UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

Exploring wind direction and SO_2 concentration by circular–linear density estimation

E. García–Portugués, R.M. Crujeiras, W. González–Manteiga

Report 11-01

Reports in Statistics and Operations Research

Exploring wind direction and SO_2 concentration by circular–linear density estimation

García–Portugués*, E. Crujeiras, R.M. González–Manteiga, W.

Abstract

The study of environmental problems usually requires the description of variables with different nature and the assessment of relations between them. In this work, an algorithm for flexible estimation of the joint density for a circular and a linear variable is proposed. The method is applied for exploring the relation between wind direction and SO_2 concentration in monitoring stations close to a power plant located in Galicia (NW–Spain).

1 Introduction

Air pollution studies require the investigation of relationships between emission sources and pollutants concentrations in nearby sites. Hence, the assessment of a relation between air pollutant concentrations and wind direction, as well as the estimation of the wind direction for the maximum pollutant concentration, is used with the aim of determining a possible emission source (see Somerville *et al.* (1996), among others).

Different statistical methods have been considered for the study of the relation between wind direction and pollutants concentration, taking into account that wind direction is a circular variable, which requires a special treatment, both for exploratory and for inferential analysis. For instance, Somerville *et al.* (1994) use circular–linear rank correlation coefficients for the association between wind direction and pollutants concentrations. Somerville *et al.* (1996) propose a regression approach based on a beta function. Johnson and Wehrly (1978) and more recently Jammalamadaka and Lund (2006), consider regression models for the pollutant concentration (linear response) over the wind direction (circular explanatory variable), constructing the regression function in terms of the sine and cosine components of the circular variable. The same authors also provide an illustrative case study on the effect of wind direction and ozone levels, considering the relation between circular and linear variables. The relation between wind direction and ozone levels is also explored using a bivariate circular–linear correlation coefficient, proposed by Mardia (1976). The appearance of circular variables in applied fields is not only reduced to environmental problems. Circular data can be

*Corresponding author. Department of Statistics and Operations Research. Faculty of Mathematics. University of Santiago de Compostela. E-mail: eduardo.garcia@usc.es

also encountered in life sciences when studying animal behaviour (direction departure in migration processes) or molecules composition (angles in their structure).

From a more technical perspective, Johnson and Wehrly (1978) and Wehrly and Johnson (1980) present a method for obtaining joint circular–linear and circular–circular densities with specified marginals, respectively. Fernández–Durán (2004) introduces a new family of circular distributions based on nonnegative trigonometric sums, and this idea is used in Fernández–Durán (2007) in the construction of circular–linear densities, adapting the proposal of Johnson and Wehrly (1978). The estimation method is illustrated with a real data example, for modelling the relation between ground–level ozone and wind direction. The introduction of nonnegative trigonometric sums for modelling the circular distributions involved in Johnson and Wehrly (1978) formulation allows for more flexible models, that may present skewness or multimodality, features that cannot be reflected through the von Mises distribution (the classical model for circular variables). In this work, we propose a procedure for modelling the relation between a circular and a linear variable through the estimation of a circular–linear density, also based on the ideas of Johnson and Wehrly (1978), but considering nonparametric kernel density estimators both in the circular and linear marginals and in the joining density function. With this approach, the lack of flexibility in the construction of circular–linear densities noticed by Fernández–Durán (2007) is overcome. Besides, the circular–linear density representation considered can be interpreted in terms of copula functions, which poses some computational advantages.

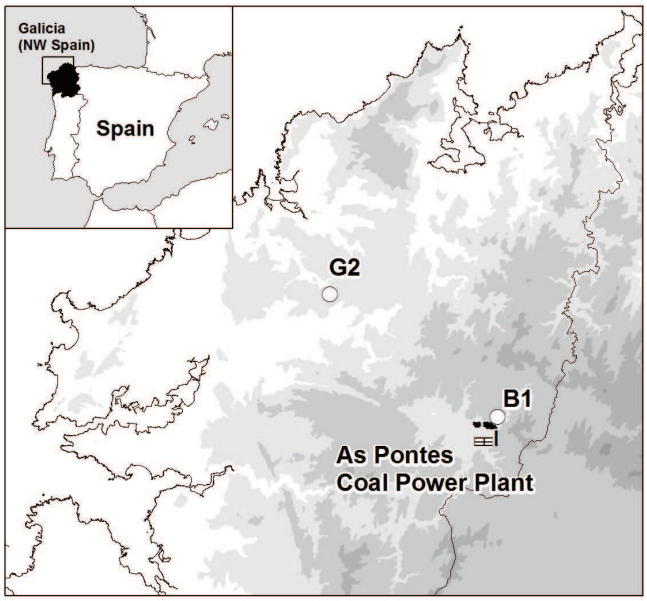

Figure 1: Locations of monitoring stations (squares) and power plant (circle) in Galicia (NW–Spain). Location of station B1: 7 °50’53”W-43 °27’14”N. Location of station G2: 8 °01’55”W-43 °33’17”N.

