PYCIRCULARSTATS: A PYTHON-BASED TOOL FOR REMOTE SENSING CIRCULAR STATISTICS AND GRAPHICAL ANALYSIS

A. Cuartero[1], *Mercedes E. Paoletti*[2], *Senior Member, IEEE,*, *P. García-Rodríguez*[3], *Juan M. Haut*[4] *Senior Member, IEEE,*

[1]Department of Graphic Expression, University of Extremadura, Cáceres, Spain.
[2]Department of Computer Architecture, Complutense University, Madrid, Spain.
[3]Department of Computer and Telematics Systems Engineering, University of Extremadura, Spain.
[4]Hypercomp, Dept. of Technology of Computers and Communications, University of Extremadura, Spain.

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9884758

ABSTRACT

Circular data, as a part of directional data engineering, is used in a wide range of fields such as Geology, Biology, Meteorology and Geomatics. It differs from traditional linear data because it is closed and has no beginning or end along the real line, i.e., circular data occurs around a circle, normally measured in degrees. Analyzing directional data, in particular circular data, requires methods that are be available in libraries with a well-known prestige as Python including SciPy, NumPy or SciKit-Learn libs. However, these libraries have a specific area of expertise and they do not combine information in a useful way for two-dimensional data analysis. In this paper, an open-source library has been implemented to be executed by the Python interpreter, called PyCircularStats. Source code: `https://github.com/mhaut/pycircularstats`

Index Terms— Analysis data; circular graphical statistics; geospatial data engineering; remote sensing

1. INTRODUCTION

There are various types of acquired data in Earth sciences. In particular, directional data [1] is characterized by being defined on a circular or spherical domain, i.e. data belongs to angular measurements, made up of an orientation or angle. Depending on whether the orientation is in space or in the plane, the directional data are classified into spherical and circular data: while spherical data [2] refers to the orientation in space and follows a Fisher distribution [3] (also called F-distribution), circular data [4] is measured in angles or directions on the unit circle and refers to the orientation in the plane, following a Von Mises distribution [5] and being the simplest kind of directional data.

We can observe the use of these types of data in fields of Earth Sciences [6] as varied as Geology [7], Geomatics, Biology [8], Oceanographic [9] or Meteorology [10] areas, in applications such as the measurement of winds, migratory movements or the analysis of striae orientation on fault planes, although directional data can be found in others fields too, as image processing [11] or even in machine learning methods [12].

However, the calculation of the descriptive and inferential statistics of directional data differs from the linear statistics due to the behavior and periodic nature of the directional data, which distinguishes this data by its unique and novel characteristics, both in terms of modeling and statistical processing. So new statistics are required since those used for linear data are inappropriate. Directional statistics, in particular circular statistics [4] deal with angular data, such as axes or vectors, to analyze the azimuth and magnitude in 2-D. In order to work with circular statistics we can found several ways and methods which allow us to use circular data. However, most of them are from a very specific area of expertise and they do not combine information in a useful way for two-dimensional data analysis or even are part of proprietary software.

In [13] *CircStat*, a circular toolbox that provides methods for the descriptive and inferential statistical analysis of directional data is presented, however this toolbox requires the use of a commercial software environment, MATLAB, involving an economic cost and limited performance. Another example is [14] where the author describes a MATLAB script that calculates descriptive statistics and performs inference on azimuthal parameters and measures of correlations between circular variables. But again, MATLAB presents several problems: it does not provide the relevant functions for circular statistics, such as algorithms to compute the probability distribution function of a Von Mises distribution [1].

On the other hand, there are several packages of directional statistics that work over the R environment, a free software tool for statistical computing and graphics which allows the use of some functions of circular statistics. Two previous packages was proposed: 1) in [15] authors presented *VecStatGraphs2D*, a R based tool for circular statistics, and in 2) [16] *VecStatGraphs3D* was described as a R based tool for spherical statistics. Nevertheless, R language focuses exclusively on the field of mathematical analysis, thus reducing flexibility and versatility in terms of use, which is a disadvantage.

