SYLWAN 155 (4): 244-252, 2011

RAFAŁ PODLASKI

Modelowanie rozkładów pierśnic drzew z wykorzystaniem rozkładów mieszanych I. Definicja, charakterystyka i estymacja parametrów rozkładów mieszanych

Modelling tree diameter distributions using mixture models
I. Definition, characteristics and parameters estimation
of mixtures distributions

ABSTRACT

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

The article presents an introduction to the theory of finite mixture distributions, discusses the ways of procedure in selecting the number and type of component distributions, the methods of assessing the initial values, and proposal of the procedure of estimating the parameters. The proposed procedure uses min.k/max.k (for k=1, 3, 6) and 0.5/1.5/average methods to select initial values for a numerical procedure (EM algorithm + Newton's method) allowing to calculate the extremum of the likelihood function. If at least two identical solutions for the extreme values are not obtained, the multistart method should additionally be applied.

KEY WORDS

finite mixture models, tree diameter distribution, maximum likelihood estimation, initial values, starting strategy

ADDRESSES

Rafał Podlaski – e-mail: r_podlaski@pro.onet.pl

Pracownia Ochrony Przyrody; Uniwersytet Jana Kochanowskiego; ul. Świętokrzyska 15; 25-406 Kielce

Wstęp

Modelowanie rozkładów pierśnic za pomocą różnych rozkładów teoretycznych ma duże znaczenie dla teorii i praktyki leśnej. Rozkłady pierśnic są wykorzystywane m.in. do konstruowania modeli wzrostu drzew oraz modeli struktury i budowy drzewostanów [Bruchwald 1999]. Analiza tego typu modeli pozwala na zrozumienie między- i wewnątrzgatunkowych oraz między- i wewnątrzgeneracyjnych zależności, co jest podstawą konstruowania i weryfikowania modeli opisujących dynamikę m.in. wielogatunkowych i wielogeneracyjnych drzewostanów [Siekierski 1991; Zasada 1995, 2000; Poznański 1997; Pretzsch 1997, 1998; Mason 2000].

Badając stopień zgodności rzeczywistych i teoretycznych rozkładów pierśnic można wykorzystać pojedyncze rozkłady, ale często zachodzi konieczność zastosowania bardziej złożonych modeli [Maltamo, Kangas 1998; Zhang i in. 2001; Liu i in. 2002]. Tego typu sytuacja występuje, jeżeli (1) rzeczywisty rozkład pierśnic jest dwu- lub wielomodalny oraz gdy (2) rzeczywisty rozkład pierśnic składa się z dwóch lub większej liczby różnych grup, które chcemy zidentyfikować (np. różnych gatunków drzew, generacji wiekowych czy klas biosocjalnych). Wówczas można zastosować rozkłady mieszane, będące mieszaniną dwóch lub większej liczby rozkładów

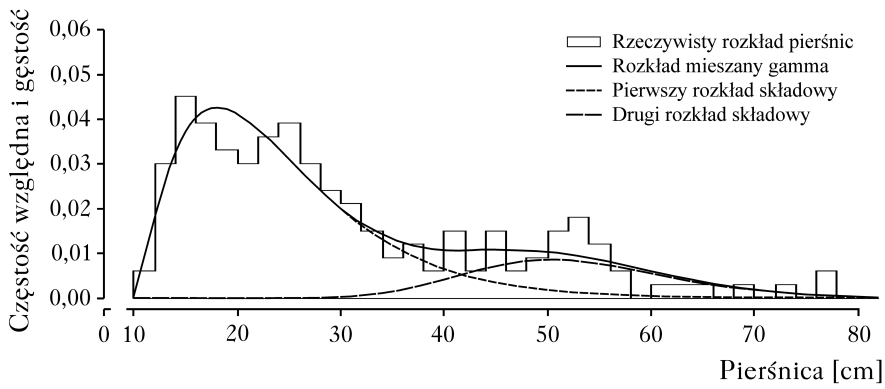
teoretycznych (np. rozkładów normalnych, Weibulla lub gamma). Jednym z podstawowych problemów występujących podczas obliczania parametrów rozkładów mieszanych jest ustalenie wartości startowych, od których najczęściej stosowany algorytm EM (ang. expectation maximisation) połączony z metodą Newtona rozpoczyna estymacje parametrów [Böhning 2000; McLachlan, Krishnan 2008]. Modele mieszane były stosowane przede wszystkim do aproksymacji rozkładów pierśnic w drzewostanach wielogatunkowych i wielogeneracyjnych [Zhang i in. 2001; Liu i in. 2002; Zhang, Liu 2006] oraz do identyfikacji różnych grup drzew [Hessenmoller, von Gadow 2001; Zucchini i in. 2001]. W Polsce modele mieszane zostały po raz pierwszy zastosowane w naukach leśnych przez Siekierskiego [1991] i Zasadę [2003, 2005].

