Asian Fisheries Science **15**(2002): 295-303
ISSN: 0116-6514
https://doi.org/10.33997/j.afs.2002.15.4.001

Asian Fisheries Society, Manila, Philippines

Nonparametric Assessment of Multimodality for Size Frequency Distributions

I.H. SALGADO-UGARTE, M. SHIMIZU, T. TANIUCHI and K. MATSUSHITA

University of Tokyo, Faculty of Agriculture
Department of Aquatic Bioscience
Yayoi 1-1-1, Bunkyo-ku, Tokyo 113
Japan

F.E.S. Zaragoza U.N.A.M. Biología
Av. Guelatao 66, Ejército de Oriente Iztapalapa 09230
A.P. 9-020, México D.F
México

Abstract

The multimodal nature of size frequency distributions is a common occurrence in fisheries analysis with modes representing potential Gaussian components in a mixed distribution. This characteristic is exploited to estimate and assess every component using any of the several parametric models that have been proposed. In general, the success of these procedures is dependent on the smoothness of the frequency distribution and the intelligent (data based and complemented with additional biological information) input of initial values. In this study we utilized nonparametric density estimators in combination with several rules for smoothing parameter (bandwidth) selection and a smoothed bootstrapping of critical bandwidths procedure to investigate multimodality of the length distribution of Japanese sea bass "suzuki" (*Lateolabrax japonicus*). These methods led to useful data-based estimations with statistically significant number of modes that produced growth estimations consistent with those obtained from fish aged by means of hard-parts (scales and otoliths) reading, specially at smaller ages. The resulting von Bertalanffy growth expressions had parameters inside the range of those reported in the literature. These nonparametric methods prove to be an alternative valuable tool for the analysis of mixed size distributions of fish.

Introduction

Researchers who work with complex distribution shapes have turned in recent years to nonparametric techniques such as nonparametric density estimation (Silverman 1986), where mixed components can be detected by identifying modes (local maxima) in the underlying distribution (Izenman and Sommer 1988). The number and location of the modes may or may not correspond with each individual component. There is a dependence upon both the

spacing of the modes and the relative shapes of the component distributions. Nevertheless, in many practical instances, the existence of more than a single mode does suggest evidence for a mixture. In the statistical literature there are several tests for detecting multimodality in a distribution. For example, the DIP test was proposed by Hartigan and Hartigan (1985) to accept or reject the unimodality hypothesis; Good and Gaskins (1980) used the penalized-llkelihood method of density estimation together with other statistical techniques; Silverman (1981a) combined kernel density estimation with a hierarchical bootstrap testing procedure. The last two combined methods are nonparametric, data-adaptive, and computationally intensive. Specific contexts often play a prominent role in relating empirical modes to plausible mixture components. The multimodal size frequency of fish may represent age groups containing important growth information.

The choice of interval width (binwidth/bandwidth) is one of the central problems in density estimation. There are several ways to select an appropriate binwidth for histograms, frequency polygons, or averaged shifted histograms and a bandwidth for kernel density estimators (KDE's). A brief review of some binwidth/bandwidth selection procedures and some programs to calculate them are contained in Salgado-Ugarte et al. (1995b). Presented in the present paper are the optimal Gaussian binwidth for histograms and frequency polygons (Scott 1979, 1985, 1992) and the optimal bandwidth for GaussianKDE's (Silverman 1986). Besides, by using the ASH-WARP technique, it is possible to calculate least squares and biased cross-validation (L2CV and BCV respectively) for kernel density bandwidth selection (Härdle 1991).

These rules in conjunction with the oversmoothed widths (Terrel 1990) represent a powerful tool for choosing the binwidth for histograms and frequency polygons and the bandwidth for kernel density estimators (Scott 1992).

The Silverman test uses the Gaussian KDE according to the following steps: identification of the critical bandwidths compatible with the hypothesis of a given number of modes; drawing of a smoothed bootstrap sample for each critical bandwidth; estimation of the corresponding densities; calculation of the significance (p value) for the number of modes as the fraction resulting from the count of estimations displaying more modes than the number indicated by the critical bandwidth used divided by the total number of repetitions (bootstrap samples). For an accounted description see Silverman (1981b, 1986) and Izenman and Sommer (1988). A computerized implementation and examples of application of this test to size frequency analysis of fish is presented in Salgado-Ugarte (1995) and Salgado-Ugarte et al. (1997). A related procedure is that of Wong (1985).

