NONPARAMETRIC METHOD: KERNEL DENSITY ESTIMATION APPLIED TO FORESTRY DATA

Rafael Romualdo Wandresen[1]*, Sylvio Péllico Netto[1], Henrique Soares Koehler[1], Carlos Roberto Sanquetta[1], Alexandre Behling[1]

[1] Federal University of Parana, Department of Forest Sciences, Curitiba, Paraná, Brazil - rafael.wandresen@gmail.com*; sylviopelliconetto@gmail.com; koehler@ufpr.br; carlossanquetta@gmail.com; alexandre.behling@yahoo.com.br

Received for publication: 02/07/2018 - Accepted for publication: 29/08/2018

Resumo

Método não paramétrico: estimativa de densidade kernel aplicada a dados de florestas. A função densidade de probabilidade pode ser ajustada por meio de métodos paramétricos ou não paramétricos. O uso do método não paramétrico é interessante e apropriado, considerando sua flexibilidade e melhor ajuste aos dados multimodais. O objetivo do presente estudo foi comparar o desempenho da distribuição não paramétrica (kernel density estimate – KDE) em relação à uma distribuição paramétrica. Foram utilizados seis conjuntos de dados florestais com características de bimodalidade e assimetria. As funções densidade de probabilidade foram estimadas para cada conjunto de dados usando o método KDE. Para avaliar a eficácia do método KDE, distribuições de probabilidade paramétricas também foram ajustadas para os mesmos dados. O teste de Kolmogorov-Smirnov foi aplicado para avaliar a aderência das distribuições paramétricas. As distribuições obtidas por meio dos dois métodos foram comparadas graficamente para identificar se o método não paramétrico e paramétrico são igualmente eficientes para obter a distribuição subjacente, especialmente para distribuições assimétricas bimodais. O método KDE é uma alternativa apropriada para descrever distribuições de probabilidade em dados florestais, especialmente quando ocorre bi - ou multimodalidade.
Palavras-chave: distribuições de probabilidade, distribuições assimétricas, bi- ou multimodalidade, cohorts.

Abstract

Probability density function can be fitted through parametric or non-parametric methods. The use of a non-parametric method is interesting and appropriate, considering its flexibility and better adjustment to multimodal data. The objective of the present study was to compare the performance of the non-parametric distribution in relation to the parametric distribution in these cases. We used six separate sets of forest database with bimodality and asymmetric characteristics. The probability density functions were estimated for each set of data using the KDE method. To evaluate the effectiveness of the KDE method, parametric probability distributions were also adjusted for the same data. The Kolmogorov-Smirnov test was applied to evaluate the goodness of fit of the parametric distributions. The distributions obtained through the two methods were compared graphically to identify if the nonparametric and parametric methods are equally efficient to obtain the underlying distribution, especially for bimodal and asymmetric distributions. The KDE method is an appropriate alternative for describing probability distributions in forest data, especially when bi- or multimodality occurs.
Keywords: Probability distributions, asymmetric distributions, bi- or multimodality, cohorts.

INTRODUÇÃO

There are several statistical tools applied to data for the formulation of forest management plans. The fitting of probability distributions to a sample of diameter or height data of trees are the most common ones. These data samples are commonly asymmetrically distributed and, in some cases, bi- or multimodality occurs.

Probability density functions can be fitted through parametric or non-parametric methods. The parametric methods assume that the data to be fitted are best adjusted to a known distribution. For example, we can estimate the mean value (μ) and variance (σ^2) from a sample and then replace them in the normal distribution function Silverman (1986). Other distributions can be fitted to the sample data, such as: Weibull 2P, Weibull 3P, Log-normal, Gamma and Beta.

Non-parametric methods are appropriate to fit some distribution data, considering its flexibility and better adjustment to multimodal data. When occurs a unimodal tendency in the database, even if the set of parametric models may be more appropriate to represent it, the non-parametric method can play an important role on prior assessment to the data distribution, so that the best probabilistic function may be selected. In addition, non-parametric method is an alternative for the detection of bi- or multimodal occurrences, caused by cohorts or other biological and abiotic effects occurring in forest populations.

Several authors have discussed the theme kernel-density estimation. Żychaluk; Patil (2008) presented a solution for the cross-validation method to select the bandwidth on data with ties. Botev *et al*. (2010) presented a new adaptive kernel density estimator based on linear diffusion processes and a new plug-in bandwidth selection method. Savchuk *et al*. (2010) proposed a new method of bandwidth selection for kernel density estimation, called indirect cross-validation. Mynbaev; Martins-Filho (2010) proposed a new method to achieve bias reduction to the classical Rosenblatt-Parzen estimator, based on imposing global Lipschitz conditions. Giné; Nickl (2010) discussed the theme of confidence bands in density estimation; an interesting idea of density estimation using convolution kernels was presented. Mammen *et al*. (2011) developed a new theorem deriving the asymptotic theory for linear combinations of bandwidths obtained from different selectors and discussed the problems of bandwidth selection in kernel density estimation. Harvey; Oryshchenko (2012) discussed kernel density estimation for time series data and improved this method. Heidenreich *et al*. (2013) reviewed methods on bandwidth selection and compared many methods by simulation. Cheng *et al*. (2017) discussed the problem of bandwidth selections methods to the situation where the data are serially dependent time series and presented a nonparametric localized bandwidth estimator for these cases. Chen (2017) presented a tutorial about kernel density estimation, provided implementations in R software and discussed some recent advances. Wu (2018) proposed a robust likelihood-based cross-validation method to select bandwidths in multivariate density estimations, obtaining better performance when data contains extreme observations and heavy-tailed distributions.

