Pakistan Journal of Statistics and Operation Research

A General Base of Power Transformation to Improve the Boundary Effect in Kernel Density without Shoulder Condition

Baker Albadareen[1*], Noriszura Ismail[2]

* Corresponding Author

1. Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia, albadareen.baker@gmail.com
2. Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia, ni@ukm.edu.my

Abstract

In this paper, a general base of power transformation under the kernel method is suggested and applied in the line transect sampling to estimate abundance. The suggested estimator performs well at the boundary compared to the classical kernel estimator without using the shoulder condition assumption. The transformed estimator show smaller value of mean squared error and absolute bias from the efficiency results obtained using simulation.

Key Words: line transect; power-transformation; kernel estimator; shoulder condition; abundance; bandwidth.

Mathematical Subject Classification: 62G07

1. Introduction

Line transect sampling is considered as a common technique to estimate population abundance (density). For this method, the study area A divides into non-overlapping parts (strips) with total length L, assuming that an observer follows each strip and records the perpendicular distance of each detected animal (object). The perpendicular distances $x_1, x_2, \dots x_n$ are a random sample that has a probability density function $f(x)$. The density of the objects is computed by $D = nf(0)/2L$, and its estimate is $\hat{D} = n\hat{f}(0)/2L$ (Buckland et al., 2001).

The intuitive condition in the line transect sampling is that the probability of detecting an object, $g(x)$, is a conditional and non-increasing function of exposing an object given that the object is far away from the line by distance x. Assuming a random sample of perpendicular distances $x_1, x_2, \dots x_n$, the probability density function $f(x)$ is related to the detection function $g(x)$ by $f(x) = g(x)/\int g(u)du$, i.e. the functions $g(x)$ and $f(x)$ have the same distribution shape (Buckland et al., 2001).

The shape of both functions $g(x)$ and $f(x)$ at $x = 0$ can be generally characterized into two types; the one that has a shoulder shape at $x = 0$ (i.e. the probability of detecting objects around the transect line is usually certain) which is equivalent to the mathematical form $f'(0) = 0$, and the one that is without a shoulder shape at $x = 0$. Several tests can be applied to examine whether the random sample of perpendicular distances satisfies the shoulder condition (see Zhang (2001)). In practice, several studies indicated that the shoulder condition is not valid for the line transect data of a particular community (see Bauer, Fromentin, Demarcq, Brisset, & Bonhommeau, 2015; Buckland, 1985).

Several approaches can be found in the literature for the estimation of $f(0)$. This article considers a common method for the estimation which is known as "the non-parametric kernel method". The method allows the data to demonstrate itself. Silverman (1986) stated and summarized the general frame of the non-parametric kernel method which is given by:

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), -\infty < x < \infty \quad (1)$$

where h is the bandwidth parameter and $K(.)$ is the kernel function.

The kernel "Rosenblatt–Parzen" density estimator is commonly applied in the literature to find a reliable estimate of $f_X(0)$; the estimator is direct, easy, and allows the sample to illustrate its characteristic density value at a chosen value of x. For the case of reducing the estimation bias, several studies were considered on the related kernel estimator. For examples, Jones, Linton, and Nielsen (1995) proposed a simple bias reduction method for density estimation, Cheng, Fan, and Marron (1997) investigated the best weight functions for local polynomial fitting at endpoints to fix the boundary correction, Mack (2002) suggested several techniques to reduce bias, Eidous (2005) proposed frequency nonparametric histogram estimators, Karunamuni and Alberts (2006) used a transformation that is easy to implement to correct bias around the boundary, Koekemoer and Swanepoel (2008) proposed a semi-parametric kernel density estimator based on transformation, Eidous (2011a) introduced an additive histogram frequency estimator based on the case of the shoulder condition doesn't valid using line transect method, Eidous (2012) proposed a new kernel estimator for abundance without the shoulder condition, Wen and Wu (2015) made an improved transformation-based kernel estimator of densities on the unit interval, and Eidous (2015) improved the histogram estimation for $f(0)$ applying line transect data with and without the shoulder condition. Recently, Albadareen and Ismail (2017) introduced several kernel estimators for $f_X(0)$, Eidous and Al-Eibood (2018) proposed a bias-corrected histogram estimator for line transect sampling, Albadareen and Ismail (2018) suggested a generalized form of Epanechnikov kernel function to the adaptive estimation of $f_X(0)$, and Albadareen and Ismail (2019) proposed a form of power-transformation to the adaptive estimation of $f_X(0)$ when the shoulder condition is violated.