Celem pracy jest (1) wprowadzenie do teorii rozkładów mieszanych, ze szczególnym uwzględnieniem różnych metod określania wartości startowych procedury numerycznej (algorytm EM oraz metoda Newtona) wykorzystywanej do estymacji parametrów rozkładów mieszanych, (2) przedyskutowanie sposobów postępowania podczas wyboru liczby i rodzaju rozkładów składowych oraz metod określania wartości startowych, a także (3) zaproponowanie procedury estymacji parametrów rozkładów mieszanych.

Podstawy teorii rozkładów mieszanych

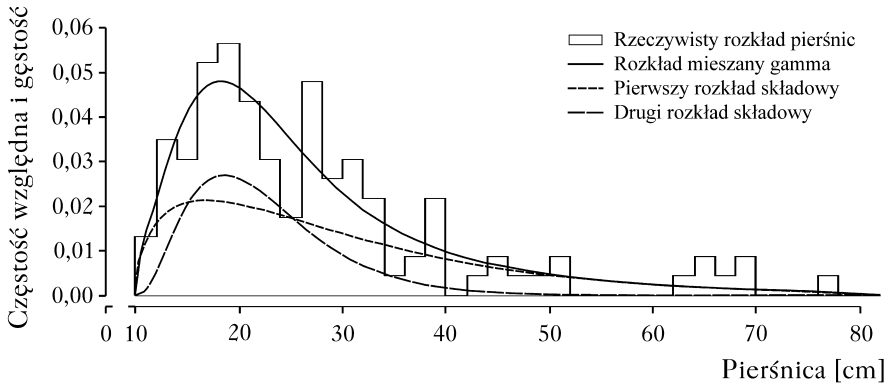
Informacje wstępne. Rozkłady mieszane składają się z dwóch lub większej liczby rozkładów składowych, zachodzących na siebie częściowo (ryc. 1) lub całkowicie (ryc. 2) i występujących w pewnej proporcji (określonej przez wagi, czyli frakcje). Suma rozkładów składowych tworzy rozkład mieszany. Stosując rozkłady mieszane należy w pierwszym etapie określić liczbę rozkładów składowych oraz wybrać określony rodzaj rozkładu teoretycznego, który będzie tworzył rozkłady składowe (może to być np. rozkład normalny, Weibulla lub gamma). W drugim etapie należy estymować zestaw wszystkich parametrów rozkładu mieszanego, dla którego uzyskuje się najlepsze dopasowanie do rozkładu rzeczywistego. Zestaw wszystkich parametrów rozkładu mieszanego zawiera parametry poszczególnych rozkładów składowych oraz proporcję, w jakiej jego komponenty występują.

Definicja i charakterystyka rozkładów mieszanych. Załóżmy, że zmienna losowa X jest określona w przestrzeni X oraz że jej rozkład losowy jest reprezentowany przez funkcję gęstości postaci:



Ryc. 1.

Rozkład mieszany złożony z dwóch rozkładów składowych gamma częściowo zachodzących na siebie
Mixture distribution composed by two gamma component distributions overlapping partly



Ryc. 2.
Rozkład mieszany złożony z dwóch rozkładów składowych gamma całkowicie zachodzących na siebie
Mixture distribution composed by two gamma component distributions overlapping completely

$$f_X(x \mid \psi) = \sum_{i=1}^{k} \pi_i f_i(x \mid \theta_i), \quad x \in X \qquad [1]$$

przy czym:

$$\psi = \begin{pmatrix} \pi_1 & \pi_2 & \dots & \pi_k \\ \theta_1^T & \theta_2^T & \dots & \theta_k^T \end{pmatrix}$$

gdzie

$i=1, 2,..., k;\ 0 \le \pi_i \le 1;\ \sum_{i=1}^{k} \pi_i = 1$

Możemy powiedzieć, że zmienna losowa X ma rozkład mieszany złożony z k rozkładów składowych oraz że $f_x(\cdot)$ jest funkcją gęstości rozkładu mieszanego, π_i są to wagi (frakcje), a $f_i(\cdot)$ – funkcje gęstości poszczególnych składników tworzących rozkład mieszany. θ_i oznacza parametry rozkładu $f_i(\cdot)$, a ψ – zestaw wszystkich parametrów rozkładu mieszanego. Indeks ^{T} oznacza macierz transponowaną.