The practical aim of this work is to explore the relation between wind direction and SO_2 levels in two monitoring stations close to a power plant located in Galicia (NW–Spain). Monitoring station loca-

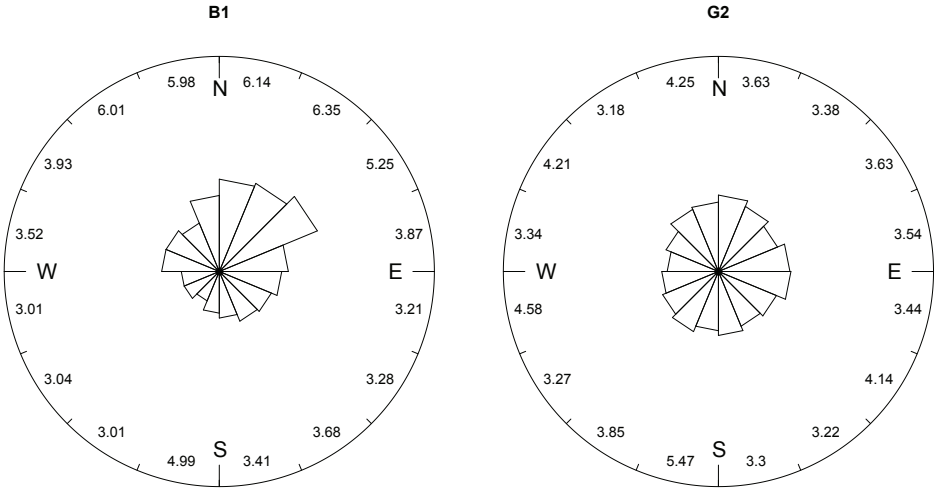

Figure 2: Rose diagrams for wind direction in station B1 (left panel) and station G2 (right panel), with average SO_2 concentrations.

tions are shown in Figure 1, around a thermal power plant. Energy is produced from the combustion of coal, which also generates pollutants as sulphur dioxide (SO_2). The analized data corresponds to measurements taken during August 2009, with one minute frequency. In Figure 2, we show the rose diagram for wind direction for data taken in station B1 (the one closest to the power plant) and station G2, in the NE direction. In the rose diagram, the average concentration for each wind direction sector is also given. It can be seen that B1 presents higher values SO_2 concentrations than G2, which seems reasonable given that G2 is 18'6 km NW far away from the power plant, and there is not clear dominant wind direction in this case. It can be also noticed that, for B1, the dominant wind direction is NE. However, the relation between wind direction and SO_2 levels is not clear from these representations, and the dependency (or lack of dependency) between them should be investigated.

This work is organized as follows. Section 2 provides some background on circular random variables and a brief review on copula methods. The algorithm for estimating a circular–linear density is detailed and discussed in Section 3. The finite sample properties of the algorithm are illustrated by a simulation study, considering parametric and nonparametric estimation methods in the circular and linear components and in the joining density. This algorithm is applied for analysing wind direction and SO_2 concentrations in Section 4. Some final comments are given in Section 5.

2 Background

As commented in the Introduction, the main goal of this work is to analyze the relation between wind direction and SO_2 concentrations in monitoring stations next to a power plant. Bearing in mind the different nature of the variables and noticing that measurements from wind direction are angles,

some background on circular random variables is introduced. This methodology will be needed in order to describe the wind direction itself and the joint relation between the two variables, through the construction of a circular–linear density introduced by Johnson and Wehrly (1978). As it has been already mentioned, the circular–linear density representation can be interpreted in terms of copulas and a brief review on these functions is also provided.

2.1 Circular and circular–linear distributions

Denote by Θ a circular random variable with support in the unit circle $\mathbb{S}^1$. A circular distribution $\Psi(\cdot)$ for Θ assigns a probability to each direction $(\cos(\theta), \sin(\theta))$ of the plane $\mathbb{R}^2$, characterized by the angle $\theta \in [0, 2\pi)$. Absolutely continuous circular distributions (with respect to the Lebesgue measure in the circumference) have an associated circular density, denoted by $\varphi(\cdot)$. The circular density must be positive and integrate one over its support, similar to linear densities. However, it must also satisfy a periodicity condition: $\varphi(\theta) = \varphi(\theta + 2\pi k)$, for all $\theta \in [0, 2\pi)$ and for all integer $k \in \mathbb{Z}$ (see Mardia and Jupp (2000)). The circular distribution $\Psi(\cdot)$ is also periodic, $\Psi(\theta + 2\pi k) = \Psi(\theta)$, $k \in \mathbb{Z}$, verifying that $\lim_{\theta \to (2\pi k)^-} \Psi(\theta) = 1$ and $\lim_{\theta \to (2\pi k)^+} \Psi(\theta) = 0$ for $k \in \mathbb{Z}$. Hence, circular distributions present a discontinuity at $2\pi k$, for $k \in \mathbb{Z}$, which represents the starting point of a new cycle. Some particular cases of circular distribution models are the uniform circular distribution and the von Mises distribution.

The uniform circular distribution in $[0, 2\pi)$ is a constant density. For $\theta \in [0, 2\pi)$:

$$\varphi_U(\theta) = \frac{1}{2\pi},\ 0 < \theta \leq 2\pi. \tag{1}$$

The von Mises distribution is the analogue of the normal distribution in circular random variables. This family of distributions, denoted by $vM(\mu, \kappa)$, is characterized by two parameters: $0 \leq \mu < 2\pi$, the circular mean and $\kappa \geq 0$, a circular concentration parameter around μ. The corresponding density function is given by:

$$\varphi_{vM}(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)},\ 0 < \theta \leq 2\pi, \tag{2}$$

with $I_0(\cdot)$ being the modified Bessel function of first kind and order zero defined by

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

The uniform circular distribution is obtained as a particular case of the von Mises family, for $\kappa = 0$. Circular density estimation can be performed by parametric methods, such as Maximum Likelihood, or using nonparametric techniques. Some estimators will be introduced in Section 3.