Also, we can found several Python based libraries that can

be useful in order to process directional data, such as SciPy[1], NumPy[2] or SciKit-Learn[3] libraries. However, these libraries and packages are specialized in a very specific field and they are used very exceptionally, therefore, they are not useful for generic analyses which involve the use of circular data.

In order to solve these limitations, *PyCircularStats* is proposed in this work. This open-source library, based on the *VecStatGraphs2D* package, is developed in Python and integrates the methods and operations of circular statistics, allowing a fast and flexible use. In order to introduce the operations of this package as well as demonstrating its great potential within the field of statistics, the new tool *PyCircularStat* is proposed. The remainder of this work is organized as follow: In section 2 a little review of circular data and statistics is described with the *PycircularStat* package, proposed as an improvement of *VecStatGraphs2D* for the statistical analysis of two-dimensional directional non-unit-length vectors. Finally, in section 3 we present a summary of the work and several future lines.

2. METHODOLOGY

2.1. Circular data

Directional statistics is a part of the statistical field that works with directions, axes and rotations in R^n. Specifically for this paper, we will focus on circular statistics. Circular data is taken around a circumference/circle usually in radians (from 0 to 2π) or in degrees (from 0°to 360°), with periodic nature and no true zero i.e. data is close without a beginning and end along the real line. This kind of data can be modeled with the Von Mises distribution, the most commonly used unimodal probability distribution for circular data whose probability density function (pdf) $p(x)$ is:

$$p(x) = \frac{1}{2\pi I_0(k)} \cdot e^{k\cos(x-\theta_1)} \tag{1}$$

with two parameters, the mode θ_1 that indicates the point $x = \theta_1$ where the pdf takes its maximum value, and the positive concentration parameter k that depends on the uniformity of distribution so $k = 0$ indicates that $p(x)$ is constant, i.e. density will be constant in any sector of the circle. However $k > 0$ indicates that $p(x)$ has a maximum value for $x = \theta_1$ and a minimum value for $x = \theta_1 + \pi$. On the other hand the normalizing factor $I_0(k)$ is an incomplete Bessel function[4].

In order to work with this kind of data, circular statistics collect several methods adapted for the processing and analysis of the circular data composed by angles, axes, vectors and rotations. Previous studies about *VecStatGraphs2D* package [16] shows that it does not only calculate high-quality graphics and statistical data but also create them, being a powerful tool for circular statistics in R language. However, the drawback of this package is that there are few applications which provide an Application Programming Interface (API) in R language. This work aims to describe the principal features of *PyCircularStats*, a new python package for the statistical analysis of 2-D directional non-unit-length vectors has been developed with the aims of update and improve *VecStatGraphs2D*, being more flexible and easy to use.

2.2. PyCircularStats

Implemented methods

PyCircularStats library implements both types of statistics, linear and circular, whose methods have been obtained from various sources [4, 18, 19] in order to create a library as complete and useful as possible. The linear statistics methods are used to analyze the module or magnitude of the vectors. The implemented methods provide descriptive statistics, including the number of elements, arithmetic mean, maximum and minimum values, population and sample variances and deviations, standard error, Skewness coefficient for establishing the level of asymmetry of the probability distribution and Kurtosis measure to calculate how flat or steep is the distribution.

On the other hand circular statistics methods are used to analyze azimuths and provide information such as the mean circular direction of a set of circular observations $\alpha_1, ..., \alpha_n$, $R = (\sum_{n=1}^{N} \cos\alpha_i, \sum_{n=1}^{N} \sin\alpha_i) = (C, S)$, the mean modules of the observation $\bar{X} = \frac{\sum_{n=1}^{N}\sqrt{(x_{2i}-x_{1i})^2+(x_{2i}-x_{1i})^2}}{N}$, the circular variance defined as $V = 1 - \tilde{R} = 1 - \frac{R}{N}$ where

