Averaged shifted histograms (ASH) or weighted averaging of rounded points (WARP): Efficient method to calculate kernel density estimators for circular data

Isaías Hazarmabeth Salgado Ugarte[1)],
Verónica Mitsui Saito Quezada[1)]
y Marco Aurelio Pérez Hernández [2)]
1) Laboratorio de Biometría y Biología Pesquera
F.E.S. Zaragoza U.N.A.M.
2) Departamento de Biología, U.A.M. Iztapalapa

Histogram Drawbacks

- **Dependency on the origin of the bins**
- **Dependency on the width and number of bins**
- **Discontinuity**
- **Fixed bindwidth**

Kernel density estimators

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

Advantages of the kernel density estimators (KDE's)

- No dependency on the origin (estimation centered at each data point).
- No discontinuity (estimation centered at each data point and use of a gradually changing weight function instead of the rectangular shape).
- Variable bandwidth implementation posible.

KDE's drawback

Large number of calculations

Approaches to overcome this problem

- Discretized estimation
- ASH-WARP method

ASH-WARP Procedure

- Binning the data
- Calculating the weights
- Weighting the bins

Circular Data I

Data points distributed around a circle occur in many applications from different disciplines as Biology, Medicine, Geology, Geography, Meteorology and Physics.

Observations of directions on a plane or in space and cyclic phenomena can be interpreted as circular (Batschelet, 1981).

The study of this information is the object of Circular Statistics.

Circular Data II

Circular data are an special type of interval scale, which not only do not have a true zero, but any designation of high or low values is arbitrary.

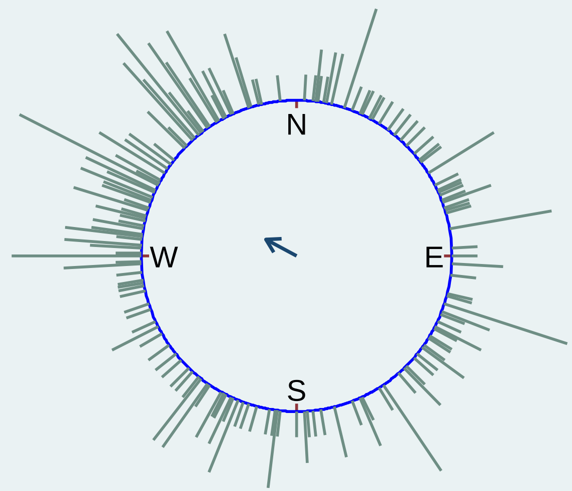
The typical example is the division of a circle in 360 equal parts (degrees): Azimutal scale.

There is no physical reason to assign 0 (or 360) to the position marked as “North” and a 270 degrees direction can not be considered larger than other of 90 degrees.

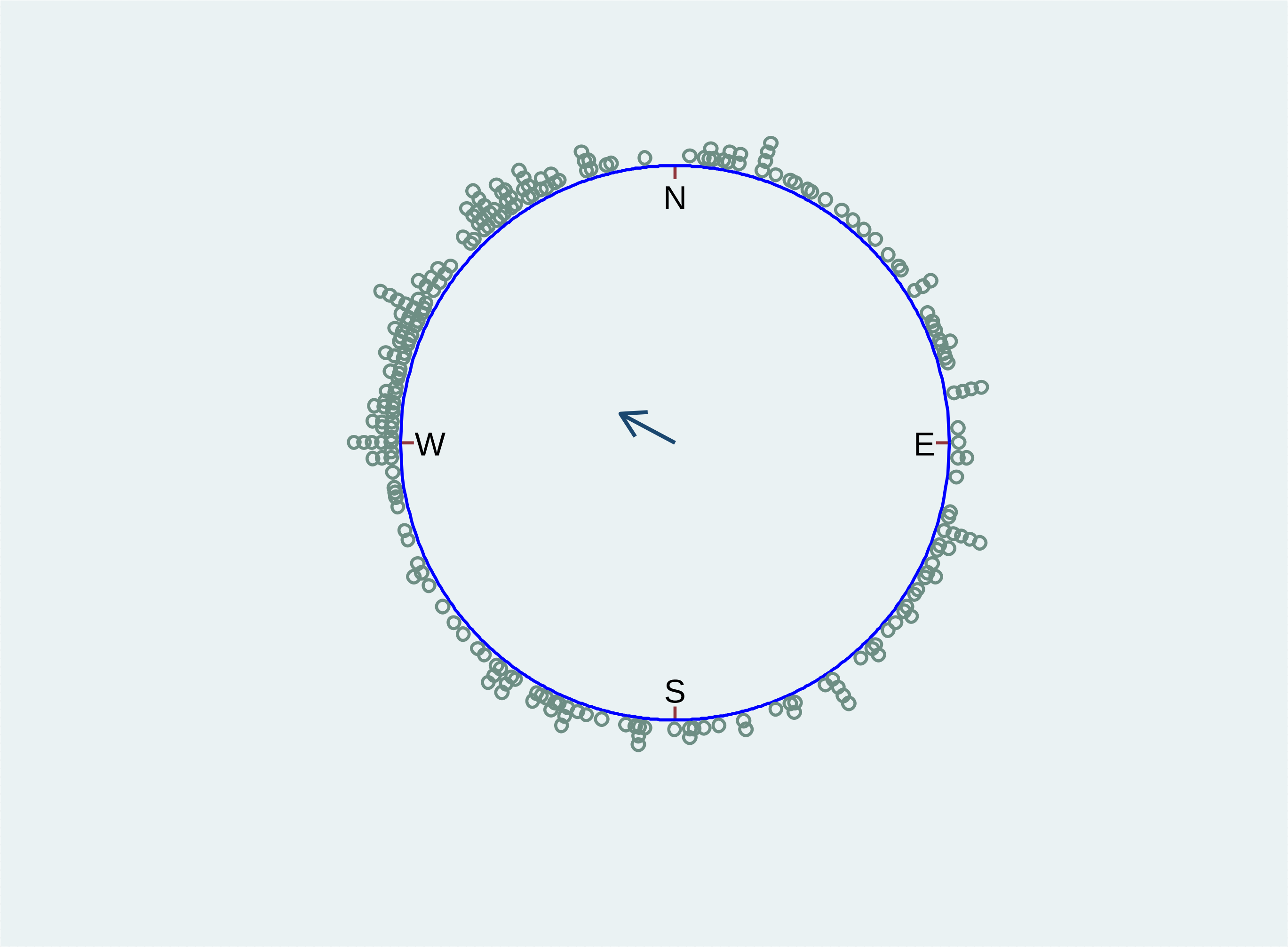
Wind direction weighted by wind force, Meteorological Station FES Zaragoza

