Proc. Indian Acad. Sci. (Math. Sci.) (2020) 130:44
https://doi.org/10.1007/s12044-020-0583-8

A matter of direction

S RAO JAMMALAMADAKA

Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA
E-mail: rao@pstat.ucsb.edu

Abstract. This paper summarizes a talk given by the author at the Indian Academy of Sciences, Bengaluru on December 12, 2019. It outlines the multiplicity of situations where measurements on directions in two, three, or more dimensions are observed, and form the basis for answering various scientific questions. After briefly outlining the novelty in dealing with such data and the need for an entirely different set of analytical tools, one of the basic questions that arises before any further inference viz. testing isotropy of circular data, is addressed. The author was initiated into this topic of directional statistics by Professor C R Rao during the 1960s and is pertinent to this occasion celebrating him.

Keywords. Directional data; examples; novel techniques; testing isotropy.

Mathematics Subject Classification. 11N37, 11A25, 11K65.

1. Introduction to directional data

In many scientific disciplines, researchers collect data that comes in the form of directions which may be represented as unit vectors, in either the plane (for 2-dimensional data), the sphere (for 3-dimensional data), or the hypersphere for dimension greater than three. Directional statistics is the study and development of statistical theory and methodology used to analyze and draw inference from such data. It may be that an ornithologist is interested in flight directions of certain species of bird as it leaves a particular area, or it may be a geologist who is researching the movement of the Earth's magnetic pole. All such investigations generate data that can be considered directional, and we need appropriate tools to extract any real meaning, as well as to quantify the uncertainty of both the observations and the conclusions.

Unlike much of the linear analogues, directional data requires special treatment due to the unique properties and features. As an example, the 2-dimensional observation on the unit circle can be represented as a unit vector, or simply as an angle on $[0, 2\pi)$, but neither representation is necessarily unique, as both depend on the choice of some appropriate *zero* direction from which to measure, as well as the sense of rotation, i.e., whether one measures going clockwise or anticlockwise as positive. Similarly, points on the unit sphere (or hypersphere) can be described in terms of two (or more) angles or unit vectors in appropriate dimensions, and are equally dependent on the choice of the zero-direction and

This article is part of the "Special Issue in Honour of Professor C R Rao on His Birth Centenary"

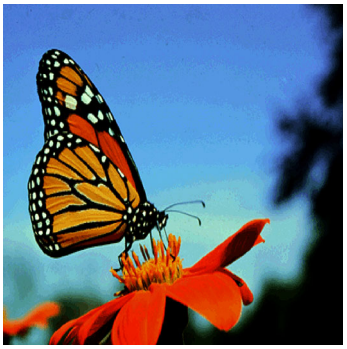
Published online: 25 July 2020

sense of rotation. The reader is referred to books by Jammalamadaka and SenGupta [4] or Mardia and Jupp [6] among others.

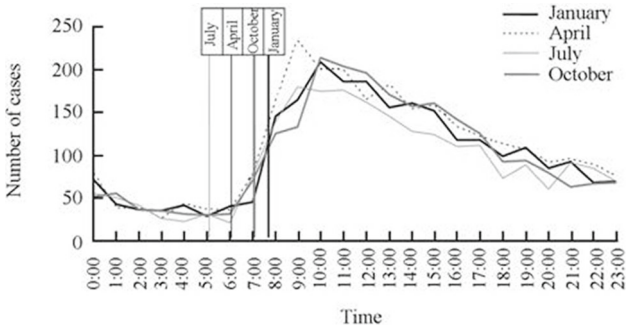
2. Some scientific questions

Consider the following scientific questions, all of which look quite unrelated!

- Did a river change its course over geological time?
- Is the ozone level in Santa Barbara affected by the pollution levels in a major city nearby?
- Migration patterns of animals, as for instance, Monarch butterflies (pictured below) that pass through Santa Barbara?



- Is there scientific evidence to support the hypothesis of continental drift?
- Is the earth's magnetic pole reversal a fact or fiction? And if true, how frequently?
- Is there a pattern to the epicenters of earthquakes, say along the fault lines?
- Is there justification to the sociological theory that more people die AFTER a major event in one's life, than BEFORE the event when they are looking forward to it?
- If a woman has to undergo certain surgery, what part of her menstrual cycle is the better time?
- Is there a particular part of the day when more heart-attacks occur (see the graph below)?



- Is a company cheating on its taxes?
- Is a student copying an essay from another source?