As we pointed in the Introduction, our goal is to explain the relation between a circular and a linear random variable (specifically, wind direction and SO_2 concentration). A circular–linear random variable (Θ, X) is supported in the cylinder $\mathbb{S}^1 \times \mathbb{R}$ or in a subset of it and a circular–linear density for (Θ, X), namely $p(\cdot, \cdot)$, must satisfy the periodicity condition in the circular argument, that is:

$$p(\theta, x) = p(\theta + 2\pi k, x), \quad \forall \theta \in [0, 2\pi),\ x \in \mathbb{R},\ k \in \mathbb{Z},$$

as well as the usual assumptions on taking nonnegative values and integrating one. Johnson and Wehrly (1978) propose a method for obtaining circular–linear densities with specified marginals. Denote by $\varphi(\cdot)$ and $f(\cdot)$ the circular and linear marginal densities, respectively, and by $\Psi(\cdot)$ and $F(\cdot)$ their corresponding marginal distributions. Let also $g(\cdot)$ be another circular density. Then:

$$p(\theta, x) = 2\pi g\left[2\pi\left(\Psi(\theta) + F(x)\right)\right]\varphi(\theta)f(x), \tag{3}$$

is a density for a circular–linear distribution for a random variable (Θ, X), with specified marginal densities $\varphi(\cdot)$ and $f(\cdot)$ (see Johnson and Wehrly (1978), Theorem 5).

From a data sample of (Θ, X), assuming that the joint density can be represented as in (3), an estimator of $p(\cdot,\cdot)$ could be obtained by the estimations of the marginals and the joining density. The circular marginal $\varphi(\cdot)$ and the joining density $g(\cdot)$ are both circular densities. Fernández–Durán (2007) proposed estimating them by nonnegative trigonometric sums, using a Maximum Likelihood method.

In the next section, we will introduce some background on copulas since expression (3) can be written in terms of these functions. We will take advantage of this relation for computational purposes.

2.2 Some notes on copulas

Copula functions are multivariate distributions with uniform marginals (see Nelsen (2006) for a complete review on copulas). One of the main results in copulas theory is Sklar's theorem which, in the bivariate case, states that if $F(\cdot,\cdot)$ is a joint distribution function with marginals $F_1(\cdot)$ and $F_2(\cdot)$ then, there exist a copula $C(\cdot,\cdot)$ such that:

$$F(x, y) = C(F_1(x), F_2(y)), \quad \forall x, y \in \mathbb{R}. \tag{4}$$

If $F_1(\cdot)$ and $F_2(\cdot)$ are continuous distributions, then $C(\cdot,\cdot)$ is unique. Conversely, if $C(\cdot,\cdot)$ is a copula and $F_1(\cdot)$ and $F_2(\cdot)$ are distribution functions, then $F(\cdot,\cdot)$ defined by (4) is a joint distribution with marginals $F_1(\cdot)$ and $F_2(\cdot)$.

If the marginal random variables are absolutely continuous, Sklar's result can be interpreted in terms of the corresponding densities. Denoting by $c(u, v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$ the copula density, the joint density of $F(\cdot,\cdot)$ in (4) can be written as

$$f(x, y) = c(F_1(x), F_2(y))f_1(x)f_2(y), \quad \forall x, y \in \mathbb{R}. \tag{5}$$

Similar to the linear case, circular–linear copulas $C_{\Theta,X}(\cdot,\cdot)$ must take into account the characteristics of the circular marginal and satisfy

$$c_{\Theta,X}(0, x) = c_{\Theta,X}(1, x), \quad \forall x \in [0, 1],$$

where $c_{\Theta,X}(\cdot,\cdot)$ is the corresponding circular–linear copula density. Note that this circular–linear copula can be linked with the circular density $g(\cdot)$ in (3) by identifying

$$c\left(\Psi(\theta), F(x)\right) = 2\pi g\left[2\pi\left(\Psi(\theta) + F(x)\right)\right]. \tag{6}$$

This means that, for each circular density $g(\cdot)$, it is possible to construct a circular–linear copula density:

$$c\left(u,v\right)=2\pi g\left[2\pi\left(u+v\right)\right],\,\forall u,v\in\mathbb{I}.$$

Hence, the circular–linear density in (3) can be written as

$$p(\theta,x)=c(\Psi(\theta),F(x))\varphi(\theta)f(x),$$

where $\varphi(\cdot)$ and $f(\cdot)$ are the circular and linear marginal densities and $c(\cdot,\cdot)$ is a copula density. An advantage of linking $g(\cdot)$ with a copula is that easy procedures for simulating circular–linear variables are possible.

3 Estimation algorithm

Recall the expression for the circular–linear density (3) introduced by Johnson and Wehrly (1978) and denote by $\{(\theta_i,x_i)\}_{i=1}^n$ a data sample from a circular–linear random variable (Θ,X). Assume that the density $p(\cdot,\cdot)$ for (Θ,X) admits a representation such as the one in (3).

In this joint circular–linear density model, three density functions must be estimated: the marginal densities $\varphi(\cdot)$ and $f(\cdot)$ (and also the corresponding distributions) and the joining circular density $g(\cdot)$. A new natural procedure for estimating $p(\cdot,\cdot)$ is given in the following algorithm:

Estimation algorithm

Step 1. Obtain estimators for the marginal densities $\hat{\varphi}(\cdot)$, $\hat{f}(\cdot)$ and the corresponding marginal distributions $\hat{\Psi}(\cdot)$, $\hat{F}(\cdot)$.

Step 2. Compute an artificial sample $\left\{2\pi\left(\hat{\Psi}(\theta_i)+\hat{F}(x_i)\right)\right\}_{i=1}^n$ and estimate the joining circular density $\hat{g}(\cdot)$.