Zgodnie z przedstawioną definicją funkcję gęstości rozkładu mieszanego, np. dla rozkładu gamma (załóżmy, że ten rozkład teoretyczny przyjmujemy za rozkład składowy) i dwóch rozkładów składowych (załóżmy, że stosujemy mieszaninę dwóch rozkładów), możemy zapisać jako:

$$f_{(gam)X}(x \mid \psi) = \sum_{i=1}^{2} \pi_i f_{(gam)i}(x \mid \theta_i) = \pi_1 f_{(gam)1}(x \mid \theta_1) + \pi_2 f_{(gam)2}(x \mid \theta_2) \qquad [2]$$

gdzie:

$f_{(gam)i}(x \mid \theta_i) = f_{(gam)i}(x \mid \alpha_i, \beta_i, \gamma_i) = \dfrac{\beta_i^{\alpha_i}(x-\gamma_i)^{\alpha_i - 1}}{\Gamma(\alpha_i)} e^{-\beta_i(x-\gamma_i)}$ – funkcje gęstości rozkładów składowych gamma (α_i – parametr kształtu, β_i – parametr skalujący, γ_i – parametr przesunięcia, $\Gamma(\cdot)$ – funkcja gamma).

Przedstawiony rozkład mieszany jest określony przez osiem parametrów, z czego sześć (α_1, β_1, γ_1, α_2, β_2 γ_2) – określa rozkłady składowe, a dwa (π_1, π_2) charakteryzują udział składników rozkładu mieszanego.

Każda funkcja gęstości rozkładu mieszanego jest całkowalna, a ponadto spełnia następujące warunki:

- $f_X(x \mid \psi) \geq 0$
- $\int_{-\infty}^{+\infty} f_X(x \mid \psi) dx = 1$

Estymacja parametrów rozkładów mieszanych. Standardową procedurą estymacji parametrów rozkładów mieszanych jest metoda największej wiarygodności (MLE; ang. maximum likelihood estimation), wykorzystująca do znalezienia maksimum funkcji wiarygodności algorytm EM połączony z metodą Newtona [Böhning 2000]. Funkcja wiarygodności dla danych pogrupowanych (dla wyróżnionych stopni grubości) przedstawiona jest w postaci [Du 2002]:

$$lL_1(\psi) = \sum_{j=1}^{l} n_j \log P_j(\psi) \qquad [3]$$

lub

$$lL_2(\psi) = -2 \sum_{j=1}^{l} n_j \log \left(\frac{P_j(\psi)}{O_j} \right), \qquad O_j = \frac{n_j}{N} \qquad [4]$$

gdzie:

$P_j(\psi)$ – teoretyczne prawdopodobieństwo zdarzenia, że dana pierśnica należy do j-tego stopnia grubości,
O_j – empiryczna częstość względna dla j-tego stopnia,
l – liczba stopni.

Znalezienie maksimum funkcji wiarygodności $lL_1(\psi)$ jest równoznaczne ze znalezieniem minimum funkcji wiarygodności $lL_2(\psi)$ i pozwala na estymację parametrów rozkładu mieszanego ψ. Funkcja wiarygodności ma na ogół bardzo skomplikowaną, wielowymiarową postać i dlatego znalezienie ekstremum globalnego wymaga zastosowania metod numerycznych. Najczęściej stosowaną metodą numeryczną jest algorytm EM połączony z metodą Newtona [McLachlan, Peel 2000]. Zbieżność procesu iteracyjnego oraz szybkość obliczeń zależy w dużym stopniu od przyjętych wartości startowych procedury numerycznej. Im wartości startowe są bardziej zbliżone do wartości, dla których funkcja wiarygodności osiąga ekstremum globalne, tym większe jest prawdopodobieństwo zbieżności procesu iteracyjnego oraz tym większa jest szybkość obliczeń. Precyzyjne wyznaczenie wartości startowych ma szczególne znaczenie w przypadku estymacji parametrów dla rozkładów mieszanych składających się z rozkładów składowych całkowicie na siebie zachodzących (ryc. 2).

