Kernel density estimation for circular data

Isaías Hazarmabeth Salgado Ugarte (1,2)
Marco Aurelio Pérez Hernández (2)

1) Laboratorio de Biometría y Biología Pesquera, Facultad de Estudios Superiores Zaragoza, UNAM
2) Departamento de Biología, Universidad Autónoma Metropolitana Iztapalapa

Trabajo realizado con el el apoyo del programa UNAM-DGAPA-PAPIME EN206213

Circular Data

Data points distributed on a circle occur in many applications of

Biology

Medicine

Geology

Geography

Meteorology

Physics

Circular Data

- A circular scale is a special type of the interval scale where not only is there no true zero but any designation of high or low values is arbitrary.

- Data from circular distributions may not be analyzed using the common (linear) statistical procedures.

- Statistical methods for describing and analizing circular data are relatively new and are still undergoing development (Batschelet, 1981; Fisher, 1989; 1993; Zar, 1999; Cox, 2005; Taylor 2008; Oliveira, *et al.* 2012)

Distribution of circular data I

Is one of the characteristics that need to be investigated in order to properly understand any batch of quantitative variables

Histograms are the traditional methods, but these nonparametric estimators have four drawbacks:

- Origin dependence
- Number and width of intervals
- Discontinuity
- Fixed interval width

Distribution of circular data II

The circular histogram equivalent diagrams share essentially the same problems of the linear versions

Additionally these estimators lose the proportionality between sector area and frequency values

Kernel density estimators for circular data I

According to Fisher (1989; 1993):

$$\widehat{f}(\theta) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\theta - \theta_i}{h}\right)$$

Based on Silverman (1986) Fisher (1989) gives an algorithm to calculate a quartic (biweight) kernel function and Cox (2001, 2004) uses this proposal in his circular Stata packages

Kernel density estimators for circular data II

For circular data it is appropriate the use of the von Mises function which is the "circular Gaussian" (Taylor, 2008)

$$\hat{f}(\theta;\nu) = \frac{1}{n(2\pi)I_0(\nu)} \sum_{i=1}^{n} exp\{\nu\cos(\theta - \theta_i)\}$$

Where $I_r(\nu)$ is the modified Bessel function of order r and the concentration parameter ν is the inverse of the smoothing parameter h. Large values of ν lead to highly variable estimations whereas small values provide oversmoothed circular densities (Oliveira, *et al*. 2012)

Circular KDE bandwidth choice I

Fisher (1989; 1993) "rule of thumb" for optimal window width using a biweight (quartic) kernel:

$$h_O = 7^{\frac{1}{2}} \left(\frac{1}{\kappa^{1/2}} \right) n^{-1/5}$$

where κ is the von Mises concentration parameter.

Circular KDE bandwidth choice II

When using a von Mises kernel the rule of thumb bandwidth adapted from Silverman (1986) minimizing the AMISE is (Taylor, 2008; Oliveira et al. 2012):

$$\hat{v}_{RT} = \left[\frac{3n\hat{\kappa}^2 I_2(2\hat{\kappa})}{4\pi^{\frac{1}{2}} I_0(\hat{\kappa})^2} \right]^{2/5}$$

Circular KDE bandwidth choice III

Besides the above described rules, in this contribution we considered as a preliminar reference two optimal and one oversmoothed rules adapted from the linear expresions by using the Batschelet's angular deviation (Batschelet, 1981):

$$\hat{\sigma}_{cir} = (\frac{180}{\pi})\sqrt{2 * (1 - r)}$$

where *r* is the length of the mean resultant vector, a measure of the data concentration named "vector strength" by Cox (1997; 2001; 2004).

Circular KDE bandwidth choice IV

The modified expresions are then as follows:

Optimal rules:

Silverman's Normal bandwidth reference rule:

$$h_o = 0.9 \min\left(\hat{\sigma}_{cir}, \frac{IQR}{1.349}\right) n^{-1/5}$$

Haerdle's Better rule of thumb:

$$h_o = 1.06 \min\left(\hat{\sigma}_{cir}, \frac{IQR}{1.349}\right) n^{-1/5}$$

Oversmoothed rule:

Scott's (Gaussian) kernel oversmoothed bandwidth:

$$h_{os} = 1.144 \hat{\sigma}_{cir} n^{-1/5}$$

Stata programs I

Syntax: (circbw.ado)

circbw v*arname* [**if** *exp*] [**in** *range*] [, **kc**(kernelcode)]

Description

circbw calculates several data-based bandwidth rules for circular variables (with azimuthal scale from 0 to 360 degrees) density estimation and reports the results in a table. It gives the rule of thumb for von Mises kernel, the Fisher rule for quartic kernel and adapted (using circular deviation) rules (oversmoothed and two optimal) for linear (Euclidean) data (as a reference). It is possible to choose the kernel function being Quartic (Biweight) the default (kernel code = 4). With duplicated orientation data it is necessary to employ only one half of the data.

Options

kc(kernelcode) set kernel (weight) function according to the following numerical codes; default is 4, Quartic(Biweight): 1 = Uniform; 2 = Triangle; 3 = Epanechnikov; 4 = Quartic (Biweight); 5 = Triweight; 6 = Gaussian; 7 = Cosine.

Remarks

circbw uses the supporting utilities i0kappa, written by Cox (2004) and i1kappa and i2kappa written by the authors based in the expressions from Fisher (1993). They are required to do the calculations.

Stata programs II

Syntax: (circkden.ado)

circkden *varname* [if *exp*] [in *range*] [,**kc**(kernelcode) **h**(#) **np**oints(#) **numo**des **mo**des **nuamo**des **amo**des **nog**raph **circg**ph **r**val(#) **fr**(#) **gs**(#) **gen**(*pdfvar degvar*) *scatter_options*]

Description

circkden calculates kernel density estimation for circular variables with azimutal scale (0 to 360 degrees) by means of a discretized procedure (Cox, 1998) and draws the result.

It is possible to choose the kernel function, to specify the smoothing parameter (half-width), the number of estimation points (at least _N) and to display a linear (default) or a circular graph. Additionally it provides modality and anti-modality information.