Step 3. Obtain the circular–linear density estimator as

$$\hat{p}\left(\theta,x\right)=2\pi\hat{g}\left[2\pi\left(\hat{\Psi}(\theta)+\hat{F}(x)\right)\right]\hat{\varphi}(\theta)\hat{f}(x). \tag{7}$$

Note that all the estimators involved in the algorithm are obtained in a strictly univariate way, which simplifies its computation. The estimation of the marginal densities in Step 1, as well as the circular joining density in Step 2, can be done by parametric methods or by nonparametric procedures. For instance, a parametric estimator for $\hat{f}(\cdot)$ (respectively, for $\hat{F}(\cdot)$) can be obtained by Maximum Likelihood, as in Fernández–Durán (2007) for nonnegative trigonometric sums. In the circular case, that is, for obtaining $\hat{\varphi}(\cdot)$ and $\hat{g}(\cdot)$, Maximum Likelihood approaches are also possible (see Jammalamadaka and SenGupta (2001), Chapter 4). These estimators are consistent, although we should restrict to parametric models. In the circular case, it is usual to consider a von Mises distribution or a mixture of von Mises distributions, although Maximum Likelihood leads to complicated computations. However, it is also feasible to build nonparametric estimators for the marginals and the joining circular density using kernel methods.

Nonparametric kernel density estimation for linear random variables was introduced by Parzen and Rosenblatt (see Wand and Jones (1995) for references on kernel density estimation) and the properties of this estimator have been well studied in the statistical literature. Consider $\{X_i\}_{i=1}^n$ a random sample of a linear variable X with density $f(\cdot)$. The kernel density estimator of $f(\cdot)$ in a point $x \in \mathbb{R}$ is given by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{8}$$

where $K(\cdot)$ is a kernel function (usually a symmetric and unimodal density) and h is the bandwidth parameter. One of the crucial problems in kernel density estimation is the bandwidth choice. There exist several alternatives for obtaining a global bandwidth minimizing a certain error criterion, usually the Mean Integrated Squared Error. Some of this methods are the plug–in rule and the least–squares cross–validatory bandwidth (see Wand and Jones (1995)).

Hall *et al.* (1987) propose a nonparametric kernel density estimator for directional data in $\mathbb{S}^q$. For the circular case ($q = 1$), denoting by Θ a random variable with density $\varphi(\cdot)$, the circular kernel density estimation from a sample $\{\Theta_i\}_{i=1}^n$ is given by

$$\hat{\varphi}_\nu(\theta) = \frac{c_0(\nu)}{n} \sum_{i=1}^{n} L\left(\nu \cos(\theta - \Theta_i)\right), \tag{9}$$

where $L(\cdot)$ is the circular kernel, ν is the circular bandwidth and $c_0(\nu)$ is a constant such that $\hat{\varphi}_\nu(\cdot)$ is a density. Some differences should be noted with respect to the linear kernel density estimator in (8). First, the kernel function $L(\cdot)$ must be a rapidly varying function, such as the exponential, quite different from the bell shaped linear kernels like Gaussian or Epanechnikov densities. Secondly, the behaviour of ν is inverse to h: in linear kernel density estimation, small values of the bandwidth h produce undersmoothed estimators (small values of ν overestimate the density), whereas large values of h give oversmoothed curves (large values of ν underestimate). See Hall *et al.* (1987) for a detailed description of the estimator and its properties.

As in the linear case, bandwidth selection is also an issue in circular kernel density estimation. Although in the linear case it is a well–studied problem, for circular density estimation there are still some open questions. Taylor (2008) proposes some automatic bandwidth selection methods as a rule of thumb based on the von Mises distribution, and the log–likelihood cross–validatory bandwidth, jointly with some other robust bandwidth selectors. Based on Taylor (2008) results, none of the selectors proposed seems to show a superior behaviour.

These two estimation alternatives (parametric and nonparametric marginals and joining density), as well as a mixed approach, considering parametric marginals and nonparametric joining density estimation, are illustrated in the following simulation study.

3.1 Some simulation results

In order to check the performance of the estimation algorithm for circular–linear densities, we reproduce the following two examples given by Johnson and Wehrly (1978).

Example 1 (Circular uniform and Normal marginal distributions)**.** *Let $\varphi(\cdot)$ denote the circular uniform density (1) and $f(\cdot) = \phi(\cdot)$ the standard Normal density. Take $g(\cdot)$, the joining density, as the von Mises density in (2) with parameters (μ, κ). The circular–linear density with margins $\varphi(\cdot)$ and $f(\cdot)$ is given by*

$$p_1(\theta, x) = \frac{1}{2\pi I_0(\kappa)} \exp\left[\kappa \cos(\theta - 2\pi\Phi(x) - \mu)\right] \phi(x), \tag{10}$$

where $\Phi(\cdot)$ denotes the standard Normal distribution.