Material and Methods

For the present study, samples from the commercial catch at Tokyo Bay of the Japanese sea bass (*Lateolabrax japonicus, insertae sedis* genus in Percoidei according to Eschmeyer et al. 1996), an important species inhabiting Japanese waters, were obtained in approximately monthly periods from Septem-

ber 1993 to May 1995. Specimens collected during the surveys done by the Laboratory of Fisheries Biology (Fisheries Department, University of Tokyo) were included. A total of 406 individuals was analyzed: 109 males, 114 females and 183 individuals of unknown sex (mainly juveniles), ranging from 162 to 664, 155 to 760, and 123 to 366 mm of standard body length respectively (Salgado-Ugarte 1995).

To avoid drawbacks of length frequency analysis of pooled data, only fish caught at the beginning of the growth season (Spring) were considered. From these, only the females (n = 31) and a subsample of the individuals of unknown sex (those aged by means of hard-part reading, n = 16) were included (Table 1).

To calculate the KDE's the programs written by Salgado-Ugarte et al. (1993, 1995a, 1995b), which include the efficient algorithms of the averaged shifted histograms (ASH) and weighted averaging of rounded points (WARP) were used. The Silverman's test was performed employing the specific programs presented in Salgado-Ugarte (1995) and Salgado-Ugarte et al. (1997). The component characterization was carried out using the routines implementing the Bhattacharya's procedure written by Salgado-Ugarte et al. (1994). A weighted nonlinear regression scheme was used to estimate the parameters of the von Bertalanffy growth function (VBGF). Weighted nonlinear regressions were performed using revised versions of the programs described in Salgado-Ugarte et al. (2000).

Results

Estimations from the histogram, frequency polygon, averaged shifted histogram and kernel density estimation all employing the corresponding optimal Gaussian width (Figs. 1 to 4) resulted to a multimodal distribution; at least three modes can be distinguished.

Least squares cross-validation provided a minimum score with a bandwidth of 6. The corresponding density estimation shown in figure 5 looks undersmoothed in comparison with the previous estimates. The Gaussian kernel density estimation using the biased cross-validation bandwidth (h = 92) is a notoriously oversmoothed unimodal distribution with the indication of an additional mode at large sizes (Fig. 6). Certainly, the distribution is not unimodal.

Table 1. Number of individuals by sampling date considered for analysis. Spring.

Sample		Sex		
Number	Date	Females	Unknown	Total
6	19/02/94	4	14	18
7	16/03/94	13	1	14
8	14/04/94	4	1	5
9	13/05/94	10	0	10
Total		31	16	47

The results of the nonparametric assessment of multimodality are presented in table 2. The value for the DIP statistics was 0.0857, and the hypothesis of unimodality was rejected. The p values from Silverman test in table 2 suggest at least 6 modes.

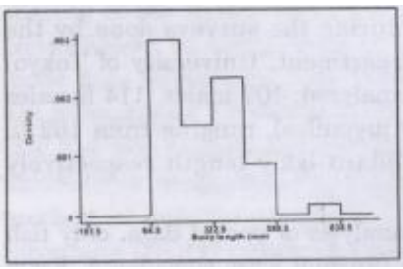

Fig. 1. Histogram using the optimal Gaussian binwidth (h = 129) and origin at 64.5

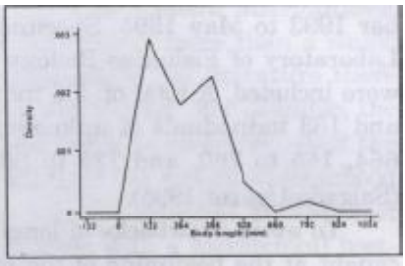

Fig. 2. Frequency polygon with the optimal Gaussian binwidth (h = 132) and origin at 0

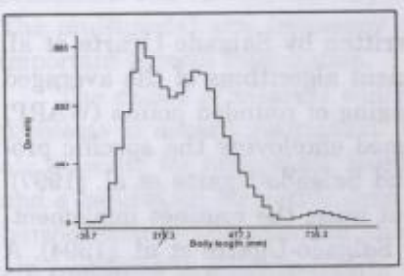

Fig. 3. Five averaged shifted histograms with the optimal Gaussian histogram binwidth (h = 129)