- meteorofesz.dta
- N total = 36,715 observations, with 18 variables
  - n = 2,219 May, 2012

Circular raw data plot ("circrplot.ado", Cox, 2004) of 240 hourly measures of wind direction (18 to 27 March, 2013). Meteorological Station, FES Zaragoza, UNAM



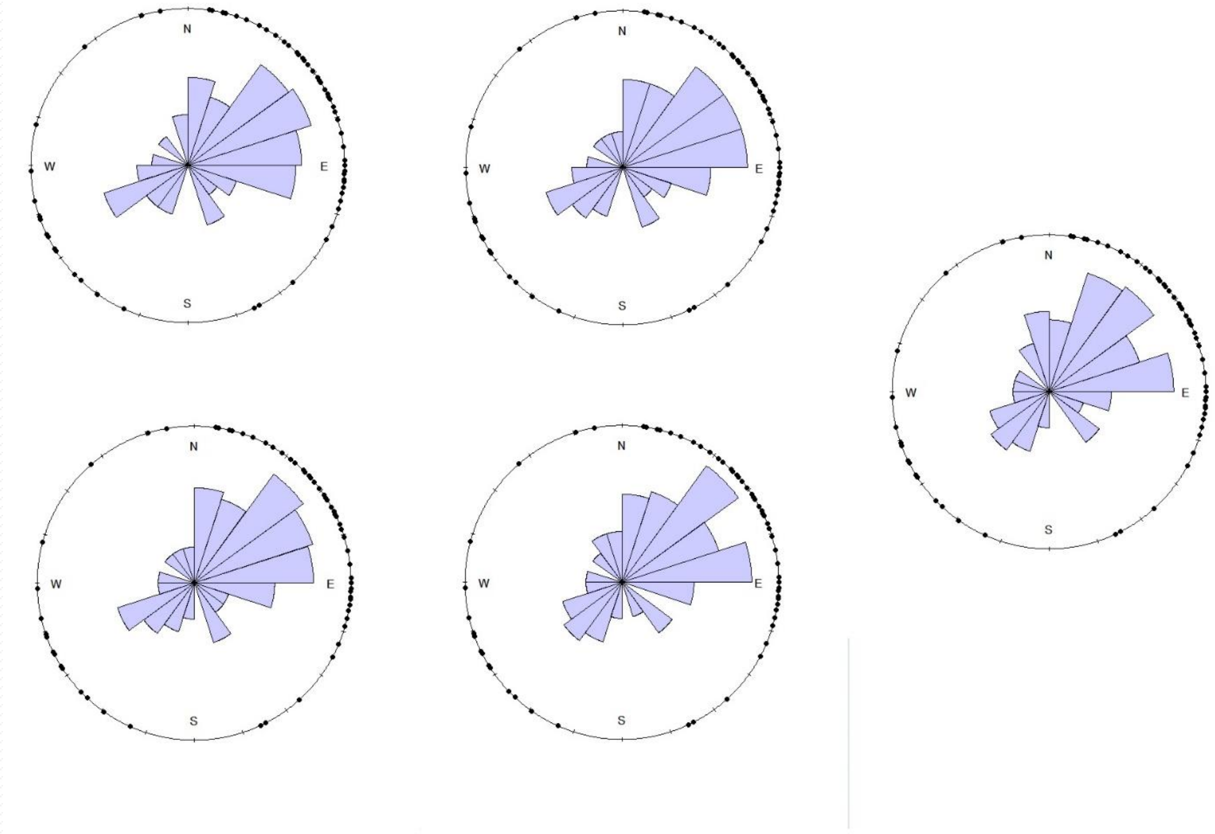
Circular dot plot ("circdplot.ado" Cox, 2004)



