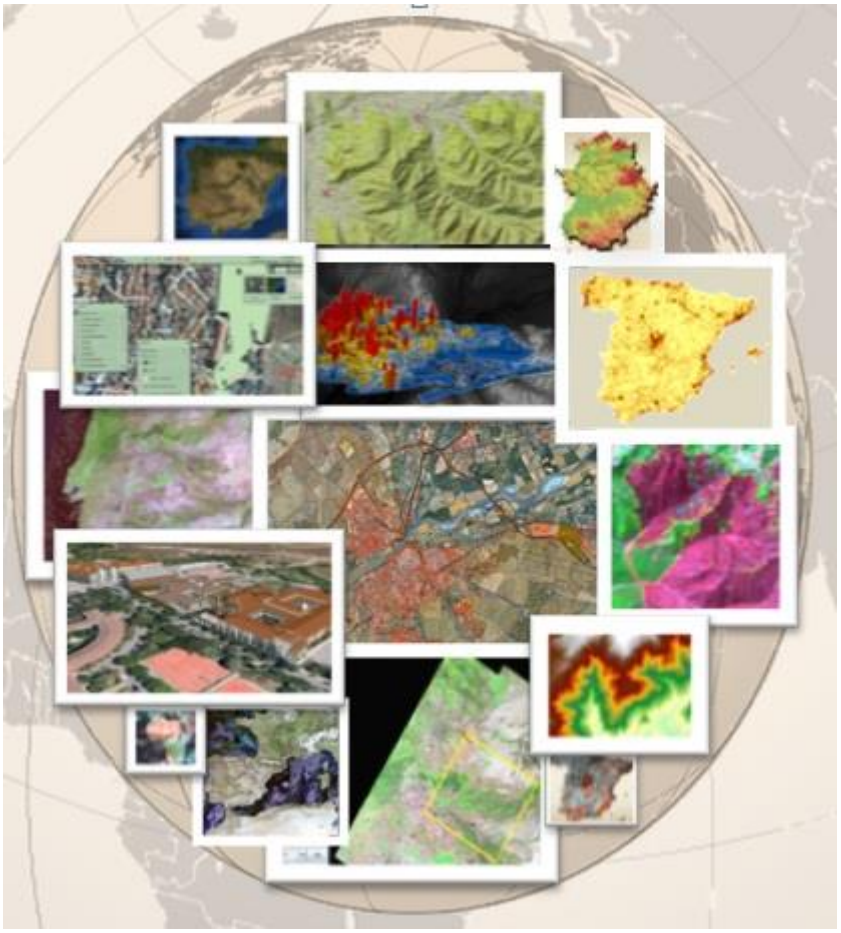
Máster en Tecnologías de la Información Geográfica: SIG y Teledetección



GEOESTADÍSTICA Y CALIDAD DE LA INFORMACIÓN

Datos circulares

María Eugenia Polo García

Índice

1. Datos circulares

La llamada **estadística circular** o el análisis estadístico de datos circulares permite el tratamiento de datos angulares. En muchos ámbitos de la ciencia se realizan mediciones angulares que a veces implican vectores (datos de viento, por ejemplo) [1] o a veces solamente direcciones, sin sentido ni módulo. Estos valores no pueden ser manejados como datos escalares sin más, pues tienen una componente de dirección que obliga a aplicar este tipo de estadística de datos para un correcto tratamiento de los mismos. Mientras que los datos lineales se presentan en la recta real, los datos circulares lo hacen en un círculo de radio 1 o **círculo unidad,** siendo el círculo unidad a los datos circulares lo que la recta real es a los datos lineales o escalares [2]. Por tanto, la estadística circular es una rama de la Estadística que está diseñada exclusivamente para el análisis de datos cíclicos denominados circulares.

Esta estadística tiene en cuenta algunas características diferenciales de este tipo de información [3]:

a) El origen de los datos circulares es una dirección arbitraria (norte geográfico, eje X…). Puede estar el 0° situado en la dirección norte o sur o en cualquier otra que consideremos. Estos datos pueden ser medidos en sentido horario o sentido contrario a las agujas del reloj. Los datos lineales, en cambio, tienen su origen en el valor 0 de la recta real, creciendo con valores positivos hasta el infinito a la derecha y con valores negativos a la izquierda.

b) No existe una relación de magnitud entre los datos; 285° no es mayor que 175° en el sentido convencional del término, mientras que en la recta real un valor de 285 sí es mayor que otro de 175.

c) Los valores máximos y mínimos son arbitrarios.

d) Las operaciones algebraicas con datos circulares se miden en una escala finita y debe reducirse al intervalo 0-360° (o los valores equivalentes en la escala que se use, por ejemplo 0-2π).

Si tenemos un dato circular de 10° y otro de 350°, por ejemplo direcciones de vientos, la media aritmética de ambos valores sería 180° (en rojo en la Figura 1). Este estadístico para datos lineales no es el apropiado para datos circulares, siendo necesario en este caso, hacer una suma vectorial, para que el resultado (0°) sea coherente.