Assume that a random sample of the line transect method with non-negative distances are $x_1, x_2, \dots, x_n$. When an asymmetric kernel function is assumed, Chen (1996) derived the classical reflection estimator of $f_X(x)$ at $x = 0$ as:

$$\hat{f}_X(0) = \frac{2}{nh}\sum_{i=1}^{n} K\left(\frac{x_i}{h}\right) \quad (2)$$

The bias and variance of the estimator (2) are:

$$Bias\left[\hat{f}_X(0)\right] = 2hf'_X(0)\int_0^\infty uK(u)\,du + h^2 f''_X(0)\int_0^\infty u^2K(u)\,du + o(h^2) \quad (3)$$

$$= 2hf'_X(0)\int_0^\infty uK(u)\,du + O(h^2) \quad (4)$$

$$Var\left[\hat{f}_X(0)\right] = \frac{4}{nh}f_X(0)\int_0^\infty K^2(u)\,du + o\left(\frac{1}{nh}\right) \quad (5)$$

The asymptotic mean squared error (AMSE) is:

$$AMSE\left[\hat{f}_X(0)\right] = \frac{4}{nh}f_X(0)\int_0^\infty K^2(u)\,du + \left(2f'_X(0)h\int_0^\infty u\,K(u)\,du\right)^2 \quad (6)$$

In this study, a general base of power-transformation of perpendicular distance under the kernel method is proposed for the population density when the shoulder condition is violated. The asymptotic theoretical properties (bias, variance, and mean squared error) of the estimator are derived and compared to the classical reflection of the kernel estimator. The efficiency results are supported by simulation studies, and the performance comparison is carried out between the proposed estimator and the classical kernel estimator.

2. Methodology

In some cases, the kernel estimator in (2) yields underestimated values and has a large negative bias under the line transect method (see Eidous (2011b)). Several bias reduction techniques were suggested in the literature, and one of the common method is the transformation approach (see Charpentier & Flachaire, 2015; Devroye & Gyorfi, 1985; Marron & Ruppert, 1994). In this article, we propose the power transformation with general base form and apply the transformation on the kernel estimator when $f'_X(0) \neq 0$. Assuming that the range of the perpendicular distances X are $0 \leq X \leq w$, the proposed transformation is $Y = a^{X/w} - 1,\ a > 1$, where a is a general base of power transformation. This function transforms the perpendicular distances by a non-decreasing function that produces estimator $f_Y(0)$. The original estimator in equation (2), which is $f_X(x)$, is applied to the original data and the transformed estimator, $f_Y(0)$, is obtained from the back-transformation such that:

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X\big(w\log_a(y+1)\big)\left(\frac{w}{(y+1)\log(a)}\right) \tag{7}$$

so that $f_Y(0) = f_X(0)\left(\frac{w}{\log(a)}\right)$, i.e. when $x = 0$ then $y = 0$.