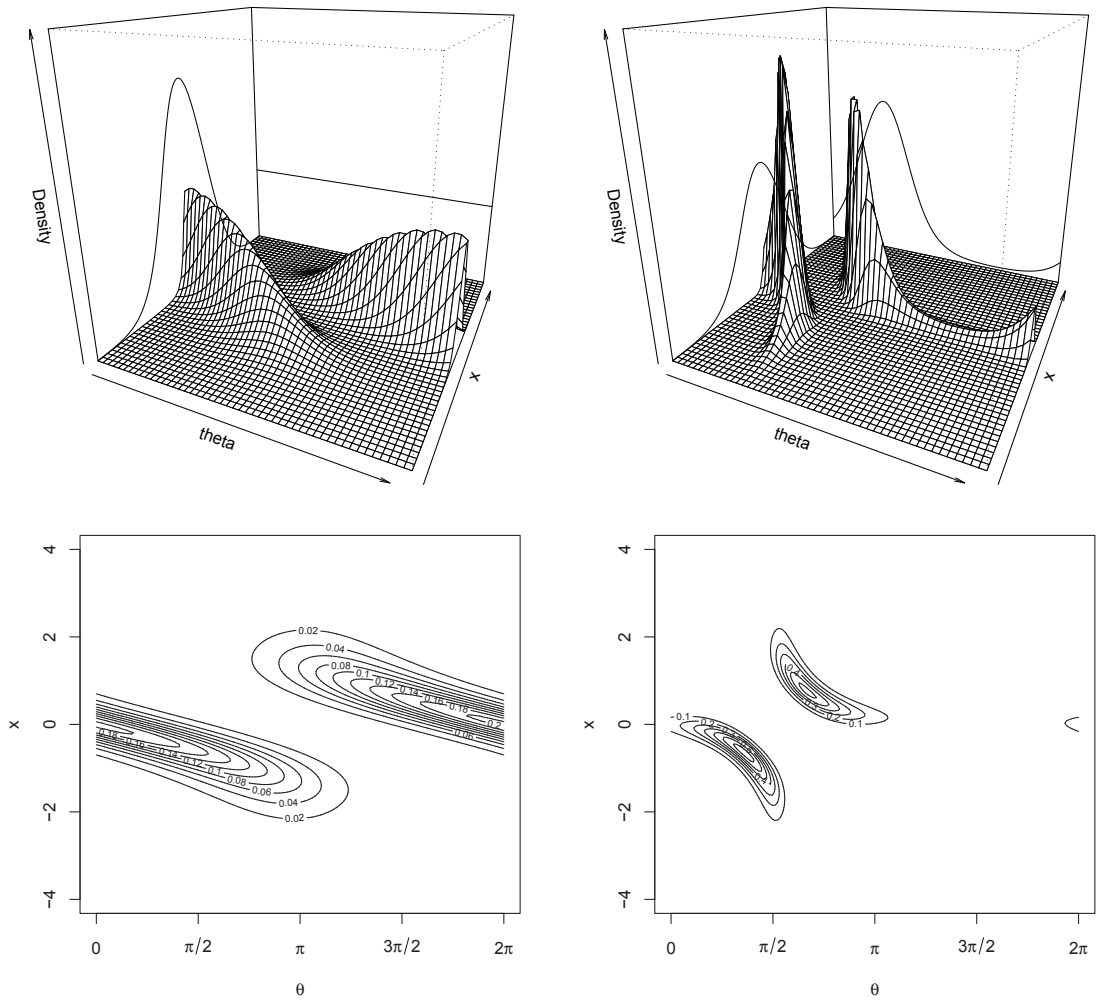

Figure 3: Left column: Example 1 density surface with parameters $\mu = \pi$ and $\kappa = 2$. Right column: Example 2 density surface with parameters $\mu = \pi$, $\kappa = 5$, $\mu_1 = \frac{\pi}{2}$ and $\kappa_1 = 2$.

Example 2 (von Mises and Normal marginal distributions)**.** *Let $\varphi_{vM}(\cdot)$ denote the von Mises marginal density (2) with parameters (μ_1, κ_1). With the same $f(\cdot)$ and $g(\cdot)$ as in the previous example, the joint circular–linear density is given by*

$$p_2(\theta, x) = \frac{1}{I_0(\kappa)} \exp\left[\kappa \cos\left(2\pi(\Psi_{vM}(\theta) - \Phi(x)) - \mu\right)\right] \varphi_{vM}(\theta)\phi(x), \tag{11}$$

where $\Psi_{vM}(\cdot)$ denotes the von Mises distribution with parameters (μ_1, κ_1).

As it has been commented in the Section 2, the formulation of the joint circular–linear density in terms of copulas simplifies the simulation of random samples. The general idea is to split the joint distribution $P(\cdot,\cdot)$ by Sklar's Theorem in a copula $C(\cdot,\cdot)$ and marginals $\Psi(\cdot)$ and $F(\cdot)$. Simulating a sample from (U,V), uniform random variables with copula $C(\cdot,\cdot)$, and applying the marginal quantiles transformations to get $(\Psi^{-1}(U), F^{-1}(V))$, we obtain a sample from distribution $P(\cdot,\cdot)$.

The simulation of (U,V) values from the copula $C(\cdot,\cdot)$ can be performed by the conditional method for simulating multivariate distributions. The conditional distribution of V given $U=u$, denoted by $C_u(\cdot)$, can be expressed in terms of the joining circular density $g(\cdot)$ as

$$C_u(v) = \frac{\partial C(u,v)}{\partial u} = \int_0^v c(u,t)dt = \int_0^v 2\pi g\left[2\pi(u+t)\right]dt,$$

where the first equality is an immediately property of copulas. For Examples 1 and 2, calculus of the latter integral leads to express $C_u(\cdot)$ as a von Mises distribution:

$$C_u(v) = \Psi_{vM}\left(2\pi v; \mu - 2\pi u, \kappa\right). \tag{12}$$

So, for simulating random samples from joint densities (10) and (11), or more generally, for a random sample from a circular–linear variable with density (3), we may proceed with the following steps:

Simulation algorithm:

Step 1 Simulate $U \sim \mathcal{U}(0,1)$, $W \sim \mathcal{U}(0,1)$.

Step 2 Calculate $V = C_u^{-1}(W)$.

Step 3 Obtain $\Theta = \Psi^{-1}(U)$, $X = F^{-1}(V)$.

In Step 1, two independent and uniformly distributed random variables are simulated. The conditional simulation method for obtaining (U,V) from the circular–linear copula $C(\cdot,\cdot)$ is performed in Step 2 (see Johnson (1987)). Finally, quantile transformations from the marginals are applied, obtaining a sample from (Θ, X) following the joint density (3). In Step 2, in our examples, the conditional distribution $C_u(\cdot)$ is related to a von Mises distribution. In the simulations, we consider, for each u

$$V = (2\pi)^{-1}\ \Psi_{vM}^{-1}\left(W; \mu - 2\pi u, \kappa\right).$$

We will apply the estimation algorithm proposed considering parametric and nonparametric estimators for the marginals and the joining density. In the parametric case, density estimators have been obtained by Maximum Likelihood, specifying the von Mises family for the circular distributions and the Normal family for the linear marginal. Nonparametric estimation has been carried out using kernel methods. The kernel density estimator in (8), with Gaussian kernel and cross–validatory bandwidth, has been used for obtaining $\hat{f}(\cdot)$. For $\hat{\varphi}(\cdot)$ and $\hat{g}(\cdot)$, the circular kernel density (9) has been implemented, with exponential kernel and least–squares cross–validatory bandwidth. In the mixed approach (parametric marginals and nonparametric joining density), Maximum Likelihood has

