## Title

Kernel regression of directional data with application to wind and wildfire data in Los Angeles County, California



## Authors

Xu, Haiyong
Schoenberg, Frederic P

## Publication Date

2007-01-29

# Kernel regression of directional data with application to wind and wildfire data in Los Angeles County, California

Haiyong Xu[1] and Frederic Paik Schoenberg[1]

Running Title: Directional kernel regression for wind and fire data.

[1] Department of Statistics, University of California, Los Angeles, CA 90095–1554, USA.

phone: 310-794-5193

fax: 310-206-5658

email: frederic@stat.ucla.edu

Postal address: UCLA Dept. of Statistics

8142 Math-Science Building

Los Angeles, CA 90095–1554, USA.

**Abstract**

This paper describes a method of kernel regression that can be used to investigate the relationship between a directional explanatory variable and a real-valued response variable. Cross-validation and bootstrap methods for obtaining sensible bandwidths and standard error estimates are also described. The proposed method is applied to wildfire and meteorological data from Los Angeles County, California, with the goal of summarizing and quantifying the impact of wind direction on the total area burned per day in wildfires. The results confirm that winds blowing from the NorthEast and East are associated with significantly higher burn areas than winds from other directions and that the daily burn area on days with winds from the NorthEast is about 4.7 times that associated with the winds from the SouthWest.

Key words: circular statistics, cross-validation, directional statistics.

# 1 Introduction.

Kernel regression is very commonly used as a means of summarizing the relationship between two real-valued variables without requiring the assumption of a parametric form for this relationship. Numerous examples of applications are described in Silverman [1]. In the context of wildfires in Los Angeles County, California, Schoenberg et al. [2] used kernel regression to show that monthly wildfire occurrence depends critically on various ecological covariates including temperature, windspeed, precipitation, relative humidity, and fuel moisture, and suggested that several of such variables have an approximately thresholded linear relationship with burn area. Similar types of analyses have been made e.g. by Flan-

nigan and Harrington [3]; see also Pyne et al. [4] and Johnson and Miyanishi [5]. While the relationship between wind direction in Southern California and wildfire activity has been well-documented, little work has aimed at quantifying this relationship. The present work extends the method of kernel regression to directional data for the purpose of describing and quantifying the association between wind direction and area burned in wildfires.

In Southern California, seasonal Santa Ana winds tend to begin each Fall just after a dry season of 6 months or more [6]. These offshore winds reach speeds exceeding 100kph at a relative humidity below 10%, and are annual events lasting several days to several weeks, creating the most wildfire-prone weather conditions in the United States [7]. The blustery, dry and warm Santa Ana winds which originate from the great basin and blow SouthWestward through the canyon towards the coast dehydrate the vegetation, heat the air and thus aggravate the dangers of wildfires [8]. While large wildfires throughout the western USA are nearly always associated with severe weather conditions, these conditions do not occur every year in most locations, whereas southern California Santa Ana winds are predictable annual events [6].

Unlike temperature, precipitation and many other meteorological variables, wind direction cannot be treated as a single real random variable. The directionality of this variable implies that standard methods for univariate or multivariate analysis cannot be used and special methods based on circular statistics are required. The goal of the present analysis is the investigation of the relationship between wind direction and burn area in Los Angeles County using a kernel regression method for directional data.

The case considered here is that of a circular explanatory variable (wind direction) and

a real-valued response variable (burn area). Methods for the reverse scenario, where the response variable is directional and the explanatory variable is real-valued, were described in Sections 12.8.1 and 12.8.2 or Mardia and Jupp [9]. Wahba [10] and Mardia et al. [11] discuss a spline approach for the case of a circular explanatory variable, but a kernel regression approach to such a case appears not to have been discussed in prior research. The current approach is similar to existing kernel density estimation methods for directional data, such as circular histograms or rose diagrams (see e.g. [9]).

The remainder of this paper is organized as follows. Section 2 describes the wildfire and wind data from Los Angeles County considered here. Kernel regression methods for its analysis, including techniques for bandwidth selection and confidence interval estimation, are described in Section 3. Section 4 contains the results, and a discussion and suggestions for future work are provided in Section 5.

# 2 Data

## 2.1 Wildfire data

Our analysis is based on the historical data of wildfire occurrences in Los Angeles County provided by the Los Angeles County Department of Public Works (LACDPW), the Los Angeles County Fire Department (LACFD). This dataset contains information on 1985 fires dating back to the year 1878 and includes the origin dates, locations, burn areas, and spatial burn patterns of recorded wildfires. Many wildfires may be missing from the catalog, but LACFD officers noted that the catalog is believed to be complete since 1950 for fires burning

at least 0.405 $km^2$ (100 acres), and previous analyses suggested that the data between 1976-2000 may be complete for fires burning at least 0.0405 $km^2$ (10 acres) [2,12]. Since wind data from Remote Automatic Weather Stations in Los Angeles County are unavailable before 1976, data from previous years were excluded from the present analysis, and our attention is restricted to the 592 fires burning at least 0.0405 $km^2$ occurring between January 1, 1976 and December 31, 2000.Such fires account for 75.5% of the available wildfire records between 1976 and 2000, and are responsible for 99.78% of the area burned in Los Angeles County wildfires during this period. For further details about this dataset as well as images of the centroid locations of these wildfires, see [12].

## 2.2 Meteorological Data

Daily meteorological observations collected from 16 Remote Automatic Weather Stations (RAWS) distributed in Los Angeles County were obtained from the USDA Forest Service since May 1975. The locations of each of the RAWS are shown in Fig. 1. Data are recorded at approximately 1:00pm every day when conditions for fire are considered to be most severe. Information on air temperature, fuel temperature, precipitation, wind speed, wind direction, relative humidity and several other variables are recorded [13]. We focus here on wind speed and wind direction data from January 1, 1976 to December 31, 2000. Each station's measurement of wind direction is recorded as an integer between 1 and 8, with 1 corresponding to winds blowing from the North-East, 2 corresponding to winds blowing from the East, etc. It should also be noted that not all of the stations contained complete data for each day covering the entire 25 years, and data were most often missing in January and February,

when wildfire incidence is typically minimal. Details on missing data in Los Angeles County can be found in [12], and for further information on RAWS data see [14,15,16].

# 3 Methodology

In order to obtain a spatial aggregate of windspeed and wind direction data for all of Los Angeles County on each day, the weighted mean wind direction was computed for each day. For any particular day, consider the wind observations from the sixteen RAWS as vectors $\vec{x}_1, ..., \vec{x}_{16}$ whose lengths representing the speed and whose angles representing the wind direction for that day. The weighted mean wind direction $\theta$ is defined as the direction of the resultant sum $\vec{x}_1 + ... + \vec{x}_{16}$ [9] . On days when data from one or more of the RAWS stations is missing, we substituted the weighted mean of the remaining observations. All of the following analysis is based on this daily weighted mean wind direction.