There are other examples about KDE applications: Kim and Scott (2012) proposed a robust method of nonparametric density estimation to avoid the contamination of the training sampling. They combined a traditional kernel density estimator with M-estimation. Wang *et al*. (2014) proposed a new learning method, denominated fast KDE, to obtain fast learning in large data sets. Chen (2017) reviewed the principal advances about kernel density estimation on modal regression, including additional algorithms and similar alternative approaches. Matioli *et al*. (2017) presented a new algorithm for clustering based on kernel density estimation.

In forestry, we can find some studies that used KDE in modeling of diameter distributions, as for example, Podlaski; Roesch (2014a) and Podlaski; Roesch (2014b).

In this study, the applications of KDE method were extended to other cases, which include six data sets with different characteristics: bimodality, slight asymmetry and strong asymmetry. The objective of the present study was to compare the performance of the non-parametric distribution with parametric distributions fitted to the previously mentioned cases. We propose the following hypotheses: "In the cases that bimodality and strong asymmetry occurs in forest data, KDE method can result in better fitting of probability density functions than parametric methods".

MATERIAL E MÉTODOS

We adopted a set of procedures for doing this research, which are summarized below and will be detailed in the following sections:

1 - We collected six data sets, with different characteristics: bimodality and asymmetry: i) S1 - Diameter at breast height (DBH) of *Araucaria angustifolia* (Bertol.) Kuntze with bimodality characteristic; ii) S2 - DBH of *Pinus elliottii* Englem with slight asymmetry; iii) S3 - Height of *Pinus elliottii* with strong asymmetry. Data with slight bimodality characteristic: iv) S4 - Periodical annual increment (PAI) of *Podocarpus lambertii* Klotzsch ex Endl. Data with strong asymmetry: v) S5 - PAI of *Araucaria angustifolia* and vi) S6 - PAI of *Blepharocalyx salicifolius* (Kunth) O. Berg. Six data sets with different characteristics were selected for this study, all from researches conducted in the state of Paraná, Brazil. The first set (S1 Dataset) is DBH (Diameter at Breast Height) of 200 trees of the *Araucaria angustifolia*, which presents bimodal characteristic. The second set (S2 Dataset) is of 250 DBH from trees of *Pinus elliottii* with slight asymmetry. The third set (S3 Dataset) is of 309 heights of *Pinus elliottii* with strong asymmetry. The fourth, fifth and sixth sets (S4, S5 and S6 Datasets) are data of average annual periodic increment of the following species: *Podocarpus lambertii* (128 trees), *Araucaria angustifolia* (829 trees) and *Blepharocalyx salicifolius* (381 trees).

2 - For each data set that was collected in step 1, we generated the histogram, after calculating the number of classes and its bin width. We also have generated the empirical cumulative distribution function (ECDF). We have called frequencies and ECDF in the histogram as observed frequencies and cumulative observed frequencies respectively.

3 - We have fitted probability density functions (PDFs) to each data set and used the observed frequencies, obtained in step 2, to fit the PDFs. The following probability density functions were used: Silva *et al*. (2003) for bimodal data S1 and S4, and Dagum distribution for light and strong asymmetric data S2, S3, S5 and S6.

4 - Kolmogorov-Smirnov's test was applied to each probability density function obtained in the step 3. The goal was to obtain the K-S statistic to evaluate the goodness of fit of all tested distributions. Kolmogorov-Smirnov's

Electronic ISSN 1982-4688
DOI: 10.5380/rf.v49 i3. 60285

test is used to compare the frequency values of the cumulative distribution function (CDF) (step 3) with the cumulative observed frequency values (ECDF) (step 2).
5 - We used the non-parametric KDE method to fit the PDFs to the six data sets. To obtain the PDFs we have used the observed collected data (step 1). The parameter h was obtained by the golden rule of Silverman (1986) and for the parameter α it was kept the default value proposed by the same author.
6 - The distributions obtained through the two methods were compared graphically to identify if the nonparametric and parametric methods are equally efficient to obtain the underlying distribution.
7 – Despite of the h and α values had been obtained in step 5, it provided a good fit, but we have added an additional step to improve it. Therefore, the estimators for h and α were optimized by a non-linear method, aiming to obtain a better underlying distribution than the one obtained in step 5.
8 - The distributions obtained through the two methods were compared graphically to identify if the nonparametric and parametric methods are equally efficient to be applied to the underlying distribution.

Histograms and ECDFs

The histograms for the six data sets were drawn using the "hist" function of the software R.