	Estimation method			Relative efficiency	
n	Parametric	Mixed	Nonparametric	Mixed	Nonparametric
50	0.00538	0.00949	0.01680	0.56744	0.32077
200	0.00128	0.00268	0.00518	0.48023	0.24831
500	0.00051	0.00121	0.00249	0.42671	0.20778
1000	0.00025	0.00066	0.00140	0.38969	0.18395

Table 1: MISE for estimating the circular–linear density in Example 1. Relative efficiencies of mixed and nonparametric estimations with respect to parametric estimation.

been used for obtaining $\hat{\varphi}(\cdot)$ and $\hat{f}(\cdot)$, whereas the circular kernel estimator has been considered for $\hat{g}(\cdot)$.

In order to check the performance of the procedure for estimating circular–linear densities, we consider the Mean Integrated Square Error in the estimation of $p(\cdot,\cdot)$:

$$\text{MISE} = \iint \mathbb{E}\left[\hat{p}(\theta,x) - p(\theta,x)\right]^2 d\theta dx.$$

The MISE is approximated by Monte Carlo simulations, taking $M = 1000$ replicates. We compare the performance of the method using parametric estimation, nonparametric estimation and a mixed approach. Four sample sizes have been used: $n = 50$, $n = 200$, $n = 500$ and $n = 1000$. In the first example, the set–up parameters are $\mu = \pi$ and $\kappa = 2$. For the second example, we take $\mu = \pi$, $\kappa = 5$, $\mu_1 = \frac{\pi}{2}$ and $\kappa_1 = 2$. In Figure 3, surface and contour plots for the first and second examples densities (left column and right column, respectively) are shown. Simulations have been also run with different parameter values, obtaining similar results.

	Estimation method			Relative efficiency	
n	Parametric	Mixed	Nonparametric	Mixed	Nonparametric
50	0.04022	0.04828	0.09773	0.83305	0.41154
200	0.01039	0.01368	0.03630	0.75950	0.28622
500	0.00425	0.00595	0.01851	0.71404	0.22960
1000	0.00213	0.00315	0.01066	0.67829	0.20056

Table 2: MISE for estimating the circular–linear density in Example 2. Relative efficiencies of mixed and nonparametric estimations with respect to parametric estimation.

Tables 1 and 2 show the simulation results for Examples 1 and 2, respectively. In all the cases, the MISE is reduced when increasing the sample size. Example 2 presents higher values for the MISE, and it is due to the estimation of a more complex structure in the circular marginal density (circular uniform in Example 1 and von Mises in Example 2). Obviously, the parametric method presents the lowest MISE values for all sample sizes in both examples, so it will be taken as a benchmark for computing the relative efficiencies of the nonparametric and mixed approaches. Relative efficiencies are obtained as the ratio between the MISE of the parametric method and the MISE of the mixed

and nonparametric procedures. The relative efficiencies (see Tables 1 and 2) are higher, in both examples, for the mixed approach, with better results for Example 2.

4 Application to wind direction and SO_2 concentration

As noticed in the introduction, the goal of this work is to explore the relation between wind incidence direction and SO_2 concentration in two monitoring stations around a power plant (see Figure 1 for stations B1 and G2 locations). SO_2 is measured in $\mu g/m^3$ and wind direction as a counterclockwise angle in $[0, 2\pi)$. With this codification, 0, $\frac{\pi}{2}$, π and $\frac{3\pi}{2}$ represent east, north, west and south direction, respectively.

The dataset contains observations recorded minutely in August 2009, but due to technical limitations in the measuring device, SO_2 is only registered when it is higher than $3\mu g/m^3$. Concentration values below this threshold are considered as non significative. Data have been hourly averaged, resulting 461 observations for B1 and 456 observations for G2. We have used a Box–Cox transformation for the SO_2 concentrations with $\lambda = -0.50$ for B1 and $\lambda = -5.65$ for G2, respectively. For the sake of simplicity, we will refer to these transformed data as SO_2 concentrations.

4.1 Data perturbation

The measurement devices, both for the wind direction and for SO_2 concentrations, did not present a sufficient precision to avoid repeated data, and this problem was inherited also for the hourly averages. The appearance of repeated measurements posed a problem in the application of the procedure, specifically, in the computation of cross–validatory bandwidths. Since repeated values in the marginals produce also repeated values in the artificial sample, data perturbation was applied to both variables. Perturbation in the linear variable, the SO_2 concentration, was carried out following Azzalini (1981). A pseudo–sample of SO_2 levels is obtained as follows:

$$\widetilde{X}_i = X_i + b\epsilon_i,$$

where X_i denote the observed values, $b = 1.3\hat{\sigma}n^{-1/3}$ and ϵ_i, $i = 1, \ldots, n$ are iid random variables from the Epanechnikov kernel in $(-\sqrt{5}, \sqrt{5})$. $\hat{\sigma}$ is a robust estimator of the variance, which has been computed using the standardized interquartile range. Azzalini (1981) shows that this choice of b for the data perturbation allows for consistent estimation of the distribution function, getting a mean squared error with the same magnitude as the one from the empirical cumulative distribution function.

