CHAPTER 1

INTRODUCTION

1.1 Introduction to the Circular Random Field and Circular Random Variables

This dissertation addresses related and practical aspects of the circular random field (CRF) including extracting the spatial correlation, modeling the spatial correlation, estimation, simulation, and plotting.

A random field (RF) is a stochastic process operating over a space of dimension ≥ 1. A CRF is defined as a RF containing circular random variables (CRVs) at multiple observation locations which are spatially correlated. With Θ the CRV and $\boldsymbol{x}$ the location, in 2-dimensional space, the CRF is the set $\{\Theta(\boldsymbol{x}), \boldsymbol{x} \in \mathrm{R}^2\}$. Circular-spatial correlation is defined here as the mean cosine of the angle between random components of directions (nonrandom component removed) vs. distance between observation locations. Spatial correlation increases as distance between observation locations decreases. Hence, random components of direction tend to be more similar as distance between observation locations decreases.

A CRV takes random directions with the total probability of all possible directions distributed on the circular support (unit circle, $[0,2\pi)$, or $[-\pi,\pi)$). The starting point of the support is the same direction as the ending point. A CRV or direction is expressed as either a scalar in units of radians or degrees (º), or as a unit vector (Chapter 4). Since trigonometric functions require angles in radian units, the input for functions of direction will be expressed in radian units with values in $[0,2\pi)$ until Chapter 5, where a new method requires values in the equivalent support of $[-\pi,+\pi)$ radians. Maps and compasses will use º units, which may be obtained from radian units by multiplying by

$180º/\pi$. On a circle, the 0s of the $[0,2\pi)$ radians, the $[-\pi,+\pi)$ radians, and the $[0,360)º$ scales are aligned. 0 radians, 0º, and the east direction will be aligned to 3 o'clock. 90º, $\pi/2$ radians, and the north direction will be aligned to 12 o'clock. π radians, 180º, and the west direction will be aligned to 9 o'clock. These scales of direction or angle are shown in Figure 1-1. Figure 1-1 is a typical plot of the probability density function (PDF) of the triangular CRV (density increases linearly toward the maximum density at 0).

Other types of directional random variables include the spherical, axial, and vectorial random variables. A spherical random variable takes random locations on a unit sphere. An axial random variable takes random axis orientations in a plane where there is no reason to distinguish a direction from its opposite. A vectorial random variable has both random direction and random magnitude. Hence, random fields may also be defined for axial, vectorial, and spherical random variables.

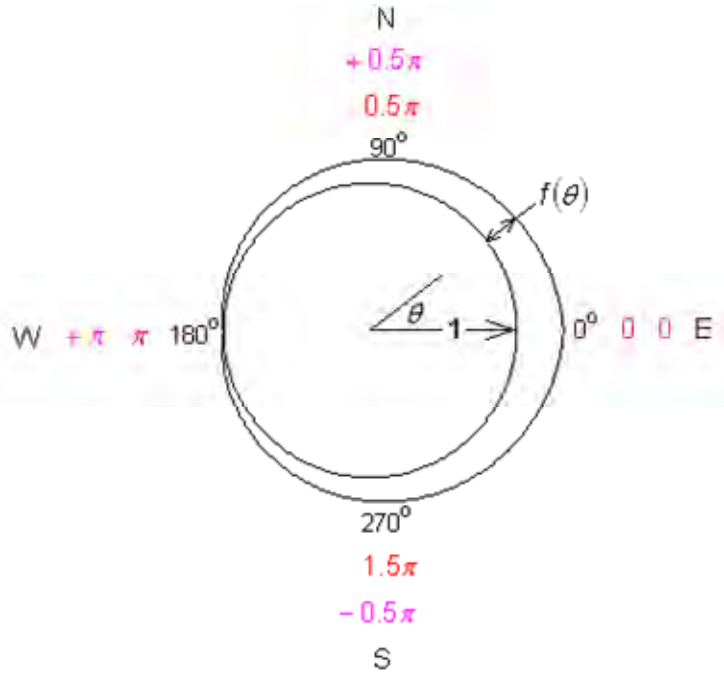


Figure 1-1. Circular PDF of the Triangular Circular Probability Distribution. The density $f(\theta)$ is often plotted on the outside of a unit circle.