Options

kc(kernelcode) set kernel (weight) function according to the following numerical codes; default is 4, Quartic (Biweight): 1 = Uniform, 2 = Triangle, 3 = Epanechnikov, 4 = Quartic (Biweight), 5 = Triweight, 6 = Gaussian, 7 = Cosine

h(#) is the smoothness parameter (half-width) in degrees. The default is 30.

npoints(#) specifies the number of equally spaced points in the range of the circular variable. At least must be equal to the number of observations (Default).

numodes displays the number of modes (maxima) in the density estimation.

modes lists the estimated values for each mode. The **numodes** option must be included first.

nuamodes displays the number of antimodes (minima) in the density estimation.

amodes lists the estimated values for each antimode. The **nuamodes** option must be included first.

circgph draws a circular graph

Stata programs III

Options with circgph

rval(#)l is a factor controlling the radius size of the circle used

frval(#)l is a factor applied to the density values in the cosine and sine transformation. It permits to stretch or compress the density values around the unit circle.

gsval(#)l is a factor controlling the size of the graph. Large values give small graphics while less than unity figures produce bigger circle graphs.

Default is 1 in all the cases. It is possible for the graphs to depart from circle by using other values. This can be corrected by using the right combination (see last two examples in the help file).

gen(denvar degvar) specifies the name of the new variables in which probability density estimates (denvar) and the equally spaced angles (degvar) are to be stored.

scatter_options are any of the options allowed with *twoway scatter;* see help for graph.

nograph suppresses the graph drawing.

Stata programs IV

Syntax: (cirkdevm.ado)

cirkdevm *varname* [if *exp*] [in *range*] [,**nu**(#) **np**oints(#) **numo**des **mo**des **nuamo**des **amo**des **nog**raph **circg**ph **r**val(#) **fr**(#) **gs**(#) **gen**(pdfvar degvar) *scatter_options*]

Description

cirkdevm calculates kernel density estimation for circular variables with azimutal scale (0 to 360 degrees) by means of a discretized procedure (Cox, 1998) and draws the result.

It uses the von Mises kernel function and it is possible to specify the smoothing parameter (nu), the number of estimation points (at least _N) and to employ a linear (default) or a circular graph. Additionally it provides modality (and anti-modality) information.

Options

nu(#) is the concentration parameter (nu) analog to h (smoothing parameter) in degrees but with inverse behavior (large values produce noisy results and viceversa). The default is 30.

npoints(#),**numodes**, **modes**, **nuamodes**, **amodes**, **circgph** and the options with **circgph: rval**, **frval**, **gsval** are the same as those of circkden, the same as **gen**(*denvar degvar*), **scatter_options** and **nograph**.

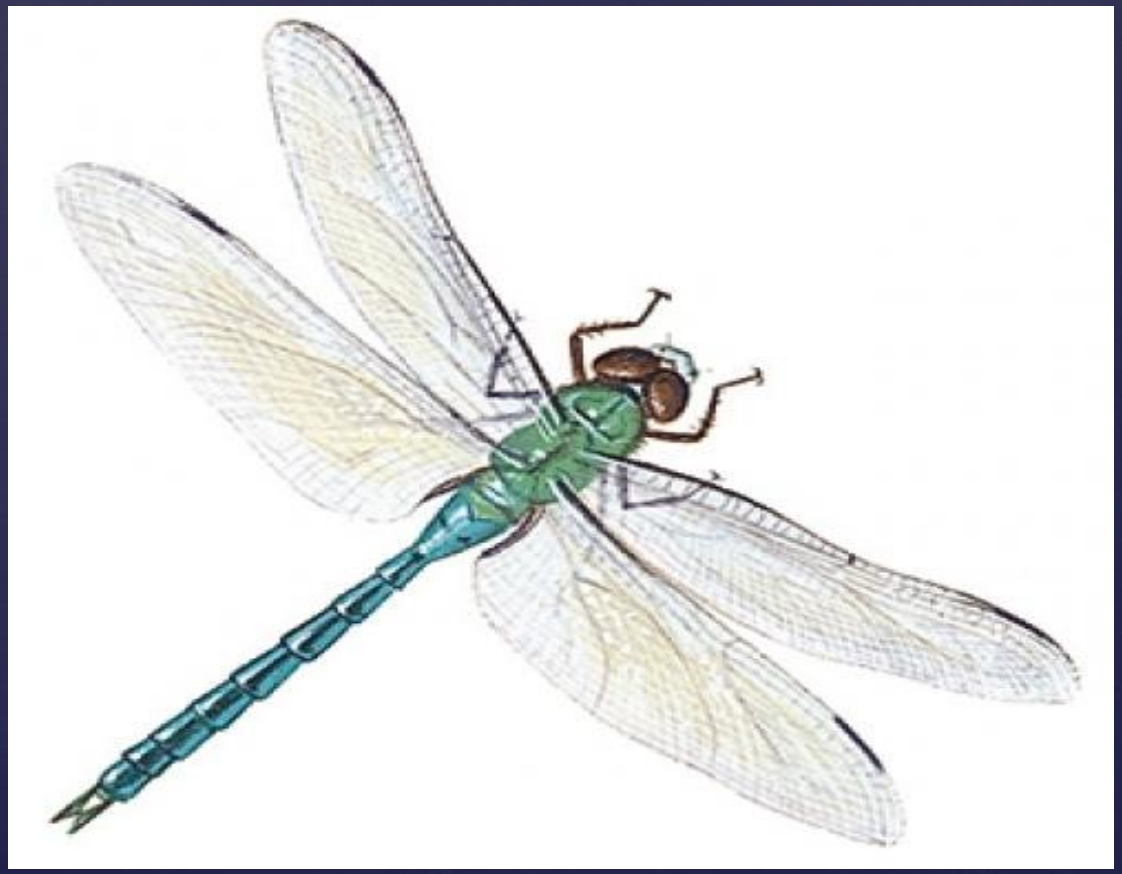
Examples: (Datasets) I

To observe the performance of the presented programs we used the following published datasets:

- Cross beds azimuths: a classical data set containing the angles of the strata measured at the Kampthi formation (Sengupta and Rao, 1966) and included in Mardia and Jupp (2000).