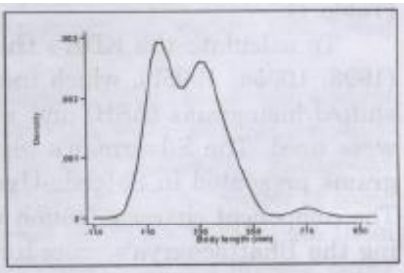

Fig. 4. Kernel density estimate using the optimal Gaussian bandwidth (h = 55)

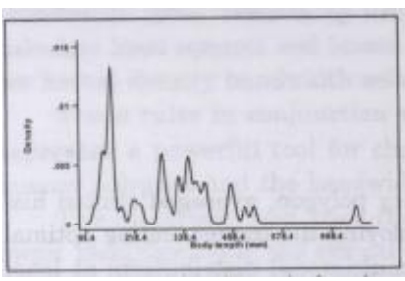

Fig. 5. Gaussian kernel density estimation using the least squares cross validation bandwidth (h = 6)

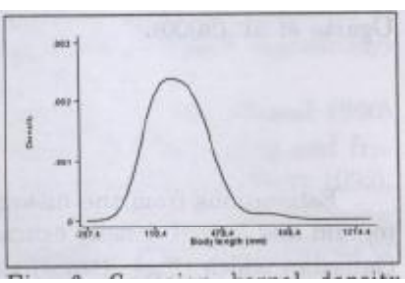

Fig. 6. Gaussian kernel density estimation using the biased cross validation bandwidth (h = 92)

Table 2. Critical KDE's bandwidths and estimated significance levels for female and unknown sex sample individuals, n = 47.

Number of modes	Critical bandwidth	p
1	84.5	0.11
2	75.8	0.01
3	31.8	0.32
4	20.12	0.32
5	16.99	0.18
6	11.96	0.48
7	8.9	0.57
8	8.73	0.35

Note: The p values were obtained from B = 100 bootstrap replications of size 47.

The density estimation with 6 modes derived from the Silverman test: a bandwidth h = (16.99+11.96)/2 » (17+12)/2 = 14.5 was considered to identify the Gaussian components in the size mix by the Bhattacharya method. The small mode at 220 mm was not included for considering, it is the result of the overlapping of the adjacent components. The first five means (Table 3) were used to estimate, by weighted nonlinear regression the von Bertalanffy growth function, resulting in the following equation:

$$L_t = 773.2976 \left[1 - e^{0.2068\ (t\ +1.0546)}\right] \qquad \text{adjusted } r^2 = 0.9986$$

The von Bertalanffy growth function estimated from length frequency analysis was compared with those (Table 4) derived from weighted nonlinear

Table 3. Component means determined from the Gaussian kernel density estimation (h = 14.5), female and individuals of unknown sex, Spring, n = 47.

Group	Estimated size	Mean (standard body length)	Standard error
0	18	149.43	4.96
1	7	286.01	6.48
2	14	351.55	7.68
3	4	452.92	8.67
4	1	501.41	15.07
5	1	754.00	14.50

Note: The sum of the estimated sizes is 45. The missing individuals are in the residual frequency after subtracting the estimated Gaussian components.

Table 4. Weighted nonlinear regression for von Bertalanffy growth function parameter estimation and growth performance index values by ageing method

Parameter	Value	Standard error	t value	$P > \lvert t \rvert$	95% Confidence interval	
a) Length frequency (LF) adjusted r^2 = 0.9986						
L_∞	773.2976	74.5034	10.379	0.000	622.8349	923.7603
K	0.2068	0.0324	6.371	0.000	0.1412	0.2723
t_0	-1.0546	0.0644	-16.380	0.000	-1.1846	-0.9245
f	3.2077					
b) Scales (SC) adjusted r^2 = 0.9990						
L_∞	900.5134	22.8443	39.420	0.000	854.4737	946.5532
K	0.1571	0.0070	22.282	0.000	0.1429	0.1713
t_0	-1.2738	0.04114	-30.730	0.000	-1.3573	-1.1902
f	3.2195					
c) Whole otoliths (WO) adjusted r^2 = 0.9918						
L_∞	785.5918	41.1347	19.098	0.000	702.3890	868.7947
K	0.1715	0.0211	8.122	0.000	0.1288	0.2142
t_0	-1.1814	0.1777	-6.648	0.000	-1.5409	-0.8219
f	3.1400					
d) Otolith sections (OS) adjusted r^2 = 0.9954						
L_∞	735.4087	23.8618	30.820	0.000	687.2537	783.5638
K	0.2515	0.0172	14.609	0.000	0.2168	0.2863
T_0	-1.0035	0.0655	-15.327	0.000	-1.1356	-0.8714
f	3.2495					