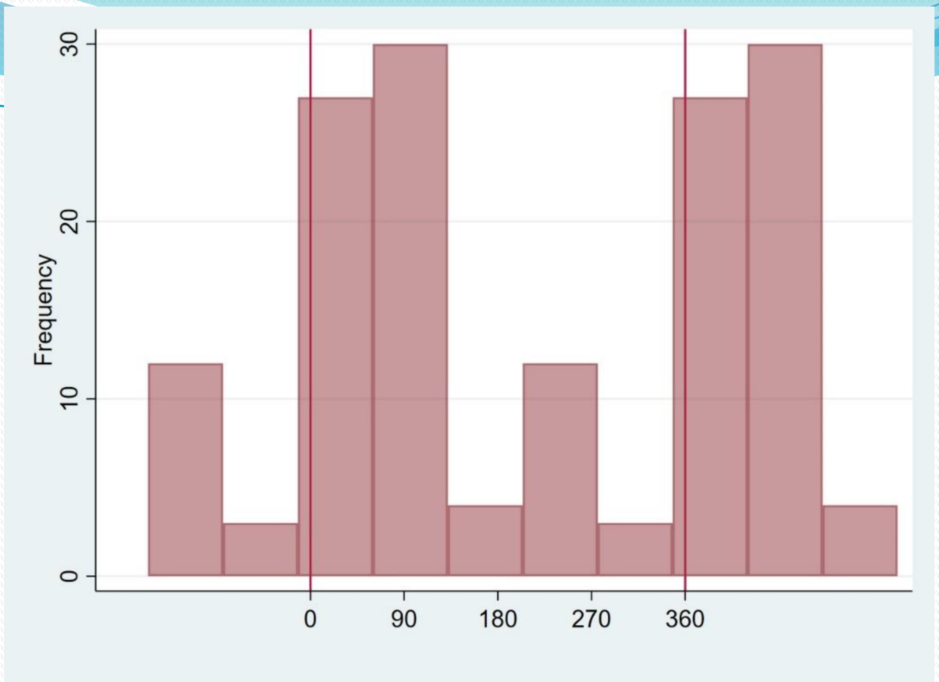
Distribution of circular data

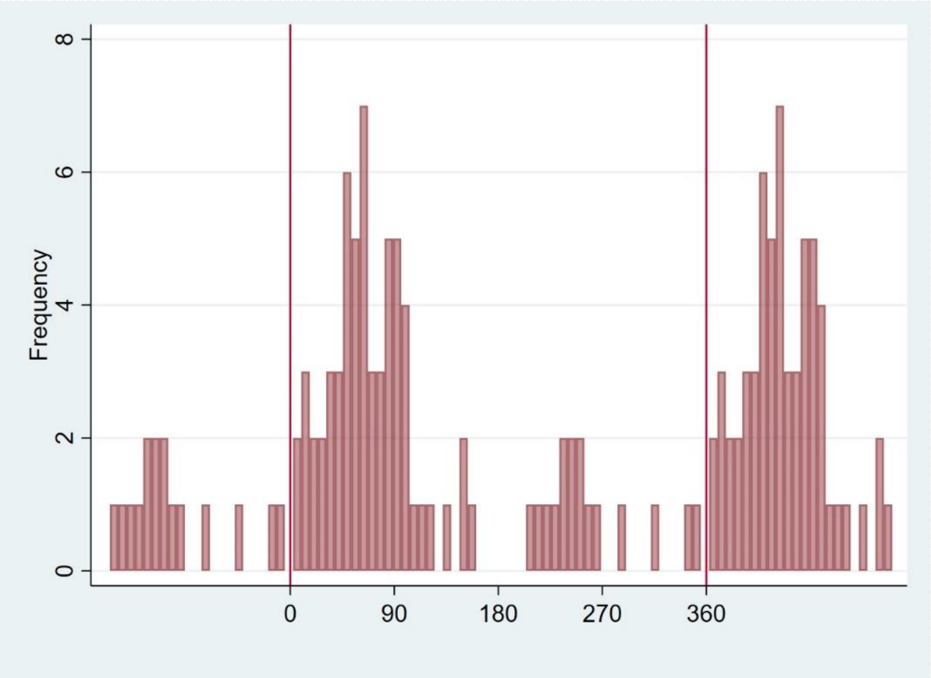
- As with the linear scales, the distribution of circular data is a characteristic that needs to be understood in order to properly interpret the data message.
- To analyze circular distributions, it is possible to use Kernel Density Estimators (Fisher, 1989; 1993) as an alternative to the Rose diagrams, that share the histogram drawbacks.

Five Rose diagrams with same intervals but different origin.



Circular histograms with five and 50 intervals; “circhistogram.ado” (Cox, 2004)





Kernel density estimator for circular data

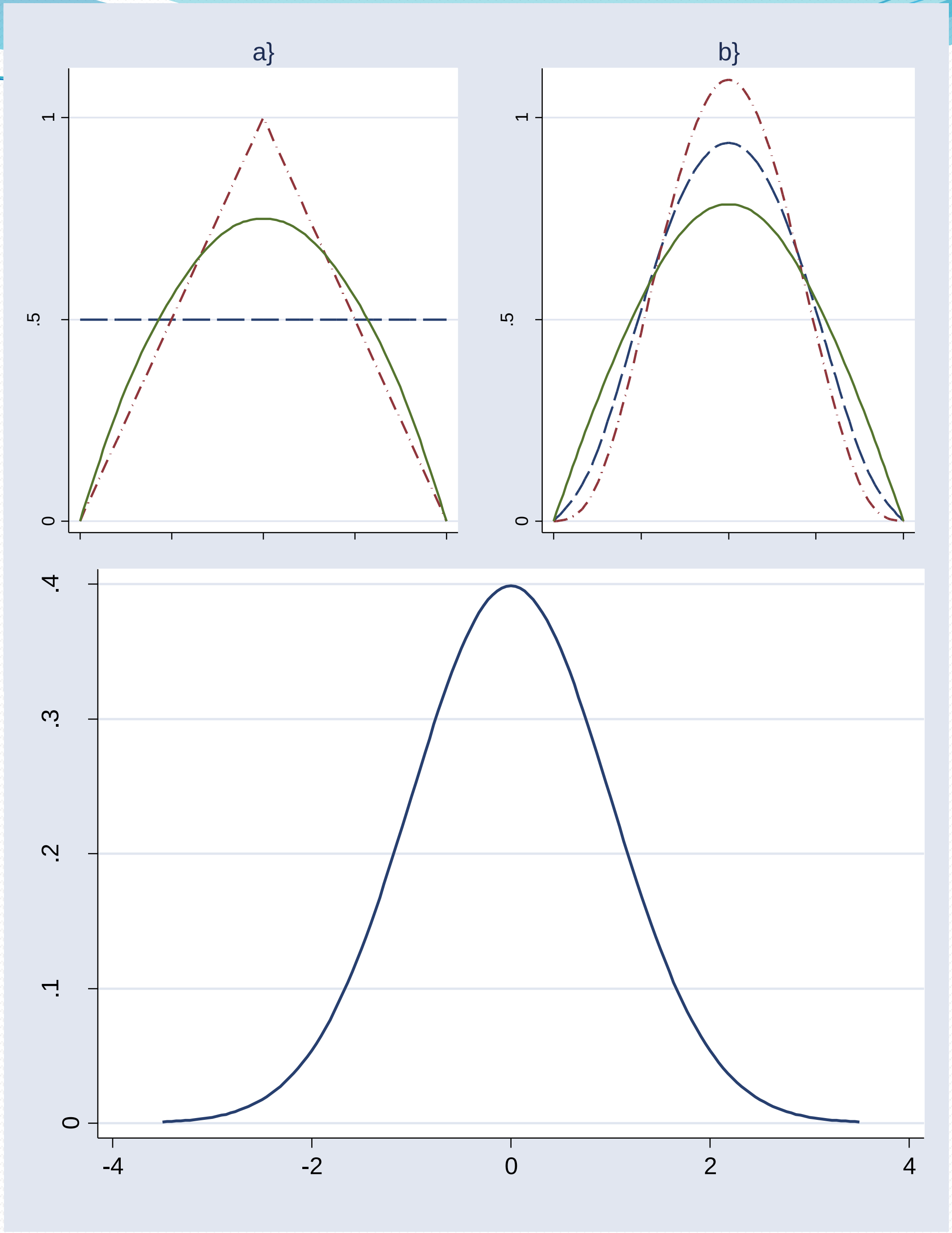
$$\widehat{f}(\theta) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\theta - \theta_i}{h}\right)$$

- h is the bandwidth or smoothness parameter
- K is the kernel (weighting) function, and
- θ is the angular (circular) variable.
- Based on Silverman (1986), Fisher (1989) gives an algorithm to calculate a quartic (biweight) kernel function. Cox (2001, 2004) uses this proposal in his circular Stata packages. It is straightforward to extend the algorithm to use other weighting functions such as the uniform, triangular, Epanechnikov, triweight, Gaussian or cosine

Some common kernel functions

Kernel	K(z)
Uniform	$\frac{1}{2} I(\|z\| \leq 1)$
Triangular (ASH)	$(1 - \|z\|) I(\|z\| \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - z^2) I(\|z\| \leq 1)$
Biweight (Quartic)	$(15/16)(1 - z^2)^2 \; I(\|z\| \leq 1)$
Triweight	$(35/32)(1 - z^2)^3 \; I(\|z\| \leq 1)$
Cosinus	$(\pi/4)\cos((\pi/2)z) I(\|z\| \leq 1)$
Gaussian	$(1/\sqrt{2\pi})\exp((-1/2)z^2$