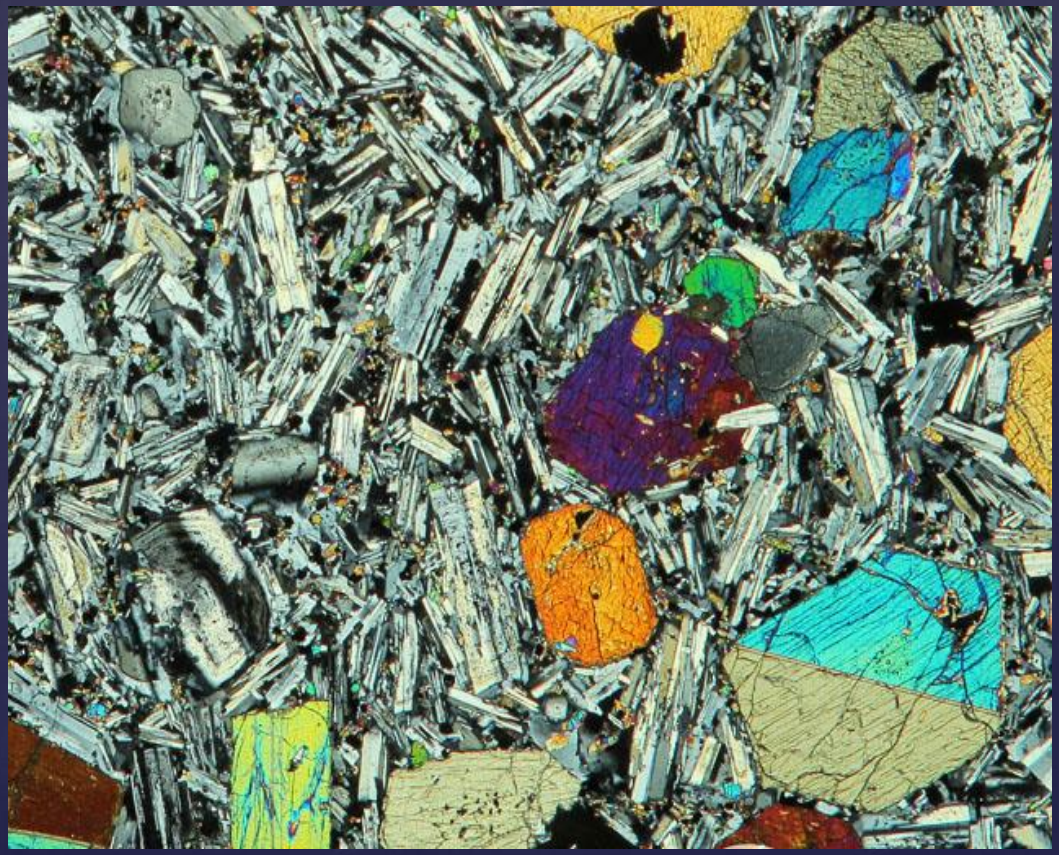
Examples: (Datasets) II

- Dragonflies orientation: Hisada (1972) presented data (n = 214) on the orientation of dragonflies respect to sun azimuth. These data were presented too by Batschelet (1981) and used by Oliveira et al. (2012).



Examples: (Datasets) III

- Long-axis orientations of feldspar laths: a data of 133 measurements of feldspar laths in basalt reported by Smith (1988) and presented by Fisher (1993).



Examples: (Datasets) IV

- Measurements of the directions taken by 76 turtles after being removed from their home territory carried out by Gould (1957) and data used later by Stephens(1969), Fisher (1989), Mardia and Jupp (2000) and Rao and SenGupta (2001).

Cross beds azimuths I

Table 1. Some practical bandwidth rules for circular data density estimation (cross bed azimuths)

von Mises rule of thumb bandwidth =	14.7246
Quartic kernel (4)	
Fisher's kappa (1.7857) bandwidth =	31.7743
Using Batschelet's angular deviation (47.1979)	
Silverman's optimal bandwidth =	31.2089
Haerdle's 'better' optimal bandwidth =	36.7571
Scott's oversmoothed bandwidth =	39.6699

Cross beds azimuths II

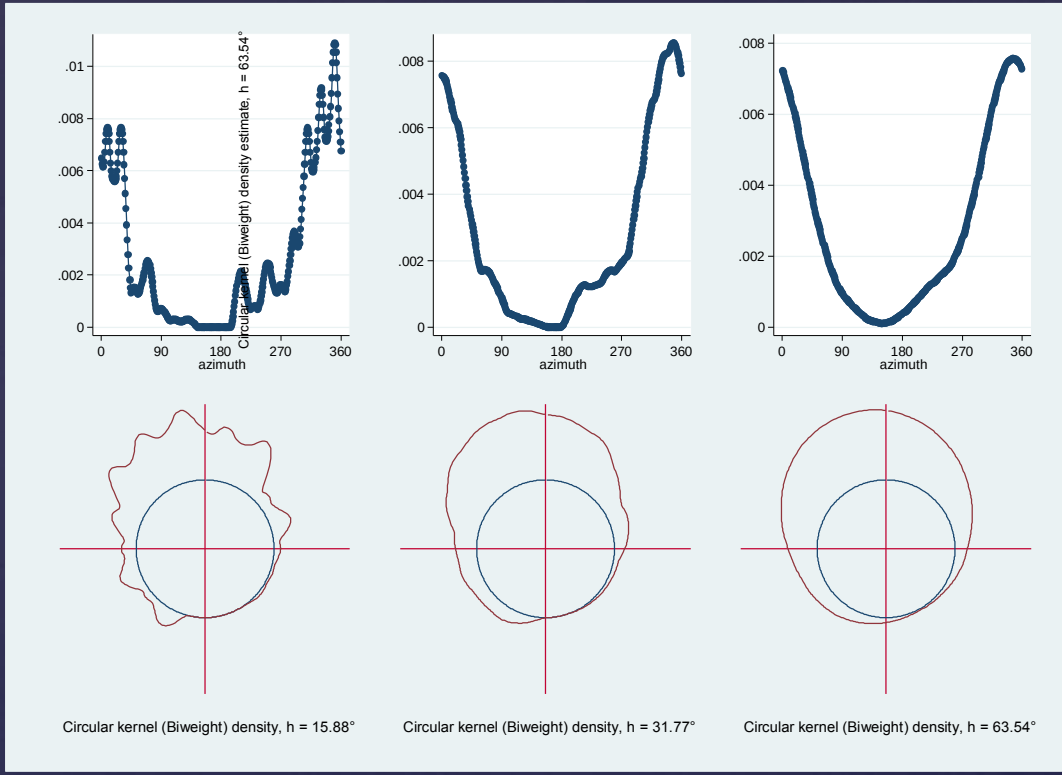


Figure 1. Circular kernel density estimators (unwrapped and circular versions) for the cross beds azimuths data with 0.5, 1 and 2 times the optimal bandwidth (31)