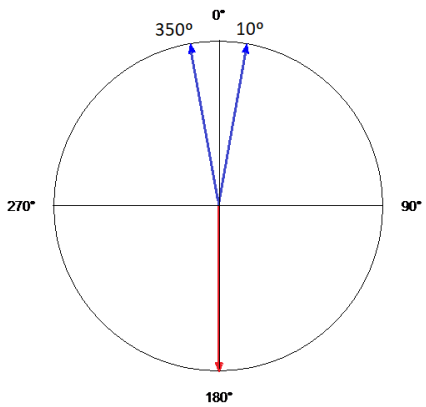


Figura 1. Media aritmética errónea de dos datos circulares (en color rojo).

Ejemplos de datos circulares los tenemos en disciplinas como la biología [4], al estudiar los vuelos de aves migratorias o la dirección del movimiento de animales bajo un determinado estímulo. En geología podemos, por ejemplo, aplicar este tratamiento de datos al estudio de las direcciones de estratos o fallas del terreno. Pero hay más disciplinas que usan datos circulares como la meteorología (direcciones de vientos predominantes), la geografía (movimientos de la población), estudio de errores de posición en cartografía [5], detección de errores en cartografía hidrológica [6] o análisis del potencial de la radiación solar [7]. En general, cualquier ciencia que trate con datos cíclicos que puedan ser representados en una rosa de los vientos (estudio espacial), pero también que pueda ser tomado como medición angular, reloj o calendario (estudio temporal). En este último caso

podemos, por ejemplo, analizar el número de asistencias médicas a un determinado centro hospitalario según las horas del día o meses del año o también el estudio de ciclos o ritmos biológicos. Estas medidas de tiempo que se representan en un reloj, son posiciones que pueden ser convertidas en ángulos. En general, son datos circulares aquellos que son cíclicos y pueden ser medidos y convertidos en grados o radianes y representados en un círculo unidad.

En la bibliografía de datos circulares se suele diferenciar entre **vectores** que tienen orientación y sentido y **ejes o datos axiales** con orientación, pero sin sentido definido, como se observa en la figura siguiente. Ejemplos del uso de vectores los tenemos en los ya mencionados estudios de direcciones de vientos, pues estos además de la dirección tienen un sentido (hacia el norte, sur…) o en el movimiento de animales. Aquí aparecen representados con una flecha en el sentido del movimiento.

Como datos axiales o ejes, mencionamos el estudio de las fracturas del terreno, que tienen dirección, pero sin sentido definido. Aquí se han representado con una línea terminada en círculos con dirección, pero sin sentido definido y, en este caso, la escala varía entre 0° y 180°, pues estos datos tienen direcciones opuestas y equivalentes. Como ejemplo, supongamos las medidas de las fracturas de un terreno representadas en la Figura 2 b.

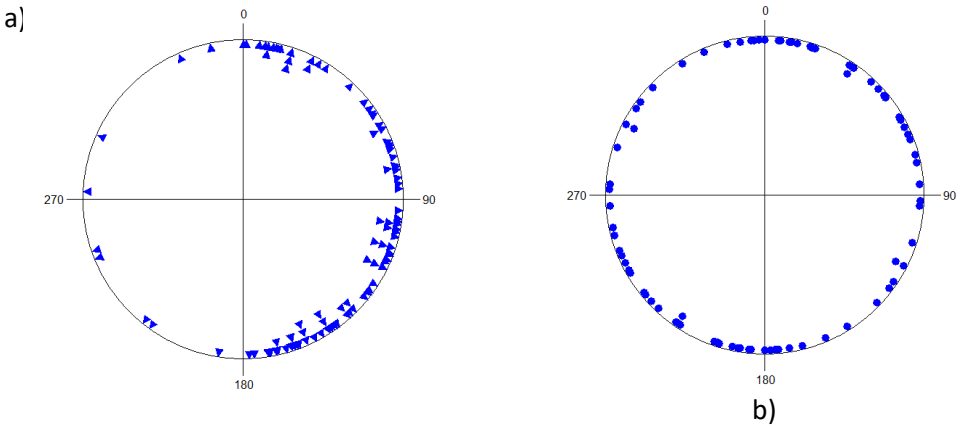


Figura 2. a) Representación de vectores (orientados con sentido). b) Representación de ejes (orientados sin sentido).

En este texto y en el tema que nos ocupa, trataremos inicialmente con **vectores**, siendo su origen (0°) la dirección norte geográfica y creciendo en sentido horario (Figura 3 a). De entre los sistemas de medidas angulares usaremos la graduación sexagesimal preferentemente sobre los radianes y la graduación centesimal, aunque esta última es habitual en mediciones topográficas. Esta explicación se corresponde a la definición de acimut topográfico.