The same problem of repeated measures occurred for wind direction. In this case, a perturbation procedure similar to the linear variable case was used, and the pseudo–sample of wind direction was obtained as

$$\widetilde{\theta}_i = \theta_i + d\varepsilon_i,$$

with θ_i denoting the wind direction measurements and ε_i, $i = 1, \ldots, n$ were independently generated from a von Mises distribution with $\mu = 0$ and $\kappa = 1$. The selection of the perturbation scale d, for circular data, has not been studied in the statistical literature, up to our knowledge. We have considered $d = n^{-1/5}$, based on the results of Liu and Yang (2008) for multivariate kernel distribution estimation. This perturbation scale solves the problem of repeated data and does not affect the underlying distribution. We have also noticed that the value of κ did not influence the perturbation result. See Section 5 for further discussion.

4.2 Analysis for station B1

The estimation procedure is first applied to B1, considering a nonparametric kernel density estimator for the SO_2 concentration, with biased cross–validatory bandwidth (see Figure 4, left plot). For the wind direction, circular kernel density estimation has been also used, with least–squares cross–validatory bandwidth (see Figure 4, right plot). See Wand and Jones (1995) for further details on bandwidth selection for linear kernel density estimators, and Hall *et al.* (1987) or Taylor (2008) for the circular case.

The joining circular density is computed using a circular kernel density estimator, with cross–validatory bandwidth (see Figure 5). It can be seen that the nonparametric estimator of $g(\cdot)$ is slightly different from the uniform circular density, represented with a dashed horizontal line in $(2\pi)^{-1}$ indicating a mild bivariate relationship between wind direction and SO_2 concentration. In order to check if the relation is significative, correlation coefficients have been computed and some tests for circular uniformity have been applied.

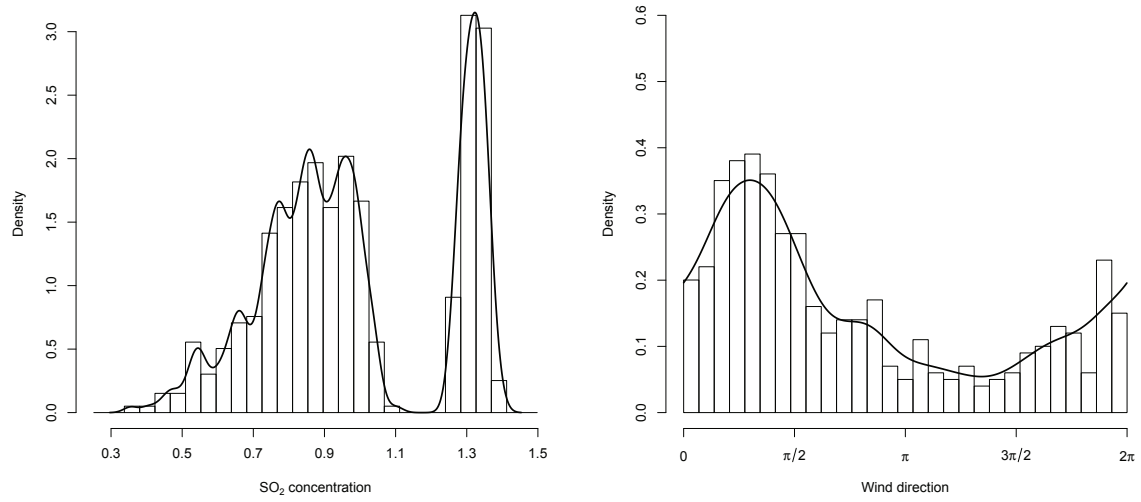

Figure 4: Monitoring station B1. Marginal densities estimators for SO_2 and wind direction, using linear and circular kernel density with cross–validatory bandwidths.

In Figure 6, the estimation of the joint density surface, with the corresponding contour plot, is shown.

Two modes can be identified, corresponding to SW direction, accordingly to Figure 2. The modes present different values of SO_2 concentrations, with the smallest one collecting lower values of SO_2.

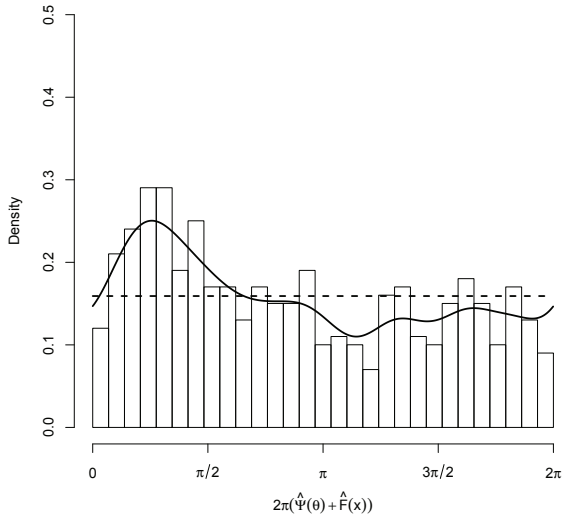

Figure 5: Monitoring station B1. Solid line: joining density estimator using circular kernel method with cross–validatory bandwidth. Dashed line: circular uniform density. Histogram for artificial sample $\{2\pi(\hat{\Psi}(\theta_i) + \hat{F}(x_i))\}_{i=1}^{n}$.