For the histograms of data set S2, S3 and S5, the number of bins was calculated using the Sturges' default method by the R software. In the case of the histograms of S1, S4 and S6, the Sturges' method in the R software did not plot well the data characteristics (resulted in non-smoothed distributions). Therefore, we have tested four other methods to choose the best one for each data set (S1, S4 and S6): Square root - where nc= $\sqrt{n}$, Doane's, Scott's, and Freedman-Diaconis' methods. The number of bins, the method applied to determine the number of the bins, bin width and initial value, respectively, for each data set was: S1 – 16, Square root, 5 cm, 10 cm; S2 – 14, Sturges', 2 cm, 6 cm; S3 – 13, Sturges, 0.5 m, 9 m; S4 – 14, Freedman-Diaconis', 0.05 cm.year^{-1}, 0.00 cm.year^{-1}; S5 – 14, Sturges' 0.1 cm.year^{-1}, 0.00 cm.year^{-1} and S6 – 16, Scott' 0.05 cm.year^{-1}, 0.00 cm.year^{-1}.

The empirical cumulative distribution function (ECDF) was generated using the "ecdf" function of R software.

Two parametric probability distribution functions were fitted to the data sets:
i) Silva's probability distribution function, Silva *et al*. (2003) proposed a mathematical segmented model, flexible enough to represent probability distributions with multimodal features.
ii) Dagum's probability distribution function, which belongs to the family of Beta II distributions, with three parameters (Type I) and with four parameters (Type II).

Silva's probability distribution function

The function proposed by these authors is generally defined as:

$$f(x)=\frac{1}{k}\begin{cases} c_1 \cdot x^d & \text{if } 0 < x < l_1 \\ a_1 \cdot x^n + a_2 \cdot x^{n-1} + a_3 \cdot x^{n-2} + ... + a_m & \text{if } l_1 \le x \le l_2 \\ \dfrac{c_2}{x^h} & \text{if } x > l_2 \\ 0 & \text{otherwise} \end{cases}$$

Where: *n*, *d* and *h* are positive integers; a_i are coefficients of the polynomial; c_1 and c_2 are real numbers; *k* is the value of the integral:

$$\int_0^{\infty}\left[c_1.x^d + \left(a_1.x^n + a_2.x^{n-1} + a_3x^{n-2} + ... + a_m\right) + c_2 / x^h\right]d_x$$

Where: *x* is the variable of interest; l_1 is the upper limit of class of which the function $c_1 \cdot x^d$ will be adjusted; l_2 is the upper limit of the last class of which the polynomial produces a good adjustment.

According to the authors the polynomial function is composed by three segments that must meet the requirements of a PDF, i.e., to form a continuous function with functional non-negative values and converging to infinity. Silva *et al*. (2003) defined some steps to prepare the polynomial function:
1. Fit the polynomial $a_1 \cdot x^n + a_2 \cdot x^{n-1} + a_3 \cdot x^{n-2} + ... + a_m$ to the frequencies of the histogram (Material and Methods – Step 2). In this work, we used the function polyfit in software R. The polynomial is identified as $g_2(x)$. The degree (*n*) of the polynomial is obtained empirically, observing the formed curve and comparing it with the real data, as presented in Silva *et al*. (2003).
2. Plot a graph of the fitted polynomial in step 1 and discard the classes where the polynomial assumes negative values, or it is contrary to the trend of the model. This step is performed to find the values of the limits l_1 and l_2 empirically, observing the curve formed by the polynomial;

3. Select functions to the regions $0 < x < l_1$ and $x > l_2$. For the first region, we should use the exponential model $g_1(x) = c_1 \cdot x^d$. For the second region, we should use a hyperbole model that converges to infinity or $g_3(x) = c_2 / x^h$. In the two data sets used in this work, the values of d and h were obtained empirically, by observing the best fit to the data. Then, we must calculate the values of $g_2(l_1)$ and $g_2(l_2)$. To compose the continuity of the function and calculate the values of c_1 and c_2, the relationships $g_2(l_1) = c_1 \cdot l_1^d$ and $g_2(l_2) = c_2 / l_2^h$ should be solved.

4. Compose the function $g(x)$ with the three other ones $g_1(x)$, $g_2(x)$ and $g_3(x)$ and calculate the integral of the resulted function: $\int_0^\infty g(x)\,dx$ to obtain the area k.

5. Multiply the function g(x) by 1/k to obtain the PDF function f(x) so that $\int_{-\infty}^{\infty} f(x)\,dx = 1$

The estimated mean of the function, proposed by Silva *et al*. (2003), is obtained by the first moment, as:

$$\bar{X}_S = E\{x\} = \int_0^{+\infty} x \cdot f(x)\,dx$$

The estimated variance of the function, proposed by Silva *et al*. (2003), is obtained by the difference between the second moment and the square of the first moment, as:

$$S_S^2 = E\left\{x^2\right\} - \left(E\{x\}\right)^2 = \left(\int_0^{+\infty} x^2 \cdot f(x)dx\right) - \bar{X}_S^2$$

The cumulative distribution function can be defined as the classic definition of CDF as:

$$F(x) = \int_0^x f(\mathrm{t})\,dt$$

The integrals were solved by numerical integration; more specifically in this work we used the "integrate" function of the R software.

Dagum's probability distribution

The Dagum's distribution belongs to the same family of Beta II distributions, with three parameters (Type I) and with four parameters (Type II), which was developed for studies of personal income in the area of economy. Its advantage is detection of extreme values in asymmetric distributions, occurring in some situations of diametric and height distributions in forest populations. The PDF, cumulative distribution, arithmetic mean and variance of this distributions are presented in Dagum (1977). Its PDF is given by:

$$f(x;a,b,p) = ap(x)^{-1}\left(xb^{-1}\right)^{ap}\left\{\left[\left(xb^{-1}\right)^a + 1\right]^{-(p+1)}\right\}$$

Where: a, b, c and p are the coefficients of the function.