Common Kernel functions: Uniform, Triangular, Epanechnikov, Biweight (Quartic), Triweight, Cosinus

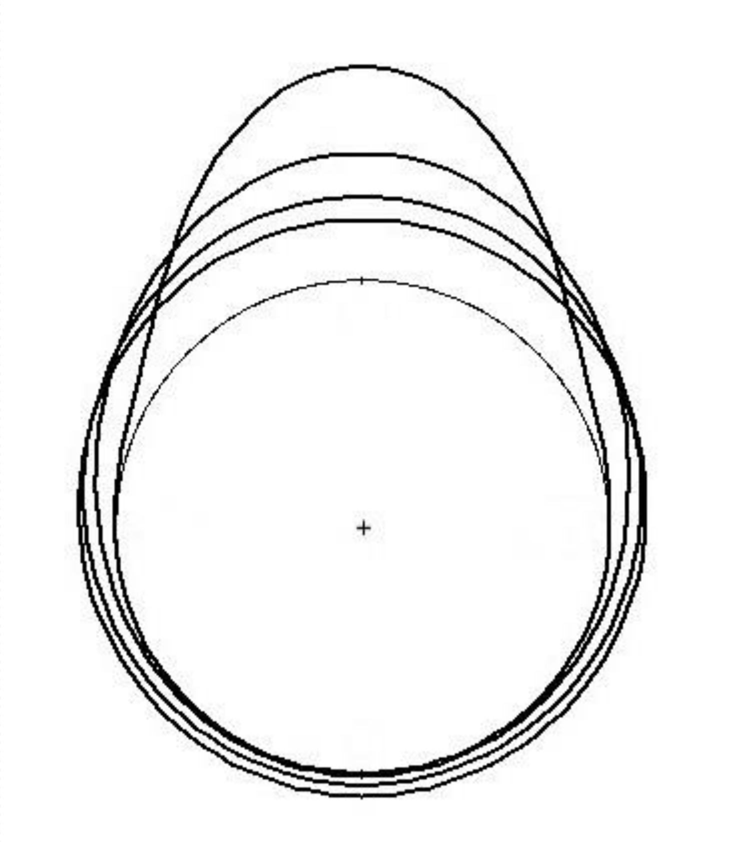


von Mises function (circular Gaussian)

- For circular data it is appropriate the use of the von Mises function which is the "circular Gaussian". According to Taylor (2008) the density estimation with this function is:

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

von Mises distributions for several κ values (5, 2, 1, 0.5)



Bandwidth choice ("circbw.ado")

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

Optimal bandwidth (*Quartic kernel*) Fisher, 1989; 1993

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

Rule of thumb adapted from Silverman (1986); Taylor (2008); Oliveira et al. (2012).

```
. use meteorofeszcor6

. sum

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 dateandtime |          0
         srd |     36,715    232.1703    304.2498          0     1364.7
         tmp |     36,715    19.79917    4.893683         .6       33.2
         hmd |     36,715    40.64518    24.47429        9.5       98.7
         wnd |     36,715     203.991     107.501          0        358
-------------+---------------------------------------------------------
         wng |     36,715    13.76004     7.10322          0         53
         wns |     36,715    4.019774    2.674511          0         18
        dias |     36,715    372.7253    181.2016         68        680

. sum wnd if dias>120 & dias < 152

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         wnd |      2,219     212.626    95.75095          0        357
```

Using “circbw.ado” (Salgado-Ugarte, et al. 2017)

```
. circbw wnd if dias>120 & dias < 152

____________________________________________________________

Some practical bandwidth rules for

circular data density estimation

============================================================

 von Mises rule of thumb bandwidth =                  5.2741

============================================================

 Quartic kernel (4)

============================================================

 Fisher's kappa (  0.6167) bandwidth =               41.3438

============================================================

 Using Batschelet's angular deviation ( 68.0530)

____________________________________________________________

Silverman's optimal bandwidth =                     34.4078

Haerdle's 'better' optimal bandwidth =              40.5247

Scott's oversmoothed bandwidth =                    43.7361

____________________________________________________________
```

Time required for calculation

- Discretized algorithm: "circkden.ado" (Salgado-Ugarte, et al. 2017)
  - 6'36" aprox.

```
. circkden wnd if dias>120 & dias < 152, h(20) numo mo
subtitle("Mayo, 2012")
```

- ASH-WARP procedure: "circwarp.ado" (Salgado-Ugarte, et al. 2018; 2021)
  - Less than 1"

```
. circwarp wnd if dias>120 & dias < 152, h(20) numo mo
subtitle("Mayo, 2012")
```

- Intel Xeon E5-1607 v4 @ 3.1GHz, 3100 Mhz,4 main processors, 4 logic processors; 8 GB RAM

Program circwarp.ado (I)

```
  circwarp --  Performs ASH-WARP kernel density estimation for circular data

Syntax

     circwarp varname [if] [in] [, Hwidth(#) Mval(number of averaged shifted histograms)
Kercode(#)
              {op gt}ype(#)  NUMOdes MOdes NUAMOdes AMOdes NOGraph rval(#) fr(#) gs(#)
              GEN(denvar degvar) PLOT(str asis)]

Description

    circwarp calculates kernel density estimators for circular variables with azimutal
scale (0 to 360 degrees) by means of the ASH-WARPing procedure (Scott, 1985, 1992; Haerdle,
1991; Salgado-Ugarte, et al. 1995) and draws the result.  It is possible to choose the
kernel function, to specify the smoothing parameter (half-width), the number of averaged
shifted histograms (10 suggested) and to employ a  linear (default) or a circular graph.

    Additionally it provides modality (and anti-modality) information if requested.
   It saves significative calculation time with big data sets.
```

Program circwarp.ado (II)