Applications of circular random variables and circular random fields include:

- Astronomy - Planet orbit inclination
- Biology - Creature migration and navigation by sun, wind, magnetic fields, etc
- Chronotherapeutics - Response to a treatment relative to the time of treatment
- Geography - Compass directions
- Geology - Crystal and fault orientation
- Geophysics - Magnetic field direction
- Meteorology - Wind direction
- Oceanography - Ocean currents
- Periodic phenomena - Births/month, deaths/month, eggs produced/month, coats sold/month, accidents per hour, accidents per month, sunspots/year, biorhythms
- Paleomagnetism – Direction of magnetism locked into lava
- Physics – Dihedral (having or formed by two planes) angles in molecules
- Rounding errors – Integer atomic weights
- Structural Geology - Fracture pattern in a region

This dissertation will treat the cardioid, triangular, uniform, von Mises, and wrapped Cauchy circular distributions in alphabetical order in all sections:

- The cardioid distribution models the direction marbles roll off when dropped on a plane inclined to the horizontal. A 0º inclination of the plane produces the circular uniform PDF.
- The triangular distribution has a PDF that increases linearly toward mean direction.
- The uniform distribution models an honest roulette wheel and provides the null model to test the alternatives of unimodal (a single cluster of directions in the data) and multimodal distributions (two or more clusters in the data).

- The von Mises distribution is practically interchangeable with the wrapped normal circular probability distribution. The "wrapped normal" distribution is obtained by "wrapping" the tails of the normal PDF around a unit circle in opposite directions. The probability density at an angle increases with each revolution of the tails by the densities of the PDF that overlap the angle. Originally, the von Mises distribution modeled experimental errors arising from determination of atomic weights. Other applications of the von Mises distribution now include the direction of the sum of unit vectors representing observations of direction or periodic phenomena. The wrapped normal distribution dominates geology and models Brownian motion on the circle. However, inference is easier with the von Mises distribution.
- The wrapped Cauchy distribution is obtained by "wrapping the tails" of the Cauchy distribution on a circle in opposing directions. The Cauchy distribution is used to indirectly simulate the von Mises distribution.

In this dissertation, an observation is a measurement of direction or a realization of a circular random variable, expressed as a unit vector or as an angle, with an angle from 0 to 360º, from 0 to 2π, or from $-\pi$ to π (see Figure 1-1 for details). The main circular statistics are based on computing with direction in unit vector format. A sample consisting of observations of direction as unit vectors is summarized as the resultant vector. The vector resultant is the sum of the unit vectors representing directions. Unit vectors are summed by attaching the tail of one vector to the head of another. The main circular statistics include the resultant vector mean direction and the resultant vector mean length.

The resultant vector mean direction, $\overline{\theta}_n$, which is the direction of the resultant vector, is the measure of central direction. Why is it necessary to use vectors to

determine the average direction? Figure 1-2 shows that the average or central direction of 15º and 345º is not the arithmetic mean = 180º as on a linear scale.

In Figure 1-2, the sum of these directions is the vector

$$(\cos(15°), \sin(15°)) + (\cos(345°), \sin(345°)) =$$
$$(\cos(15°), \sin(15°)) + (\cos(15°), -\sin(15°)) =$$
$$(2\cos(15°), 0)$$

which has a direction of 0º. Hence, the average direction is 0º. The extensive use of trigonometry distinguishes circular statistics from the statistics of linear random variables.

The other main circular statistic is the resultant vector mean length $\overline{R}_n$. With n the number of observations of direction, the resultant vector mean length $\overline{R}_n$ of n observations of direction is 1/n times the vector resultant length R_n. It is a measure of concentration about the mean direction, where the sense of concentration is the opposite the sense of variability (a measure of spread). When variability increases, concentration decreases and vice versa. If all n observations have the same direction, the variability is zero, the resultant vector length $R_n = n$ (the unit vector observations of direction added tail to head are aligned and n long), and the resultant vector mean length $\overline{R}_n = \frac{1}{n} R_n = 1$, which is the theoretical maximum. When direction takes random values, the variability is greater than 0, $R_n < n$, and $\overline{R}_n = \frac{1}{n} R_n < 1$. If n is even, and the angles between all pairs of adjacent observations of direction are equal, the variability (spread) is the theoretical maximum, the horizontal and vertical components of the unit vectors cancel, $R_n = 0$, $\overline{R}_n = \frac{1}{n} R_n = 0$, and the resultant vector mean direction $\overline{\theta}_n$ is undefined.