Its cumulative distribution is given by

$$F(x;a,b,p) = \left[1 + \left(xb^{-1}\right)^{-a}\right]^{-p} \qquad x > 0 \quad a,\ b,\ p > 0$$

Its arithmetic mean is given by

$$\mu_D = \left[-\frac{b}{a}\frac{\Gamma\left(-\frac{1}{a}\right)\Gamma\left(\frac{1}{a}+p\right)}{\Gamma(p)}\right] \qquad a > 1$$

Undefined i.o.c

Where: Γ is the gamma function.

Its variance is given by

$$\sigma_D^2 = -\frac{b^2}{a^2}\left\{2a\frac{\Gamma\left(\frac{2}{a}\right)\Gamma\left(\frac{2}{a}+p\right)}{\Gamma(p)} + \left[\frac{\Gamma\left(\frac{1}{a}\right)\Gamma\left(\frac{1}{a}+p\right)}{\Gamma(p)}\right]^2\right\} \qquad a > 2$$

Undefined i.o.c.

Kernel density estimation – KDE method

KDE is a nonparametric technique for the estimation of the density function f(x) of a random variable X, with all the properties of a probability density function. The model is described in Parzen (1962). The Gaussian function was selected as the kernel function in this work, therefore the KDE model can be described as:

FLORESTA, Curitiba, PR, v. 49, n. 3, p. 561 - 570, jul/set 2019
Wandresen, R. R. *et.al.*
Electronic ISSN 1982-4688
DOI: 10.5380/rf.v49 i3. 60285

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-X_i)^2}{2h^2}}$$

Where: $\hat{f}(x)$ is the function that estimates the density function *f*(*x*); n is the number of sample units; *h* is the window or parameter of smoothness; *x* is the point where it is evaluated the density function; *Xi* is a value of the random sample distribution $X_1, X_2, \ldots, X_n$. The smoothness parameter *h* was optimized using the golden rule of Silverman. One of the implementations of KDE method was proposed by Silverman (1986) and is defined as:

$$\hat{f}_S(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h\lambda_i}K\left(\frac{x-X_i}{h\lambda_i}\right)$$

Where: $\hat{f}_S(x)$ is the function that estimates the density function *f*(*x*); λ_i is the location bandwidth factor.

The locality factor acts to make $h\lambda_i$ small around the mode region and large on the tails of the distribution. According to Silverman (1986) λ_i is given by:

$$\lambda_i = \left(\frac{\tilde{f}(X_i)}{g}\right)^{-\alpha}$$

Where: the function $\tilde{f}(X_i)$ is a pilot function generated by fixed KDE; α is the sensitivity parameter with occurrence in the interval $0 \le \alpha \le 1$; g is the geometric mean of $\tilde{f}(X_i)$, $g = \left[\prod_{i=1}^{n}\tilde{f}(X_i)\right]^{1/n}$. The parameter α controls the influence of λ_i in the estimation and the parameter h and can be obtained as described previously.

According to Nadaraya (1964), cited by Altman; Léger (1969), the cumulative function of the kernel probability estimation can be defined as:

$$\hat{\mathrm{F}}(x) = \frac{1}{n}\sum_{i=1}^{n}L\left(\frac{x-X_i}{h}\right)$$

Where: *L* is a cumulative positive kernel distribution function and *L*(*x*) is defined as $(x) = \int_{-\infty}^{x}K(t)dt$.

Extending the concept described by Nadaraya (1964) in the last expression, we can rewrite the function of variable kernel cumulative probability estimator, including the location factor λ_i, for the case in which the kernel is a function of a Gaussian distribution, as:

$$\hat{f}_s(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}h\lambda_i}e^{-\frac{(x-X_i)^2}{2(h\lambda_i)^2}}$$

This makes it a sum of the Gaussian distribution functions with mean X_i and standard deviation $h\lambda_i$. It can be observed that, unlike the fixed kernel density estimation, the standard deviation of Gaussian functions varies for each sample. Knowing that the function of accumulated distribution is defined by the integration of the distribution function, we get:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}F_{\mu=X_i;\sigma=h\lambda_i}$$

In a similar way to the fixed kernel density estimation method, the integration that is inside the summation in last expression, is nothing more than a function of accumulated Gaussian distributions with mean Xi and standard deviation $h\lambda_i$. Therefore, in this case we can say that the estimate of the cumulative distribution is the sum of cumulative distributions of all Gaussian curves defined for each sample unit Xi.

The inverse of the cumulative probability can be obtained by the sum of the integral of the known probability. For example, to obtain the value of x, whose cumulative probability is equal to 0.45, we have:

$$\hat{F}(x) = 0.45 \;\therefore 0.45 = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}h\lambda_i}e^{-\frac{(x-X_i)^2}{2(h\lambda_i)^2}} \;\therefore 0.45 = \frac{1}{n}\sum_{i=1}^{n}F_{\mu=X_i;\sigma=h\lambda_i}$$

It is not possible to obtain an analytical solution for the integral of last expression, so it is necessary to apply a numerical solution for an x value of a given cumulative probability. One of the possible algorithms that can be applied to solve the cumulative probability is presented in Brent (2013), included in software R.