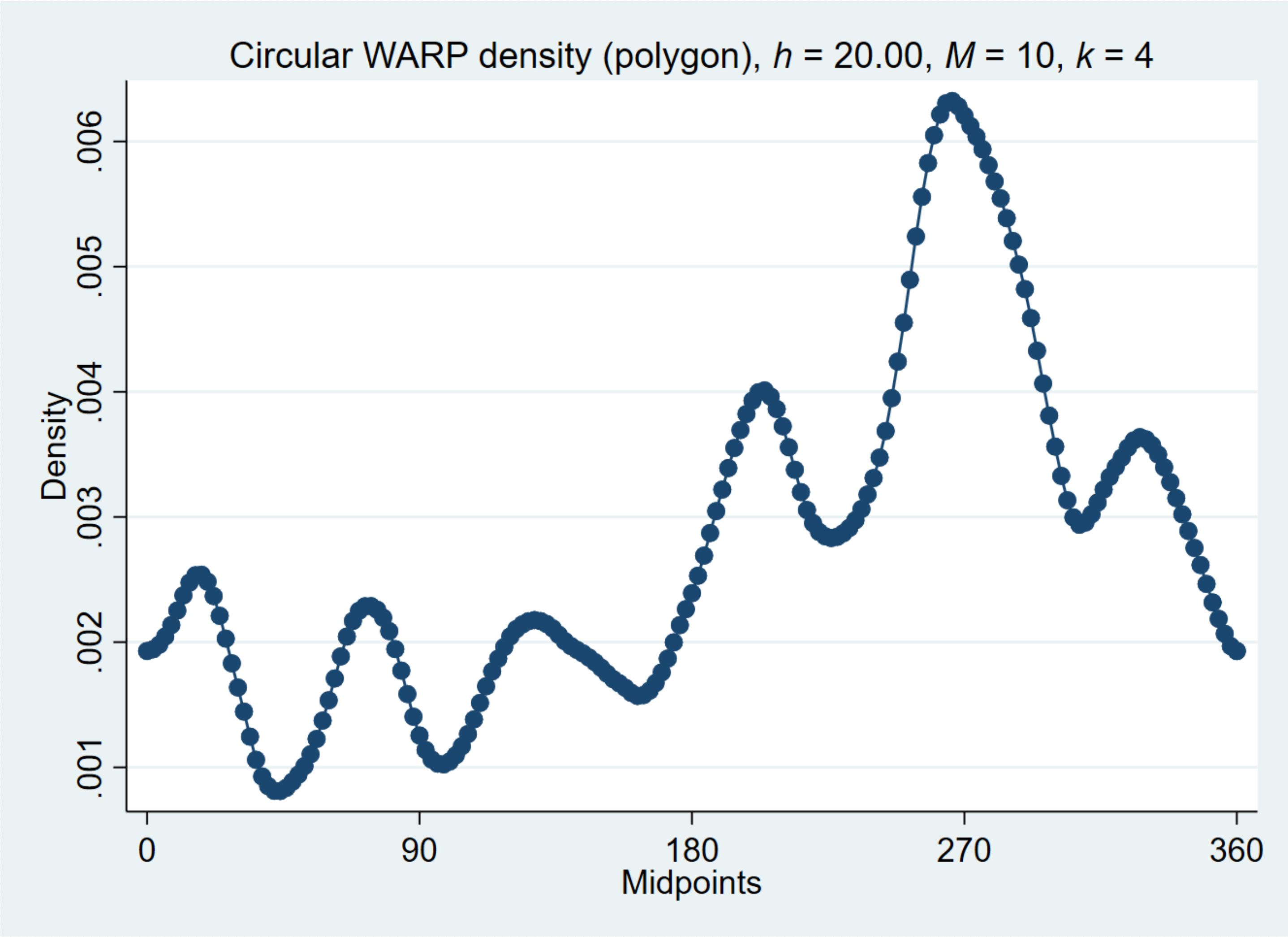
Options

- hwidth(#) is the smoothness parameter (half-width) in degrees. The default is 30.
- mval(#) specifies the number of averaged shifted histograms used to calculate the density
  estimations. A number of 10 (default) is suggested.
- kercode(#) set kernel (weight) function according to the following numerical codes (default is 4):
  1 = Uniform
  2 = Triangle
  3 = Epanechnikov
  4 = Quartic (Biweight)
  5 = Triweight
  6 = Gaussian
- gtype permits to chose the resulting graphical display according to the following numerical codes
  (defalut is 1):
  1 = Polygon
  2 = Step (histogram like)
  3 = Circular
- 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.

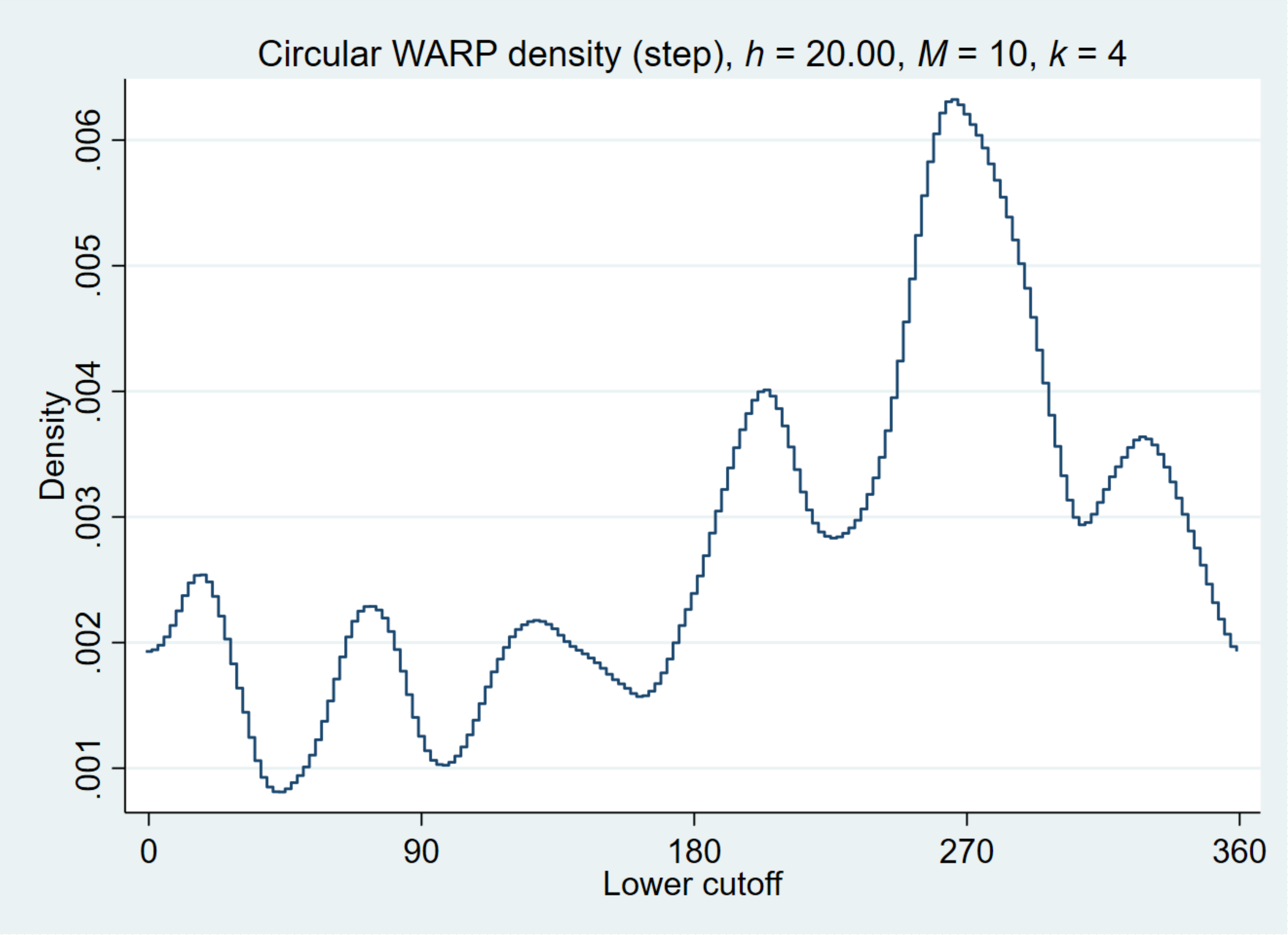
Program circwarp.ado (III)