In the line transect method, the estimation of $f_X(x)$ is required at $x = 0$. For this proposed method (estimation by back-transformation), an equivalent estimation value $\hat{f}_Y(0)\left(\frac{\log(a)}{w}\right)$ is substituted. To obtain the estimation value at $y = 0$, the kernel estimator with respect to Y in equation (2) can be applied, and the transformed kernel estimator $\hat{f}_Y(0)$ is:

$$\hat{f}_Y(0) = \frac{2}{nh}\sum_{i=1}^{n} K\left(\frac{y_i}{h}\right), \qquad y_i = a^{x_i/w} - 1 \tag{8}$$

If the kernel function $K(u)$ is assumed to follow a Gaussian density, the estimated values obtained from (2) and (8) converge to zero as $x > w$, when considering that w is large (such as $w \geq \max(x_i) + 4h$). The density value of $K\left(\frac{x \mp x_i}{h}\right)$ disappear when $|x \mp x_i| > 4h$.

The bias and variance of $\hat{f}_Y(0)$ are:

$$Bias\big[\hat{f}_Y(0)\big] = 2hf_Y'(0)\int_0^\infty uK(u)\,du + h^2 f_Y''(0)\int_0^\infty u^2K(u)\,d + o(h^2) \tag{9}$$

$$= 2h\left(\left(\frac{w}{\log(a)}\right)^2 f_X'(0) - \left(\frac{w}{\log(a)}\right)f_X(0)\right)\int_0^\infty uK(u)\,du + O(h^2) \tag{10}$$

$$Var\big[\hat{f}_Y(0)\big] = \frac{4}{nh}f_Y(0)\int_0^\infty K^2(u)\,du + o\left(\frac{1}{nh}\right) \tag{11}$$

$$= \frac{4}{nh}\left(\frac{w}{\log(a)}\right)f_X(0)\int_0^\infty K^2(u)\,du + o\left(\frac{1}{nh}\right) \tag{12}$$

The asymptotic mean squared error is obtained by assuming the small terms $o(.)$ and $O(.)$ to be zero,

$$AMSE\big[\hat{f}_Y(0)\big] = \frac{4}{nh}\left(\left(\frac{w}{\log(a)}\right)f_X(0)\right)\int_0^\infty K^2(u)\,du + \left[2h\left(\left(\frac{w}{\log(a)}\right)^2 f_X'(0) - \left(\frac{w}{\log(a)}\right)f_X(0)\right)\int_0^\infty uK(u)\,du\right]^2 \tag{13}$$

It should be noted that $Var\big[\hat{f}_Y(0)\big] \leq Var\big[\hat{f}_X(0)\big]$ if $\frac{w}{\log(a)} \leq 1$. The value of a that produces a smaller theoretical variance is defined as $\frac{w}{\log(a)} \leq 1$, i.e. $a \geq e^w$, under the constrain $a > 1$. Without loss of generality, the base value $a = e^w$ will be assumed throughout this article.

3. Simulation

The theoretical asymptotic value in (13) is derived based on a large sample assumption. The simulation study is carried out to compare and examine the proposed estimator $\hat{f}_Y(0)$ with the classical kernel estimator $\hat{f}_X(0)$ using different small sample sizes, which are $n = 50, 100, \text{and } 500$. The efficiency measurements are the relative bias $RB = \{E\big[\hat{f}(0) - f(0)\big]\}/f(0)$ and the relative mean error $RME = \sqrt{MSE\big[\hat{f}(0)\big]}/f(0)$.

Random samples from two common density families are generated. These families are the reference densities when the shoulder condition is violated. Four different detection functions are also chosen for each model which cover wide possibility of density shapes. The two density models considered are:

a) Beta (BE) model (Eberhardt, 1968)
The detection function is $g(x) = (1-x)^\beta,\ 0 \leq x \leq \omega,\ \beta \geq 1$, and $f(x) = (1+\beta)(1-x)^\beta,\ 0 \leq x \leq \omega,$ $\beta \geq 1$. The density parameter values $\beta = 3.0, 4.0, 5.0 \text{ and } 6.0$ are chosen with the truncation point $\omega = 1$ for these models.

b) Negative exponential model (Gates, Marshall, & Olson, 1968)
The detection function is $g(x) = e^{-\beta x}, \beta > 0, 0 \le x \le w$, and $f(x) = \beta e^{-\beta x},\ 0 \le x \le w$. The density parameter values $\beta = 1.5, 2.0, 2.5,$ and 3.0 are chosen with the truncation point $\omega = 3.0$ for these models.

