is a very interesting result because the kernel density estimation is characterised by high flexibility (e.g. Buch-Larsen et al. 2005; Podlaski and Roesch 2014).

The problem of highly skewed and heavy-tailed distributions can be circumvented by data transformations. Procedures of this kind are used, among others, in the analysis of variance and in the regression models (Box-Cox, etc.). However, there are some possible drawbacks of these methods (Garay et al. 2016): (1) transformations reduce information on the underlying data generation scheme, (2) parameters may lose interpretability on a transformed scale and (3) transformations are usually not universal and often vary with the data set. Hence, in the case of modelling the highly skewed and heavy-tailed DBH distributions, it is necessary to seek flexible theoretical models.

6 Conclusions

This study has revealed that the GSM model is flexible and accurate when modelling the highly skewed and heavy-tailed DBH distributions of two-generation stands. The GSM model precisely separates older and younger tree generations; it is useful in smoothing the small local DBH maxima. A simulation study has shown that the MDGR procedure was slightly more precise than the MH method. The DBH random variates, generated with the use of these methods from the GSM model, represented all tree generations that are significant from a biological point of view. The high structural diversity of patches of natural, near-natural and managed forests, especially with shade-tolerant species, should stimulate further research related to the analysis of empirical DBH distributions in the context of the GSM model.

Table 3 The *p* values for the combined Anderson-Darling *k*-sample test comparing DBH distributions within drawn DBH samples grouped in two blocks; DBHs drawn using the MDGR procedure were grouped in the first block and DBHs drawn using the MH method were grouped in the second block

Plot	Samples of 100 DBHs	Samples of 250 DBHs	Samples of 500 DBHs
RS01	0.3108	0.4980	0.0046
RS02	0.0015	0.1982	0.1135
RS03	0.0001	0.0022	0.0002
RS04	0.2442	0.0356	0.1094
RS05	0.0187	0.0095	0.0001
RS06	0.5551	0.7420	0.0815
RS07	0.2261	0.2445	0.8061
RS08	0.1229	0.1769	0.8838
RS09	0.0053	0.0047	0.0386
RS10	0.3001	0.0964	0.4761
BMS01	0.8851	0.6236	0.1553
BMS02	0.0687	0.4404	0.7016
BMS03	0.9036	0.3028	0.6201
BMS04	0.8380	0.4023	0.5346
BMS05	0.4505	0.9964	0.4287
UID01	0.4340	0.2244	0.0851
UID02	0.3522	0.1016	0.3428
UID03	0.0435	0.7581	0.6164
UID04	0.7125	0.2086	0.3848
UID05	0.0101	0.6657	0.0647

Table 4 Fractions of the DBHs >70 cm, assessing the frequency of trees representing the older generation in the sample; in all the investigated plots, the older generation was composed of trees with DBHs above 70 cm

Plot	Samples of 100 DBHs		Samples of 250 DBHs		Samples of 500 DBHs	
	MDGR procedure[a]	MH method[b]	MDGR procedure	MH method	MDGR procedure	MH method
RS01	0.86	0.80	0.96	0.94	1.00	1.00
RS02	0.86	0.92	0.98	1.00	1.00	1.00
RS03	0.84	0.78	1.00	0.96	1.00	1.00
RS04	0.74	0.66	0.96	0.96	1.00	1.00
RS05	0.96	0.96	0.98	1.00	1.00	1.00
RS06	0.68	0.64	1.00	0.88	1.00	1.00
RS07	0.54	0.34	0.76	0.68	1.00	0.96
RS08	0.52	0.30	0.88	0.62	0.98	0.84
RS09	0.58	0.46	0.82	0.76	0.98	0.96
RS10	0.50	0.28	0.84	0.70	0.90	0.92
BMS01	0.54	0.50	0.82	0.66	0.98	0.90
BMS02	0.58	0.44	0.82	0.76	0.98	0.98
BMS03	0.64	0.48	0.86	0.86	0.98	0.96
BMS04	0.52	0.44	0.72	0.72	0.96	0.88
BMS05	0.42	0.32	0.80	0.60	0.98	0.96
UID01	0.84	0.84	1.00	1.00	1.00	1.00
UID02	0.60	0.28	0.94	0.76	1.00	0.88
UID03	0.54	0.30	0.78	0.48	0.98	0.80
UID04	0.56	0.48	0.92	0.84	1.00	0.96
UID05	0.62	0.26	0.76	0.58	0.98	0.80

[a,b] For the characteristics of the MDGR procedure and the MH method, refer to Table 2

Acknowledgments The author wishes to thank the editor and the referees for valuable and pertinent comments.

Compliance with ethical standards

Funding This work was supported by the Polish Ministry of Science and Higher Education (public resources dedicated to research in 2010–2013, Grant No. N N309 044138).

References

Ahrens JH, Dieter U (1974) Computer methods for sampling from gamma, beta, Poisson and binomial distributions. Computing 12:223–246

Ahrens JH, Dieter U (1982) Generating gamma variates by a modified rejection technique. Commun ACM 25:47–54

Buch-Larsen T, Nielsen JP, Guillen M, Bolance C (2005) Kernel density estimation for heavy-tailed distribution using the Champernowne transformation. Statistics 39:503–518. doi:10.1080/02331880500439782

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38

Denslow JS, Ellison AM, Sanford RE (1998) Treefall gap size effects on above- and below-ground processes in a tropical wet forest. J Ecol 86:597–609

Diebolt J, Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. J R Stat Soc Ser B 56:363–375

Garay AM, Lachos VH, Lin TI (2016) Nonlinear censored regression models with heavy-tailed distributions. Stat Interface 9:281–293