Estymując parametry należy, o ile jest to możliwe, funkcje gęstości rozkładów składowych przedstawić w formie, dla której wartości startowe mogą być ustalone na podstawie wartości aproksymowanych danych rzeczywistych [Du 2002; McLachlan, Krishnan 2008]. Jest to procedura optymalna, ponieważ dla tej postaci funkcji gęstości rozkładów składowych wartości startowe są generowane ze zbioru danych rzeczywistych. Parametrami rozkładu normalnego są średnia i odchylenie standardowe (μ_i, σ_i), a parametrami rozkładu Weibulla i gamma – parametr kształtu i parametr skalujący (α_i, β_i). Dlatego na podstawie analizy histogramu danych rzeczywistych można generować wartości startowe dla normalnych rozkładów składowych (czyli dla μ_i, σ_i), natomiast nie jest to możliwe dla rozkładów składowych Weibulla i gamma (określonych na podstawie α_i, β_i). W przypadku tych dwóch ostatnich, parametry α_i, β_i można zastąpić parametrami μ_i, σ_i (dla rozkładu Weibulla i gamma znając α_i, β_i można obliczyć μ_i, σ_i i odwrotnie). Po tym przekształceniu również dla rozkładu Weibulla i gamma można generować wartości startowe ze zbioru danych rzeczywistych. Z tego względu w programach obliczających parametry

rozkładów mieszanych dla rozkładów składowych Weibulla i gamma jako wartości startowe wprowadza się przybliżone wartości μ_i, σ_i, a nie α_i, β_i [Haughton 1997]. W tej sytuacji wektor parametrów $\theta_i=(\alpha_i, \beta_i)$ jest przedstawiany jako $\theta_i^*=(\mu_i=f_\mu(\alpha_i, \beta_i), \sigma_i=f_\sigma(\alpha_i, \beta_i))$, gdzie f_μ i f_σ są funkcjami pozwalającymi na obliczenie μ_i, σ_i przy wykorzystaniu wartości parametrów α_i, β_i [Du 2002]. I tak np. dla rozkładu mieszanego złożonego z rozkładów składowych gamma (ryc. 1) odpowiednie wektory parametrów przyjmują następujące wartości: $\theta_1=(\alpha_1=2{,}3306; \beta_1=6{,}0499)$, $\theta_2=(\alpha_2=18{,}6662; \beta_1=2{,}2854)$ czyli $\theta_1^*=(\mu_1=14{,}10; \sigma_1=9{,}236)$, $\theta_2^*=(\mu_2=42{,}66; \sigma_2=9{,}874)$. Natomiast udział rozkładów składowych wynosi $\pi_1=0{,}7929$; $\pi_2=0{,}2071$, a z kolei parametr $\gamma_1=\gamma_2=9{,}9$.

Dla rozkładów mieszanych złożonych z dwóch rozkładów składowych najczęściej stosowane są następujące metody określania wartości startowych [Böhning i in. 1994; Seidel i in. 2000a, b; Du 2002; Zasada, Cieszewski 2005; Podlaski 2010]:

– min.*k*/max.*k* – algorytm numeryczny rozpoczyna estymację parametrów od wartości startowych μ_1^0=min.*k*; $\sigma_1^0=s$ i $\pi_1^0=0{,}5$ dla pierwszego rozkładu składowego oraz μ_2^0=max.*k*; $\sigma_2^0=s$ i $\pi_2^0=0{,}5$ dla drugiego rozkładu składowego (min.*k*, max.*k*, *s* są to odpowiednio najmniejsza i największa *k*-ta pierśnica oraz odchylenie standardowe wszystkich pierśnic zbioru danych rzeczywistych). Za min.*k* i max.*k* przyjmowana jest najpierw min.1 i max.1, czyli pierśnica najmniejsza i największa, następnie min.3 i max.3, czyli pierśnica trzecia „od dołu" i trzecia „od góry" i wreszcie min.6 i max.6, czyli pierśnica szósta „od dołu" i szósta „od góry". Po wykonaniu trzech estymacji za maksimum globalne przyjmujemy wartość, która była największa dla funkcji wiarygodności $lL_1(\psi)$ i która przynajmniej dwukrotnie wystąpiła w zbiorze wyników obliczeń.
– 0,5/1,5/średnia – estymacja parametrów rozpoczyna się od wartości startowych $\mu_1^0=0{,}5m$; $\sigma_1^0=s$ i $\pi_1^0=0{,}5$ dla pierwszego rozkładu składowego oraz $\mu_2^0=1{,}5m$; $\sigma_2^0=s$ i $\pi_2^0=0{,}5$ dla drugiego rozkładu składowego (*m* jest to średnia pierśnica obliczona dla całego zbioru danych rzeczywistych).
– średnia/odchylenie standardowe/ustalone — scenariusz ten realizowany jest w dwóch etapach. W pierwszym etapie estymacja parametrów rozpoczyna się od wartości startowych $\mu_1^0=m_{G1}$; $\sigma_1^0=s_{G1}$ i $\pi_1^0=w_{G1}$ dla pierwszego rozkładu składowego oraz $\mu_2^0=m_{G2}$; $\sigma_2^0=s_{G2}$ i $\pi_2^0=w_{G2}$ dla drugiego rozkładu składowego, przy czym wartości m_{G1}, s_{G1}, m_{G2} są ustalone (tzn. nie są zmieniane w trakcie obliczeń). W drugim etapie estymacja parametrów rozpoczyna się od wartości startowych $\mu_1^0=m_{G1}$, $\sigma_1^0=s_{G1}$ i $\pi_1^0=\hat{\pi}_{G1}$ dla pierwszego rozkładu składowego oraz $\mu_2^0=m_{G2}$, $\sigma_2^0=s_{G2}$ i $\pi_2^0=\hat{\pi}_{G2}$ dla drugiego rozkładu składowego, przy czym wartości $\hat{\pi}_{G1}$, $\hat{\pi}_{G2}$ zostały estymowane w pierwszym etapie i teraz są ustalone. W obu etapach *m*, *s* i *w* to odpowiednio średnia, odchylenie standardowe i wagi dla grup *G*1 i *G*2, które chcemy zidentyfikować stosując rozkład mieszany (np. dla różnych gatunków drzew, generacji wiekowych czy klas biosocjalnych). Przybliżone wartości *m*, *s* i *w* określamy (1) na podstawie analizy histogramu danych rzeczywistych i (2) wykorzystując dodatkowe wiadomości o badanym drzewostanie (np. analizując znany z przeprowadzonych badań udział różnych gatunków drzew, generacji wiekowych czy klas biosocjalnych). W metodzie tej najpierw estymowane są wagi i jedno odchylenie standardowe przy ustalonych pozostałych parametrach, a następnie określane są średnie i odchylenia standardowe przy ustalonych wcześniej parametrach. Stosowana jest też procedura „odwrotna" – najpierw estymowane są średnie i odchylenia standardowe przy ustalonych wagach, a następnie wagi i jedno odchylenie standardowe przy ustalonych, wcześniej estymowanych parametrach.

– metoda wielopunktowa – estymacja parametrów rozpoczyna się od wartości startowych $\mu_1^0=u_1$, $\sigma_1^0=s$, $\pi_1^0=0{,}5$ dla pierwszego rozkładu składowego oraz $\mu_2^0=u_2$, $\sigma_2^0=s$, $\pi_2^0=0{,}5$ dla drugiego rozkładu składowego (u_1, u_2 to wartości pierśnicy wygenerowane ze zbioru danych rzeczywistych). Po zakończeniu estymacji dla u_1, u_2 rozpoczyna się ponownie estymację dla u_1, u_3, następnie dla: u_1, u_4; u_1, u_5; ... u_1, u_{10}; ... u_2, u_{10}; ... u_9, u_{10}. Łącznie dla dziesięciu punktów pokrywających przestrzeń danych (u_1, u_2, ... u_{10}) zostaje wykonanych 45 estymacji. Punkty pokrywające przestrzeń danych są wyznaczane w następujących miejscach: $u_1=d_{\min}$, $u_2=d_{R(1\cdot N/9)}$, ..., $u_j=d_{R((j-1)\cdot N/9)}$, ..., $u_{10}=d_{\max}$ (u_1 i u_{10} to odpowiednio najmniejsza i największa pierśnica ze zbioru danych rzeczywistych, $R(\cdot)$ – liczba naturalna (N), otrzymana przez zaokrąglenie liczby rzeczywistej oznaczonej kropką w nawiasie ($\cdot$), $R(\cdot)$ – numer pierśnicy wybranej spośród wszystkich pierśnic uszeregowanych od najmniejszej do największej (np. może to być 8. lub 129. pierśnica), N – liczba wszystkich pierśnic ze zbioru danych rzeczywistych). Po wykonaniu wszystkich estymacji za maksimum globalne przyjmujemy wartość, która była największa dla funkcji wiarygodności $lL_1(\psi)$ i przynajmniej dwukrotnie wystąpiła w zbiorze wyników obliczeń.