[1]Information available on its official web site: http://docs.scipy.org/doc/scip/reference/index.html. The scipy.stats module is of particular interest, specializing in statistical functions (http://docs.scipy.or/doc/scipy/reference/tutoria/stats.html)

[2]Information available on its webpage, http://www.numpy.org/. This is a mathematical package which contains some statistical functions. It is the perfect complement for SciPy

[3]Specialized library in Machine Learning, http://scikit-learn.org/stable/

[4]Bessel functions (understood as cylinder functions of the first type or Bessel functions of the first type) can be defined, for a real number n, via the Taylor series:

$$J_n(z) = \left(\frac{z}{2}\right)^n \sum_{k=0}^{\infty} (-1)^k \frac{\left(\frac{z^2}{4}\right)^k}{k!\Gamma(k+n+1)}$$

where Γ is the Gamma-function $\Gamma(z)$ that extends the values of the factorial $z!$ to any complex number z, being J_n an analytic function. The modified Bessel function of the first kind $I_n(z)$ can be defined for a real number n as:

$$I_n(z) = \left(\frac{z}{2}\right)^n \sum_{k=0}^{\infty} \frac{\left(\frac{z^2}{4}\right)^k}{k!\Gamma(k+n+1)}$$

That could be defined by the contour integral when $\nu \in \mathbb{C}$ is a complex parameter as:

$$I_\nu(z) = \frac{1}{2\pi i} \oint e^{\frac{z}{2}\frac{t+1}{t}} t^{-\nu-1} dt$$

Finally, incomplete Bessel functions of zero order are based on the modified Bessel function and can be defined as [17]:

$$K_\nu(x, y) = \int_1^{\infty} e^{-xt-\frac{y}{t}} t^{-\nu-1} dt$$

$0 \leq \tilde{R} \leq 0$ and them $0 \leq V \leq 0$[5], circular standard deviation $s = \sqrt{-2\ln(1-V)} = \sqrt{-2\ln(\tilde{R})}$, Von Mises parameters equivalent to the normal distribution in the linear analysis $g(x;\theta_1,k) = \frac{1}{2\pi I_0(k)} e^{k\cos(x-\theta_1)}$, also the Skewness for the circular asymmetry $b = \frac{1}{N}\sum_{n=1}^{N}\sin 2(\alpha_i - \bar{\alpha})$[6], and finally the Kurtosis as a measure of the shape $k = \frac{1}{N}\sum_{n=1}^{N}\cos 2(\alpha_i - \bar{\alpha})$[7].

Also the proposed *PyCircularStats* package includes methods for the uniformity analysis of circular data, useful for knowing if the sample of the data is distributed uniformly around the circle or, otherwise, it has some direction. The uniformity analysis of the data distribution is performed using Rayleigh [4] ($z = nr^2$ where n is the size of the problem and r the mean of the angles) and Rao[8] [21] ($U = \frac{1}{2}\sum_{n=1}^{N}|d_i - \lambda|$ with $d_i = \alpha_{i+1} - \alpha_i$, $d_N = 360° - (\alpha_N - \alpha_1)$ and $\lambda = \frac{360°}{N}$) tests.

Input Data Format

Input files must be plain text files with two or four columns separated by tabs. As in [15], the format of the input data in two dimensions can be of three types:

1. Cartesian coordinates: Four columns, the first and second columns define the (x_o, y_o) origin and the last two columns the (x_d, y_d) destination, i.e. origin node and end node respectively.

2. Polar coordinates: Two columns, where the first is the module and the second the azimuth.

3. Incremental data: Two columns, the first is the increase of $x(\Delta x)$ and the second the increase of $y(\Delta y)$.

By default, the angles are in sexagesimal system and the geographical reference system is 0°/North, increasing clockwise.