Bandwidth selection

The efficiency of the kernel estimator is based on the value of bandwidth. Since the mean squared error of the kernel estimator in equation (2) does not produce large variability by applying different symmetric kernel functions such as Gaussian, Epanechnikov, and biweight (Ghosh, 2018), the kernel function $K(u)$ is assumed to follow the standard normal distribution for comparison purposes.

For our study, recommended bandwidth approaches are applied for the original data and the transformed data. The two estimators considered are:

- Estimator 1 (Est1): The kernel estimator given by equation (2) is considered using the original data with the bandwidth method recommended by Silverman (1986), which minimizes the mean integrated squared error; $h = 1.06\,\hat{\sigma}\,n^{-\frac{1}{5}}$, where $\hat{\sigma} = \sqrt{\sum_{i=1}^{n} x_i^2/n}$. Although these bandwidth value was computed based on a reference density has a shoulder (i.e. the half-normal density), it is better to compute h based on another density hasn't a shoulder as it is assumed with the proposed estimator (i.e. the negative exponential function).

- Estimator 2 (Est2): The proposed estimator given by equation (8) is considered using the transformed data with the bandwidth method that minimizes $AMSE\left[\hat{f}_Y(0)\right]$. The bandwidth is $h = \left(\frac{\left(\left(\frac{w}{\log(a)}\right)f_X(0)\right)\int_0^\infty K^2(u)du}{2n\left[\left(\left(\frac{w}{\log(a)}\right)^2 f_X'(0) - \left(\frac{w}{\log(a)}\right)f_X(0)\right)\int_0^\infty uK(u)du\right]^2}\right)^{1/3}$. A suitable reference density is assumed to substitute the estimated values of $f_X(0)\ and\ f_X'(0)$ for the case that the shoulder condition is violated. The reference density is the negative exponential function (see Al-Bassam & Eidous, 2018; Mack & Quang, 1998; Silverman, 1986), such that $\left(\hat{f}_X(0) = \frac{1}{\bar{x}}\right)$ and $\left(\hat{f}_X'(0) = \frac{-1}{\bar{x}^2}\right)$. The bandwidth of estimator 2 is $h = \left(\frac{\left(\left(\frac{w}{\log(e^w)}\right)\left(\frac{1}{\bar{x}}\right)\right)\left(\frac{1}{4\sqrt{\pi}}\right)}{2n\left[\left(\left(\frac{w}{\log(e^w)}\right)^2\left(\frac{-1}{\bar{x}^2}\right) - \left(\frac{w}{\log(e^w)}\right)\left(\frac{1}{\bar{x}}\right)\right)\left(\frac{1}{\sqrt{2\pi}}\right)\right]^2}\right)^{1/3}$.

4. Simulation results

Table 1 and Table 2 provide the simulation results. The transformed estimator (Est2) shows smaller absolute relative bias and relative mean error than the traditional kernel estimator (Est1) under both families. Likewise, the relative mean errors of estimator 2 (Est2) decrease as the sample sizes increase, i.e. (Est2) provides a more consistent fit asymptotically as illustrated in Figure 1.