- How far do mosquitoes (pictured below if you have not come across one!) travel from their breeding ground, say a pool of water?

- *Glaucoma*: Ophthalmologists often measure the "intraocular pressure" in the eye as an indicator of Glaucoma and diagnose. Recently, the author has been collaborating on a project at the University of Pittsburgh, on a study involving Glaucoma.
- More sophisticated measurements involve the OCT (Optical Coherence Tomography) data which gives measurements of thickness of NRR (Neuro Retinal Rim) and the RNFL (Retinal Neuro Fibre Layer) *around* the eye ball. Can we use these circular curves as predictors of Glaucoma?

A common theme to *all* these questions, is that the empirical/statistical evidence for answering them, comes in the form of data, which is a set of "directions" – hence the title of this presentation, "A Matter of Direction!"

3. A general introduction

In many physical and natural sciences like geology, biology, ecology, the basic measurements are quite often, directions. In practice, these directions are in 2-dimensions or in 3-dimensions, but could be in higher dimensions as in some of the questions raised in the earlier section.

A direction in 2-*dimensions* (Figure 1) can be represented as

- as an angle $0° \leq \alpha < 360°$,
- or as a point on the circumference of a circle, whose radius can be taken to be 1 for convenience,
- or as a vector of unit length $(x = \cos\theta,\ y = \sin\theta)$ or even as a complex number $z = e^{i\theta}$ of unit modulus.

Such a set of observations are referred to as "*circular data*."
We are all familiar with statements like "the wind is blowing in a north-easterly direction". But if one wants to be precise, one may assign a number from 0° to 360°, to such a direction.

Similarly, a 3-*dimensional* direction (Figure 2) can be represented by 2 angles

commonly called the 'longitude (W–E)' and 'latitude (S–N)', and referred to as "spherical data".

A 3-*dimensional* direction can also be represented in several equivalent ways, e.g.,

- in terms of 2 angles, $(-\pi \leq \phi < \pi,\ -\pi/2 \leq \theta < \pi/2)$, called the 'longitude (W–E)' and 'latitude (S–N)' respectively

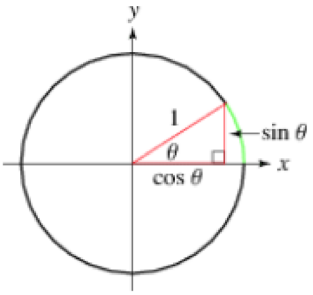


Figure 1. Direction in 2-dimensions.

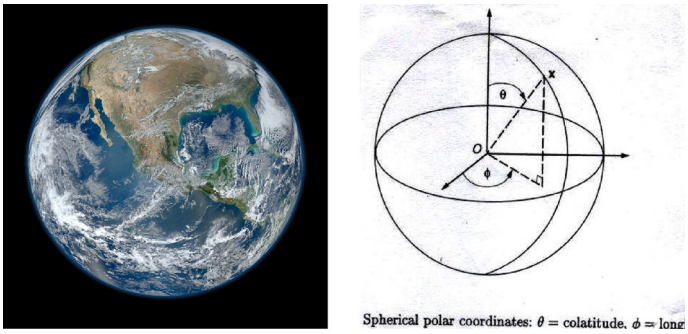


Figure 2. Direction in 3-dimensions.

- as a unit vector – with rectangular coordinates,

$$\mathbf{x} = (x = \cos\theta, \quad y = \sin\theta \cdot \cos\phi, \quad z = \sin\theta \cdot \sin\phi).$$

Directional observations in 3-dimensions are referred to as "spherical data" (Figure 3).

4. Directional measurements: what makes them different?

For observations on the real line (like heights, weights etc.), or when dealing with multivariate data, there is a natural zero and an ordering.

But in dealing with points on the circumference of a circle, both the starting point and which way we order things, are both arbitrary!!

In *linear* statistics, typically one uses the familiar sample (arithmetic) mean $\bar{X}$ and variance S^2 as measures of location and variation, as well as other measures for assessing other properties.

However, these "standard measures" are of no use when dealing with circular data – as the examples in Figures 3 and 4 illustrate.

- Thus *none* of the usual measures like the sample mean, the sample variance, moments, mgf, t- and F-tests, etc. are appropriate for directional data, because they all are highly dependent on how one assigns the values to the given set of directions!
- We need an entirely new set of tools – descriptive measures (corresponding to $\bar{X}$, S^2 etc.), models, sampling distributions, inference – both parametric and nonparametric.
- Of course, this is not meant to be a tutorial on how to do this, but just to bring awareness to this issue, namely that when one measures directions, one needs to think differently.