Graphical analysis

The proposed *PyCircularStats* package includes features to plot not only vectors with their corresponding modules but also to plot statistics. The library can plot the following graphics: 1) *Q-Q plot*, a Quantile-Quantile plot which determines how close is the distribution of a set of data with an ideal distribution or compare the distribution of two set of data; 2) azimuth data distribution, plotting each azimuth as a stacked point within a unit circle, with the mean azimuth and confidence interval; 3) circular histograms over a polar-area diagram, where sectors have the same angular width and the radius is proportional to the relative frequency of azimuths in the sector; 4) points distribution plotting a polar diagram which represents the information through points; 5) vector distribution in polar coordinates; 6) spatial vector distribution with the corresponding module and azimuth, and finally 7) density maps that plots the vector density, represented by a color scale.

The graphics generated by this library are perfectly exportable and these can be edited by third parties software as for example GIMP (see Figure 1).

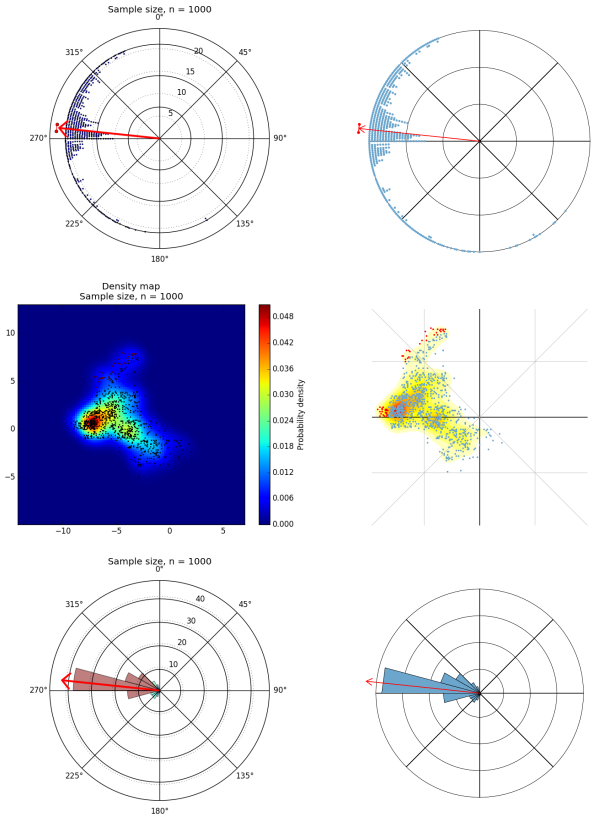

Fig. 1: Visualization of azimuths (first row), density point maps (second row) and azimuth histograms (third row) in *PycircularStat* (left column) and *VecStatGraphs2D* (right column).

3. CONCLUSIONS

A Python library has been developed to perform statistical analysis of circular data. This library includes linear analysis, as well as angular descriptive statistics, in addition it includes

[5]We must take into account that $\tilde{R} = \sqrt{(\tilde{C}^2 + \tilde{S}^2)}, 0 \leq \tilde{R} \leq 0$ where $\tilde{C} = \frac{1}{N}\sum_{n=1}^{N}(\cos\alpha_i), \tilde{S} = \frac{1}{N}\sum_{n=1}^{N}(\sin\alpha_i)$

[6]A value close to 0 shows symmetry around the mean direction. In addition, a standard measurement of the mean is defined as [4]:

$$b_0 = \frac{R_2\sin(\bar{\alpha}_2 - 2\bar{\alpha})}{(1-R)^{\frac{2}{3}}}$$

[7]Alternatively, a standard mean of kurtosis is defined as follows [4]:

$$k_0 = \frac{R_2\cos(\alpha_2' - 2\bar{\alpha}) - R^4}{(1-R)^2}$$

[8]The U-distribution calculated by Rao is computationally very complex. Thus, normally tables are used instead of the complete distribution [20].

some inferential tests for circular distributions. Another feature of this library is the ability to generate graphs based on these methods. These graphics are easily editable from other software. On the other hand, a tool has been created, therefore the circular statistics is carried out through data files from remote sensors as satellites. With this work not only a library of circular statistics that generates graphics has been managed to provide to the community, in addition an API has been provided to be integrated into any application that has a compatible API, so it can be used and maintained in a simple way.