- nograph(nograph) suppresses the graph drawing.
- 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.
- PLOT(str asis) are any of the options allowed with graph, twoway; see help for graph.
- Options for graph type 3 (circular)
- rval is a factor controlling the radius size of the circle used.
- frval is a factor applied to the density values in the cosine and sine transformation. It permits
  to stretch or compress the density values arround the circle.
- gsval is a factor controlling the size of the graph. Large values give small graphics while less
  than unity figures produce bigger circular graphs.
- Defaults are 1 in all cases. It is possible for the graphs to depart from circle by using other
  values. This can be corrected by using the right combination.

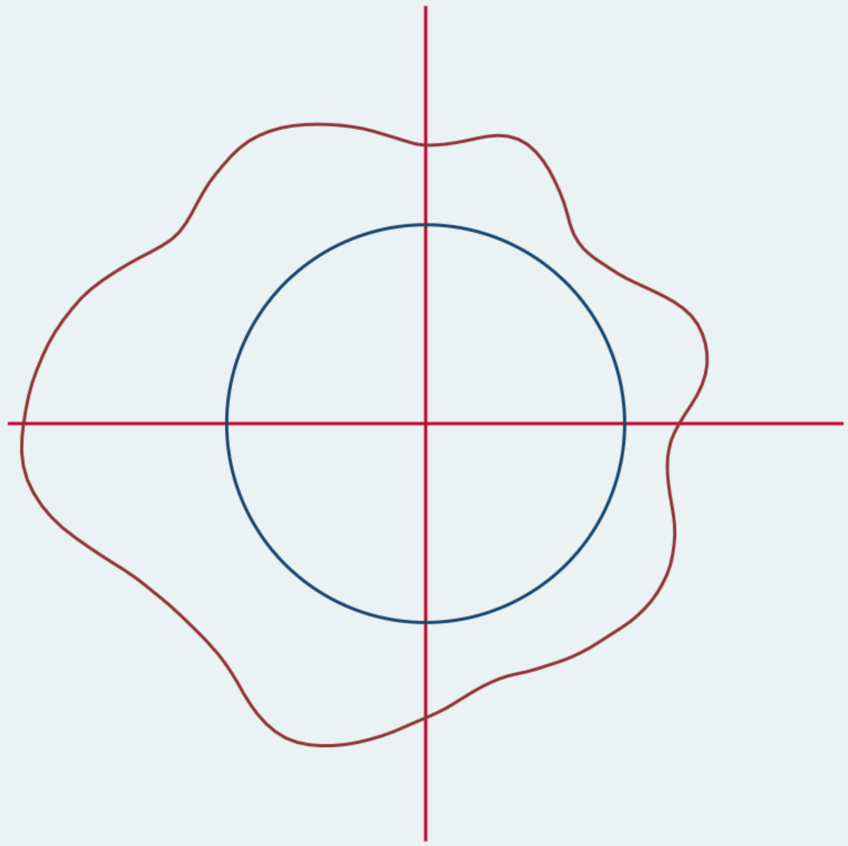
ASH-WARP density estimation, Quartic kernel, (linear; gt= 1. polygonal)



ASH-WARP density estimation, Triangular kernel (linear; gt = 2. step)



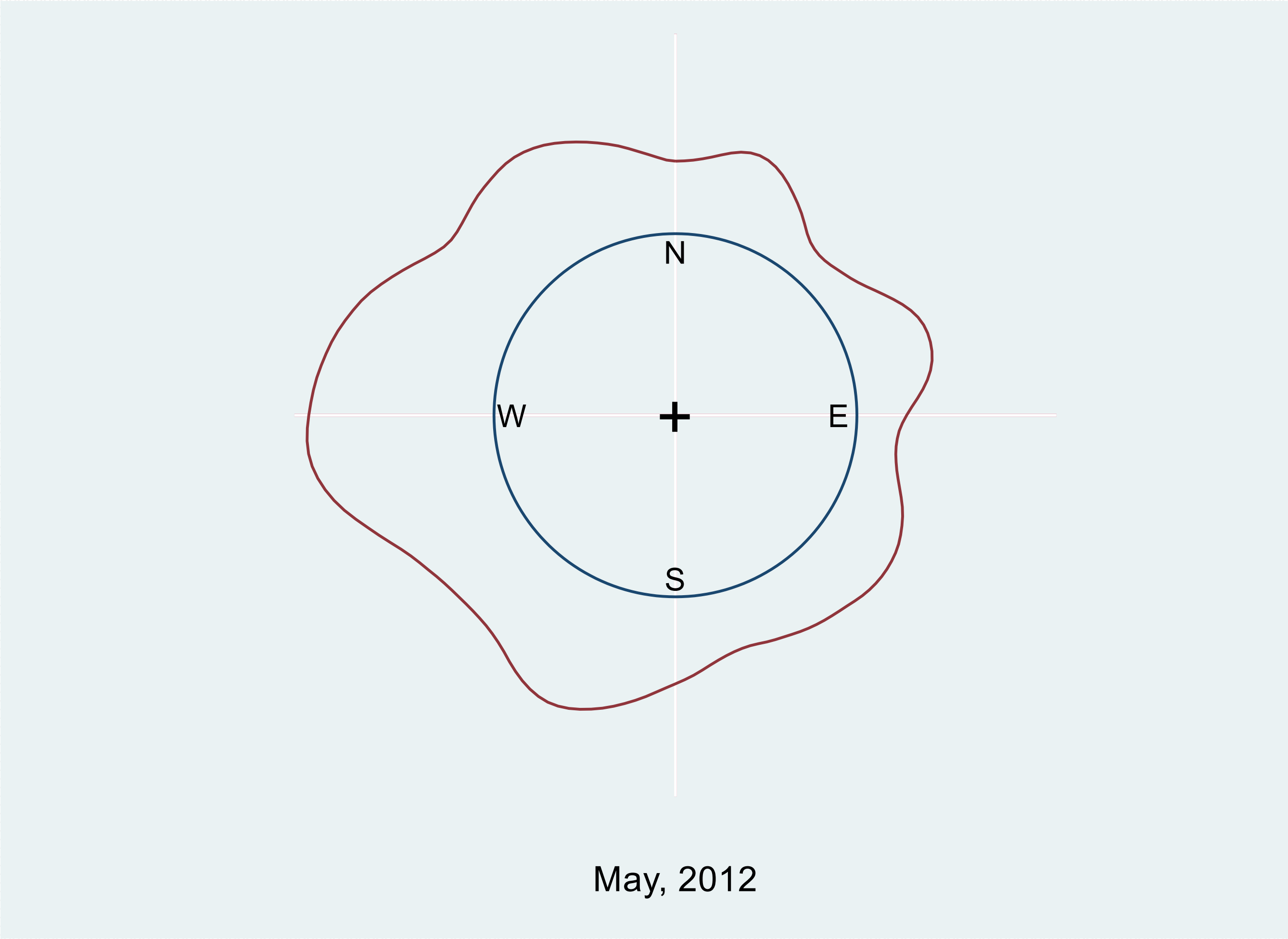
ASH-WARP density estimation, Quartic kernel (gt = 3. circular)