Cross beds azimuths II

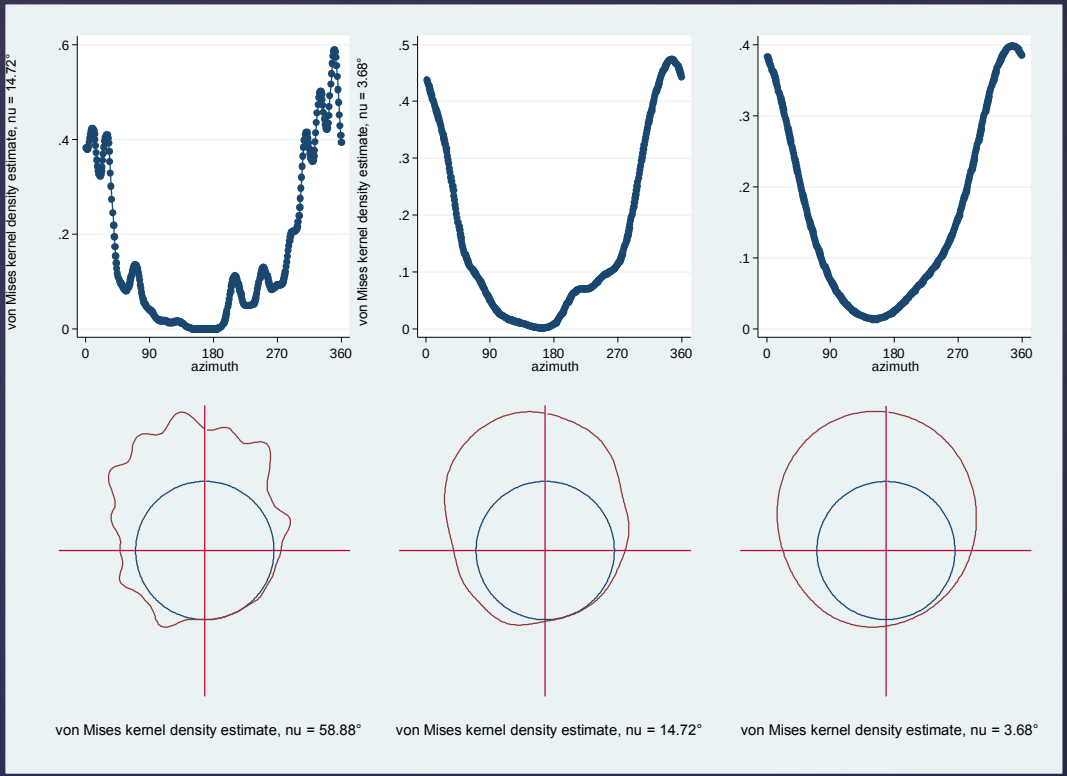


Figure 2. von Mises kernel circular density estimation for the azimuths of crossbedding data from the Kampthi formation with 2^2, 1 and 1/(2^2) times the "rule of thumb" bandwidth (14.72).

The optimal values present approximately the same smoothness and in this case seem to be the appropriate ones. To obtain the same roughness in the estimations may be necessary to multiply the von Mises plug-in rule by bigger numbers (it is suggested the use of powers of 2).

Dragonfly orientations I

Table 2. Some practical bandwidth rules for circular data density estimation (dragonfly orientation measurements)

von Mises rule of thumb bandwidth =	34.4314
Quartic kernel (4)	
Fisher's kappa (6.3741) bandwidth =	23.4521
Using Batschelet's angular deviation (23.2154)	
Silverman's optimal bandwidth =	20.5056
Haerdle's 'better' optimal bandwidth =	24.1510
Scott's oversmoothed bandwidth =	27.2096

To calculate the bands with circbw only one half of the data set is considered (those less than 180). Both, the quartic and the von Mises kernel density estimation show clearly the bimodal characteristic of this data set.

It is interesting to note (Table 2.) that the modified linear rules gives values similar to those obtained by means of more elaborated methods (Oliveira et al. 2012).

Dragonfly orientations II

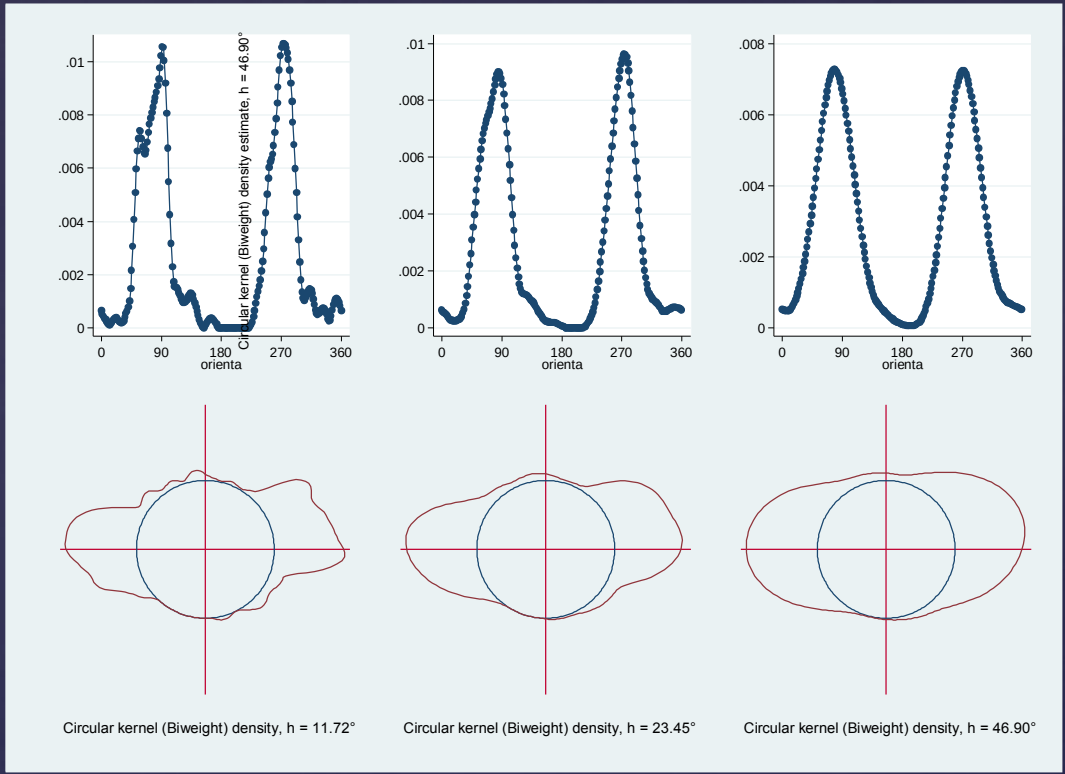


Figure 3. Quartic kernel circular density estimation for the orientation data of dragonflies relative to the sun's azimuth.