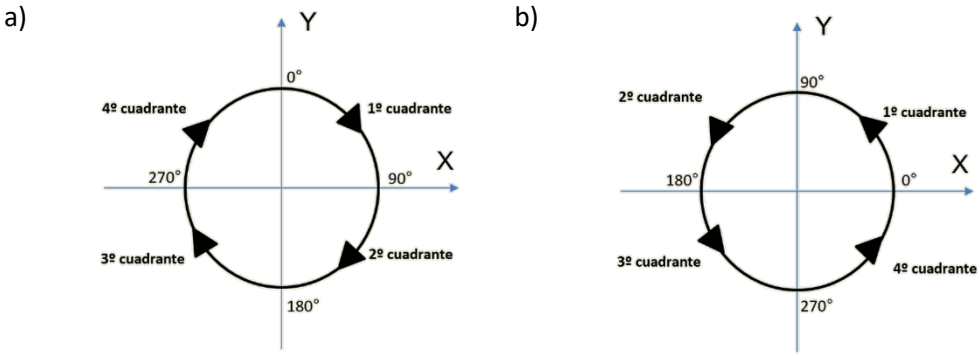


Figura 3. Cuadrantes y sentidos angulares: a) topográfico y b) matemático.

Si trabajamos en un círculo unidad, cualquier dato circular está definido por su ángulo, dado que el radio es la unidad. Para pasar de coordenadas polares (distancia y ángulo) a

coordenadas rectangulares (X, Y), usamos las siguientes expresiones según las relaciones que aparecen en la Figura 4.

En cuanto a la representación gráfica de los vectores, optaremos generalmente por una línea terminada en flecha (nodo final) apuntando con un acimut determinado. El vector bidimensional se calcula a partir de las diferencias de coordenadas entre la posición final (*Xf, Yf*) y la inicial (*Xi,Yi*) de los puntos considerados (coordenadas cartesianas). Con estos incrementos de coordenadas se obtiene el **módulo** (*d*) y un **acimut** (θ) de cada vector (coordenadas polares), según las expresiones siguientes y la Figura 4:

$$d = \sqrt{(Xf - Xi)^2 + (Yf - Yi)^2} \qquad \theta = arctg\left(\frac{Xf - Xi}{Yf - Yi}\right)$$

En relación al cálculo del **acimut**, hay que indicar que es un ángulo plano, medido en sentido horario desde la dirección norte geográfico hasta el vector formado por la posición medida o inicial y la verdadera o posición final, tal como se indica en el gráfico adjunto.

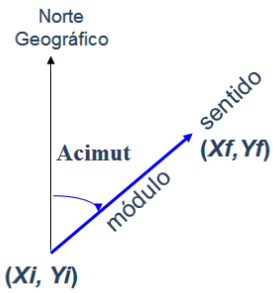


Figura 4. Vector con módulo y ángulo (acimut).

El aspecto clave de este concepto es que tenemos acceso a las propiedades espaciales de los vectores, que son: **módulo** (propiedad métrica), **ángulo** (propiedad métrica) y **sentido** (propiedad topológica). Estas propiedades permitirán analizar características, como la isotropía, que no eran determinadas en el caso de datos lineales o escalares.

Aunque la mayoría de los ejemplos que nos ocupan en este tema se refieren a movimientos angulares en el espacio, recordemos que la estadística circular también se aplica a eventos cíclicos en el tiempo que se repitan a lo largo de las horas del día, los días de la semana, los días del año o los meses del año. En este artículo [8] se analizan las denuncias por ruido recibidas en Ostrava, (República Checa), donde por ley, el ruido excesivo entre las 10 de la noche y las 6 de la mañana está sujeto a una multa. En este otro estudio [9] se usa la estadística circular para analizar los modelos de tiempo en el campo de la ciberseguridad con incidentes de infección de malware.

2. Medidas de tendencia central

Los principales estadísticos de tendencia central de una serie de observaciones angulares $\theta_1, \theta_2, \theta_3, \ldots, \theta_n$, son [10-13]:

- **Dirección media** ($\bar{\theta}$). Obtenida mediante la suma vectorial de todas las observaciones de la muestra. Representa el ángulo medio del vector resultante (*R*) de la suma vectorial citada. Su expresión, en la que hay que tener en cuenta el cuadrante en que nos encontremos, es:

$$\bar{\theta} = arctg\frac{S}{C}$$

Donde: $S = \sum_{i=1}^{n} seno\theta_i \qquad C = \sum_{i=1}^{n} \cos\theta_i$

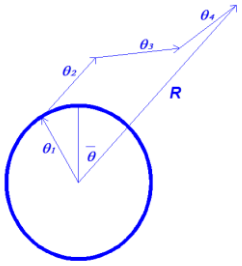


Figura 5. Cálculo de la dirección media.