Optimization of the parameters h and α

The shape of the probability distribution function obtained by variable kernel density estimation method is defined by the parameter of smoothness h and the sensitivity parameter α. In this work, we have used the

smaller difference between the two cumulative distributions, the one obtained with the KDE function and the other with the observed data. We did the minimization of the maximum difference between the observed and estimated accumulated frequencies through a nonlinear optimization process, varying the *h* and α parameters, to optimize them and improve the results of the underlying distribution. The optimal function present in software R was used for nonlinear optimization, using simplex algorithm.

Kolmogorov-Smirnov's test

The Kolmogorov-Smirnov's test was used in this work to test the goodness of fit of the probability distribution functions obtained by the parametric method. This test consists in comparing the maximum difference between the observed cumulative frequency (by the ECDF) and the estimated cumulative frequency (by the Silva's CDF, the Dagum's CDF), divided by the number of observations, as:

$$D = \left[\max \left| F_0(X) - F_e(X) \right| \right] / n$$

Where: D is the Kolmogorov-Smirnov's statistic; F_0 (X) is the observed accumulated frequency; F_e (X) is the expected accumulated frequency estimated by the used model; *n* the number of observations.

In cases when n is greater than 50, the critical value (D_C) of the test is obtained by:

$$D_{C95} = 1.36 / \sqrt{n} \text{ for } (1-\alpha) = 95\% \text{ and } D_{C99} = 1.63 / \sqrt{n} \text{ for } (1-\alpha) = 99\%$$

RESULTS

Figure 1 shows the graphs of the six data sets (S1, S2, S3, S4, S5 and S6) fitted by the non-parametric KDE method (Figure 1 S1-1, S2-1, S3-1, S4-1, S5-1 and S6-1) and by the probability density functions: Silva's (Figure 3 S1-2 and S4-2) and Dagum's (Figure 3 S2-2, S3-2, S5-2 and S6-2).

Table 1 displays the values of the means and variances calculated with the non-parametric KDE method and the results of mean and variance obtained from data. In addition, this Table displays the result of the fitting to the six sets of data and the results of Kolmogorov-Smirnov's tests for each data set for Silva's and Dagum's probability density functions.

The procedures used to estimate the parameters h and α for each data set followed the steps presented in section materials and methods, expect for the data set (B), in which it was suppressed the optimization of the parameters h and α. In this case, the parameter h was calculated by the golden rule of Silverman (1986). This was done because, with the optimization of these parameters, the estimated PDF by KDE method was very sensitive and did not represent well the data.

The results plotted in Figure 1 confirmed that the non-parametric variable kernel density estimation method is as good as the parametric methods for fitting probability distributions to forest data sets. In the cases of data with bimodality characteristics, the KDE method fitted better to PDF than Silva's PDF. Furthermore, the KDE method, unlike parametric methods, has the advantage that is not necessary to plot the histogram to fit the PDF.

DISCUSSION

Several parametric models have been used to describe the frequency of diameters in forest. The main ones are: Normal, Log-normal, Weibull, Beta, Gamma and Johnson's SB, owing to their flexibility. In this research, we have applied the Silva's and Dagum's probability distributions for modeling diameter, height and periodical annual increment data (Figure 3). Adjusting PDFs using the parametric method depends on a defined form of the distribution, often resulting in non-smoothed adjustments, because the distribution of many variables of the forest is most often not smooth.

Although the application of parametric probability distribution functions is fundamental for forest yield prediction, the occurrence of multimodality limits its use in modeling forest dynamics. Even considering the existence of continuous probability functions in the literature for multimodality cases (SILVA *et al*., 2003), their adjustment to various situations caused by cohorts have shown to be not very satisfactory.

FLORESTA, Curitiba, PR, v. 49, n. 3, p. 561 - 570, jul/set 2019
Wandresen, R. R. *et.al.*
Electronic ISSN 1982-4688
DOI: 10.5380/rf.v49 i3. 60285