Dragonfly orientations III

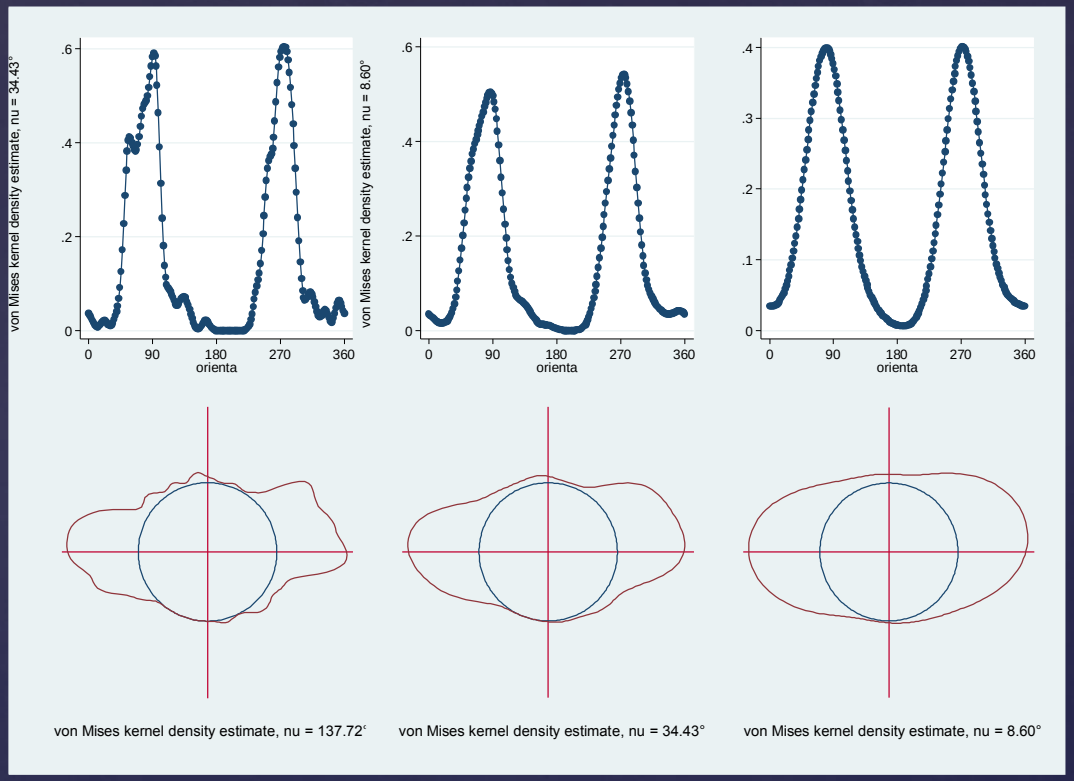


Figure 4. von Mises kernel circular density estimation for the orientation data of dragonflies relative to the sun's azimuth.

Long-axis orientations of feldspar laths I

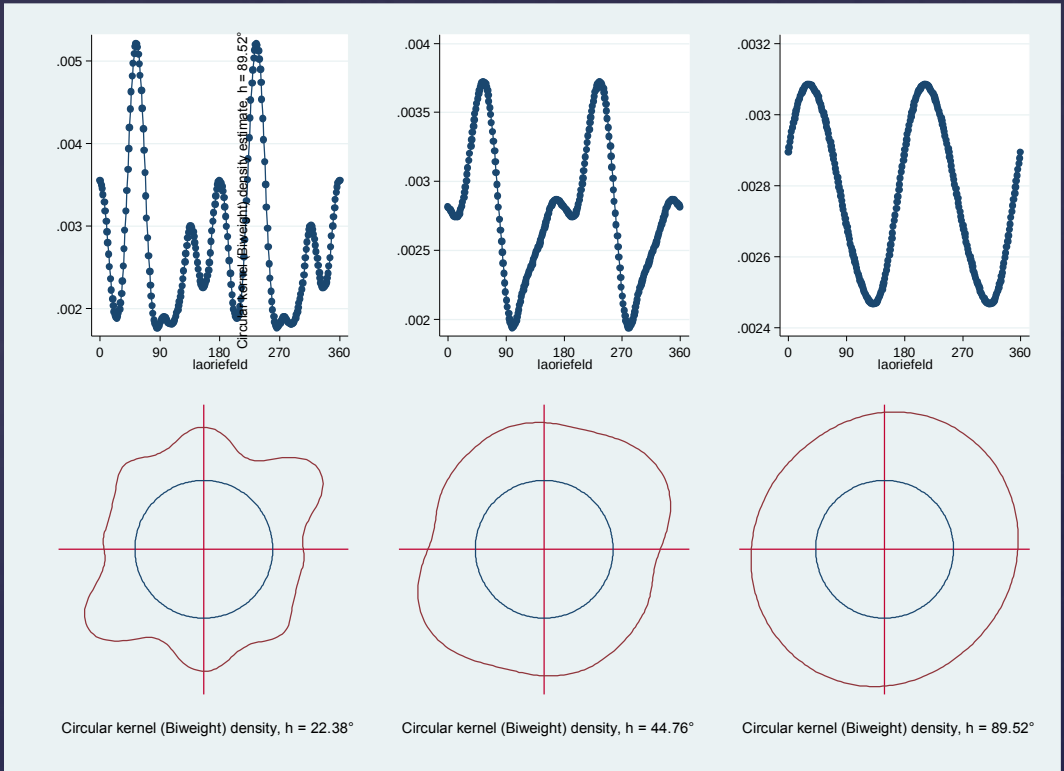


Figure 5. Quartic kernel circular density estimation for the long-axis orientation of feldspar laths.

To analyze this data set a group variable ("half") has been added in order to select only one half of the duplicated measured orientations.