Note: Standard errors, P values, CI's and correlations are asymptotic approximations. The f values were calculated using total length values (cm) predicted by the regression with standard body length

regression of the mean length at age data from hard parts readings (Table 5). The growth curves are presented in figure 8.

Discussion

In fisheries studies, the use of modes in size frequency distributions of aquatic organisms have been advocated as an attempt to identify groups of fish with similar age. This would be the case if the sample of size is unbiased and

Table 5. Number of individuals by sampling date and estimated age by hard part reading. The mean body length by age is included in the Totals column.

	Sample number				
Age	6	7	8	9	Total (Mean)
a) Scales (SC)					
0	14	3	0	0	17 (168.18)
1	4	4	2	1	11 (257.10)
2	0	5	2	7	14 (361.29)
3	0	2	0	2	4 (462.00)
11	0	0	1	0	1 (760.00)
Total	18	14	5	10	47
b) Whole otoliths (WO)					
0	3	1	0	0	4 (164.00)
1	13	6	2	3	24 (230.96)
2	0	4	2	5	11 (362.00)
3	0	0	0	1	1 (394.00)
6	0	0	0	1	1 (454.00)
15	0	0	1	0	1 (760.00)
Total	16	11	5	10	42
c) Otolith sections (OS)					
0	17	1	0	0	18 (158.39)
1	1	8	2	3	14 (301.86)
2	0	3	2	6	11 (392.55)
5	0	0	0	1	1 (454.00)
14	0	0	1	0	1 (760.00)
Total	18	12	5	10	45

Note: Sample numbers correspond to the following dates (1994): 6 = 19/02; 7 = 16/03; 8 = 14/04; 9 = 13/05.

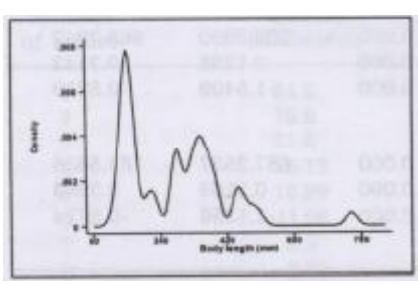

Fig. 7. Gaussian kernel density estimate using the bandwidth suggested by the Silverman test (h = 14.5).

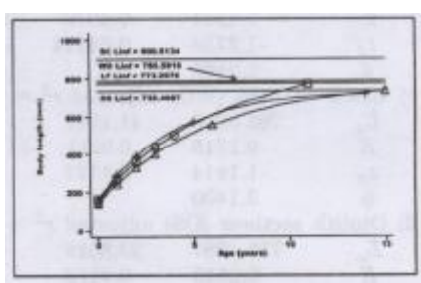

Fig. 8. Comparison of the von Bertalanffy growth functions estimated by length frequency (LF, circles), scales (SC, squares), whole otoliths (WO, triangles) and otolith sections (OS, plus signs).

the species under analysis reproduces during a relatively short span of time at regular periods (King 1995). There are several factors affecting the occurrence of modes in size data such as the sample size, the distance between adjacent means, the shape of the distribution, and the magnitude of the correspondent variances (Caddy 1986; Erzini 1990). Schnute & Fournier (1980) remark that length-frequency analysis tends to lump the final age-classes together if they are in close proximity or contain small percentages of fish. In such cases it may be impossible to distinguish the final ages, and the best approach may be to assume that all fish beyond a certain age comprise a single group.

For the length distribution of *Lateolabrax japonicus* analyzed in this study, comparing the mean lengths at age (Table 3) with those estimated by the hard-part readings (Table 5) it is remarkable that the nonparametric assessment of multimodality provided a density estimation with the modes corresponding correctly to ages 0, 1 and 2 occurring in the Spring samples. The identification of the groups of ages 1 and 2 was not evident in any of the KDE's using the rules for bandwidth selection. The undersmoothed L2CV and oversmoothed BCV density estimations emphasize the high variability of the former, the conservative tendency of the later (Härdle 1991; Scott 1992) and the usefulness of the test of multimodality in dealing with small samples.