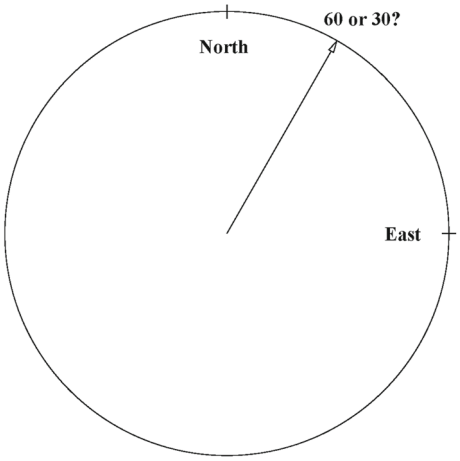


Figure 3. Value depends on choice of origin, North or East and the sense of rotation.

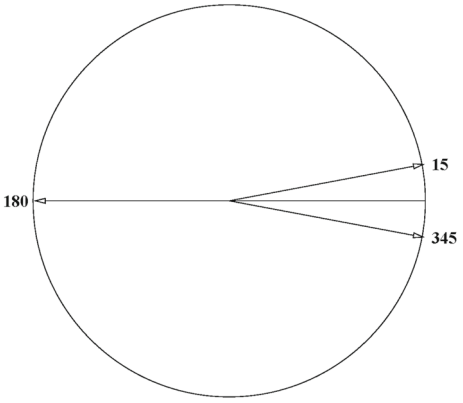
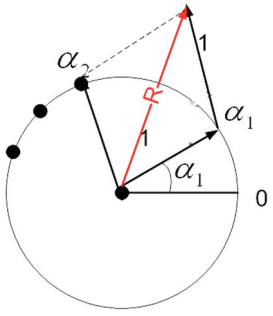


Figure 4. The arithmetic mean of 15 and 345 degrees, points the wrong way!

4.1 *Alternate measures and models*

A good way to treat the angular data is as unit vectors pointing in the specific directions, and obtain the "resultant vector" of such vectors



Then a "circular mean" is defined as the "direction" that the resultant vector points to. It depends only on the data, and is immune to what we call the zero or, the sense of rotation, i.e., has the properties one wants in a location measure.

In addition, the length of the resultant vector, $0 \leq R \leq n$ serves as a measure of *concentration* of the data (and is used in place of the more familiar, S^2).

5. Back to the scientific questions and how they relate to directions

We now revisit the various scientific questions we raised and show how they relate to directional data. On each of these topics, there is considerable scientific literature and even monographs, but we will refer the interested reader to the two books cited earlier, occasionally giving one or two citations.

5.1 *Directional data in 2-dimensions or circular data*

- Geologists actually measure the direction of flow of a river from sedimentary rocks, whose age they can determine —- called the "paleo-current analysis" (see [7]), a collaboration between the geologist Sengupta of the Geological Studies Unit at the ISI and the author. This is how the field of directional data got started in India, at the suggestion and encouragement of Professor C R Rao.
- Similarly, scientists who work in Ornithology (the "bird-scientists") record the direction of flight of birds to see where they are heading (see [1]).

Such measurements involve 2-dimensional directions – thus, it is "A Matter of Direction!"

5.2 *Directional data on a sphere*

On the other hand, when discussing issues like the Earth's magnetic pole etc., we are dealing with 3-dimensional directions.

- Earth's N–S magnetic pole is completely reversed, once in a while! Not overnight while you are sleeping, but over a period of 1,000 years! It takes place once every 450,000 years, on an average. We may be due for one, since the last pole-reversal took place over 780,000 years ago!! (see [2]).
- Matching the paleo-magnetic directions in different continents provides further evidence of continental drift (besides matching paleo-botany).

It is again "A Matter of Direction!"

5.3 *Periodic phenomena*

Any periodic phenomenon with a known period, such as a day, a month, a year, etc. can be represented on the circumference of a circle.

- Diurnal variations: time of heart attacks, time for taking a medicine.
- Timing a surgery say for a breast cancer patient, — chronobiology and chrono-therapy (see [3]).
- Birth-death cycle.

It is again "A Matter of Direction!"

5.4 *"First significant digit phenomena" and cheating on taxes*

When one looks at data that has a large variation (small numbers, mixed with large, and very large numbers), and pays attention to the "First Significant Digit (FSD)", i.e., the first non-zero digit, this can be 1, 2, …, 9.