Long-axis orientations of feldspar laths II

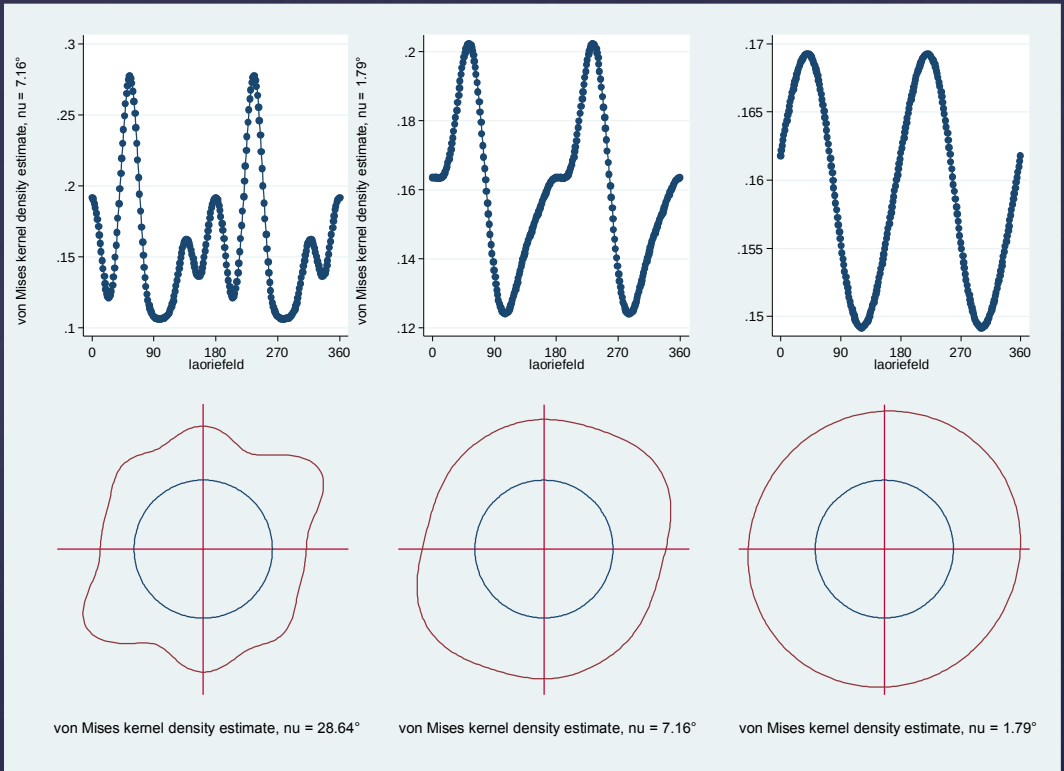


Figure 6. von Mises circular density estimation for the long-axis orientation of feldspar laths.

Direction of turtles I

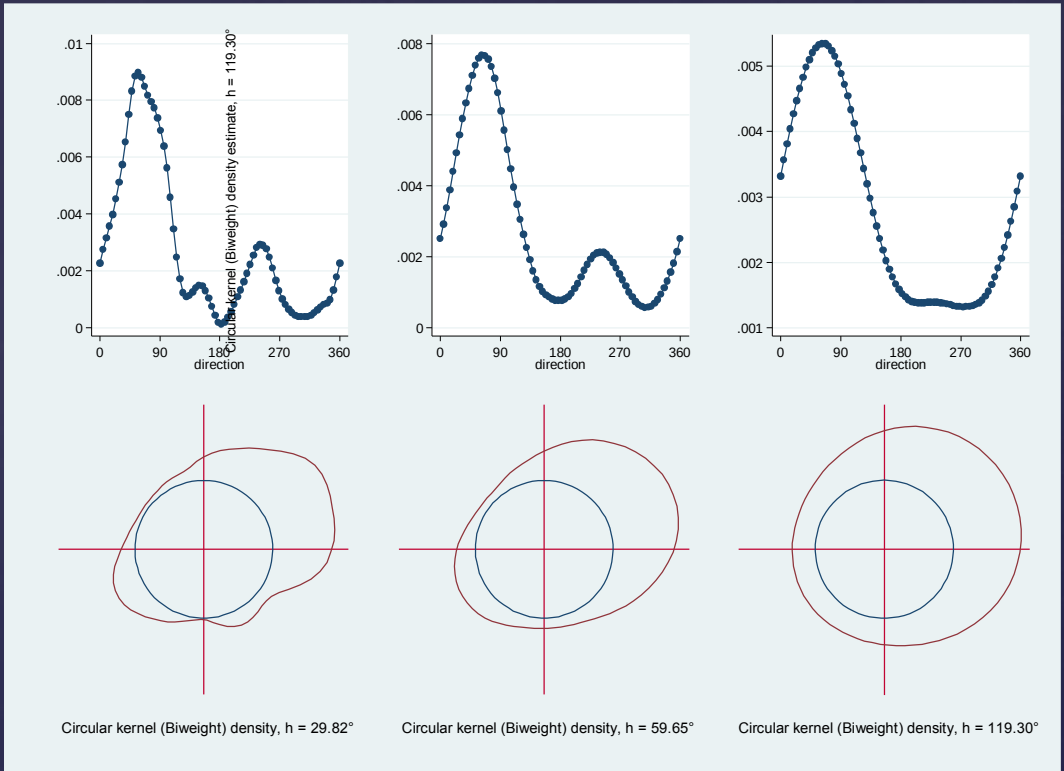


Figure 7. Quartic kernel circular density estimation for the direction taken by turtles after being removed from their home range.