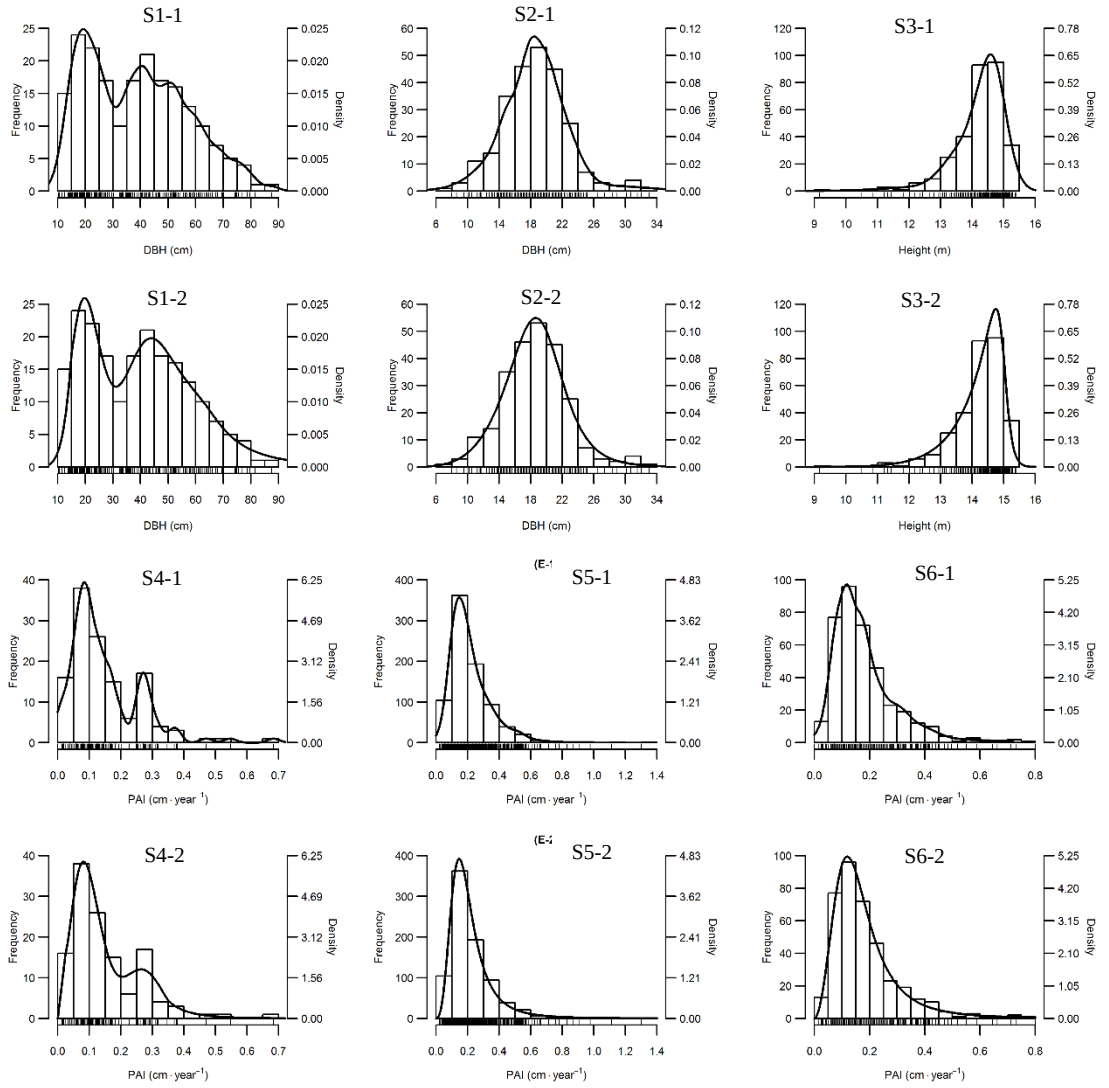

Figure 1. PDFs fitted by parametric and non-parametric methods on different forestry data.
Figura 1. FDPS ajustadas por métodos paramétricos e não-paramétricos à diferentes dados florestais.
Where: S1= *Araucaria angustifolia*, S1-1 fitted by KDE and S1-2 fitted by *PDF* of Silva. S2 = *Pinus elliottii*, S2-1 fitted by KDE and S2-2 fitted by *PDF* of Dagum. S3= *Pinus elliottii*, S3-1 fitted by KDE and S3-2 fitted by *PDF* of Dagum. S4 = *Podocarpus lambertii*, S4-1 fitted by KDE and S4-2 fitted by *PDF* of Silva. S5 = *Araucaria angustifolia*, S5-1 fitted by KDE and S5-2 fitted by *PDF* of Dagum. S6 = *Blepharocalyx salicifolius*, S6-1 - fitted by KDE and S6-2 fitted by *PDF* of Dagum.

Recently, the authors of this manuscript studying distributions of mean annual periodic increments of native species in the Mixed Tropical Forest found the occurrence of multimodality in the data of some species. This reality is not commonly assessed, but becomes relevant in studies of forest dynamics, since the arithmetic mean of these increments is commonly used, ignoring completely the existence of such events, caused basically by oscillations in the incidence of light inside the forest, when openings of vital space are caused by fallen trees or other occurrences of clearings.

Non-parametric methods are applied less frequently, but the results of KDE were appropriated and showed to be very flexible under different conditions (Figure 1). The flexibility of this method, according to Silverman (1986), occurs because, when using the procedure of non-parametric methods, there is no presumption of the distribution form, because that is determined by the original data. A potential application of nonparametric methods in forestry characterizes by its inherent sensibility for fitting to different distribution forms, for example, sample data with strong asymmetry, multimodality and the occurrence of cohorts.

Table 1. Fitted parameters of the probability density functions, mean and standard deviation of data and mean and standard deviation of estimated PDFs on different forestry data.

Tabela 1. Parâmetros ajustados das funções densidade de probabilidade, média e desvio padrão dos dados e média e desvio padrão das PDFs estimadas em diferentes dados florestais.