- This is related to a fact on directional distributions which says that when a random variable on the real line which has a large variation, is wrapped around a circle of unit circumference, it has a circular uniform distribution on [0, 1).
- As a result, the relative frequency of these FSDs is quite non-uniform, with 1 *occurring nearly* 6 *or* 7 *times more often than* 9! The chance of seeing the first 3 numbers viz. 1, 2, and 3 is almost 60%!! (called the Benford's law). A quick argument in support is as follows:
  FSD of a value X, say $\mathrm{FSD}(X) = 3$ for instance, when
  $$3 \times 10^k \leq X < 4 \times 10^k \quad \text{for some integer} k$$
  or
  $$\log 3 + k \leq \log(X) < \log 4 + k \quad \text{for some integer } k$$
  or
  $$\log 3 \leq \log(X) \bmod(1) < \log 4$$
  which is approximately $(\log 4 - \log 3) = \log(4/3)$ because of the circular uniform distribution mentioned above. More generally,
  $$\mathrm{Prob}(\mathrm{FSD}(X) = k) = \log\left(\frac{k+1}{k}\right), \quad k = 1, \ldots, 9.$$
- This is one check that can be used to verify if a set of figures submitted by an individual or corporation, are genuine!, or if someone is cooking the books with regard to, say their taxes!!

5.5 *Plagiarism and latent semantic analysis*

One way to summarize a student essay, or any given text (like Shakespeare's drama) is to look at the frequency of different words used in the text, which has been called "latent semantic analysis." Such a vector of frequencies can be "normalized" to become a vector of unit length, which can then be treated as a high-dimensional direction.

- Any 2 texts become 2 different points on the surface of such a hyper-sphere, and one can look at how far apart they are, to judge the similarity of these texts.
- Two texts are compared by taking the cosine of the angle between the directional values corresponding to these texts.

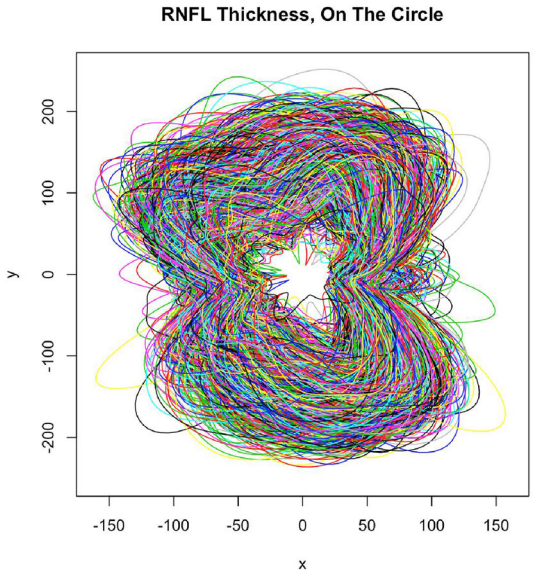
It is again "A Matter of Directions!"

5.6 *Curves around a circle and Glaucoma*

At the L.V. Prasad Eye Institute in Hyderabad, they have collected OCT data for many patients that walk in, both for normal eyes and for those with Glaucoma. Data consists of close to 5,000 circular curves on NRR thickness and RNFL thickness, besides various covariates like gender, age, etc.

(1) First we group these curves to detect any association with the disease or its progression.
(2) This "functional clustering" was done in a couple of ways: (i) by fitting a Fourier series of an appropriate degree and using the coefficients, and (ii) treating each curve as a mixture of von Mises distributions and looking at the parameter-set for each curve.

Just to give an idea, this is what the jumble looks like for the RNFL thickness (each color represents a different cluster).



It is again "A Matter of Directional Measurements!"

6. Testing isotropy as an essential first step, and a random walk problem

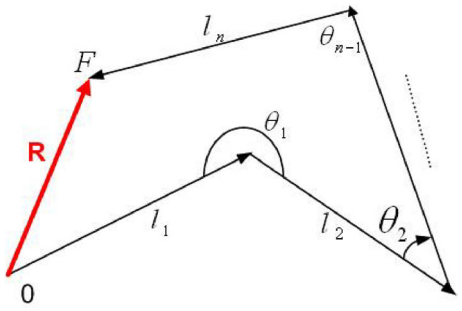
When one measures directions, say the flights of birds, and comes to a statistician to figure out what their average direction is: the statistician has to pause, and ask the all-important question, "is there really a preferred direction to these birds, or are they merely flying in *all* possible directions?" (what we might call a case of "isotropy"!)