A systematic introduction to kernel methods is given in the monograph of Silverman [1]. In the elementary case of estimating a univariate density $f(x)$ using real-valued observations $x_1, ..., x_n$, the kernel estimator with kernel $k$ is defined by

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

where $h$ is the bandwidth. Usually the kernel function $k$ is assumed to be a symmetric probability density function such as the Gaussian density. Kernel density estimation extends readily to circular data, though care must be taken in the selection of the kernel function. Since for any $\theta$ and any integer $j$, the angles $\theta$ and $\theta + 2j\pi$ are identical, a circular probability density function may be desired for the kernel $k$. One possible option raised by Mardia and

Jupp [9] is the von Mises distribution which has the probability density function

$$g(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)},$$

where $I_0$ denotes the modified Bessel function of the first kind and order 0, which is defined by

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} e^{\kappa \cos(\theta)} d\theta.$$

The parameter $\mu$ is the directional center of the distribution and the parameter $\kappa$ is known as the concentration parameter. The concentration parameter can be used to control the degree of smoothing in circular kernel density estimation and is analogous to the bandwidth parameter except that larger values of $\kappa$ lead to less smoothing and smaller values of $\kappa$ yield more smoothing.

The focus of our analysis is to quantitatively describe the association between wind direction and wildfire burn area using directional kernel regression. As with kernel density estimation, kernel regression can be extended from the case of a real-valued explanatory variable to the case of a directional explanatory variable by choosing a circular density function $g$ such as the von Mises distribution as the kernel. Given $n$ observations of a directional explanatory variable $\theta_1, ..., \theta_n$ such as daily mean wind direction and a real-valued response variable $y_1, ..., y_n$ such as daily burn area, suppose that $y_i = m(\theta_i) + \epsilon_i$, where $\epsilon_i$ are independent and identically distributed random variables with zero mean. Then the kernel regression estimate of $m(\theta)$ may be given by

$$\hat{m}(\theta; \kappa) = \frac{n^{-1} \sum_{i=1}^{n} y_i g(\theta - \theta_i; 0, \kappa)}{n^{-1} \sum_{i=1}^{n} g(\theta - \theta_i; 0, \kappa)}. \tag{1}$$

Such an estimate is a natural analog of the Nadaraya-Watson estimator described in [1].

Choosing the degree of smoothing is of crucial importance in both density estimation and kernel regression. Various methods for selecting the smoothing parameter subjectively or automatically exist [1,17] and may be extended to the case of circular kernel regression. For instance, when kernel-smoothing real-valued data using a Gaussian kernel, Silverman [1] recommends a bandwidth of $0.9\hat{\sigma}n^{-1/5}$ and the value $h_S = 1.06\hat{\sigma}n^{-1/5}$ was suggested in [18]. (In both cases, $\hat{\sigma}$ is typically given by the minimum of the sample standard deviation and the interquartile range divided by 1.34.) They can be connected to the parameter $\kappa$ in the von Mises distribution using the following well-known connection between the von Mises distribution and the Gaussian distribution property of the von Mises distribution (see [9]): if $\theta$ is distributed according to the von Mises distribution centered at $\mu$ and with parameter $\kappa$, then

$$\kappa^{-1/2}(\theta - \mu) \rightarrow_D N(0,1) \tag{2}$$

as $\kappa \rightarrow \infty$. This suggests choosing the concentration parameter $\kappa$ via

$$\kappa = \frac{1}{h_S^2}. \tag{3}$$

Another option is to select $\kappa$ by least-squares cross-validation (LCV). As in the real-valued case, one may choocse $\kappa$ to minimize the function

$$CV(\kappa) = n^{-1} \sum_{j=1}^{n} \left[ y_j - \hat{m}^{-j}(\theta_j; \kappa) \right]^2,$$

where $\hat{m}^{-j}$ are the leave-one-out estimators:

$$\hat{m}^{-j}(\theta_j; \kappa) = \frac{\sum\limits_{i \neq j} y_i g(\theta_j - \theta_i; 0, \kappa)}{\sum\limits_{i \neq j} g(\theta_j - \theta_i; 0, \kappa)}.$$

It is well-known that kernel estimates with bandwidths selected by LCV can be inconsistent under a variety of circumstances [19,20]. In particular, for discrete data with multiple repeated values, such as the Los Angeles County wind direction data considered here, cross-validation tends to suggest bandwidths that smooth too little, i.e. $\kappa$ will tend to be too large. Hence values of $\kappa$ chosen according to (3) may be preferred in this case.

Confidence intervals for $m(\theta)$ may readily be obtained by resampling, as follows. After randomly assigning a mean wind direction to each day by sampling with replacement from the empirical distribution of mean wind directions, one may obtain an estimate of $m(\theta)$ using kernel regression. This procedure is repeated 1000 times and the confidence band is generated by connecting the 0.025 and 0.975 quantiles. The resulting 95% confidence bounds represent range within which $\hat{m}$ would be expected to fall, assuming no relationship between wind direction and wildfire burn area exists.

Circular kernel regression may also be used in order to assess the extent to which the observed relationship between wind direction and burn area may be confounded by other ecological variables, For instance, one may suspect that large burn areas occur when temperatures are high and when relative humidities are low (see e.g. [3]), and such days may be more prevalent on days with certain wind directions. Hence we also investigate the relationships between wind direction and air temperature and relative humidity using kernel-smoothed scatterplots.

# 4 Results

The seasonal patterns of Los Angeles wind direction from 1976-2000 are highlighted in Figure 2, which shows a circular kernel density estimate of daily mean wind direction. In Fig. 2, the estimated density function of the wind direction is shown separately for each month. The results indicate that the annual wind directions can generally be divided into three groups. During October through April, the wind direction is roughly uniformly distributed in all directions with only a slightly higher prevalence of winds blowing from the South and SouthWest. By contrast, winds blowing from the SouthWestern coast dominate in June, July and August. May and September can be viewed as transition months between the other two groups. Fig. 2 confirms that the Santa Ana winds, which originate from the NorthEastern great basin and blow through the canyon, occur mostly during Fall, Winter and Spring, tending to peak in occurrence in December.

Fig 3 displays a scatter plot of daily mean wind direction versus burn area. One sees that 7 of the 13 largest wildfires that occurred during this period were associated with winds blowing generally from the NorthEast. Indeed, the largest fire in this catalog began on October 23, 1978, burning 52.3 sq. miles, and it occurred when the mean windspeed was 58 mph and the mean wind direction was NorthEast.