W metodzie pierwszej i czwartej szukane parametry estymujemy odpowiednio 3 i 45 razy. Postępujemy w ten sposób, gdyż nie zawsze każda z wyznaczonych wartości startowych prowadzi do znalezienia ekstremum globalnego i obliczenia prawidłowych parametrów rozkładu mieszanego. Powtórzenie procesu estymacji dla różnych wartości startowych pozwala na weryfikację obliczeń i uniknięcie sytuacji, kiedy znajdowane jest ekstremum lokalne lub punkt przegięcia. Zatrzymanie procesu estymacji w tych punktach uniemożliwia estymację prawidłowych parametrów rozkładu mieszanego. Szczególnie użyteczna jest metoda czwarta, ponieważ parametry rozkładu mieszanego są obliczane 45 razy i jeżeli przy różnych wartościach startowych otrzymamy ten sam wynik, to z dużym prawdopodobieństwem możemy stwierdzić, że znaleziono ekstremum globalne. Parametry określone dla maksimum globalnego są szukanymi parametrami rozkładu mieszanego. Teoria rozkładów mieszanych jest szczegółowo przedstawiona w literaturze angielskojęzycznej [McLachlan, Basford 1988; Lindsay 1995; McLachlan, Peel 2000; McLachlan, Krishnan 2008].

Dyskusja

Jednym z podstawowych problemów związanych z zastosowaniem rozkładów mieszanych jest określenie liczby i rodzaju rozkładów składowych. W przypadku rozkładów jednomodalnych do określenia liczby rozkładów składowych stosuje się często odpowiednie testy statystyczne [McLachlan, Peel 2000]. Podczas modelowania rozkładów pierśnic na ogół wystarczą rozkłady mieszane składające się z dwóch rozkładów składowych [Zhang i in. 2001; Liu i in. 2002; Zasada, Cieszewski 2005; Zhang, Liu 2006; Podlaski 2010]. Zastosowanie rozkładów mieszanych zwiększa dokładność aproksymacji, a ponadto sugeruje, że badana populacja składa się z podpopulacji. Należy jednak pamiętać, że jeżeli nie ma teoretycznego uzasadnienia, że badana populacja jest złożona z podpopulacji, to fakt, że zastosowano np. rozkład mieszany składający się z dwóch populacji składowych, nie zawsze świadczy o istnieniu dwóch podpopulacji [Zasada, Cieszewski 2005]. Dlatego najpierw, na podstawie teoretycznych wiadomości, należy określić liczbę rozkładów składowych, a dopiero później można stosować rozkłady mieszane do identyfikacji różnych podpopulacji. Najczęściej jako rozkłady składowe stosowane są rozkłady normalny, Weibulla i gamma [Zhang i in. 2001; Liu i in. 2002; Zasada, Cieszewski 2005; Zhang, Liu 2006; Podlaski 2010]. Zwłaszcza rozkłady Weibulla i gamma są bardzo uniwersalne i użyteczne.

Nawet pojedyncze rozkłady Weibulla i gamma mogą być stosowane do modelowania pierśnic w drzewostanach o różnym składzie gatunkowym i o różnej budowie [np.: Zasada 1995, 2000; Rymer-Dudzińska, Dudzińska 1999; Podlaski 2006]. W przypadku rozkładów jednomodalnych, w których rozkłady składowe całkowicie zachodzą na siebie, często nie można uzyskać zbieżności procesu iteracyjnego. Rozkłady składowe Weibulla, w porównaniu do rozkładów gamma, częściej generują problemy z uzyskaniem zbieżności procedury numerycznej [McLachlan, Krishnan 2008]. Na uzyskanie zbieżności wpływa również liczba parametrów rozkładów składowych i dlatego dla rozkładu Weibulla i gamma należy zastosować następującą strategię: (1) za parametr przesunięcia γ_i przyjąć stałą wartość, np. γ_i=min–0,1; gdzie *min* to najmniejsza pierśnica [Podlaski, Zasada 2008], (2) wycentrować dane rzeczywiste, czyli od każdej wartości rzeczywistej odjąć γ_i, (3) dopiero dla tak przygotowanych danych obliczać parametry.