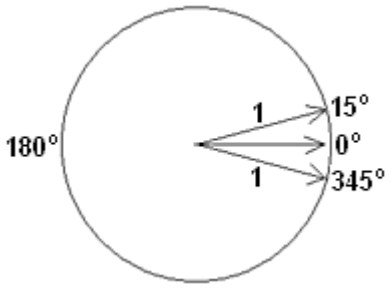


Figure 1-2. The Arithmetic Mean of 180º Does Not Point in the Central Direction of 0º.

The parameter of the circular distribution corresponding to the resultant vector mean length of the sample is the population resultant vector mean length ρ. The effects of ρ on the sample observations of direction and the sample mean resultant vector (sample resultant vector scaled by 1/n) are illustrated in Figure 1-3. The sample observations are indicated by tan arrows and the sample mean resultant vectors by black arrows. Circles with unit radius are over plotted in black to indicate a distance of 1. Going left-to-right in Figure 1-3, the population resultant vector mean length ρ increases, concentration about the mean direction increases, and the length of the mean resultant vector of the sample tends to increase. In the right hand plot with ρ =0.99, the length of the sample mean resultant vector gets close to 1, but is not exactly 1 as can be seen in the zoom view on the right.

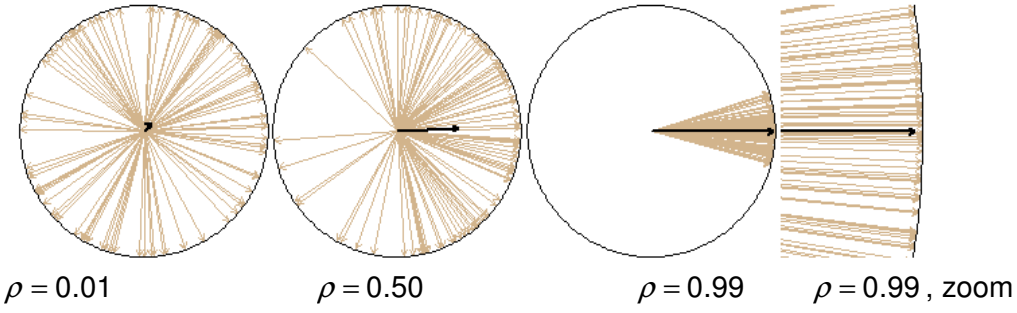


Figure 1-3. The Effect of the Population Resultant Vector Mean Length ρ on the Sample Mean Resultant Vector (Black) of a Sample (Tan) from the von Mises Circular Distribution. As ρ goes toward 1, concentration about the mean direction increases and the length of the mean resultant vector goes toward 1.