- **Mediana ($\tilde{\theta}$).** En datos lineales la mediana se corresponde con el valor central de una serie de elementos ordenados de forma creciente. Esta definición, por la naturaleza de los datos circulares, no es aplicable a este tipo de datos, dado que el origen de medida de ángulos es arbitrario. La forma de calcular la mediana es minimizando la función siguiente:

$$d(\theta) = \pi - \frac{1}{n}\sum_{1}^{n}\left|\pi - \left|\theta_i - \theta\right|\right|$$

Además, para distribuciones multimodales el valor de la mediana puede no ser único.

- **Moda.** Al igual que en datos lineales es el valor que más se repite en la muestra de datos.

3. Medidas de dispersión

Al igual que ocurría con los datos lineales, las medidas de dispersión cuantifican la separación de valores respecto a las medidas de tendencia central complementando la información proporcionada por estas. Destacamos:

- **Longitud resultante media** ($\bar{R}$). Obtenido al dividir el vector resultante R entre el número de observaciones.

$$R = \sqrt{C^2 + S^2} \qquad \bar{R} = \frac{R}{n}$$

Lógicamente, si trabajamos con vectores unitarios su valor oscila entre 0 y 1. Si es 1 implica que todos los vectores son coincidentes en su dirección; en cambio, un valor nulo puede ser el resultado de una distribución uniforme pero pueden darse situaciones, donde

el valor de $\overline{R}$ es muy pequeño y la distribución dista bastante de ser uniforme, como se aprecia en la figura siguiente:

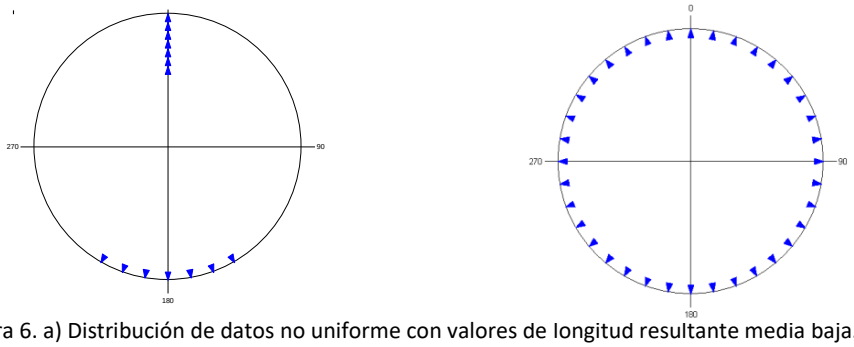


Figura 6. a) Distribución de datos no uniforme con valores de longitud resultante media baja. b) Distribución de datos uniforme.

- **Varianza circular de la muestra** (*V*). En similitud con la varianza de datos lineales, es un indicador de dispersión angular de forma que cuanto más pequeño sea este valor, más concentrados están los datos. Su valor oscila entre 0 y 1, pero al igual que ocurría con la longitud resultante media, *V*=1, no implica una distribución uniforme. Su expresión es:

$$V = 1 - \overline{R}$$

- **Desviación estándar circular de la muestra** (υ): Es resultado de una raíz cuadrada por analogía a lo que ocurre en datos lineales aunque en datos circulares queda definida por:

$$\upsilon = \left[-2\log(1-V)\right]^{1/2}$$

- **Dispersión circular de la muestra** (δ): En el caso de una distribución uniforme este valor es infinito. Valores próximos a cero implican una alta concentración de datos.
- **Parámetro de concentración** (κ): Se llama de von Mises y mide la variación de la distribución en relación con un círculo perfecto (distribución uniforme) en la distribución del mismo nombre. Cuando este parámetro κ tiende a cero, la distribución converge a la distribución uniforme; en caso de tender a infinito, la distribución se concentra en la dirección media del vector resultante.

También se encuentran definidos en datos circulares los coeficientes de asimetría y curtosis.