- For instance, when we see mosquitoes buzzing around us, not sure if they have any preferred direction!
- Or when particles (or planets at the time of "Big Bang") collide and bounce off in all possible directions.

The big question is whether there is any consistent pattern to the observed directions? Consider the related problem of what has been called a "Pearson's Random Walk".

(1) A ("fully") drunk starts at the bar or origin say "0″, heads off in a random direction and walks l_1 units of length (l_1 steps, or l_1 feet).
(2) Then, falters (falls down) and heads in an arbitrary new direction, i.e., takes a turn/twist through a random angle θ_1 at that point.

(3) After walking l_2 more units of length, s/he takes another random turn through angle θ_2. This is repeated in n stretches of lengths $l_1, l_2, \ldots, l_n$ with random (isotropic or uniform) changes in direction, of $\theta_1, \ldots, \theta_{n-1}$ in between these stretches.



If such a random walk ends up at the point F after these n stretches, how far is s/he from the starting point, 0? That is, what can one say about the distribution of $R =$ the distance between 0 to F?.

Some related history. Sir Ronald Ross, who in 1902, won a Nobel Prize in medicine for his work on malaria vector, wanted to know how far the mosquitoes would travel from their breeding ground – say a pool of water, and how their density would decrease as the distance from their breeding ground increased.

- Mosquitoes buzzing off in random directions – analogous to drunks leaving the bar,
- 2-dimensional projection of their flying paths provide reasonably good approximation.

He posed this question in 1904 to Karl Pearson, the pre-eminent statistician of that time. Pearson, after having tried his hand at it for a while, posed this problem, in *Nature* (July 1905). It came to be known as the "Pearson's Random Walk problem" even though he neither came up with the question nor the answer!!

Related problems. This topic has many other interesting and amusing connections, like

- in chemistry, this is related to the length a polymer chain from end to end (like your key-chain), when it has n links of lengths $l_1, l_2, \ldots, l_n$, and
- in physics, one may ask: *Are* 2 *violins twice as loud as one*? Sound waves can be represented as vectors with an amplitude and phase, and their phase-difference determines if they reinforce or weaken each other (we all know of noise-canceling headphones!).

6.1 *Some answers*

The answers – especially the *exact* ones, are not as elegant as the problem itself, and the density function of R is given in terms of the Bessel functions, as follows (see Section 3.2.1 of Jammalamadaka and SenGupta [4] for details):

$$\int_0^\infty rt J_0(rt) \prod_{i=1}^n J_0(l_i t) \mathrm{d}t \quad \text{for } 0 \le r \le \sum_{i=1}^n l_i .$$

After some calculations, for the simple case where each step is of unit length, one can figure out how far the mosquitoes will be. Table 1 gives some critical values.

Table 1. Some critical values.

n	95th percentile of R
20	7.7
50	12.2
100	17

With $n = 100$ steps of unit length, say a foot each, if they are flying in random directions at each step, 95% of the mosquitoes will be within 17 feet from where they started!

$$E(R) \approx 9 \quad \text{and} \quad P(R > 17) = 0.05.$$

This could be a "a statistical test for drunk driving": Release him/her if, in 100 steps of 1 foot each, they can walk >17 feet away from where they started!!

Acknowledgements

The author like to thank Professors Partha Majumdar and B L S Prakasa Rao for the invitation to give this talk and to present the material in the form of this short article.

References

[1] Batschelet E. Circular Statistics in Biology (1981) (London: Academic Press)
[2] Fuller M, Laj C and Herrero-Bervera E, The reversal of earth's magnetic field, *Amer. Sci.* **84** (1996) 552–561
[3] Hrushesky W J M, Circdian Cancer Therapy (1994) (CRC Press, Cambridge)
[4] Jammalamadaka S Rao and SenGupta A, Topics in Circular Statistics (2001) (World Scientific Press, Singapore)
[5] Jammalamadaka S Rao, An R-package called *CircStats* for doing much of this type of statistical analysis and inference, that was developed at UCSB, is freely available online (http://jammalam.faculty.pstat.ucsb.edu/html/books/circstat.htm) (2001)
[6] Mardia K V and Jupp P E, Directional Statistics (2000) (John Wiley, New York)
[7] Sengupta S and Rao J S, Statistical analysis of cross-bedding azimuths from the Kamthi formation around Bheemaram, Pranhita: Godavari valley, *Sankhya B* **28** (1966) 165–174