Direction of turtles II

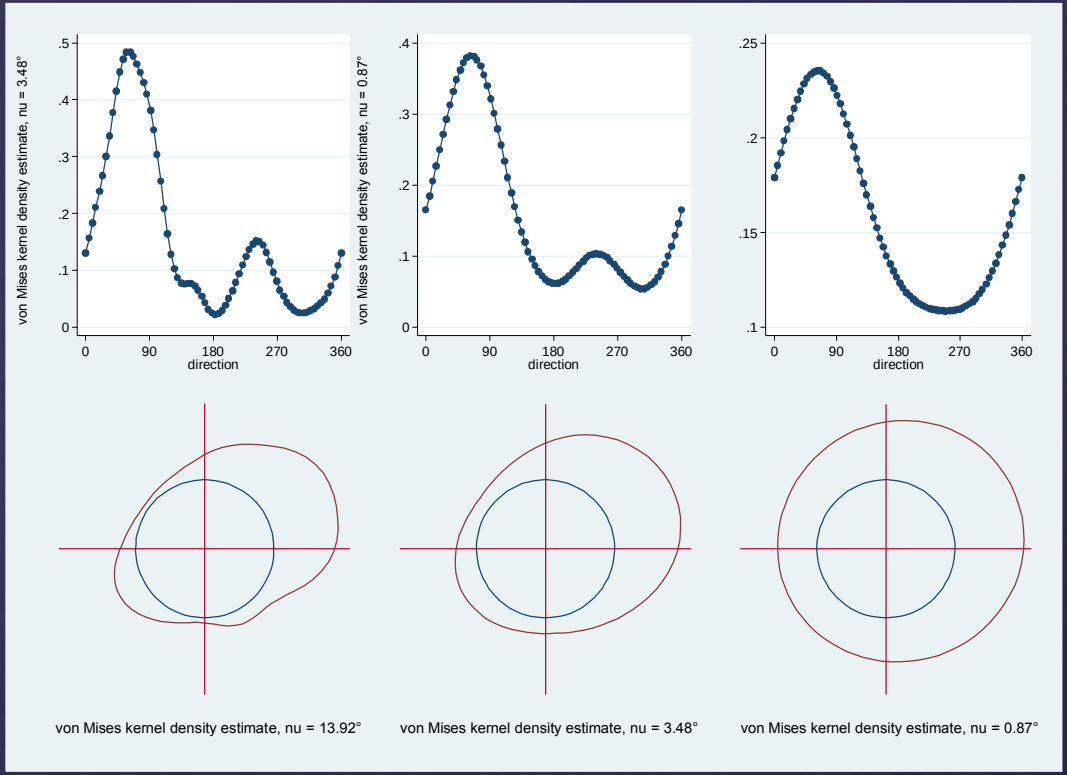


Figure 8. von Mises circular density estimation for the direction taken by turtles after being removed from their home range.

Example of circgph.ado

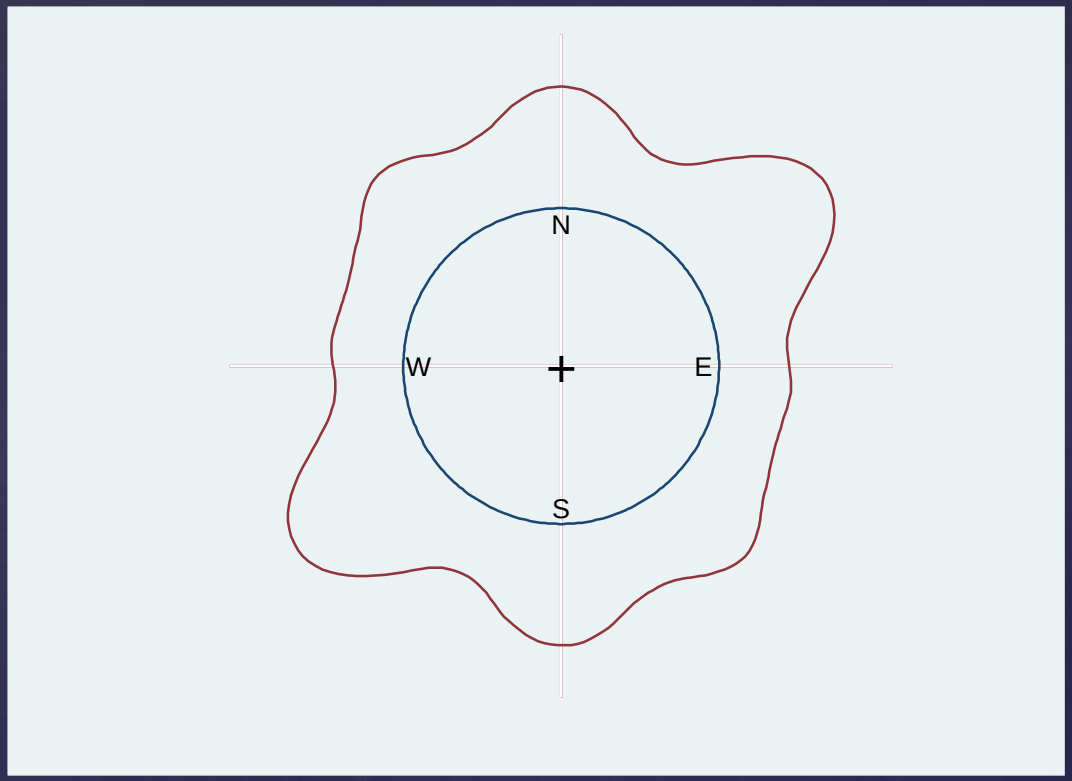


Figure 9. A customized circular density estimation graph with the cardinal points indicated and axis replaced by a central cross.

```
. use fisherfeldes
. circkden lao, h(22) gen(den deg)

. circgph den deg,  xline(0, lc(white)) yline(0, lc(white)) text(0 0
"+", size(huge)) text(.9 0 "N") text(0 .9 "E") text(-.9 0 "S") text(0
-.9 "W")
```

Example with reduced *n*

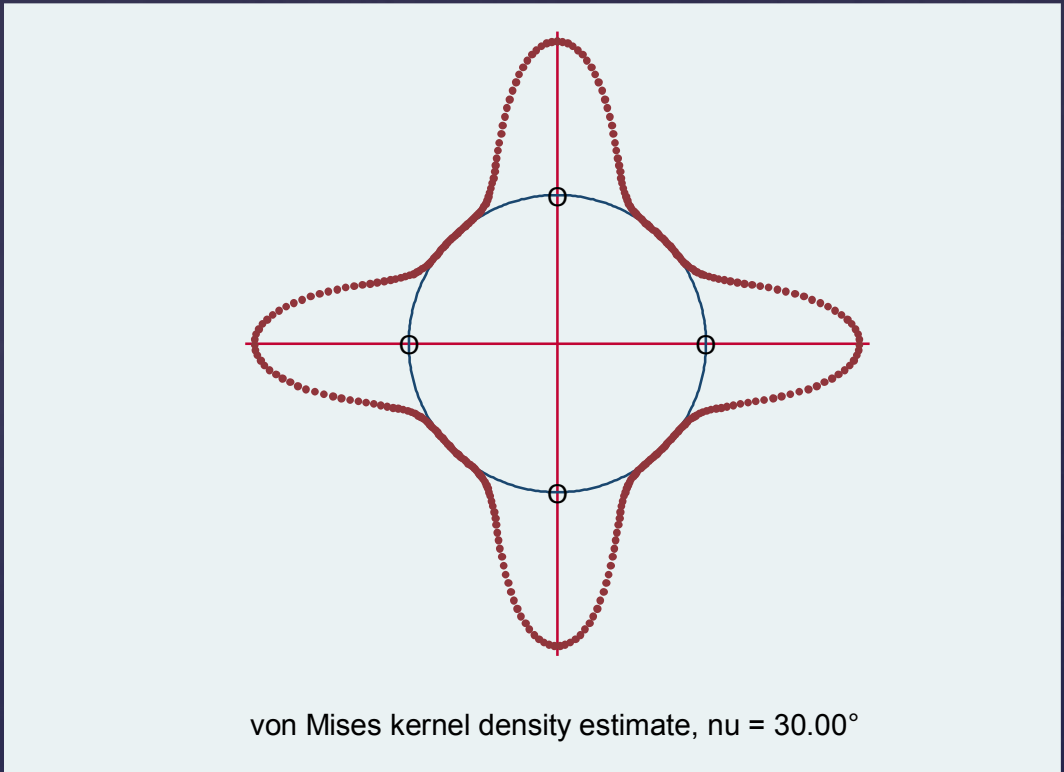


Figure 10. Circular density estimation with von Mises kernel for 0, 90, 180 and 270 angles.

```
. cirkdevm angle,  np(360) circgph ms(o) msize(small) text(1 0
"O") text(0 1 "O") text(-1 0 "O") text(0 -1 "O")
```