The association between the wind direction and wildfire hazard is more obviously illustrated by the smoothed curve in Fig 4. The estimated burn area associated with winds from the NorthEast is 0.179 square miles per day, and which is 4.72 times that associated with winds from the SouthEast ($\hat{m} = 0.038$ sq. miles per day). The difference is clearly statistically significant, as shown by the 95% confidence bounds obtained by resampling,

representing the null hypothesis that wind direction has no impact on wildfire hazard. Note that the 95% bounds are not quite circular because of the non-uniformity of the observations: more days are observed with winds from the SouthWest than those from the NorthEast, for instance, the confidence bounds corresponding to the SouthWest are somewhat smaller as a result.

One may question the extent to which the association between wind direction and burn area may be confounded by other covariates known to be related to burn area, such as temperature and relative humidity. Fig. 5 displays a scatterplot of the weighted mean wind direction versus daily air temperature (averaged over the available RAWS stations) and Fig 6 displays the corresponding kernel regression estimate of the relationship between these two variates. One sees from Fig. 6 that higher temperatures are associated with winds from the SouthWest: indeed, the fitted temperature associated with SouthWestern winds is $72.5^o$F and the fitted temperature associated with NorthEastern winds is $66.2^o$F. Since this association is in the opposite direction of that between wind direction and burn area, the correlation observed between wind direction and burn area certainly cannot be explained by the correlation between wind direction and temperature.

Similarly, Figures 7 and 8 show the scatterplot and corresponding kernel regression estimate of the relationship between mean wind direction and daily relative humidity (again averaged over the available RAWS stations). The association between NortheEastern winds and dry air is readily apparent and highly significant. This is consistent with well-known meteorological principles, since winds from the SouthEast and South tend to bring more moisture while the Santa Ana winds from the NorthEast tend to bring dry air into the

Los Angeles basin [8]. Hence the relationship between wind direction and burn area may be expected to be seriously confounded by relative humidity. Nevertheless, winds from the NorthEast tend to be associated with higher burn area rates even when relative humidity is held fixed. Fig. 9a displays the kernel regression estimate of the relationship between mean wind direction and wildfire area burned on days when mean relative humidity is less than 15% and Fig. 9b presents the same result for days when mean relative humidity is between 15% and 30%. In agreement with Figure 4, both plots indicate that NorthEastern winds are associated with significantly higher burn area rates, for given relative humidity conditions.

# 5 Discussion

The fact that in Los Angeles County, winds blowing from the NorthEast are associated with higher wildfire burn area rates than those blowing from other directions is not surprising, and is in agreement with existing knowledge about Santa Ana winds. The main contribution of circular kernel regression in the present context is the ability to quantify that such wind conditions are associated with approximately 4.7 times as much wildfire burn area per day, compared with the more common winds from the SouthWest. This relationship is not confounded by temperature, which has the reverse relationship with wind direction, and remains statistically significant even when controlling for relative humidity. These results suggest that more accurate assessments of wildfire hazard may be obtained by incorporating wind direction into a model for predicting burn area, such as that in [12] or [21].

Several shortcomings in the present analysis must be emphasized. It should be noted that the response variable studied here is really not daily burn area but rather the area burned by

wildfires *ignited* on the given day. The latter may actually be more relevant to firefighters, urban planners, and others concerned with conditions under which large wildfires initiate. Whereas the daily amount of burn from pre-existing wildfires may depend more heavily on firefighting practices and other management decisions, it should also be noted that such practices are also related to the probability that a small ignition becomes a large wildfire.

The current work assesses mean wind direction and burn area over all of Los Angeles County; a more localized analysis may be a subject for future research. It is possible, for instance, that certain wind directions at particular RAWS stations within Los Angeles County are associated with higher burn areas, compared with other stations. Further, on days when weather observations were missing from particular stations, wind data from existing stations were averaged in computing the daily mean wind direction in the present analysis. Such missing observations may have corresponded to extreme wind conditions at those locations, in which case their inclusion might have had an impact on our results. Since most of these missing observations occurred on cool days in January and February when conditions were ill-suited for wildfires, however, it seems unlikely that these missing observations would have very significantly altered the present results.

## Acknowledgements

Thanks to Larry Bradshaw at the USDA Forest Service for generously providing us with RAWS data and helping us to process it. Thanks also to James Woods, Roger Peng, and members of the LACFD and LADPW (especially Mike Takeshita, Frank Vidales, and Herb Spitzer) for sharing their data and expertise.

# References


Silverman BW. *Kernel Density Estimation and Data Analysis.* Chapman and Hall: London, 1986.

Schoenberg FP, Peng R, Huang Z, Rundel P. Detection of nonlinearities in the dependence of burn area on fuel age and climatic variables. *Int. J. Wildland Fire*, 2003; 12(1): 1Ñ10

Flannigan MD, Harrington JB. A study of the relation of meteorological variables to monthly provincial area burned by wildfire in Canada (1953-80). *Journal of Applied Meteorology* 1988; 27:441-452.

Pyne SJ, Andrews PL, Laven RD. *Introduction to Wildland Fire*, 2nd edition. Wiley: New York, 1996.

Johnson EA, Miyanishi K. *Forest Fires: Behavior and Ecological Effects.* Academic Press: San Diego, 2001.

Keeley JE. Impact of antecedent climate on fire regimes in coastal California. *Int. J. Wildland Fire* 2004; 13, 173-182

Schroeder MJ, Glovinsky M, Hendricks V, Hood F, Hull M, Jacobson H, Kirkpatrick R, Krueger D, Mallory L, Oertel A, Reese R, Sergius L, Syverson C. *Synoptic Weather Types Associated with Critical Fire Weather.* Institute for Applied Technology, National Bureau of Standards, U.S. Department of Commerce, Washington, D.C., 1964.

Keeley JE, Fotheringham CJ. Impact of past, present, and future fire regimes on North American Mediterranean shrublands. In: *Fire and climatic change in temperate ecosystems of the western Americas*, T.T. Veblen, W.L. Baker, G. Montenegro and T.W. Swetnam, editors. Springer: NY, 2003; 218–262.

Mardia KV, Jupp PE. *Directional statistics.* Wiley: New York, 2000.

Wahba G. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, Philadelphia, 1990.

Mardia KV, Goodall CR, Walder A. Distributions of projective invariants and model-based machine vision. *Adv. Appl. Prob.* 1996; 28, 641–661.

Peng RD, Schoenberg FP, Woods J. A space-time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association* 2005; 100(469): 26–35.

Warren JR, Vance DL. *Remote Automatic Weather Station for Resource and Fire Management Agencies.* United States Department of Agriculture Forest Service Technical Report INT-116, Intermountain Forest and Range Experiment Station, 1981.

Andrews PL, Bradshaw LS. *FIRES: Fire Information Retrieval and Evaluation System - a program for fire danger rating analysis.* Gen. Tech. Rep. INT-GTR-367. Ogden, UT; U.S. Department of Agriculture, Forest Service, Intermountain Research Station, 1997, 1–64.