In relation to growth function estimation (VBGF), the weighted nonlinear regressions were statistically significant (P values < 0.05), but only the equation estimated from scales produced statistical significance (t values with $P < 0.05$) for all the three parameter estimates (L_∞, K and t_0). The growth function calculated from length frequency analysis had the wider confidence intervals for the equation parameters (Table 4). In figure 8, it will be noted that the growth curve from length frequency analysis were very close at smaller ages to the curve from scale reading, which means that inside the age range of 0 to 5 they can be used to produce essentially the same results. However, the growth expressions obtained from otoliths (whole and sectioned) were very similar to the length frequency equation as the estimated growth parameters were contained in the corresponding confidence intervals of the otolith growth functions (Table 4). The weighted nonlinear regression attenuated the influence of estimated ages of 15 and 14 (by whole and sectioned otoliths, respectively) for the fish with a body length of 760.

It is not possible to make direct comparisons with other studies because none discriminated sexes in the growth estimations. However it is worth mentioning that the values obtained in the present study are in accordance with the general equations reported for the species at different Japanese localities ranging from (L_∞, K and t_0, respectively) 1112.503, 0.1577, -0.3115 at Wakasa Bay (Kuwatani 1962) to 741.0699, 0.1930, -0.6595 at Sendai Bay (Hatanaka and Sekino 1962). The growth performance index (Pauly and Munro 1984) values (Table 4), were inside of the range (3.14 to 3.35) reported in *L. japonicus* section in FishBase (Froese and Pauly 2001). Assuming the same size frequency structure and repeating it during four years, we used the link means routine of FISAT (Gayanilo et al. 1993) with the four first age groups from table 3 arriving at the following results: L_∞ = 797.85; K = 0.21. These values are reasonably close to those from the nonlinear weighted regression and inside the confidence intervals for the quoted parameter estimates found in this work.

Conclusions

For the analyzed data set (female and unknown sex individuals collected in the Spring months), the modes suggested by the multimodality test were in good agreement with the means length values from ages 0 to 2. Depending on the hard part, the larger modes corresponded to fish of 3, 5 or 6 years old besides the largest individual with estimated age from 11 to 14. The limited number of specimens made it difficult to draw firm conclusions for older fish. Clearly, the necessity of larger samples including older fish was perceived.

As a general conclusion we emphasize the fact that if the sample of lengths contains information on the cohorts, the use of KDE's in combination with optimal, oversmoothed and cross-validation bandwidth rules, in addition to nonparametric assessment of multimodality procedures such as the DIP and over all the Silverman's tests, has proved to be a valuable alternative procedure for its extraction. In our opinion, these sets of nonparametric computing intensive statistical procedures have an enormous potential to provide useful guidelines for bandwidth selection leading to density estimations with a significant number modes. More research on the subject is guaranteed. The computer programs for kernel density estimators calculation, multimodality assessment and Gaussian characterization are available from the first author.

Acknowledgments

The first author is grateful to the Ministry of Education, Science and Culture of Japan and to the Universidad Nacional Autónoma de México (F.E.S. Zaragoza, D.G.A.P.A. and D.G.A.P.A.-P.A.P.I.I.T. project IN217596) for their support. We are also grateful to the two anonymous referees for their valuable suggestions and corrections on the manuscript.

References

Caddy, J.F. 1986. Size frequency analysis in stock assessment ¾some perspectives, approaches and problems. *Proceedings of the 37th Annual Gulf and Caribbean Fisheries Institute*: 212-238.

Erzini, K. 1990. Sample size and grouping of data for length-frequency analysis. *Fisheries Research* 9; 355-366.

Eschmeyer, W.N., C. J. Ferraris, Jr., M. Dang Hoang, and D.J. Long. 1996. *A Catalog Of The Species Of Fishes*. California Academy of Sciences, Golden Gate Park, San Francisco, California. http://www.calacademy.org/research/ichthyology/species/

Froese, R. and D. Pauly (Eds.) 2001. FishBase. World Wide Web electronic publication. www.fishbase.org, 07 June 2001.