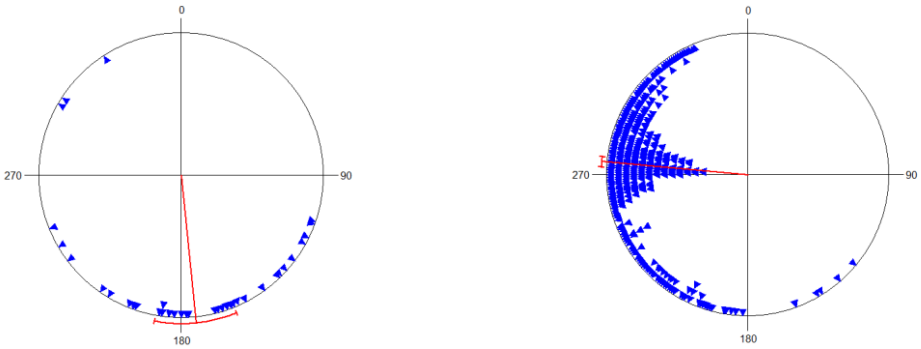
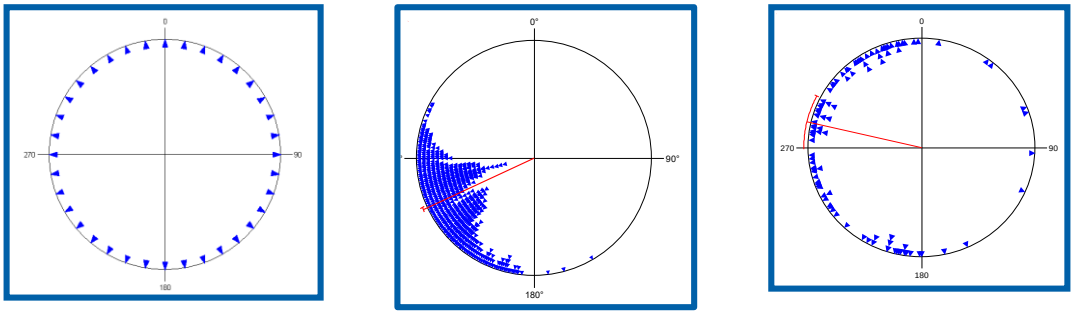


Figura 7. Diferentes concentraciones de datos circulares con indicación del acimut medio o dirección media y el intervalo de confianza en color rojo.

En la figura siguiente se muestran tres configuraciones diferentes de concentración de datos, con los resultados de los estadísticos explicados anteriormente. Vemos que, en el primer conjunto, la dirección media y la desviación circular no pueden ser calculadas.



	Datos1	Datos2	Datos3
Dirección media	No definido	244.97°	283.19°
Longitud resultante media	0	0.93	0.53
Varianza circular	1	0.06	0.47
Desviación circular	No definido	20.85°	64.75°
Parámetro von Mises	0	8.03	1.24
Dispersión circular	infinito	0.13	1.56

Figura 8. Ejemplos de datos circulares en tres configuraciones diferentes.

4. Funciones de distribución de probabilidad de datos circulares

Las distribuciones de probabilidad de datos circulares más comunes son la distribución uniforme y la von Mises.

- **Distribución uniforme (*Uc*)**

En esta distribución todas las direcciones entre *0* y *2π* radianes son igualmente probables. La función de densidad de probabilidad atiende a la expresión:

$$f(\theta) = \frac{1}{2\pi} \qquad 0 \le \theta \le 2\pi$$

Mientras que la función de distribución aparece en la ecuación siguiente.

$$F(\theta) = \frac{\theta}{2\pi} \qquad 0 \le \theta \le 2\pi$$

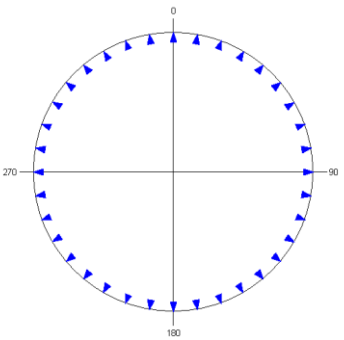


Figura 9. Distribución uniforme de datos circulares.

La dirección media de la población no está definida, la longitud resultante media poblacional es igual a cero y la dispersión circular es infinita. Un ejemplo de distribución uniforme de datos lo tenemos en la figura adjunta.

Esta distribución será frecuentemente la hipótesis nula sobre la cual se contrastarán otras alternativas de distribución de datos.

- **La distribución von Mises**

Esta distribución, simétrica y unimodal, es el modelo más habitual de distribución en muestras unimodales de datos circulares, siendo esta distribución en el círculo, análoga a la distribución Normal en el caso de variables lineales. En la figura siguiente aparece representada la distribución de probabilidad de von Mises.

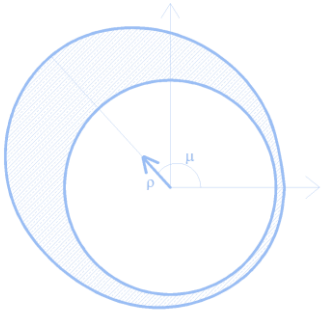


Figura 10. Distribución von Mises.

Existen otras distribuciones para datos circulares, como la distribución cardioide, wrapped o wrapped normal. Se remite al lector a la bibliografía indicada al final del tema para ampliar conceptos.

5. Representación gráfica

Existen diversas formas de representar los datos circulares según sean vectores o ejes.

La representación más simple es la del dibujo de los **datos brutos** (Figura 11 a) . Este tipo de gráfico es muy útil para una primera aproximación de la distribución de la variable.

Otra forma de representación es usando **histogramas circulares**. En este caso, el ancho de clases se expresa como sectores circulares de un número determinado de grados. El sector puede estar calculado teniendo el radio o el área proporcional a la frecuencia de los datos.