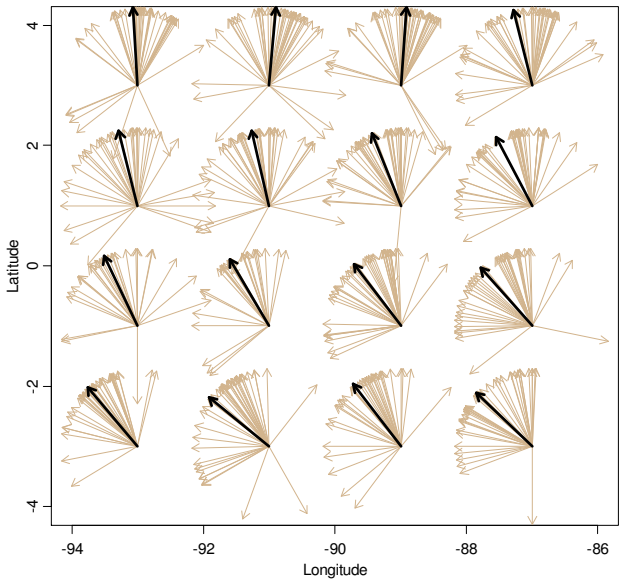
1.2 A Motivational Example

Figure 1-4 shows traditional arrow plots of ocean wind data as circular-spatial data (top), which is the focus of the dissertation, and vector-spatial data (bottom). The data are plotted as tan colored arrows and the means as black arrows. The data were freely extracted from the Comprehensive Ocean Atmosphere Data Set (Chapter 2, Subsection 2.2.1) at http://iridl.ldeo.columbia.edu/SOURCES/.COADS/.mean/ for 1980 to 1990, December to March, and for the area of latitude -3° N to +3° N and longitude -93° E to -87° E. Note that -3° N means 3° south of the equator, and -93° E means 93° west of the Greenwich prime meridian. The data contain 1934 observations of month, year, longitude, latitude, and east and north components of wind velocity. In the vector-spatial plot, the mean resultant vectors were computed from the average horizontal and vertical velocity components by location. The circular-spatial data were obtained from the vector-spatial data by scaling the vector observations to unit length loosing the magnitude information. In the circular-spatial plot, the mean resultant vectors of the circular-spatial data were computed by location, and scaled to unit length. The difference between differently computed means is 9.96° at -87° E and +3° N. Average wind direction is changing smoothly in the south-north direction, rotating clockwise as latitude increases and evidencing a global trend.

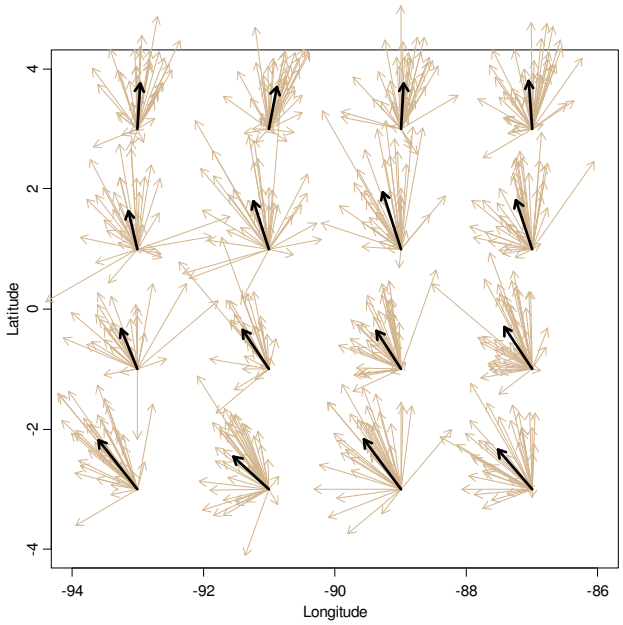
1.3 Problem Description

The problems addressed in this dissertation include:

- How may circular-spatial data be efficiently interpolated based on spatial correlation? Jammalamadaka and SenGupta summarized many expressions of nonspatial circular correlation (2001, Chapter 8). How could spatial correlation be extracted from circular-spatial data and modeled to be useful for the interpolation



Circular-spatial data and mean directions as unit vectors



Vector-spatial data and mean resultant vectors

Figure 1-4. Circular and Vector Spatial Data and Their Means for the Direction the Ocean Wind Blows Toward. Data from 1980 to 1990, December to March, in the Area of Latitude -3° N to +3° N and Longitude -93° E to -87° E. At each sampling location, the raw data are indicated by tan arrows and the means by black arrows.

of direction between observation locations?

- How can a CRF be simulated based on input specifications of a circular probability distribution and spatial properties such as the distance at which CRVs are no longer correlated, and what are the properties of the simulated CRF?
- The intelligibility of arrow plots (Figure 1-4) decreases as data density and random variation increase. Intelligibility is also affected by missing values. How can circular-spatial data, interpolations of circular-spatial data, or simulations of a CRF be plotted as a heat map that is not color discontinuous at any direction, e.g., where the color encoding direction at 0º and 360º are the same? How can these data be plotted with high data density such that both large scale and small scale directional structure can be easily recognized?

1.4 Literature Review

1.4.1 Brief History of Circular Statistics

Circular statistics, the statistics of direction, is a relatively new statistical domain as indicated by some history extracted from Fisher (1993, chap. 1). Circular-spatial statistics is very new. In 1767, John Mitchell, FRS (Fellow of the Royal Society), tested the hypothesis that the distribution of angular separations of stars is uniform. He determined that the number of close stars were too many to support this hypothesis. In 1802, John Playfair noted that directional data should be analyzed differently from linear data, recommending that average direction be the direction of the resultant vector. In 1858, Florence Nightingale, chief nurse in the British Army during the Crimean War, created the rose diagram (for example, see Figure 1-5, a rose plot of ocean wind direction) displaying the effect of sanitation vs. month of year, saving thousands of lives in military hospitals. In 1880, Lord Rayleigh created a statistical test for the hypothesis