Bradshaw LS, Brittain S. FireFamily Plus: fire weather and fire danger climatology at

your fingertips. in *3rd Symposium on Fire and Forest Meteorology, Long Beach, CA.* American Meteorological Society, Boston, 1999.

Bradshaw LS, McCormick E. *FireFamily Plus User's Guide, Version 2.0.* USDA Forest Service, 2000.

Hardle W. *Applied Nonparametric Regression.* Cambridge University Press: Cambridge, 1992.

Scott DW. *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley: New York, 1992.

Scott DW, Factor LE. Monte Carlo study of three data based nonparametric density estimators. *Journal of the American Statistical Association* 1981; 76: 9–15.

Schuster EF, Gregory CG. On the non-consistency of maximum likelihood parametric density estimators. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, W. Eddy, ed., 1981.

Schoenberg FP, Chang C, Keeley J, Pompa J, Woods J, Xu H. A critical assessment of the burning Index in Los Angeles County, California. *International Journal of Wildland Fire*, 2007; in press.

List of Figures and their Captions:

Figure 1: Locations of the 16 Remote Automatic Weather Stations in Los Angeles County.

Figure 2: Kernel density estimation of weighted mean wind direction for each month ($\kappa$ is calculated by (3), and the values for January through December, respectively, are 3.98, 4.91, 5.45, 4.70, 10.27, 16.04, 18.58, 15.44, 8.74, 5.50, 4.38, 2.34 ).

Figure 3: Scatterplot of burn area versus weighted mean wind direction.

Figure 4: Kernel regression of burn area versus weighted mean wind direction ($\kappa$=2.3 by LCV).

Figure 5: Scatterplot of temperature versus weighted mean wind direction.

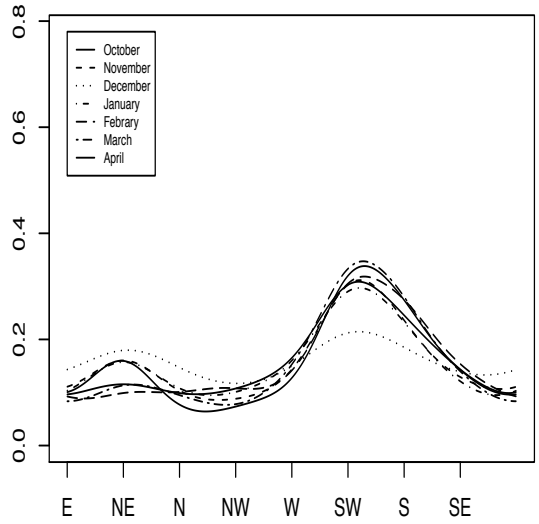
Figure 6: Kernel regression of temperature versus weighted mean wind direction ($\kappa$=51 by LCV).

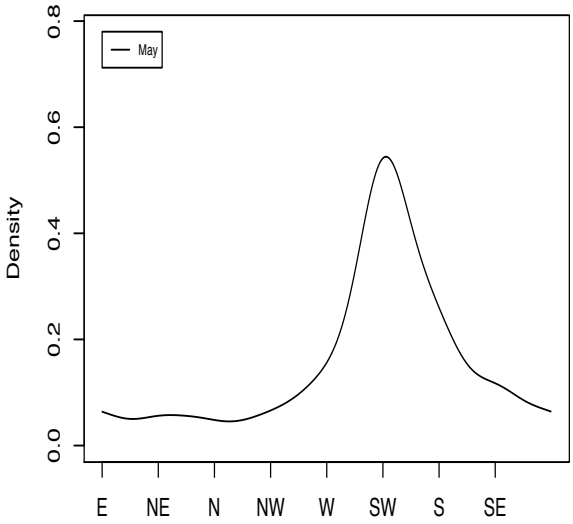
Figure 7: Scatterplot of relative humidity versus weighted mean wind direction.

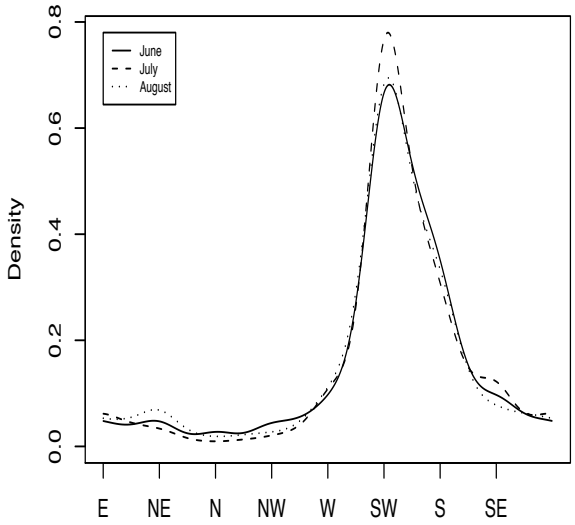
Figure 8: Kernel regression of relative humidity versus weighted mean wind direction ($\kappa$=49 by LCV).

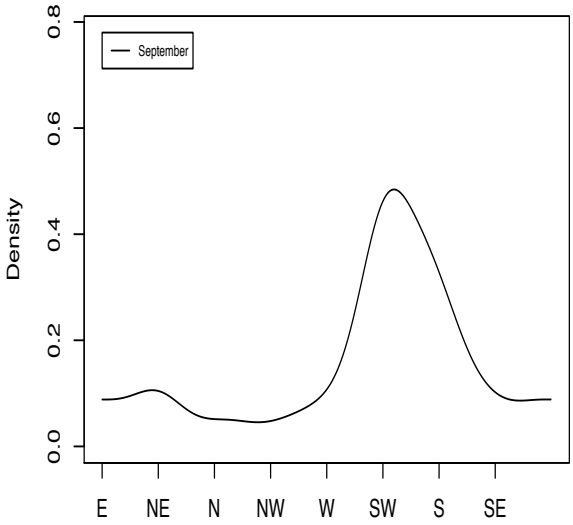
Figure 9: Kernel regression of burn area versus weighted mean wind direction when relative humidity is (a) less than 15% ($\kappa$=0.8 by LCV); (b) between 15% and 30% ($\kappa$=0.9 by LCV).