Podczas estymacji parametrów rozkładów mieszanych, dla których rozkłady składowe tylko częściowo zachodzą na siebie, na ogół każda metoda określania wartości startowych pozwala na otrzymanie prawidłowych wyników [McLachlan, Krishnan 2008]. W przypadku stosowania rozkładów mieszanych do aproksymacji jednomodalnych rozkładów konieczne jest zastosowanie odpowiednich metod określania wartości startowych [Karlis, Xekalaki 2003]. W prezentowanej pracy przedstawiono dwie najpopularniejsze metody określania wartości startowych (min.*k*/max.*k* i 0,5/1,5/średnia) oraz po jednym przykładzie metod bardziej skomplikowanych (średnia/odchylenie standardowe/ustalone i metoda wielopunktowa). W literaturze można spotkać jeszcze wiele innych metod: m.in. procedury graficzne, scenariusze bazujące na analizie skupień oraz na metodzie składowych głównych [Leroux 1992; Böhning i in. 1994; McLachlan, Krishnan 2008].

Metoda min.*k*/max.*k* i 0,5/1,5/średnia są najprostsze, ale nie zawsze pozwalają na uzyskanie zbieżności procesu iteracyjnego i znalezienie ekstremum globalnego. Scenariusz z dwoma etapami nakłada ograniczenia na estymowane parametry. Ich negatywną konsekwencją może być zbieżność procedur numerycznych do lokalnych ekstremów lub punktów przegięcia. Dlatego podstawowym warunkiem stosowania tego typu ograniczeń jest wybór takich wartości startowych, dla których funkcja wiarygodności znajduje się bardzo blisko ekstremum globalnego. W przeciwnym wypadku zastosowane ograniczenia mogą uniemożliwić prawidłowe obliczenie parametrów rozkładów mieszanych [Zasada, Cieszewski 2005; Podlaski 2010]. Metoda wielopunktowa jest najdokładniejsza, ale równocześnie najbardziej pracochłonna.

Podsumowanie

Po przeanalizowaniu zalet i wad podstawowych metod określania wartości startowych w procesie estymacji parametrów rozkładu mieszanego proponuje się (1) wybrać wartości startowe metodą min.*k*/max.*k* (dla *k*=1, 3, 6) i 0,5/1,5/średnia, (2) zastosować standardową procedurę obliczania parametrów (algorytm EM + metoda Newtona), (3) jeżeli nie uzyskano przynajmniej dwóch takich samych rozwiązań dla wartości ekstremalnych, to należy wykorzystać dodatkowo metodę wielopunktową, (4) porównać uzyskane wyniki, co umożliwi zlokalizowanie ekstremum globalnego i w konsekwencji obliczenie parametrów rozkładu mieszanego. Proponowana procedura nie jest pracochłonna i może być stosowana do modelowania rzeczywistych rozkładów pierśnic reprezentujących drzewostany o różnej budowie.

Literatura

Böhning D. 2000. Computer-Assisted Analysis of Mixtures and Applications. Chapman & Hall/CRC, Boca Raton.

Böhning D., Dietz E., Schaub R., Schlattmann P., Lindsay B. 1994. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Ann. Inst. Statist. Math. 46: 373-388.

Bruchwald A. 1999. Dendrometria. Wydawnictwo SGGW, Warszawa.

Du J. 2002. Combined Algorithms for Constrained Estimation of Finite Mixture Distributions with Grouped Data and Conditional Data. Praca magisterska, McMaster University.

Haughton D. 1997. Packages for estimating finite mixtures: a review. Am. Stat. 51: 194-205.

Hessenmoller D., von Gadow K. 2001. Beschreibung der Durchmesserverteilung von Buchenbeständen mit Hilfe der bimodalen WEIBULLfunktion. Allg. Forst- und Jagdzeitung 172: 46-50.

Karlis D., Xekalaki E. 2003. Choosing initial values for the EM algorithm for finite mixtures. Comput. Statist. Data Anal. 41: 577-590.

Leroux B. G. 1992. Consistent estimation of a mixing distribution. Ann. Statist. 20: 1350-1360.

Liu C., Zhang L., Davis C. J., Solomon D. S., Gove J H. 2002. A finite mixture model for characterizing the diameter distribution of mixed-species forest stands. For. Sci. 48: 653-661.

Maltamo M., Kangas A. 1998. Methods based on k-nearest neighbor regression in estimation of basal area diameter distribution. Can. J. For. Res. 28: 1107-1115.

Mason E. G. 2000. Evaluation of a model of beech forest growing on the West Coast of the South Island of New Zealand. New Zeal. J. For. 44: 26-31.

McLachlan G. J., Basford K. E. 1988. Mixture Models: Inference and Applications to Clustering. M. Dekker, New York.

McLachlan G. J., Krishnan T. 2008. The EM algorithm and Extensions. Wiley, Hoboken.

McLachlan G. J., Peel D. 2000. Finite Mixture Models. Wiley, New York.