Gayanilo, F.C., P. Sparre and D. Pauly. 1993. *The FiSAT User's Guide*. F.A.O. Rome: 4-5, 4-6.

Good, I.J. and R.A. Gaskins. 1980. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, 75: 42-73.

Härdle, W. 1991. *Smoothing Techniques. With Implementations in S*. Springer-Verlag. New York

Hartigan, J.A. and P.M. Hartigan. 1985. The Dip test of unimodality. *The annals of Statistics*, 13: 70-84.

Hatanaka, M. and K. Sekino. 1962. Ecological studies on the Japanese sea-bass, *Lateolabrax japonicus* - II. Growth. *Bulletin of the Japanese Society of Scientific Fisheries*, 28(9): 857-861. (in Japanese with English summary).

Izenman, A.J., and C. Sommer. 1988. Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association*, 83(404): 941-953.

King, M. 1995. *Fisheries Biology, Assessment and Management*. Fishing News Books. Oxford, 341 p.

Kuwatani, Y. 1962. Suzuki o taisho to suru gyosho no sogoteki kenkyu (The synthetic study on the fish bank for the Japanese sea bass, *Lateolabrax japonicus*). *Bulletin of the Kyoto Prefecture Fisheries Research Laboratory*, 8: 1-129. (in Japanese).

Pauly, D. and J.L. Munro. 1984. Once more on growth comparison in fish and invertebrates. *Fishbyte* 2(1): 21.

Salgado-Ugarte, I.H. 1995. Nonparametric methods for fisheries data analysis and their application in conjunction with other statistical techniques to study biological data of the Japanese sea bass *Lateolabrax japonicus* in Tokyo Bay. Unpublished Ph.D. Dissertation. University of Tokyo, Faculty of Agriculture, Department of Aquatic Bioscience, 389 p.

Salgado-Ugarte, I.H., M. Shimizu, and T. Taniuchi. 1993. Exploring the shape of univariate data using kernel density estimators. *Stata Technical Bulletin* 16: 8-19.

Salgado-Ugarte, I.H., M. Shimizu, and T. Taniuchi. 1994. Semi-graphical determination of Gaussian components in mixed distributions. *Stata Technical Bulletin* 18: 15-27.

Salgado-Ugarte, I.H., M. Shimizu, and T. Taniuchi. 1995a. ASH, WARPing, and kernel density estimation for univariate data. *Stata Technical Bulletin* 26: 2-10.

Salgado-Ugarte, I.H., M. Shimizu, and T. Taniuchi. 1995b. Practical rules for bandwidth selection in univariate density estimation. *Stata Technical Bulletin* 27: 5-19.

Salgado-Ugarte, I.H., M. Shimizu, and T. Taniuchi. 1997. Nonparametric assessment of multimodality for univariate data. *Stata Technical Bulletin* 32: 27-35.

Salgado-Ugarte, I.H., J. Martínez-Ramírez, J.L. Gómez-Márquez and B. Peña-Mendoza. 2000. Some programs for growth estimation in fisheries biology. *Stata Technical Bulletin* 53: 35-47.

Schnute, J. and D. Fournier. 1980. A new approach to length-frequency analysis: growth structure. *Canadian Journal of Fisheries and Aquatic Sciences*, 37: 1337-1351.

Scott, D.W. 1979. On optimal and data-based histograms. *Biometrika*, 66: 605-610.

Scott, D.W. 1985. Frequency polygons: Theory and application. *Journal of the American Statistical Association*, 80(390): 348-354.

Scott, D.W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons.

Silverman, B.W. 1978. Choosing the window width when estimating a density. *Biometrika*, 65: 1-11.

Silverman, B.W. 1981a. Density estimation for univariate an bivariate data. In *Interpreting Multivariate Data*, ed. V. Barnett, 37-53, Chichester, John Wiley and Sons.

Silverman, B.W. 1981b. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, B*, 43: 97-99.

Silverman, B.W. 1986. *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Terrell, G.R. 1990. The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, 85(410): 470-477.

Wong, M.A. 1985. A bootstrap testing procedure for investigating the number of subpopulations. *Journal of Statistical Computation and Simulation*. 22: 99-112.

Manuscript received 28 October 1999; Accepted 13 June 2002