Circular WARP density, h = 20.00°, M = 10, k = 4

ASH-WARP density estimation, Quartic kernel (gt = 3. circular)

```
. circwarp wnd if dias>120 & dias < 152, h(20) subtitle("May, 2012") gt(3) 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") gen(dc mc)
```



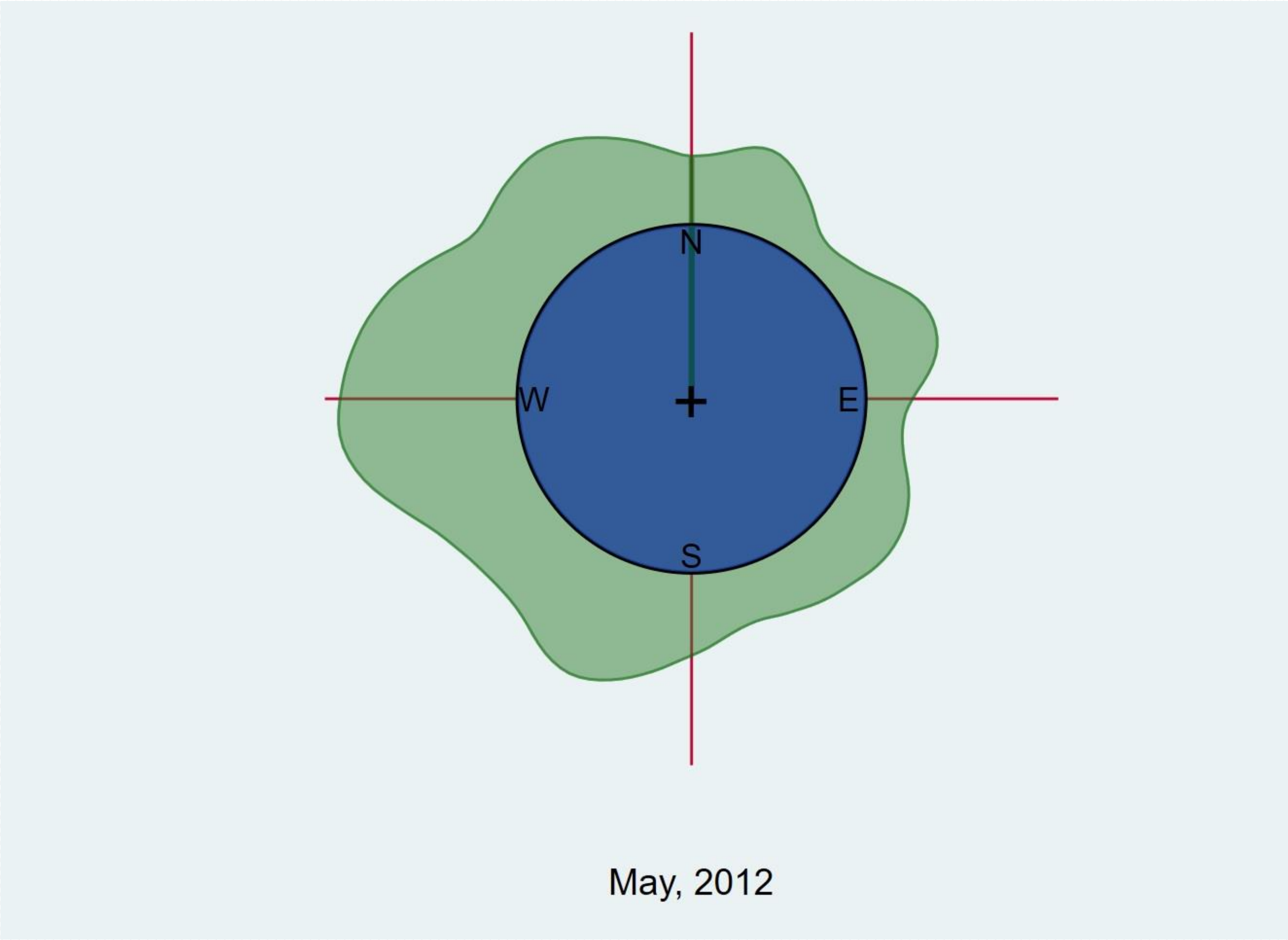
"circgph2.ado"

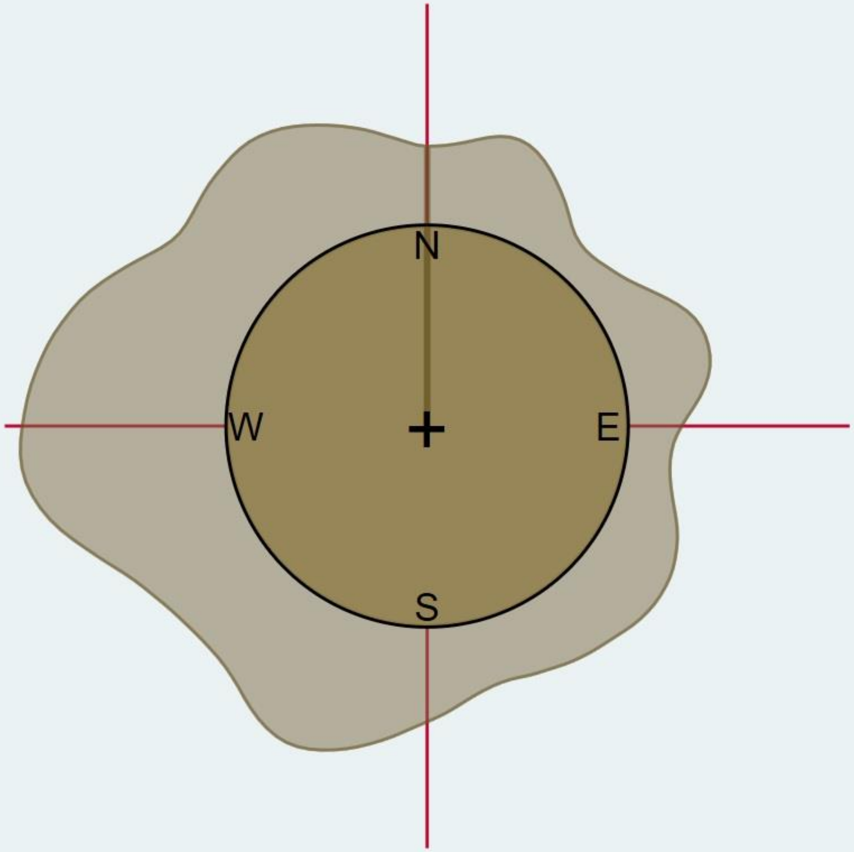
- Generates degree and density values for graphing density around the circle