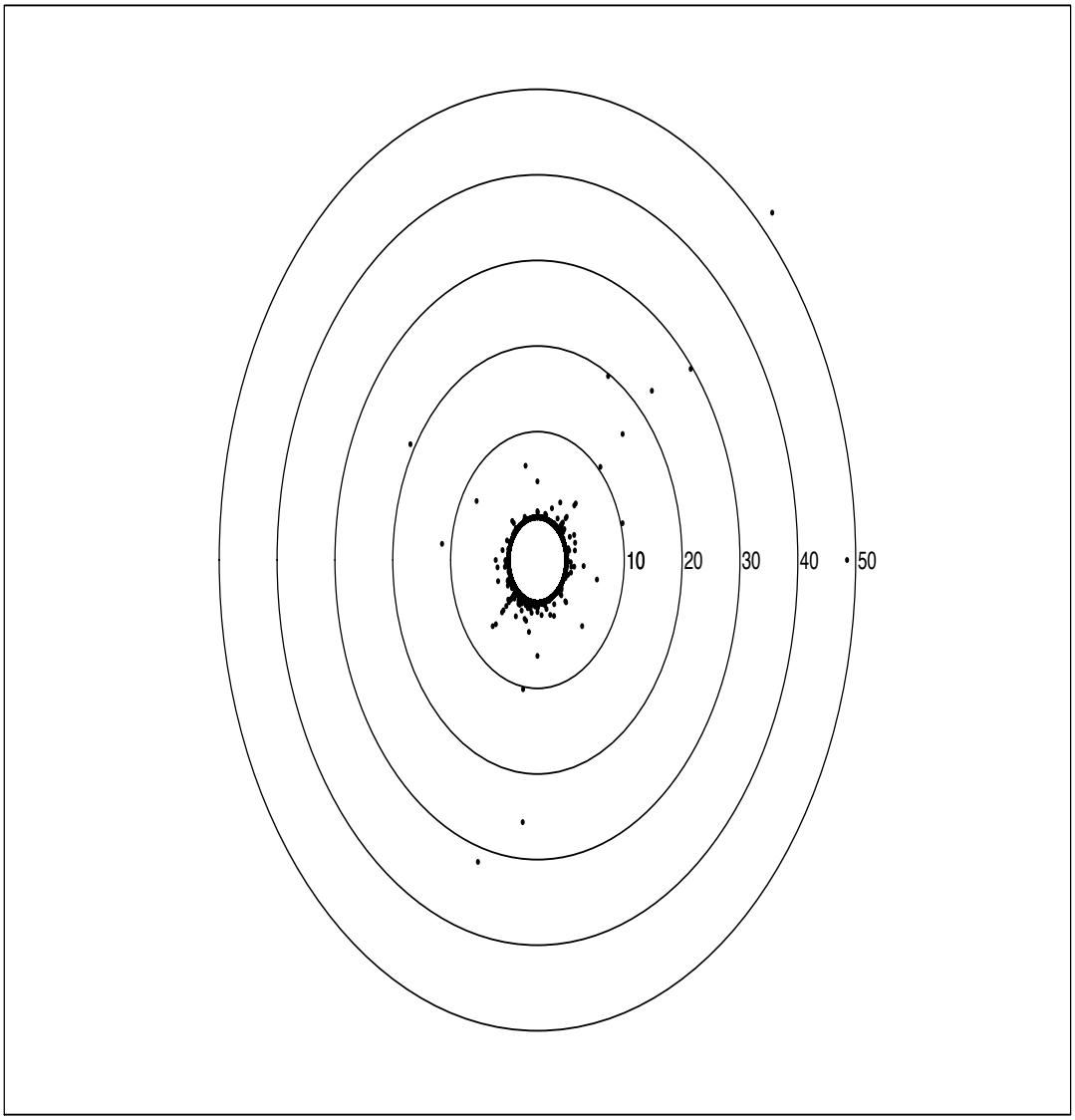

Latitude
Longitude

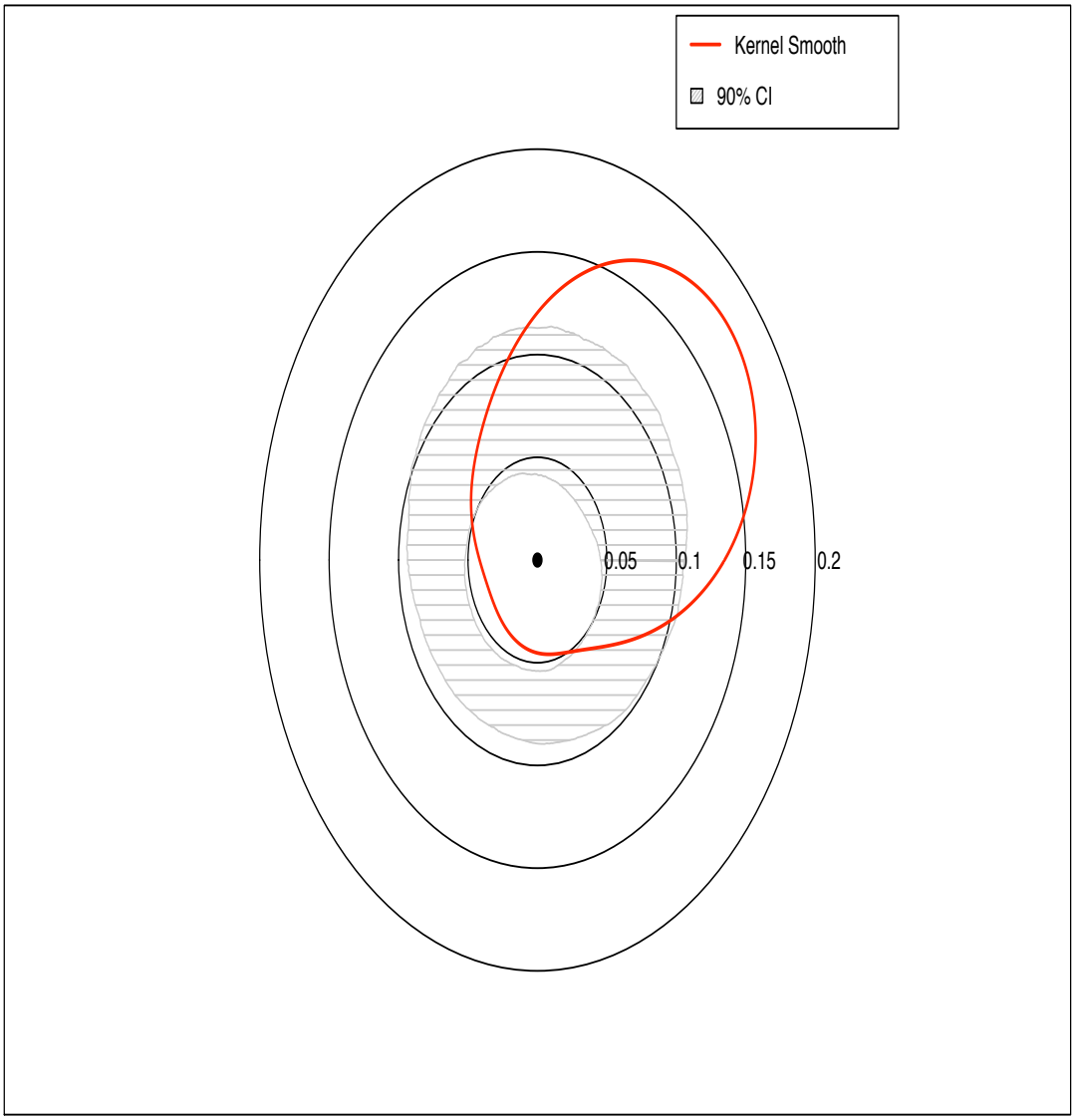

October
November
December
January
Febrary
March
April
0.0
0.2
0.4
0.6
0.8
E
NE
N
NW
W
SW
S
SE