of the uniform circular distribution vs. the unimodal alternative. In 1918, von Mises defined the circular normal, or von Mises distribution, which is a basis of parametric statistical inference for circular data. In 1938, Reiche introduced what is now called the CUSUM chart, which plots cumulative vector direction and average magnitude, to indicate when a sufficient amount of vectorial data has been acquired. In 1939, Krumbein introduced the transformation of axial to vectorial data for analysis, and back transformation to axial results. The paper by Watson and Williams (1956) about statistical inference for the mean and variability of a sample from the von Mises distribution and methods for comparing two or more samples started a period of significant theoretical development. Following developments of the 1960s, Mardia (1972) published a comprehensive account of methods for display, summarization, goodness of fit, and parametric/nonparametric analyses of circular data. Batschelet (1981) studied methods for bio-circular data analysis. Large sample theory was introduced about a decade after Mardia's book. Developments in circular correlation and regression, time series analysis, large sample and bootstrap methods, and nonparametric density estimation are found in Jupp and Mardia (1989). McNeill (1993) extended geostatistics to circular data. Thus, most of the theoretical developments in the field of circular statistics are relatively recent. Additional past contributors are listed in Mardia (1972).

The latest books on circular statistics include those written by Fisher (1993), Mardia and Jupp (2000), and Jammalamadaka and SenGupta (2001).

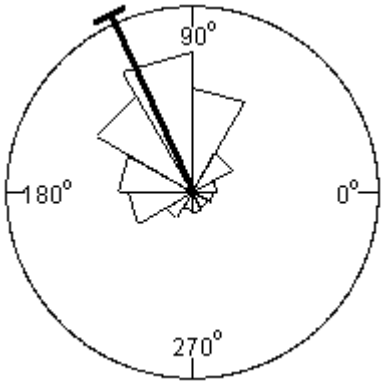


Figure 1-5. Rose Plot of the Circular Data Derived from the Data of Figure 1-4. The angle of the wedge is the bin width and the area of a wedge is proportional to the bin count. The heavy radial line indicates the mean of 106.3° and the short arc indicates the 95% confidence interval of (104.8°,107.8°).

1.4.2 Literature Review for Imaging Circular-Spatial Data

Sources were examined for examples of imaged circular-spatial data including:

- Visualization displays of computational fluid dynamics (CFD) software:
  - FLUENT (FLUENT 2008) – Software for simulation of fluid flow, heat and mass transfer, and related phenomena involving turbulence, reactions, and multiphase (liquid and gas) flow.
  - FIELDVIEW (Intelligent Light 2008) – Post-processing software for identification of important flow features and characteristics in simulations, and for interactive exploration of results.
  - Ensight (CEI 2008) – General tools for visualizing complex datasets.
- Software for the analysis of circular data:
  - Axis (Pisces Conservation Ltd 2003) – Implements the principal graphical methods and statistical tests described by Fisher (1993) for the analysis of circular data.

- CircStats (Lund and Agostinelli 2007) – This R package implements the graphical methods and statistical tests described by Jammalamadaka and SenGupta (2001) for the analysis of circular data.
- Oriana 2 (Kovach Computing 2004) – Calculates statistics, tests, and correlations for circular data. Graphics include the rose diagram, linear and circular histograms, the arrow plot with arrow length as frequency or magnitude, stacked raw data plots, and circular QQ plots.
- Surfer 8 (Scientific Software Group 2008) – Converts vector-spatial data into contour, surface, wireframe, vector, and shaded relief maps.
- Vector Rose 3.0 (Zippi 2001) – Calculates circular statistics, tests, and graphics (including the rose diagram and the circular histogram) for circular data.

None of these software packages provide a method of imaging circular-spatial data similar to the new circular dataimage of Chapter 2.