- `circgph2 dc mc, gen(cosdg sindg cosd sind)`

- `two area cosdg sindg,  bc(blue) || area cosd sind,/// aspect(1) ysc(r(-2.1 2.1) off fill) xsc(r(-2.1 2.1) off /// fill) color(dkgreen%50) legend(off) yline(0) xline(0) /// ylab(, nogrid) plotregion(margin(zero) style(none)) ///`
- `|| line cosdg sindg, lc(black)`

Using area plots and transparency capabilities





May, 2012

CIRCULARKDE: Stata module to perform kernel density estimation for circular data

- This set of Stata programs allows to calculate KDE's for circular data based on previous algorithms by Fisher (1989; 1993), Cox (1997; 2001; 2004) D.W. Scott (1985; 1992; 2015), W. Härdle (1990) and Salgado-Ugarte et. al. (1995; 2018).

- https://ideas.repec.org/c/boc/bocode/s458922.html

CIRCULARKDE module contents

- circbw.ado
- circkden.ado
- cirkdevm.ado
- circgph.ado
- circwarp.ado

- circgph2.ado, not included (here presented)

Acknowledgements

- E. Batschelet,
- N.I. Fisher,
- N.J. Cox,
- B. Silverman,
- D.W. Scott, and
- W. Härdle

for having provided the basis for our algorithms.

References (I)

- Batschelet, E. (1979). Introduction to Mathematics for Life Scientists. 3d. Ed. Springer-Verlag, Heidelberg, Germany: 643 p.
- Batschelet, E. (1981). *Circular Statistics in Biology*. Academic Press Inc. London, United Kingdom, 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. *North American Stata Users Group Meeting Proceedings*. March 2001. Boston, EUA.
- Cox, N.J. (2004). Circular statistics in Stata, revisited. *United Kingdom Stata Users' Meeting Proceedings*. June 2004. London, United Kingdom.
- Fisher, N.I., (1993). *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge, United Kingdom, 277 p.
- 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: *Animal Orientation and Navigation* (S.R. Galler, K. Schmidt-Koenig, G.J. Jacobs y R.E. Belleville eds). National Aeronautic and Space Administration, Washington, USA: 511-522.
- Oliveira, M., Crujeiras R.M. y Rodríguez-Casal, A. (2012). A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics and Data Analysis,* 56(2012): 3898-3908.
- Salgado-Ugarte, I.H. (2002). *Suavización no paramétrica para análisis de datos*. FES Zaragoza y DGAPA, UNAM, México, 189 p.
- 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. & M.A. Pérez-Hernández, 2017. Estimación de densidad por núcleo (kernel) para datos circulares: 518-526. In: Rodríguez-Yam, G.A., F.J. Ariza-Hernández, B.R. Pérez-Salvador & F. Ulín-Montejo (Eds.), *Aportaciones recientes a la estadística en México*. Asociación Mexicana de Estadística, Instituto Nacional de Estadística y Geografía. INEGI, Aguascalientes, México. ISBN: 978-607-503-067-2.
- Salgado-Ugarte, I., R. Rivera-Reyes, A. Monroy-Ata, and V.M. Saito-Quezada (2015). Distribución de la dirección del viento en la FES Zaragoza analizada mediante estimadores de densidad por kernel circulares. In: Resúmenes del 11vo Congreso de Investigación de la FES Zaragoza, UNAM, CDMX, México.

References (II)

- Salgado-Ugarte, I.H., V.M. Saito-Quezada & M.A. Pérez-Hernández, 2018. Averaged shifted histograms (ASH) or weighted averaging of rounded points (WARP), efficient methods to calculate kernel density estimators for circular data: 89-96. In: Martínez-Martínez, A.F., L. Naranjo-Albarrán, P. Pérez-Rodríguez, L.J. Rodríguez-Esparza & C.E. Rodríguez-Hernández-Vela (Eds.), *Memorias del XXXI Foro Internacional de Estadística y del XXXII Foro Nacional de Estadística*. Instituto Nacional de Estadística y Geografía, Asociación Mexicana de Estadística. INEGI, Aguascalientes, México: 227 p.
- Salgado-Ugarte, I.H., V.M. Saito-Quezada, & M.A. Pérez-Hernández, (2021). "CIRCULARKDE: Stata module to perform kernel density estimation for circular data". Statistical Software Components S458922. Boston College Department of Economics.
- Salgado-Ugarte, I.H., Shimizu, M. y Taniuchi, T. (1993). Exploring the shape of univariate data using kernel density estimators. *Stata Technical Bulletin*, 16: 8-19.
- Salgado-Ugarte, I.H., Shimizu M. & Taniuchi, T. (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, New York.
- SenGupta, S. y Rao, J.S. (1966). Statistical analysis of cross-bedding azimuths from the Kamthi formation around Bheemaram, Pranhita: Gosvari Valley. Sankhya: *The Indian Journal of Statistics, Series B*, 28: 165-174.
- Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman y Hall, London, UK:
- StataCorp, (2013). *Stata: Release 13. Statistical Software*. StataCorp LP, College Station, USA.
- StataCorp, (2019). *Stata: Release 16. Statistical Software*. StataCorp LP, College Station, USA.
- Stephens, M.A. (1969). *Techniques for directional data*. Technical Report #150, Dept. of Statistics, Stanford University, Stanford, CA, USA (23, 102, 241).
- 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. New Jersey. 663 p.

Thank you very much

FES Zaragoza UNAM Campus, Mexico City (satellite Google view); Circle indicate Meteorological Station position.