Conclusions I

- With distributions having opposite modes the plug-in rule performs poorly to represent the distribution. Fisher and adapted linear rules produce results with distribution graphs presenting clearly the bimodal distribution. The inconvenience of the Taylor plug in rule is surpassed if only one half of the data are used to calculate the bandwidths.
- Being able to see both the linear and circular graphs it is possible to state that is easier to recognize multimodality with the linear version than with the circular. Nevertheless, we consider that the circular representation of the data gives a more clear image of the angular information.
- It is interesting to note that in general, the performance of the linear adapted rules (by using a circular variability measure instead of the standard deviation) is almost the same as the Fisher's rule and is better when analyzing circular data with opposed modes (axial data).

Conclusions II

- Due to the inverse quadratic relationship between *nu* and *h* (Taylor, 2008), in order to obtain similar smoothing between von Mises and the other kernels results we suggest to employ powers of two multipliying *nu* (2^2 and $1/2^2$ as first instances) instead of the simple double and half multiples.
- The use of kernel density estimators (in linear unwrapped and circular versions in combination with circular raw data plots to explore circular data) is a very important procedures set that permits to easily recognize distribution characteristics such as modality and bias.
- Our programs permit to estimate densities even with a few data points, so it is possible to represent the kernel shape for every individual observation.

Acknowledgements

- Salgado-Ugarte I.H. received support from the Departmento de Biología, Universidad Autónoma Metropolitana during his stay as a titular professor (Dr. Ramón Riba y Nava Esparza professorship) and from DGAPA-FES Zaragoza PAPIME PE206213 (Universidad Nacional Autónoma de México).

- M. Oliveira (Department of Statistics and Operations Research, University of Santiago de Compostela, Spain) kindly gave us advice and sent some of the data files presented in her works.

References:

- Batschelet, E. 1981. *Circular Statistics in Biology*, Academic Press. London, 371 p.
- Cox, N.J. 1997. Circular statistics in Stata. *Proceedings of the 3rd UK User Group Meeting*, London
- Cox, N.J. 2001. Analysing circular data in Stata. Proceedings of the 3rd North American User Group Meeting, Boston, USA: 4 p.
- Cox, N.J. 2004. Circular statistics in Stata, revisited. Proceedings of the 10th UK Users Group Meeting, London, UK: 4 p.
- Davis, J.C. 2002. Statistics and Data Analysis in Geology. 3rd edition, John Wiley & Sons, New York: 656 p.
- Fisher, N.I. 1989. Smoothing a sample of circular data. *Journal of Structural Geology*, **11**(6): 775-778.
- Fisher, N.I. 1993. *Statistical Analysis of Circular Data*. Cambridge University Press. Great Britain, 277 p.
- Fox, J. 1990. Describing univariate distributions. In: *Modern Methods of Data Analysis*, eds. J. Foxy J.S. Long, 58-125. Newbury Park, CA: Sage publications.
- Gould, E. (1957) Orientation in box turtles, Terrapene c. Carolina (Linneaus). *The Biological Bulletin*, 112: 336-348.
- Hisada, M., 1972. Azimuth orientation of the dragonfly (Sympetrum) In: S.R. Galler, K. Schmidt-Koenig, G.J. Jacobs y R.E. Belleville (eds) *Animal Orientation and Navigation*. National Aeronautic and Space Administration,Washington, EUA: 511-522.
- Mardia, K.V. and P.E. Jupp, 2000. *Directional Statistics*. John Wiley & Sons, Ltd., New York, U.S.A. 430 p.
- Mosqueda-Romo, N.A. & I.H. Salgado-Ugarte, 2011. Updated programs of improved Stata ado-files for nonparametric smoothing. *Proceedings of the 2011 Mexican Stata Users Group meeting, May 12, 2011*, Institute for Economic Research, National Autonomous University of Mexico, Mexico.
- Oliveira, M., R.M. Crujeiras & A. Rodríguez-Casal, 2012. A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics and Data Analysis*, **56**(2012): 3898-3908.
- Rao, J.S. & A. Sengupta, 2001. *Topics in Circular Statistics*. World Scientific Publishing, Singapore: 5.
- Salgado-Ugarte, I.H. 2009. Some improved Stata ado-files for nonparametric smoothing procedures. *Proceedings of the 2009 Mexican Stata Users Group meeting, April 23, 2009*, Universidad Iberoamericana, Mexico
- Salgado-Ugarte, I.H. & Pérez-Hernández, M.A. 2003. Exploring variable bandwidth kernel density estimators. *Stata Journal*.
- Salgado-Ugarte, I.H., M. Shimizu, y 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 & T. Taniuchi, 1995. Practical rules for bandwidth. *Stata Technical Bulletin.* 27: 5-19.
- Scott, D.W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Nueva York.
- Sengupta, S. & J.S. Rao, 1966. Statistical analysis of crossbedding azimuths from the Kamthi formation around Bheemaram, Pranhita-Godavari valley. *Sankhya* Ser. B, 28, 165-174.
- Silverman, B.W. 1986. *Density estimation for statistics and data analysis*. Chapman & Hall. London.
- Smith, N.M. 1988. Reconstruction of the Tertiary drainage systems of the Inverell region. Unpublished B.Sc.(Hons.) thesis, Department of Geography, University of Sidney, Australia.
- Stephens, M.A. 1969. *Techniques for directional data*. Technical Report # 150, Dept. of Statistics, Stanford University, Stanford, CA: 23, 102, 141.
- Taylor, C.C. 2008. Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis*, **52**(7): 3493-3500.
- Zar, J.H. 1999. *Biostatistical Analysis*. 4th ed. Prentice Hall, Upper Saddle River, U.S.A. 592-663.