Table 1. Simulation results of negative exponential family

		$n = 50$		$n = 100$		$n = 500$	
β	Estimator	RB	RME	RB	RME	RB	RME
1.5	Est1	-0.3445	0.3577	-0.3067	0.3146	-0.2419	0.2452
	Est2	-0.1488	0.2930	-0.1113	0.2253	-0.0670	0.1408
2	Est1	-0.3638	0.3753	-0.3264	0.3349	-0.2594	0.2621
	Est2	-0.1536	0.2799	-0.1072	0.2236	-0.0773	0.1398
2.5	Est1	-0.3645	0.3760	-0.3389	0.3464	-0.2672	0.2697
	Est2	-0.1479	0.2703	-0.1382	0.2259	-0.0783	0.1381
3	Est1	-0.3638	0.3759	-0.3388	0.3462	-0.2680	0.2708
	Est2	-0.1466	0.2664	-0.1337	0.2212	-0.0780	0.1368

Table 2. Simulation results of the beta model

		$n = 50$		$n = 100$		$n = 500$	
β	Estimator	RB	RME	RB	RME	RB	RME
3	Est1	-0.2483	0.2691	-0.2231	0.2374	-0.1706	0.1767
	Est2	-0.1019	0.2606	-0.0802	0.2176	-0.0547	0.1320
4	Est1	-0.2695	0.2868	-0.2430	0.2556	-0.1878	0.1930
	Est2	-0.1015	0.2510	-0.0901	0.2106	-0.0530	0.1263
5	Est1	-0.2862	0.3040	-0.2577	0.2696	-0.1984	0.2027
	Est2	-0.1092	0.2595	-0.0861	0.2047	-0.0503	0.1255
6	Est1	-0.2891	0.3063	-0.2682	0.2789	-0.2089	0.2133
	Est2	-0.1042	0.2545	-0.0923	0.2124	-0.0584	0.1271

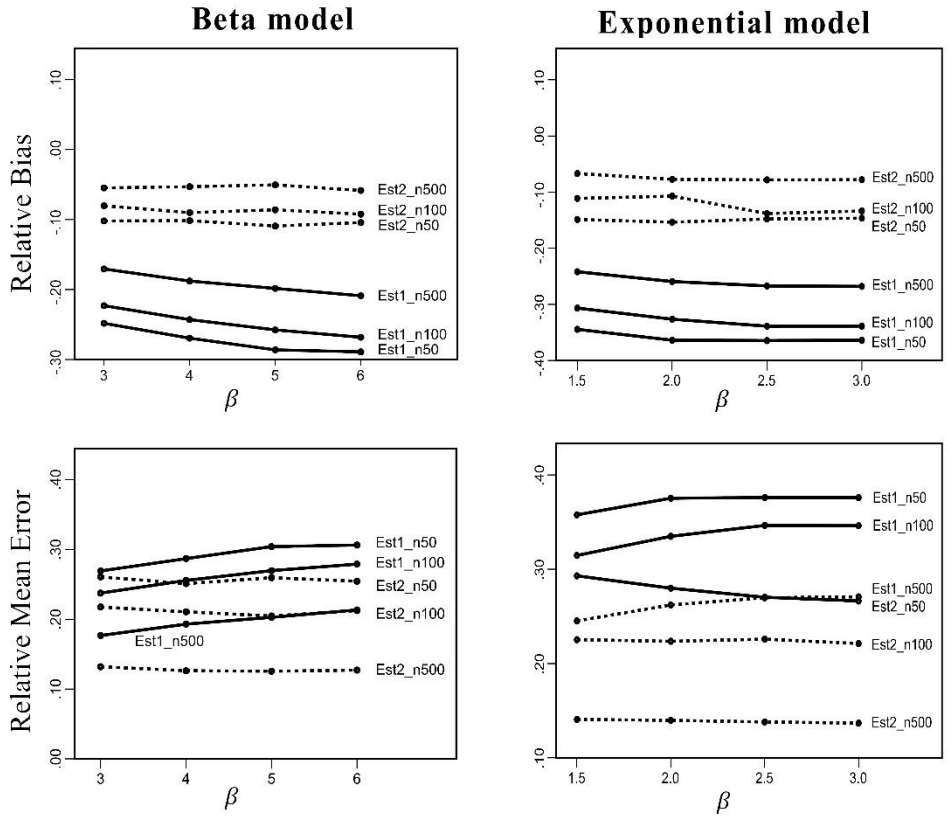

Figure 1. RB and RME of the simulation results of the Beta and the negative exponential model

5. Conclusion

This article proposed an adaptive method of the kernel estimator to estimate the population abundance (density) at the boundary under the line transect method. A general base of power-transformation is suggested to improve the estimation efficiency when the shoulder condition is violated. The proposed transformation estimator presents more efficient and consistent results than the traditional kernel estimator. The asymptotic bias, variance and mean squared error of the proposed estimator are also derived. The simulation results show that the proposed estimator is more efficient than the traditional kernel estimator.