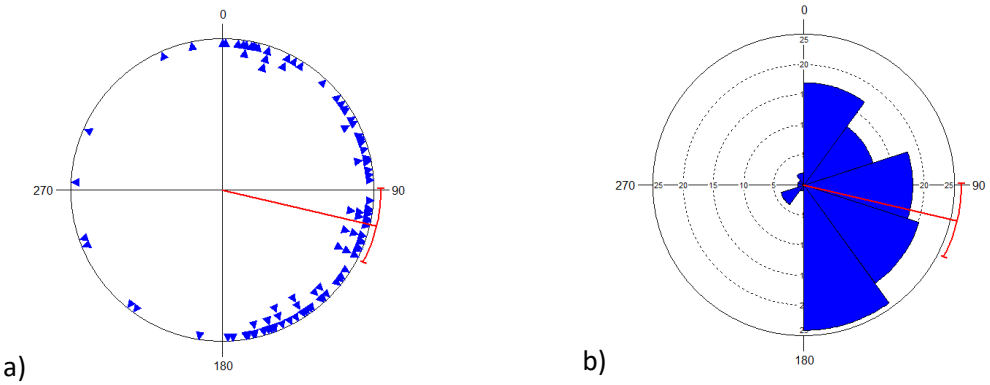


Figura 11. a) Representación de vectores. b) Mismos datos como histogramas de datos circulares.

Aunque la estadística de datos circulares trabaja con vectores unidad, se pueden representar datos de vectores con módulos distintos a la unidad en los llamados **gráficos circular-lineal**. En la figura siguiente (Figura 12) se llevan a un mismo origen todos los vectores donde se puede visualizar el módulo y la dirección y sentido de cada uno de ellos (caso a). El caso b) es igual al caso a) pero solo se representan los nodos finales de los vectores por un punto.

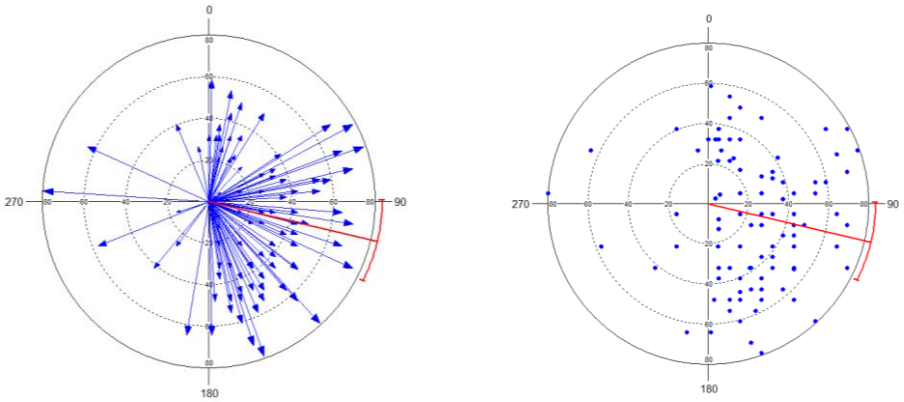


Figura 12. a) Representación de dos variables (circular-lineal) con flechas. b) Representación de dos variables (circular-lineal) con puntos.

Se debe recordar que los valores de dirección media e intervalo de confianza pueden no ser representativos debido a una baja concentración de datos o a distribuciones específicas, como ocurre en la figura siguiente. El caso a) muestra una baja concentración de datos en zonas diametralmente opuestas que genera un intervalo de confianza para el acimut medio de casi la circunferencia completa. El caso b) se corresponde con una distribución bimodal, con la misma incertidumbre.

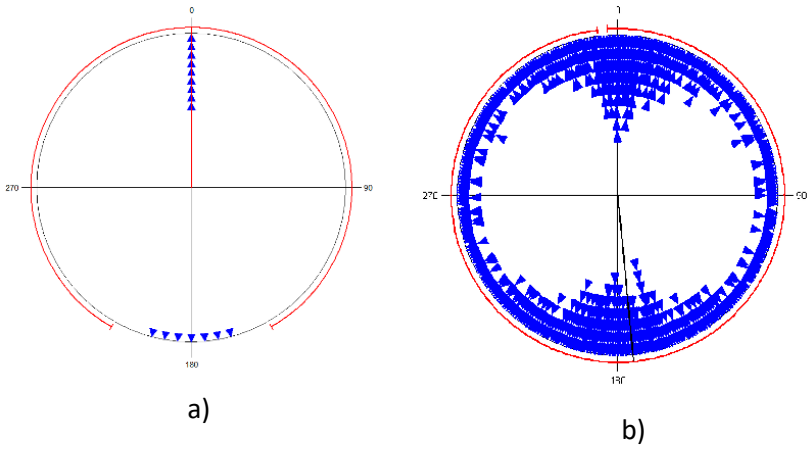


Figura 13. Valores de dirección media e intervalo de confianza no representativos dada la distribución de los datos.