1.4.3 Literature Review for Circular-Spatial Correlation

Bivariate or multivariate data involving CRV is common. However, the study of association or correlation is newer than the relatively new area of circular statistics. Further, the study of circular-spatial correlation is newest. Jammalamadaka and SenGupta (2001) described several methods for computing the association and correlation of nonspatial CRV and circular data. These include:

- The population circular correlation coefficient

  $\rho_c(\alpha,\beta)=\frac{\mathrm{E}\{\sin(\alpha-\mu)\sin(\beta-\nu)\}}{\sqrt{Var(\sin(\alpha-\mu))Var(\sin(\beta-\nu))}}$ with E the expectation operator, angle α, $\mu=\mathrm{E}\{\alpha\}$, angle β, $\upsilon=\mathrm{E}\{\beta\}$, and *Var* the variance.

- Parametric cases of ρ_c involving specific circular probability distributions.

- The sample circular correlation coefficient, $r_{c,n} = \frac{\sum_{i=1}^{n} \sin(\alpha_i - \overline{\alpha})\sin(\beta_i - \overline{\beta})}{\sqrt{\sum_{i=1}^{n} \sin^2(\alpha_i - \overline{\alpha})\sin^2(\beta_i - \overline{\beta})}}$

  with n the sample size, sample $(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)$, and $\overline{\alpha}$ and $\overline{\beta}$ the sample mean directions.

- The nonparametric version of $r_{c,n}$ with α_i replaced by $\frac{rank(\alpha_i)}{n} 2\pi$, and β_i replaced by $\frac{rank(\beta_i)}{n} 2\pi$.

1.4.4 Literature Review for Kriging of Circular-Spatial Data

1.4.4.1 Terminology

Kriging is a body of techniques for predicting spatially correlated data. Figure 1-6 shows a heatmap before and after kriging. The name of the technique is derived from Daniel Krige, a South African mining geologist, who originated the method. Kriging uses the measurements, their distances apart, and a model of their spatial dependence based on the variogram or covariogram. The covariogram is the graph of the mean covariance between observations a distance *d* apart vs. *d*. The variogram is the graph of the mean squared difference of observations a distance *d* apart vs. *d*. In general, the variogram is less sensitive to minor departures from the assumption that the process mean is independent of location than the covariogram. The data are called isotropic, as opposed to anisotropic, when the spatial dependence is independent of the direction in which measurements are taken, and dependent on the distance *d* only.

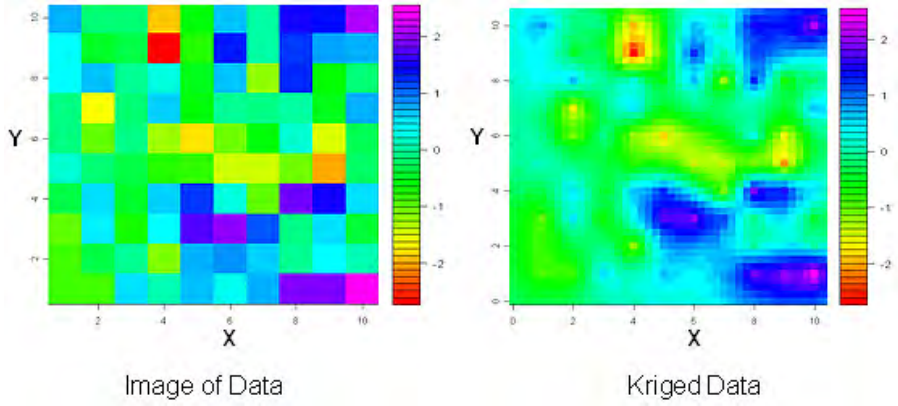


Figure 1-6. Kriging, the Estimation of Spatial Data Based on Spatial Correlation.

1.4.4.2 Literature Review for Kriging of Circular-Spatial Data

A significant body of literature exists for the kriging of circular and vector spatial data. Quimby (1986) estimated an n-vector at a location assuming each location has a different mean and variance, and using the multivariate auto covariance-cross covariance matrix. Young (1987) showed that kriging is applicable to 3D vectors describing rock fracture orientation assuming each location has the same unknown mean, and using a scalar variogram function of vectors. Young's method is evaluated using cross validation. Schaeben, Boogaart, and Apel (2001) predicted the polar unit vector at a given location, using multivariate variograms and covariance functions, assuming a constant mean, and defining different types of isotropy, which lead to different simplifications of the general cross-covariance function and kriging procedures. A measure of confidence in the estimate was given. Boogaart and Schaeben (2002a) extended prediction to direction, axis, or orientation by embedding a sphere/hemisphere in a real vector space. Boogaart and Schaeben (2002b) predicted rotation by embedding the rotations in a real vector space with assumptions of isotropy.