Function	Parameter	Coefficient	Function	Parameter	Coefficient
(A-1) Kernel	H	2.83	(D-1) Kernel	H	0.015
	A	0.00		A	0.090
(A-2) Silva	N	10	(D-2) Silva	N	10
	D	5		D	1
	H	6		H	6
	a_1	1.30E-13		a_1	5.70 E+07
	a_2	-6.73E-11		a_2	-2.12E+08
	a_3	1.52E-08		a_3	3.39E+08
	a_4	-1.97E-06		a_4	-3.04E+08
	a_5	1.62E-04		a_5	1.67E+08
	a_5	-8.75E-03		a_6	-5.76E+07
	a_7	3.13E-01		a_7	1.24E+07
	a_8	-7.28		a_8	-1.55E+06
	a_9	104.39		a_9	1.00E+05
	a_{10}	-827.57		a_{10}	-2.51E+03
	a_{11}	2760.49		a_{11}	3.60E+01
	c_1	2.27E-05		c_1	640.89
	c_2	6.92E11		c_2	8.09E-03
	K	991.76		K	6.34
	l_1	15		l_1	0.025
	l_2	70		l_2	0.35
(B-1) Kernel	H	1.10	(E-1) Kernel	H	0.036
	A	0.50		A	0.04
(B-2) Dagum	K	0.54	(E-2) Dagum	K	1.18
	A	10.48		A	2.95
	B	20.34		B	0.17
(C-1) Kernel	H	0.30	(F-1) Kernel	H	0.024
	A	0.27		A	0.40
(C-2) Dagum	K	0.15	(F-2) Dagum	K	1.00
	A	118.13		A	2.84
	B	15.00		B	0.15

Function	D	D_{C95} α=0.05	Function	D	D_{C95} α=0.05
(S1-2) Silva	0.045	0.096	(S4-2) Silva	0.029	0.120
(S2-2) Dagum	0.014	0.086	(S5-2) Dagum	0.035	0.047
(S3-2) Dagum	0.035	0.077	(S6-2) Dagum	0.013	0.071

Data	Mean	Variance	Data	Mean	Variance
S1	38.75cm	327.93cm^2	S4	0.15cm	1.260cm^2
S2	18.59cm	17.16cm2	S5	0.22cm	1.940cm^2
S3	14.21m	0.70m^2	S6	0.18cm	0.014cm^2

Function	Mean	Variance	Function	Mean	Variance
(S1-1) Kernel	38.75cm	335.96cm^2	(S4-1) Kernel	0.14cm	1.29cm^2
(S1-2) Silva	41.69cm	460.47cm^2	(S4-2) Silva	0.15cm	1.18cm^2
(S2-1) Kernel	18.59cm	19.03cm^2	(S5-1) Kernel	0.22cm	2.07cm^2
(S2-2) Dagum	18.57cm	17.06cm^2	(S5-2) Dagum	0.23cm	3.36cm^2
(S3-1) Kernel	14.21m	0.80m^2	(S6-1) Kernel	0.18cm	0.015cm^2
(S3-2) Dagum	14.21m	0.65m^2	(S6-2) Dagum	0.19cm	0.029cm^2

The results plotted in Figure 3 and 4 have confirmed that the non-parametric KDE method is as good as the parametric method for fitting probability distributions to forest data sets. Advantages and disadvantages of application of the KS test and how it is affected by the data distribution are discussed in Zeng *et al*. (2015). Therefore, we decided to compare the two methods (parametric and non-parametric) only graphically. In the cases of data with bimodality the KDE method fitted better than Silva's PDF. Furthermore, the KDE method, unlike parametric methods, does not require to draw the histogram to fit the PDF.

The mean and variance in the case of kernel estimator resulted as meaningful as those obtained from the best parametric models fitted to the same data.

KDE method proves to be an effective solution for fitting distribution data of variables with multimodality, unimodal distributions with extreme asymmetry, or the case when the distribution approximates to normality. The achievement of good results with the application of non-parametric methods in forestry was also observed by other researchers.

Electronic ISSN 1982-4688
DOI: 10.5380/rf.v49 i3. 60285

Droessler; Burk (1989) evaluated non-parametric methods to obtain smoother curves in modeling the diameter distribution of Pinus resinosa, using hypothetical populations from permanent plots of a forest area. The authors have observed that in extreme bimodal distribution, the nonparametric methods performed better than the Weibull distribution. They also mentioned that nonparametric methods described every detail in any DBH distribution, whereas the Weibull distribution could not do the same. However, for the cases in which the DBH distribution was approximately unimodal, the Weibull distribution fitted adequately.

Podlaski; Roesch (2014a) evaluated two-component mixtures of either the Weibull and the Gamma distributions for fitting DBH on forest stands with mixed-species and compared the results with KDE estimates. The result of the fitting of mixture Weibull and Gamma models in this case resulted in a similar precision, slightly less accurate estimate was obtained with the KDE estimator. In the research conducted by Podlaski; Roesch (2014b), application of KDE on DBH distributions in stands, where two cohorts occurred, showed good results when compared to mixture-probability density functions.

Considering the results observed in this research and in the other already mentioned, it was possible to smooth the fitted distributions when using KDE estimator, consequently approximating them to the underlying population data.

CONCLUSÕES

- The non-parametric KDE method appears as an alternative for fitting probability distributions of diameter, height and periodical annual increment data, widely used for characterization of forest populations.
- The KDE estimator has proved to be a sensitive method and an alternative for the detection of bi- or multimodal occurrences, caused by cohorts or other biological and abiotic effects occurring in forest populations.