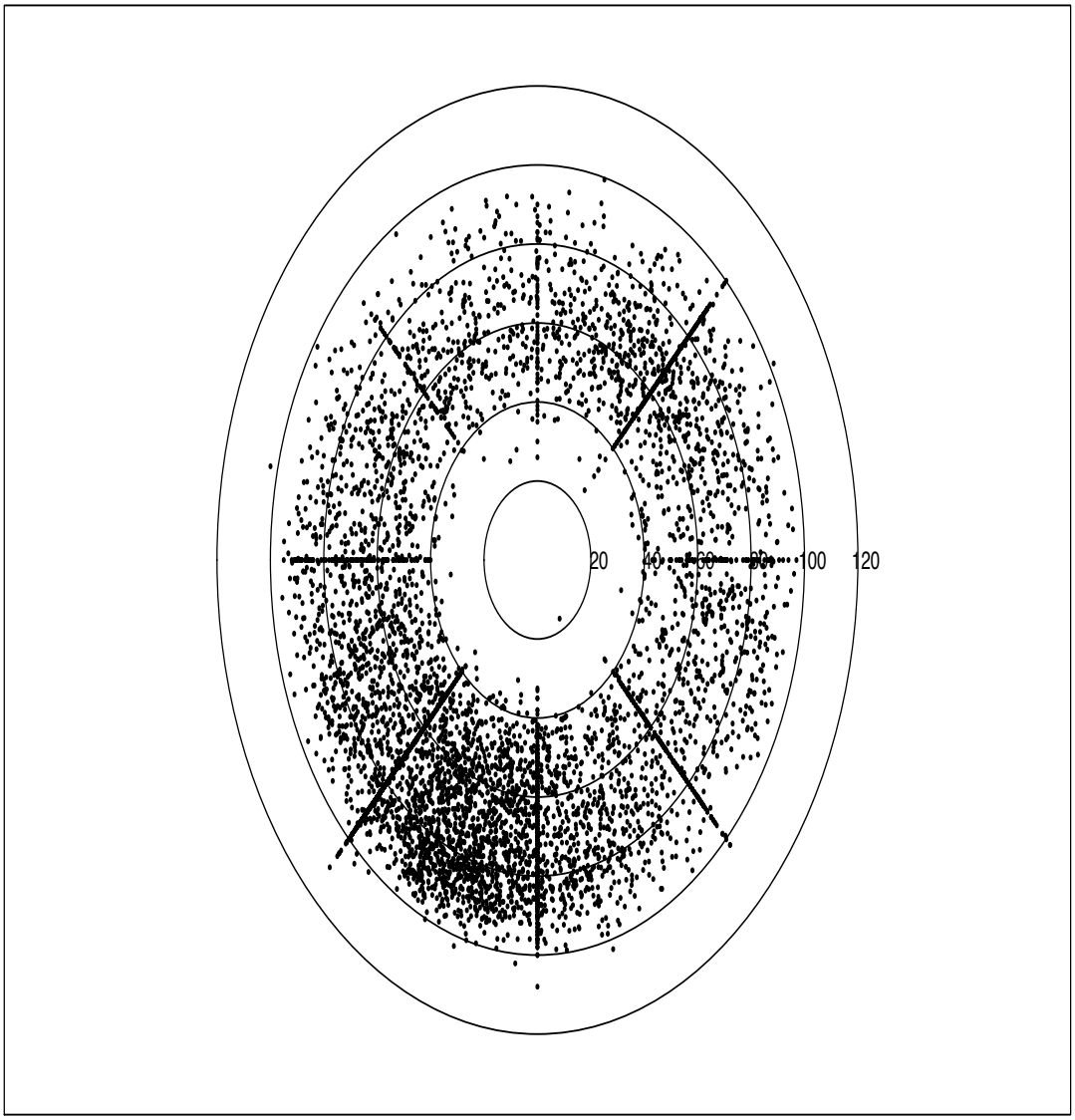

May
Density
0.0
0.2
0.4
0.6
0.8
E
NE
N
NW
W
SW
S
SE

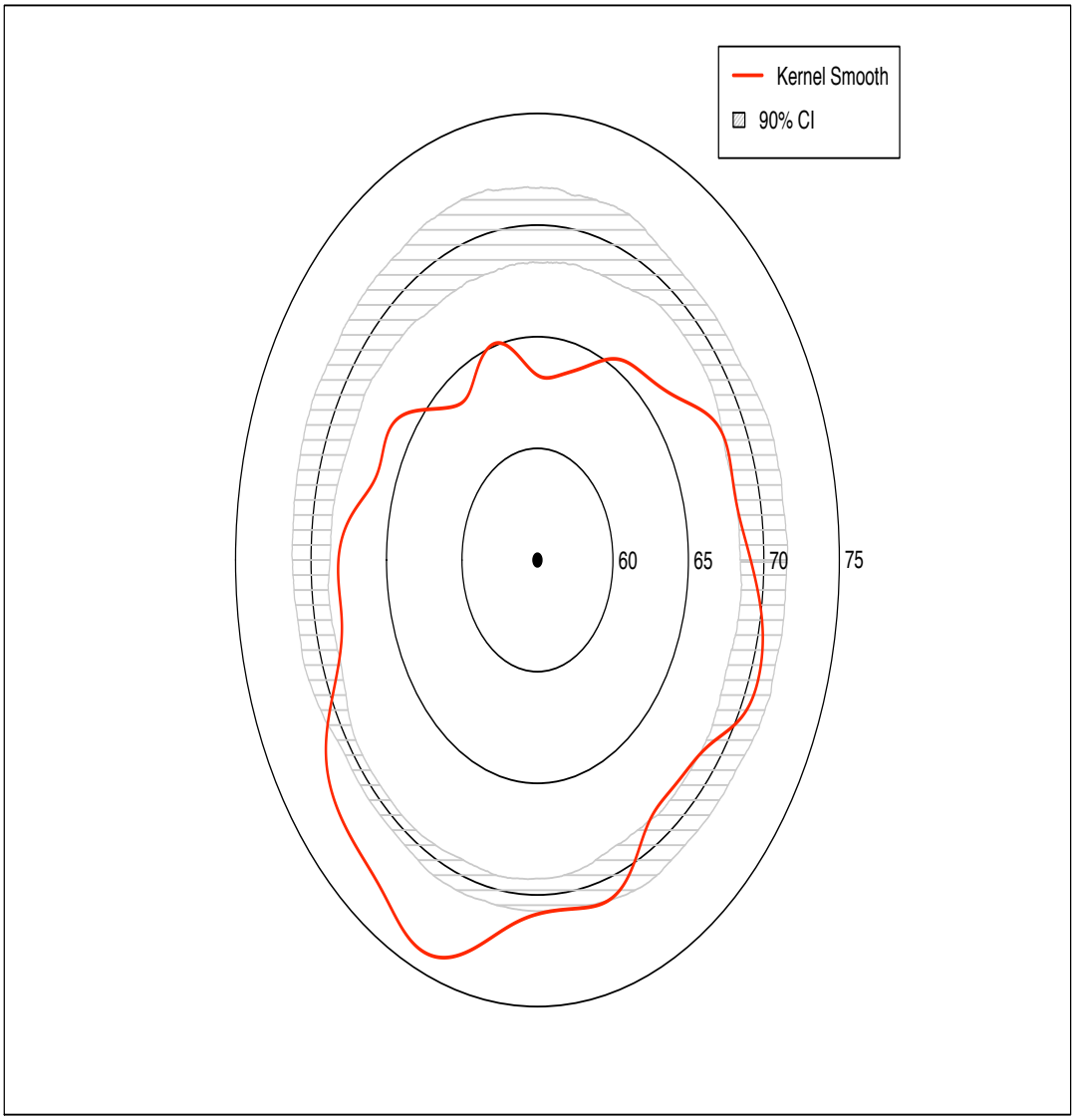

June
July
August
Density
0.0
0.2
0.4
0.6
0.8
E
NE
N
NW
W
SW
S
SE

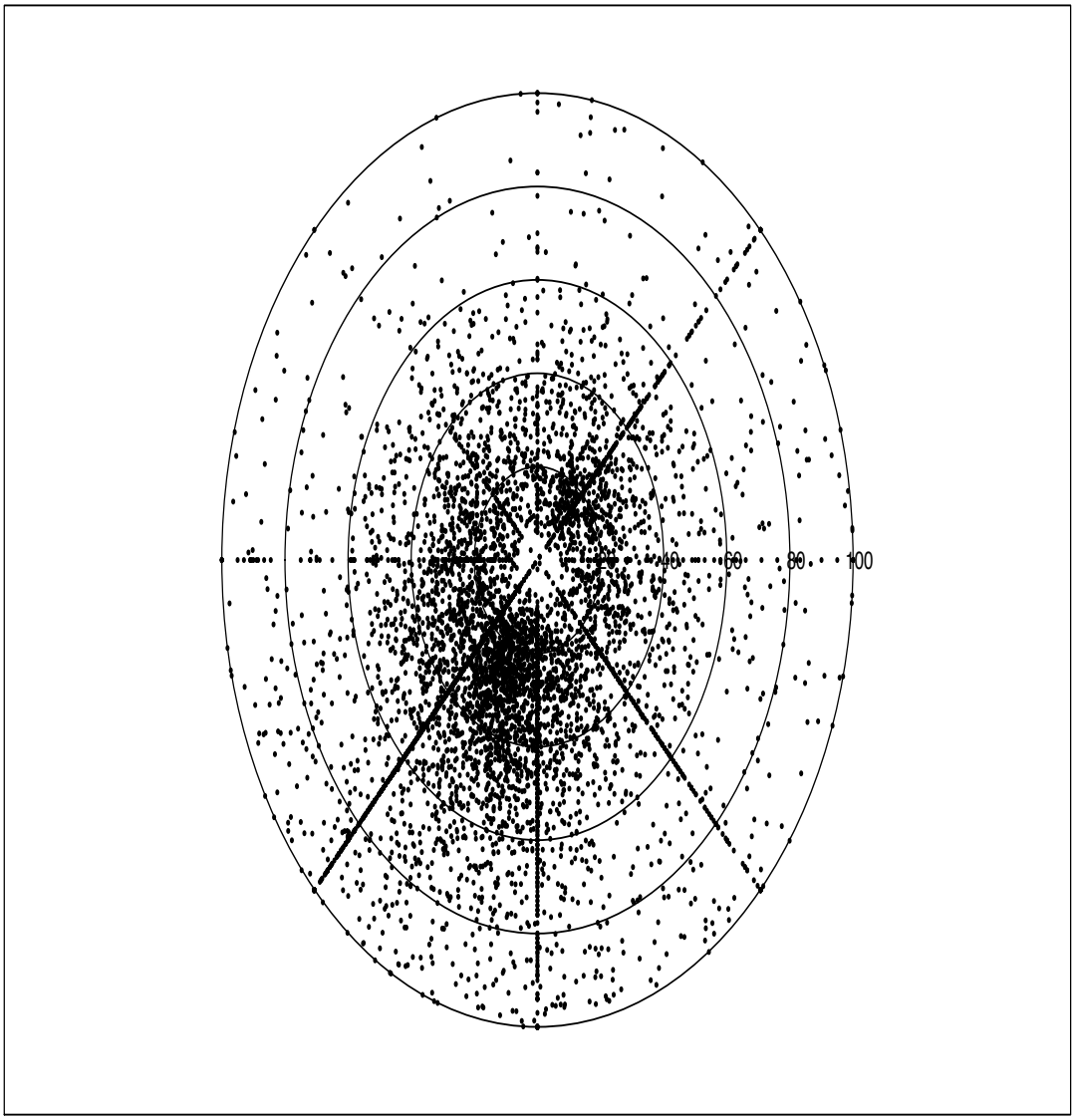