Podlaski R. 2006. Suitability of the selected statistical distributions for fitting diameter data in distinguished development stages and phases of near-natural mixed forests in the Świętokrzyski National Park (Poland). For. Ecol. Manage. 236: 393-402.

Podlaski R. 2010. Two-component mixture models for diameter distributions in mixed-species, two-age cohort stands. For. Sci. 56: 379-390.

Podlaski R., Zasada M. 2008. Comparison of selected statistical distributions for modelling the diameter distributions in near-natural *Abies-Fagus* forests in the Świętokrzyski National Park (Poland). Eur. J. For. Res. 127: 455-463.

Poznański R. 1997. Typy rozkładu pierśnic a stadia rozwojowe lasów o zróżnicowanej strukturze. Sylwan 141 (3): 37-44.

Pretzsch H. 1997. Analysis and modelling of spatial stand structures. Methodological considerations based on mixed beech-larch stands in Lower Saxony. For. Ecol. Manage. 97: 237-253.

Pretzsch H. 1998. Structural diversity as a result of silvicultural operations. Lesnictvi-Forestry 44: 429-439.

Rymer-Dudzińska T., Dudzińska M. 1999. Analiza rozkładu pierśnic w drzewostanach bukowych. Sylwan 143 (8): 5-24.

Seidel W., Mosler K., Alker M. 2000a. A cautionary note on likelihood ratio tests in mixture models. Ann. Inst. Statist. Math. 52: 481-487.

Seidel W., Ševčíková H., Alker M. 2000b. On the Power of Different Versions of the Likelihood Ratio Test for Homogeneity in an Exponential Mixture Model. Diskussionsbeiträge zur Statistik und Quantitativen Ökonomik 92. Universität der Bundeswehr, Hamburg.

Siekierski K. 1991. Three methods of estimation of parameters in the double normal distribution and their applicability to modeling tree diameter distributions. SGGW-AR For. 42: 13-17.

Zasada M. 1995. Ocena zgodności rozkładów pierśnic w drzewostanach jodłowych z niektórymi rozkładami teoretycznymi. Sylwan 139 (12): 61-69.

Zasada M. 2000. Ocena zgodności rozkładów pierśnic drzew drzewostanów brzozowych z niektórymi rozkładami teoretycznymi. Sylwan 144 (5): 43-48.

Zasada M. 2003. Możliwość zastosowania rozkładów mieszanych do modelowania rozkładów pierśnic drzew w naturalnych klasach biosocjalnych. Sylwan 147 (9): 27-37.

Zasada M., Cieszewski C. J. 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. Manage. 204: 145-158.

Zhang L. J., Gove J. H., Liu C., Leak W. B. 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.

Zhang L. J., Liu C. 2006. Fitting irregular diameter distributions of forest stands by Weibull, modified Weibull, and mixture Weibull models. J. For. Res. 11: 369-372.

Zucchini W., Schmidt M., von Gadow K. 2001. A model for the diameter-height distribution in an uneven-aged beech forest and a method to assess the fit of such models. Silva Fenn. 35: 169-183.

SUMMARY

Modelling tree diameter distributions using mixture models I. Definition, characteristics and parameters estimation of mixtures distributions

Single distributions can be used to examine the degree of compliance of empirical and theoretical distributions of breast height diameters, however sometimes it becomes necessary to apply more complex models. The aim of the paper is (1) to introduce to the theory of finite mixture models, with a particular focus on different methods of determining the initial values of the numerical procedure (EM algorithm + Newton's method) used for the parameters estimation, (2) to discuss the ways of procedure in selecting the number and type of component distributions and the methods of assessing the initial values, and (3) to propose the procedure of estimating mixture models parameters.

Mixture distributions composed of two component distributions are most often used in forest sciences. The following component distributions are most frequently used: normal, Weibull and gamma distributions.

To estimate the parameters of mixture distributions, the following procedure is recommended: (1) to select initial values using the method min.k/max.k (for k=1, 3, 6) and 0.5/1.5/average, (2) to apply a standard procedure for calculating the parameters (EM algorithm + Newton's method), (3) if at least two identical solutions for the extreme values are not obtained, the multistart method should additionally be applied, (4) to compare the findings obtained by the method min.k/max.k (for k=1, 3, 6) and 0.5/1.5/average, and possibly, by the multistart method allowing to locate the global extremum and in consequence, to calculate the parameters of mixture distributions. The proposed procedure is not very labour-intensive and can be used for modelling the actual distribution of breast height diameters representing stands of different structure.