6. Test de bondad de ajuste

La aplicación de diversos test estadísticos a las mediciones realizadas permitirá analizar la distribución y características de las observaciones e inferir su comportamiento. Antes de aplicar cualquier test estadístico en busca de las características de la muestra de datos o de la población, es interesante realizar un análisis exploratorio consistente en el dibujo de los datos brutos.

Existen pruebas para valorar de forma gráfica el ajuste de los datos a determinadas distribuciones como los gráficos del tipo qq-plot. Si los puntos se encuentran dibujados siguiendo una línea aproximada de 45°, pasando por el origen, los datos se adaptan al modelo que se quiere testar, en el ejemplo de la figura siguiente, la distribución uniforme.

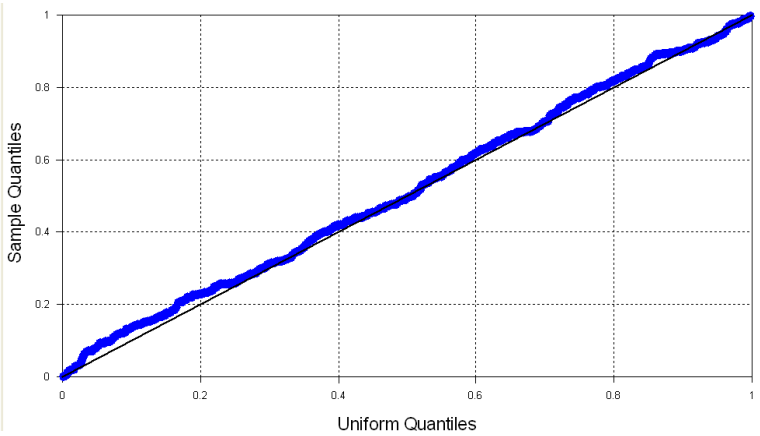


Figura 14. Gráfico qq-plot.

Una de las pruebas más comunes para testar la uniformidad de los datos circulares es el **test de Rayleigh.**

Supongamos de entrada que no conocemos la dirección media especificada, que será lo más común. La hipótesis de uniformidad se rechaza si la longitud resultante media de la muestra es muy grande. El estadístico a calcular es:

$$Z = n\overline{R}^2$$

Siendo la probabilidad *P* de aceptar la hipótesis de uniformidad, en caso de ser n≥50:

$$P = \exp(-Z)$$

Este test supone que un valor de longitud resultante media más larga implica una concentración mayor en torno a la media y por tanto menos probabilidad de que los datos estén uniformemente distribuidos.

El **test de Rao de espaciado de datos** considera también como hipótesis nula que los datos están distribuidos uniformemente, con la salvedad de buscar si el espaciado entre puntos adyacentes es aproximadamente igual en todo el círculo. Para una distribución uniforme el espaciado entre puntos debería ser 360°/n. Si el espaciado existente se desvía mucho de este valor, la probabilidad de que los datos pertenezcan a una distribución uniforme se reduce. Se necesita para ello calcular el valor de la ecuación siguiente.

$$L = \frac{1}{2}\sum_{i=1}^{n}\left|T_i - \frac{2\pi}{n}\right|$$

El valor de T_i aparece en la ecuación siguiente:

$$T_i = \theta_{(i)} - \theta_{(i-1)} \quad i = 1,\ldots,n-1, \quad T_n = 2\pi - \left(\theta_{(n)} - \theta_{(1)}\right)$$

Existen otras pruebas como el test de Kuiper, el test de Watson, test de Hodges-Ajne, entre otros, para determinar si la muestra sigue una determinada distribución, generalmente la uniforme o de von Mises [2].

7. Ejemplos de aplicación

Como ejercicio práctico vamos a instalar un paquete de R llamado ***VecStatGraphs2D***, que permite analizar vectores en 2D desde un punto de vista numérico y gráfico. Las funciones a utilizar y la descripción de las mismas aparecen en el script denominado ***script_vectorial.R.*** Como conjunto de datos usaremos 6 ficheros en formato txt de la carpeta ***vientos*** que se corresponden con los datos de las distribuciones de vientos de los días 225, 250, 275, 300, 315 y 350 del año 2002, en una determinada zona del planeta, según se explica en el artículo siguiente que se encuentra subido al campus virtual:

P. G. Rodríguez, M. E. Polo, A. Cuartero, Á. M. Felicísimo, and J. C. Ruiz-Cuetos, "VecStatGraphs2D, A Tool for the Analysis of Two-Dimensional Vector Data: An Example Using QuikSCAT Ocean Winds," *IEEE Geoscience and Remote Sensing Letters,* vol. 11 (5), pp. 921-925, 2014.

Las dos columnas que aparecen en los citados ficheros se corresponden con los valores de incremento de x e incremento de y, respectivamente de los datos de viento.