Acknowledgements

This publication has been supported by the Consejería de Economía, Ciencia y Agenda Digital of the Junta de Extremadura, the European Regional Development Fund (ERDF) of the European Union through grant reference GR21040 and the Programa Estatal de I+D+i Orientada a los Retos de la Sociedad, Ministerio de Economía y Competitividad (PID2019-105221RB-C42) AEI /10.13039/501100011033.

4. REFERENCES

[1] Martin H. Trauth. *Statistics on Directional Data*, pages 263–277. Springer, Berlin, Heidelberg, 2007.

[2] Peter E. Jupp. *Spherical Statistics*, chapter . American Cancer Society, 2006.

[3] Ronald Aylmer Fisher. On a distribution yielding the error functions of several well-known statistics. *Proceedings of the International Congress of Mathematics*, 2:805–813, 1924.

[4] N.I. Fisher. *Statistical Analysis of Circular Data*. Statistical Analysis of Circular Data. Cambridge University Press, 1995.

[5] M. S. Bingham and K. V. Mardia. Maximum likelihood characterization of the von mises distribution. In Ganapati P. Patil, Samuel Kotz, and J. K. Ord, editors, *A Modern Course on Statistical Distributions in Scientific Work*, pages 387–398, Dordrecht, 1975. Springer.

[6] Graham J. Borradaile. *Statistics of Earth Science Data: Their Distribution in Time, Space and Orientation*. Springer-Verlag Berlin Heidelberg, 2003.

[7] John C. Davis. *Statistics and Data Analysis in Geology*. John Wiley & Sons, Inc., New York, NY, USA, 1990.

[8] E. Batschelet. *Statistical methods for the analysis of problems in animal orientation and certain biological rhythms*. A.I.B.S. monograph. American Institute of Biological Sciences, 1965.

[9] J.A. Bowers, I.D. Morton, and G.I. Mould. Directional statistics of the wind and waves. *Applied Ocean Research*, 22(1):13 – 30, 2000.

[10] Y. Chikuse. *Statistics on Special Manifolds*. Lecture Notes in Statistics. Springer New York, 2012.

[11] N. Nikolaidis and I. Pitas. Nonlinear processing and analysis of angular signals. *IEEE Transactions on Signal Processing*, 46(12):3181–3194, Dec 1998.

[12] Suvrit Sra. Directional Statistics in Machine Learning: a Brief Review. ., 2016.

[13] Philipp Berens. CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, 31(10):1–21, 2009.

[14] Thomas A. Jones. Matlab functions to analyze directional (azimuthal) data-i: Single-sample inference. *Comput. Geosci.*, 32(2):166–175, 2006.

[15] P. G. Rodriguez, M. E. Polo, A. Cuartero, A. M. Felicisimo, 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*, 11(5):921–925, May 2014.

[16] A. Cuartero, M. E. Polo, P. G. Rodríguez, Á. M. Felicísimo, and J. C. Ruiz-Cuetos. The use of spherical statistics to analyze digital elevation models: An example from lidar and aster gdem. *IEEE Geoscience and Remote Sensing Letters*, 11(7):1200–1204, July 2014.

[17] Frank E. Harris. Incomplete bessel, generalized incomplete gamma, or leaky aquifer functions. *Journal of Computational and Applied Mathematics*, 215(1):260 – 269, 2008.

[18] S. R. Jammalamadaka, A. Sengupta, and A. Jammalamadaka, S. R. Sengupta. *Topics in circular statistics (Vol. 5)*. World Scientific, 2001.

[19] K. V. Mardia and P. E. Jupp. *Directional statistics (Vol. 494)*. John Wiley & Sons, 2009.

[20] Gerald S. Russell and Daniel J. Levitin. An expanded table of probability values for rao's spacing test. *Communications in Statistics - Simulation and Computation*, 24(4):879–888, 1995.

[21] E. Batschelet. *Circular Statistics in Biology*. Mathematics in biology. Academic Press, 1981.