Acknowledgments

The authors would like to thank the editor and the reviewer for their valuable comments and constructive reviews to further improve this paper. The authors are also grateful and would like to acknowledge the financial support granted by the Ministry of Higher Education (MOHE) and Universiti Kebangsaan Malaysia (UKM) in the form of research grants (FRGS/1/2019/STG06/UKM/01/5 and GUP-2019-031).

References

1. Al-Bassam, M., & Eidous, O. M. (2018). Combination of parametric and nonparametric estimators for population abundance using line transect sampling. *Journal of Information and Optimization Sciences*, *39*(7), 1449–1462. https://doi.org/10.1080/02522667.2017.1367510
2. Albadareen, B., & Ismail, N. (2017). Several new kernel estimators for population abundance. *AIP Conference Proceedings*, *1830*(1), 80018. https://doi.org/10.1063/1.4981002
3. Albadareen, B., & Ismail, N. (2018). Adaptive kernel function using line transect sampling. *AIP Conference Proceedings*, *1940*, 020112. https://doi.org/10.1063/1.5028027
4. Albadareen, B., & Ismail, N. (2019). An Adaptation of Kernel Density Estimation for Population Abundance using Line Transect Sampling When the Shoulder Condition is Violated. *International Journal of Innovative Technology and Exploring Engineering*, *9*(2), 3494–3498. https://doi.org/10.35940/ijitee.B6582.129219
5. Bauer, R. K., Fromentin, J.-M., Demarcq, H., Brisset, B., & Bonhommeau, S. (2015). Co-Occurrence and Habitat Use of Fin Whales, Striped Dolphins and Atlantic Bluefin Tuna in the Northwestern Mediterranean Sea. *PLOS ONE*, *10*(10), e0139218. https://doi.org/10.1371/journal.pone.0139218
6. Buckland, S. T. (1985). Perpendicular Distance Models for Line Transect Sampling. *Biometrics*, *41*(1), 177. https://doi.org/10.2307/2530653
7. Buckland, Stephen T, Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. (2001). *Introduction to distance sampling: estimating abundance of biological populations* (1st ed.). London: Oxford University Press. Retrieved from https://global.oup.com/academic/product/introduction-to-distance-sampling-9780198509271?q=9780198509271&cc=my&lang=en
8. Charpentier, A., & Flachaire, E. (2015). Log-Transform Kernel Density Estimation of Income Distribution. *L'Actualité Économique*, *91*(1–2), 141–159. https://doi.org/10.7202/1036917ar
9. Chen, S. X. (1996). A Kernel Estimate for the Density of a Biological Population by Using Line Transect Sampling. *Applied Statistics*, *45*(2), 135. https://doi.org/10.2307/2986150
10. Cheng, M. Y., Fan, J., & Marron, J. S. (1997). On automatic boundary corrections. *Annals of Statistics*, *25*(4), 1691–1708. https://doi.org/10.1214/aos/1031594737
11. Devroye, L., & Gyorfi, L. (1985). *Nonparametric Density Estimation: The L 1 View*. *Journal of the Royal Statistical Society. Series A (General)* (1st ed., Vol. 148). New York: John Wiley and Sons. https://doi.org/10.2307/2981908
12. Eberhardt, L. L. (1968). A Preliminary Appraisal of Line Transects. *The Journal of Wildlife Management*, *32*(1), 82. https://doi.org/10.2307/3798239
13. Eidous, O. M. (2005). Frequency Histogram Model For Line Transect Data With And Without The Shoulder Condition. *Journal of the Korean Statistical Society*, *34*(1), 49–60. Retrieved from http://www.koreascience.or.kr/article/JAKO200516610508354.page
14. Eidous, O. M. (2011a). Additive histogram frequency estimator for wildlife abundance using line transect data without the shoulder condition. *Metron*, *69*(2), 119–128. https://doi.org/10.1007/BF03263552
15. Eidous, O. M. (2011b). Variable location kernel method using line transect sampling. *Environmetrics*, *22*(3), 431–440. https://doi.org/10.1002/env.1082
16. Eidous, O. M. (2012). A new kernel estimator for abundance using line transect sampling without the shoulder condition. *Journal of the Korean Statistical Society*, *41*(2), 267–275. https://doi.org/10.1016/j.jkss.2011.09.004
17. Eidous, O. M. (2015). Nonparametric Estimation of f(0) Applying Line Transect Data with and without the Shoulder Condition. *Journal of Information and Optimization Sciences*, *36*(4), 301–315. https://doi.org/10.1080/02522667.2013.867726
18. Eidous, O. M., & Al-Eibood, F. (2018). A bias-corrected histogram estimator for line transect sampling. *Communications in Statistics - Theory and Methods*, *47*(15), 3675–3686.