McNeill (1993) introduced kriging of circular random variables via trigonometric based circular statistics, assuming a common circular probability distribution, isotropic spatial correlation, and a variogram as a function of cosines.

1.4.5 Literature Review for Simulation of Circular-Spatial Data

A RF is a stochastic process operating over a space. A Gaussian RF (GRF) is a RF in which the random variables follow the multivariate normal distribution with covariance depending on distance between locations of the random variables. References include Gneiting and Schlather (2004), Lantuejoul (2002), and Schlather (1999). The function grf in the R package geoR (Ribeiro and Diggle 2001) generates simulations of GRFs for a given covariance model. The function GaussRF in the R package RandomFields (Schlather 2001) generates spatial GRFs and spatial-temporal GRFs.

1.5 Dissertation Overview

Chapter 2 extends the graphical methods of spatial statistics. It details a new method for circular-spatial data that produces a continuous image with high resolution such that directional structure can be simultaneously recognized on both local and global scales.

Chapter 3 introduces a new graphical method called the cosineogram (graph of cosines) and related theory for the extraction of spatial correlation from circular-spatial data in the form required by the circular kriging method of Chapter 4. The empirical cosineogram plots the mean cosine of the angle between random components of direction a distance d apart vs. d. The cosineogram is replaced with a fitted positive definite function to achieve optimal fit of estimated direction to the actual, but

unobserved and unknown direction. Three main positive definite functions from linear kriging are adapted to the cosine behavior of the CRF. Additional functions are identified in Appendix M.

Chapter 4 provides a detailed linear-algebraic and trigonometric derivation of an estimator of direction in an isotropic CRF (correlation same in all directions). Building on the work of McNeill, the derivation proceeds without assuming the variogram as a function of cosines and avoids the Taylor series approximation. This is accomplished by minimization of the mean squared length of the error between the estimator and the actual, but unknown and unobserved direction. This derivation produces a new expression of circular-spatial correlation as the mean cosine of the angle between random components of direction observed at a distance *d* apart. Optimality of the estimator is proved. A computationally efficient form of the estimator is derived. Chapter 4 also derives a first order estimator of the imprecision of the direction estimator, correcting the result of McNeill. The interpolation is called “exact” in the sense that, although undesirable in the presence of noise, the estimate at a location where direction is observed equals the observed direction, and the imprecision or variability of the estimate goes to zero as distance to an observation location goes to zero.

In Chapter 5, the ideas of GRFs are extended to CRFs. A method is provided to simulate a CRF with a specified circular probability distribution from a GRF with a specified spatial covariance model. Some properties of the simulated CRF are argued and others involving one or two nonclosed form transformations are characterized. Figure 1-7 summarizes circular-spatial methods of Chapters 1-5. Chapter 6 provides a comprehensive example, which shows each step of Figure 1-7, and connects the chapters.