September
Density
0.0
0.2
0.4
0.6
0.8
E
NE
N
NW
W
SW
S
SE

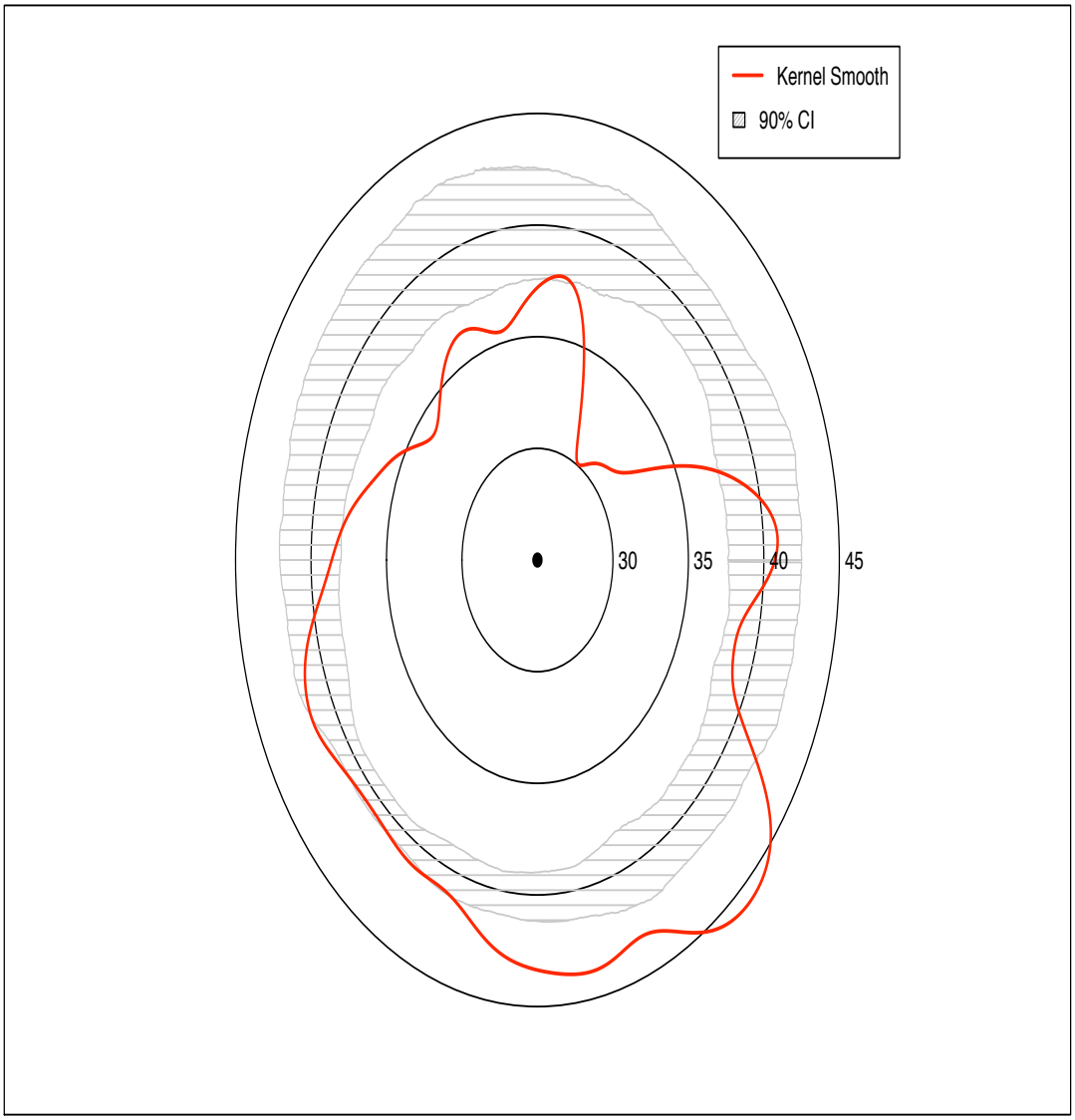

10
20
30
40
50

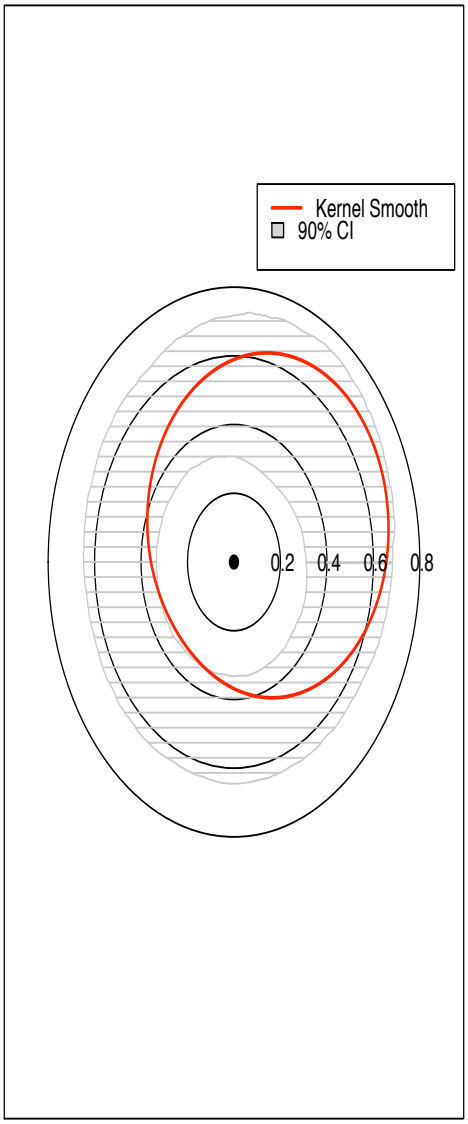

Kernel Smooth
90% CI
0.05
0.1
0.15
0.2