https://doi.org/10.1080/03610926.2017.1361987
19. Gates, C. E., Marshall, W. H., & Olson, D. P. (1968). Line Transect Method of Estimating Grouse Population Densities. *Biometrics*, *24*(1), 135. https://doi.org/10.2307/2528465
20. Ghosh, S. (2018). *Kernel Smoothing: Principles, Methods and Applications* (1st ed.). New York: Chapman & Hall. https://doi.org/10.1002/9781118890370
21. Jones, M. C., Linton, O., & Nielsen, J. P. (1995). A simple bias reduction method for density estimation. *Biometrika*, *82*(2), 327–338. https://doi.org/10.1093/biomet/82.2.327
22. Karunamuni, R. J., & Alberts, T. (2006). A locally adaptive transformation method of boundary correction in kernel density estimation. *Journal of Statistical Planning and Inference*, *136*(9), 2936–2960. https://doi.org/10.1016/j.jspi.2004.12.014
23. Koekemoer, G., & Swanepoel, J. W. H. (2008). Transformation Kernel density estimation with applications. *Journal of Computational and Graphical Statistics*, *17*(3), 750–769. https://doi.org/10.1198/106186008X318585
24. Mack, Y. P. (2002). Bias-corrected confidence intervals for wildlife abundance estimation. *Communications in Statistics - Theory and Methods*, *31*(7), 1107–1122. https://doi.org/10.1081/STA-120004909
25. Mack, Y. P., & Quang, P. X. (1998). Kernel Methods in Line and Point Transect Sampling. *Biometrics*, *54*(2), 606. https://doi.org/10.2307/3109767
26. Marron, J. S., & Ruppert, D. (1994). Transformations to Reduce Boundary Bias in Kernel Density Estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(4), 653–671. https://doi.org/10.1111/j.2517-6161.1994.tb02006.x
27. Silverman, B. W. (1986). *Density estimation: For statistics and data analysis*. *Density Estimation: For Statistics and Data Analysis* (1st ed.). London: Chapman & Hall. https://doi.org/10.1201/9781315140919
28. Wen, K., & Wu, X. (2015). An Improved Transformation-Based Kernel Estimator of Densities on the Unit Interval. *Journal of the American Statistical Association*, *110*(510), 773–783. https://doi.org/10.1080/01621459.2014.969426
29. Zhang, S. (2001). Generalized likelihood ratio test for the shoulder condition in line transect sampling. *Communications in Statistics - Theory and Methods*, *30*(11), 2343–2354. https://doi.org/10.1081/STA-100107690