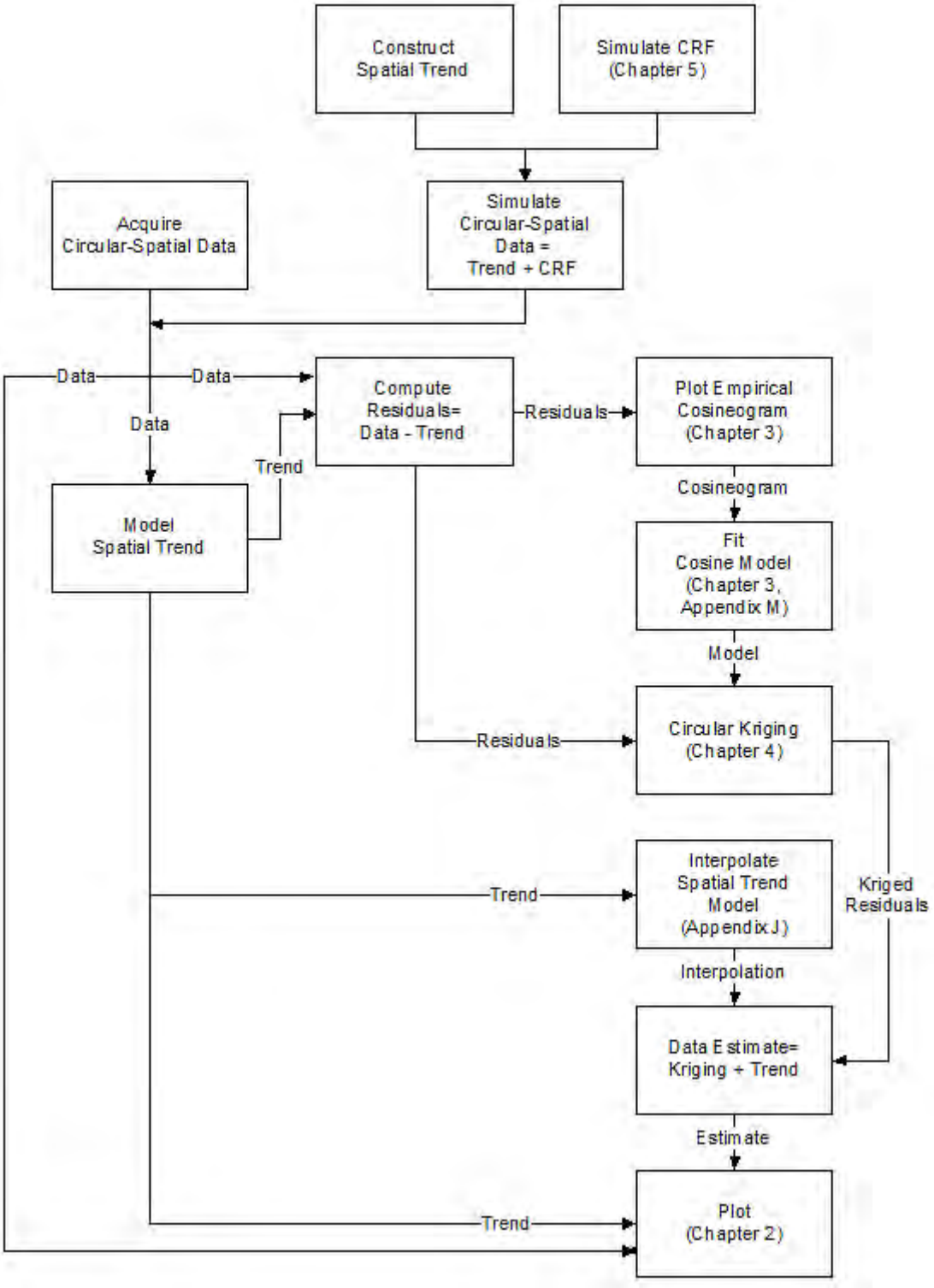


Figure 1-7. Flow Chart of Methods for Circular-Spatial Data.

Additional details are given in the appendices. Appendix A summarizes the mathematical notation used. Appendix B organizes the linear algebra theory and subordinate proofs required by the circular kriging derivations of Chapter 4.

Appendices C – M are relevant to the simulation of circular random fields of Chapter 5. Appendices C and D continue the qualitative evaluations of CRFs of Chapter 5, Section 5.5. Circular CDFs are derived in Appendix E for support $[0,2\pi)$, verified by integration in Appendix F, and modified for the equivalent rotated support $[-\pi,+\pi)$ in Appendix G. Rotation of the support from $[0,2\pi)$ to $[-\pi,+\pi)$ is required to map standard normal random variables to a CRV with mean direction 0 using the method of Chapter 5. Appendix H corrects a form of the wrapped Cauchy CDF, evaluates three forms of the CDF, and selects the form for implementation in the R package CircSpatial that is simple and does not have numerical issues at extreme low variability. Appendix I derives the inverse CDF of the triangular circular probability distribution. It is required to simulate triangular CRFs. The inverse CDFs of the cardioid, von Mises, and wrapped Cauchy circular distributions are nonclosed form transformations. Hence, Appendix M characterizes the spatial dependence of CRFs simulated by the method of Chapter 5.

Appendix J documents the R software package CircSpatial, which covers all the chapters and details a method of interpolation of global trend models based on circular-spatial data. Estimated direction is obtained by adding the kriging interpolation to the global trend model interpolation. Appendices K and L contain the R function code of the R package CircSpatial and the R command line input used to produce many of the figures in the dissertation. Appendix N has graphics for CRV and circular data introduced in Chapter 1 including a new cylindrical display of the probability density function of CRV.