REFERÊNCIAS

ALTMAN, N.; LÉGER, C. Bandwidth selection for kernel distribution function estimation. **Journal of Statistical Planning and Inference**, Salt Lake City, v. 46, p. 195-214, 1969.

BOTEV, B.Z.I.; GROTOWSKI, J.F.; KROESE, D.O. Kernel density estimation via diffusion. **The Annals of Statistics**, Rockville, v. 38, p. 2916-2957, 2010.

BRENT, R.P. **Algorithms for minimization without derivatives**. New Jersey: Prentice-Hall, 2013, 206 p.

CHEN, Y.C. A tutorial on kernel density estimation and recent advances. **Biostatistics & Epidemiology**, Tehran, v. 1, n. 1, p. 161-187, 2017.

CHEN, Y.C. Modal regression using Kernel density estimation: A review. **WIREs Computational Statistics**, New Jersey, 2017. Published online: DOI: 10.1002/wics.1431

CHENG, T.; GAO, J.; ZHANG, X. Nonparametric localized bandwidth selection for Kernel density estimation. **Econometric Reviews**, London, 2017. Published online: DOI: http://dx.doi.org/10.1080/07474938.2017.1397835

DAGUM, C. A new model of personal income distribution: Specification and estimation. Economic Apliquée, v. 30, p. 413-437, 1977.

DROESSLER, T.D.; BURK, T.E. A test of nonparametric smoothing of diameter distributions. **Scandinavian Journal of Forest Research**, Olso, v. 4, p. 407-415, 1989.

GINÉ, E.; NICKL, R. Confidence bands in density estimation. **The Annals of Statistics**, Rockville, v. 38, n. 2, p. 1122-117, 2010.

HARVEY, A.; ORYSHCHENKO, V. Kernel density estimation for time series data. **International Journal of Forescasting**, New York, v. 28, p. 3-14, 2012.

HEIDENREICH, N.; SCHINDLER, A.; SPERLICH, S. Bandwidth selection for kernel density estimation: a review of fully automatic selectors. **AStA Advances in Statistical Analysis**, New York, v. 97, p. 403-433, 2013.

KIM, J.; SCOTT, C.D. Robust kernel density estimation. **Journal of Machine Learning Research**, Brookline, v. 13, p. 2529-2565, 2012.

MAMMEN, E.; MIRANDA, M.D.M.; NIELSEN, J.P.; SPERLICH, S. Do-validation for kernel density estimation. **Journal of the American Statistical Association**, New York, v. 106, n. 494, p. 651-660, 2011.

MATIOLI, L.C.; SANTOS, S.R.; KLEINA, M.; LEITE, E.A. A new algorithm for clustering based on kernel density estimation. **Journal of Applied Statistics**, London, 2017. Published online: DOI: http://dx.doi.org/10.1080/02664763.2016.1277191

MYNBAEV, K.; MARTINS-FILHO, C. Bias reduction in kernel density estimation via Lipschitz condition. **Journal of Nonparametric Statistics**, London, v. 22, n. 2, p. 219-235, 2010.

PARZEN, E. On estimation of a probability density function and mode. **Annals of Mathematical Statistics**, New York, v. 33, p. 1065-1076, 1962.

PODLASKI, R.; ROESCH, F.A. Aproksymacja rozkładów pierśnic drzew w dwugeneracyjnych drzewo− stanach za pomocą rozkładów mieszanych. III. Estymatory jądrowe a rozkłady mieszane. **Sylwan**, Warszawa, v. 158, p. 414−422, 2014b.

PODLASKI, R.; ROESCH, F.A. Modelling diameter distributions of two-cohort forest stands with various proportions of dominant species: A two-component mixture model approach. **Mathematical Biosciences**, New York, v. 249, p. 60-74, 2014a.

SAVCHUK, O.Y.; HART, J.D.; SHEATHER, S.J. Indirect cross-validation for density estimation. **Journal of the American Statistical Association**, New York, v. 105, p. 415-423, 2010.

SILVA, E.Q.; PÉLLICO NETTO, S.; MACHADO, S.A.; SANQUETTA, C.R. Função densidade de probabilidade aplicável à ciência florestal. **Floresta**, Curitiba, v. 3, p. 285-294, 2003.

SILVERMAN, B.W. **Density estimation for statistics and data analysis**. New York: Chapman & Hall, 1986, 176 p.

WANG, S.; WANG, J.; CHUNG, F. Kernel density estimation, kernel methods, and fast learning in large data sets. **IEEE Transactions on Cybernetics**, New York, v. 44, p. 1-20, 2014.

WU, X. Robust Likelihood cross-validation for kernel density estimation. **Journal of Business & Economic Statistics**, London, 2018. Published online: DOI: https://doi.org/10.1080/07350015.2018.1424633

ZENG, X.; WANG, D.; WU, J. Evaluating the three methods of goodness of fit test for frequency analysis. **Journal of Risk Analysis and Crisis Response**, v. 5, n. 3, p. 178-187, 2015.

ŻYCHALUK, K.; PATIL, P.N. A cross-validation method for data with ties in kernel density estimation. **Annals of the Institute of Statistical Mathematics**, New York, v. 60, p. 21-44, 2008.