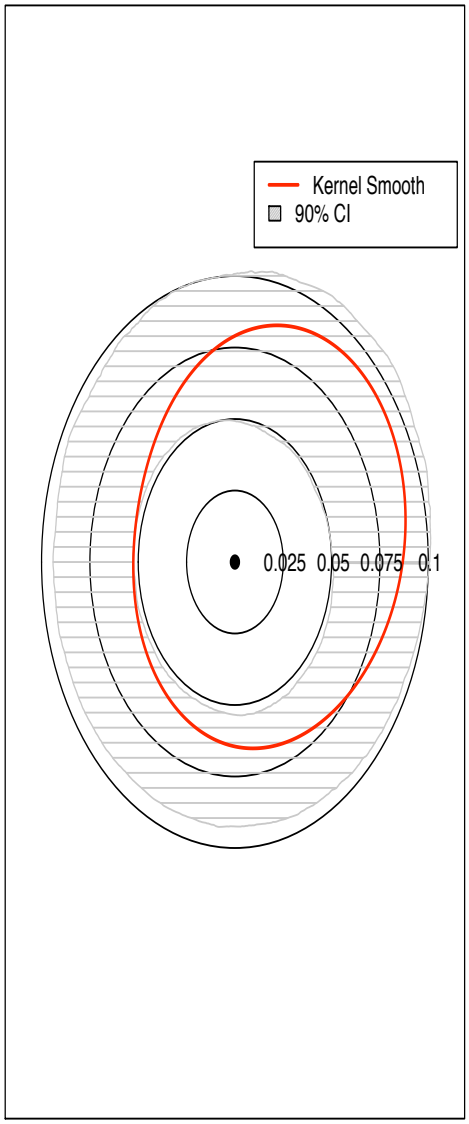

20
40
60
80
100
120


Kernel Smooth
90% CI
60
65
70
75


40
60
80
100


Kernel Smooth
90% CI
30
35
40
45


Kernel Smooth
90% CI
0.2
0.4
0.6
0.8


Kernel Smooth
90% CI
0.025
0.05
0.075
0.1