Otro ejemplo de aplicación puede consultarse en el siguiente trabajo, que también se encuentra subido al campus virtual:

Polo, M.-E.; Pozo, M.; Quirós, E. (2018). Circular Statistics Applied to the Study of the Solar Radiation Potential of Rooftops in a Medium-Sized City. Energies, 11 (10), 2813.

En este trabajo se aplica la estadística circular al estudio del potencial de la radiación solar en los tejados de la ciudad de Cáceres. Se analiza la orientación de los tejados en tres zonas concretas: a) en el centro de la ciudad, b) en una zona residencial a las afueras y c) en un polígono industrial. Las orientaciones de los tejados, según la zona, son las que aparecen en la Figura 15. Las tres zonas presentan un entramado urbano diferente que se refleja en la orientación de los tejados.

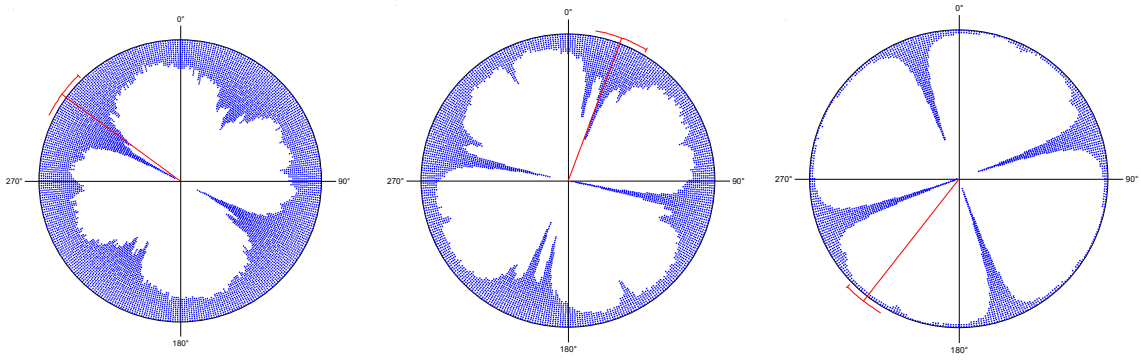


Figura 15. Orientación de los tejados en a) zona residencial en el centro, b) zona residencial en las afueras y d) polígono industrial.

8. Bibliografía

[1] P. G. Rodríguez, M. E. Polo, A. Cuartero, Á. M. Felicísimo, and J. C. Ruiz-Cuetos, "VecStatGraphs2D, A Tool for the Analysis of Two-Dimensional Vector Data: An Example Using QuikSCAT Ocean Winds," *IEEE Geoscience and Remote Sensing Letters,* vol. 11, pp. 921-925, 2014.

[2] J. L. Pérez Bote, *Introducción a la Estadística Circular*. Cáceres: Universidad de Extremadura. Servicio de Publicaciones, 2019.

[3] A. Pewsey, M. Neuhäuser, and G. D. Ruxton, *Circular statistics in R*: Oxford University Press, 2013.

[4] M. Mena, *Aplicaciones de estadística circular a problemas de ciencias naturales*. Buenos Aires: Akadia, 2004.

[5] M. E. Polo and Á. M. Felicísimo, "Full Positional Accuracy Analysis of Spatial Data by Means of Circular Statistics," *Transactions in GIS,* vol. 14, pp. 421–434, 2010.

[6] E. Quiros, M. E. Polo, and A. M. Felicisimo, "Detection and Labeling of Sensitive Areas in Hydrological Cartography Using Vector Statistics," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 54, pp. 189-196, 2016.

[7] M.-E. Polo, M. Pozo, and E. Quirós, "Circular Statistics Applied to the Study of the Solar Radiation Potential of Rooftops in a Medium-Sized City," *Energies,* vol. 11, p. 2813, 2018.

[8] J. Horák and L. Orlíková, "Circular Statistics for Directional and Temporal Data : Case Studies of Lineaments and Noise Violations Offence," in *2019 International Conference on Military Technologies (ICMT)*, 2019, pp. 1-6.

[9] L. Pan, A. Tomlinson, and A. A. Koloydenko, "Time Pattern Analysis of Malware by Circular Statistics," in *2017 ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)*, 2017, pp. 119-130.

[10] N. I. Fisher, *Statistical analysis of circular data*, 2nd ed. Cambridge, UK: Cambridge University Press, 1995.

[11] K. V. Mardia and P. E. Jupp, *Directional Statistics*. Chichester, UK: Wiley, 2000.

[12] S. R. Jammalamadaka and A. SenGupta, *Topics in circular statistics* vol. 5. Singapore: World Scientific Publishing, 2001.

[13] M. E. Polo García and Á. M. Felicísimo, "Propuesta de metodología para el análisis del error de posición en bases de datos espaciales mediante estadística circular y mapas de densidad," *Geofocus,* vol. 8, pp. 281-296, 2008.