Gehringer KR, Turnblom EC (2014) Constructing a virtual forest: using hierarchical nearest neighbor imputation to generate simulated tree lists. Can J For Res 44:711–719. doi:10.1139/cjfr-2014-0020

Gove JH, Ducey MJ, Leak WB, Zhang L (2008) Rotated sigmoid structures in managed uneven-aged northern hardwood stands: a look at the Burr type III distribution. Forestry 81:161–176. doi:10.1093/forestry/cpm025

Gratzer G, Canham CD, Dieckmann U, Fischer A, Iwasa Y, Law R, Lexer MJ, Spies T, Splechtna B, Szwagrzyk J (2004) Spatio-temporal development of forests—current trends in field methods and models. Oikos 107:3–15. doi:10.1111/j.0030-1299.2004.13063.x

Jamshidian M, Jalal S (2010) Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. Psychometrika 75:649–674. doi:10.1007/s11336-010-9175-3

Jasra A, Holmes CC, Stephens DA (2005) Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Stat Sci 20:50–67. doi:10.1214/088342305000000016

Kong A, Liu JS, Wong WH (1994) Sequential imputations and Bayesian missing data problems. J Am Stat Assoc 89:278–288

Kou SC, Zhou Q, Wong WH (2006) Equi-energy sampler with applications in statistical inference and statistical mechanics (with discussion). Ann Stat 34:1581–1619. doi:10.1214/009053606000000515

Lawton RO, Putz FE (1988) Natural disturbance and gap-phase regeneration in a wind-exposed tropical cloud forest. Ecology 69:764–777

Lehmann EL, Casella G (1998) Theory of point estimation. Springer, New York

Liu JS (2001) Monte Carlo strategies in scientific computing. Springer, New York

Lopez-de-Ullibarri I (2015) Bandwidth selection in kernel distribution function estimation. Stata J 15:784–795

MacEachern SN, Clyde M, Liu JS (1999) Sequential importance sampling for nonparametric Bayes models: the next generation. Can J Stat 27:251–267. doi:10.2307/3315637

Matuszkiewicz JM (2008) Zespoły leśne Polski. Państwowe Wydawnictwo Naukowe, Warszawa

Merganič J, Sterba H (2006) Characterisation of diameter distribution using the Weibull function: method of moments. Eur J For Res 125:427–439. doi:10.1007/s10342-006-0138-2

Podlaski R (2008) Dynamics in central European near-natural *Abies–Fagus* forests: does the mosaic-cycle approach provide an appropriate model? J Veg Sci 19:173–182. doi:10.3170/2008-8-18350

Podlaski R (2011a) Modelowanie rozkładów pierśnic drzew z wykorzystaniem rozkładów mieszanych I. Definicja, charakterystyka i estymacja parametrów rozkładów mieszanych. Sylwan 155(4):244–252

Podlaski R (2011b) Modelowanie rozkładów pierśnic drzew z wykorzystaniem rozkładów mieszanych II. Aproksymacja rozkładów pierśnic w lasach wielopiętrowych. Sylwan 155(5):293–300

Podlaski R, Roesch FA (2014) Modelling diameter distributions of two-cohort forest stands with various proportions of dominant species: a two-component mixture model approach. Math Biosci 249:60–74. doi:10.1016/j.mbs.2014.01.007

Podlaski R, Żelezik M (2012) Ocena kondycji modrzewia *Larix decidua* Mill. subsp. *polonica* (Racib.) Domin i innych gatunków drzew na Chełmowej Górze w Świętokrzyskim Parku Narodowym. Sylwan 156(3):170–181

Porté A, Bartelink HH (2002) Modelling mixed forest growth: a review of models for forest management. Ecol Model 150:141–188. doi:10.1016/s0304-3800(01)00476-8

Pretzsch H (2010) Forest dynamics, growth and yield. From measurement to model. Springer, Berlin

R Development Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna Available from www.R-project.org

Robert CP (1996) Mixtures of distributions: inference and estimation. In: Gilks WR, Richardson S, Spiegelhalter DJ (eds) Markov chain Monte Carlo in practice. Chapman and Hall/CRC, New York, pp 441–464

Robert CP, Casella G (2004) Monte Carlo statistical methods. Springer, New York

Roesch FA, Coulston JW, van Deusen PC, Podlaski R (2015) Evaluation of image-assisted forest monitoring: a simulation. Forests 6:2897–2917. doi:10.3390/f6092897

Scholz FW, Stephens MA (1987) *K*-sample Anderson-Darling tests. J Am Stat Assoc 82:918–924

Scholz FW, Zhu A (2016) kSamples: k-sample rank tests and their combinations. R package version 1:2–3 http://CRAN.R-project.org/package=kSamples. Accessed 1 March 2016

Thompson JR (2000) Simulation: a modeler's approach. John Wiley & Sons, New York

Venturini S (2015) GSM: gamma shape mixture. R package version 1(3):2 http://CRAN.R-project.org/package=GSM. Accessed 1 March 2016

Venturini S, Dominici F, Parmigiani G (2008) Gamma shape mixtures for heavy-tailed distributions. Ann Appl Stat 2:756–776. doi:10.1214/07-AOAS156

Zasada M (2013) Evaluation of the double normal distribution for tree diameter distribution modeling. Silv Fenn 47,id 956:**17 p**. doi: 10.14214/sf.956

Zasada M, Cieszewski CJ (2005) A finite mixture distribution approach for characterizing tree diameter distributions by natural social class in pure even-aged Scots pine stands in Poland. For Ecol Manag 204:145–158. doi:10.1016/j.foreco.2003.12.023

Zhang LJ, Gove JH, Liu C, Leak WB (2001) A finite mixture of two Weibull distributions for modeling the diameter distributions of rotated-sigmoid, uneven-aged stands. Can J For Res 31:1654–1659. doi:10.1139/x01-086