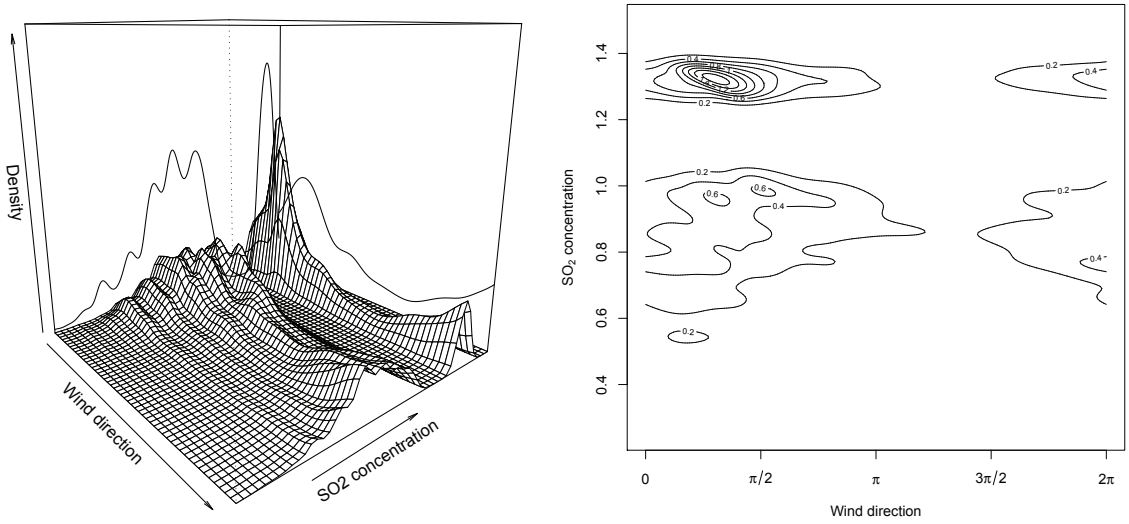

Figure 6: Monitoring station B1. Joint density estimator for wind direction and SO_2 concentrations.

Correlation coefficients between circular and linear variables were introduced by Mardia (1976). A first coefficient ρ_{cl} is defined as a multiple correlation coefficient between a the linear variable and the sine and cosine components of the circular one. Computing the circular–linear correlation coefficient for wind direction and SO_2 concentration, $\rho_{cl} = 0.1516$ is obtained. Mardia (1976) introduced

Test	Statistic	p–value
Kuiper	2.8196	< 0.01
Watson	0.6425	< 0.01
Rayleigh	0.1552	< 0.01
Rao	140.85	< 0.05

Table 3: Monitoring station B1. Circular uniformity tests for joining density.

another coefficient based on ranks, namely D_n, with values between 0 and 1, the lowest indicating independence. A test for independence, that is, taking as null hypothesis $D_n = 0$, is based on the asymptotic distribution of a rescaled coefficient, given by a χ^2_2 distribution. The rank correlation coefficient was also computed for our data, with value $D_n = 0.1422$. Performing the test for independence, we obtain a p–value smaller than 0.01. Hence, the hypothesis of independence was rejected.

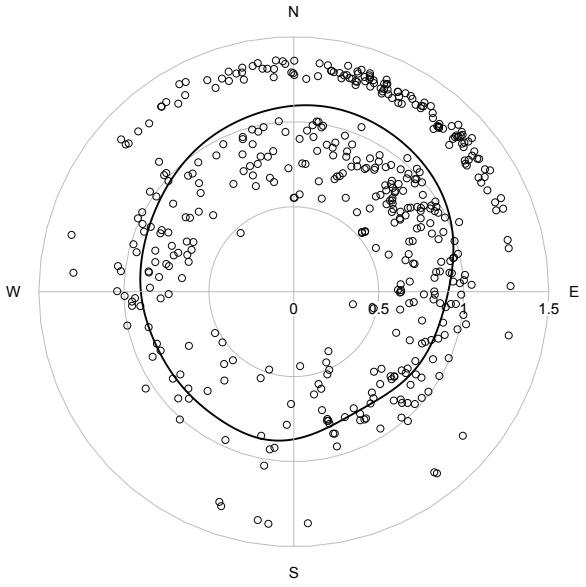

Figure 7: Monitoring station B1. Nadaraya–Watson estimator (solid line) with circular explanatory variable for SO_2 over wind direction, with cross–validatory bandwidth. SO_2 concentrations are shown in the horizontal axis.

Under formulation (3) for the joint circular–linear density, independence is equivalent to take $g(\cdot)$ as a circular uniform density. Taking into account Step 2 in the estimation algorithm (see Section 3), we construct the artificial sample $\left\{2\pi\left(\hat{\Psi}(\theta_i) + \hat{F}(x_i)\right)\right\}_{i=1}^{n}$, where $\hat{\Psi}(\cdot)$ and $\hat{F}(\cdot)$ are the circular and linear kernel distribution estimators, for the wind direction and the SO_2 concentration, respectively. Based on these artificial data, we have applied some classical uniformity test (see Jammalamadaka and SenGupta(2001)). The histogram for the artificial data can be seen in Figure 5. Results of these tests, taking circular uniform distribution as null hypothesis, are presented in Table 3. Note that,

the p–value is smaller than 0.05 (significance level) for all the tests. We can conclude that there is evidence of relation between wind direction and SO_2 concentrations.

The exploratory analysis has been also completed with the estimation of the regression function of the SO_2 over the wind direction, using a nonparametric kernel regression. In particular, a Nadaraya–Watson regression estimator, considering a circular explanatory variable with linear response, has been computed. The adaptation of local polynomial fitting to circular explanatory variables has been proposed by Di Marzio *et al.* (2009) and Nadaraya–Watson regression can be interpreted as a locally–constant polynomial fitting. Bandwidth selection for nonparametric regression has been done using a least–squares cross–validation criterion.

In Figure 7, the circular dispersion plot of SO_2 with respect to wind direction is shown, jointly with the Nadaraya–Watson regression estimator. As it has been noticed for Figure 6, there are two modes in the SW direction. It can be clearly seen that the regression function is not constant, confirming the dependence between SO_2 and wind direction. We have also tried nonparametric regression estimators that do not take into account the circular character of the wind direction, obtaining quite similar results except in directions close to the fixed circular origin, as expected.

4.3 Analysis for station G2

Station G2 is almost 20 km apart from the power plant, in the NW direction, and it shows a quite different behaviour from B1. The joining density estimation is shown in Figure 8, applying the proposed estimation algorithm with marginal kernel density estimators.

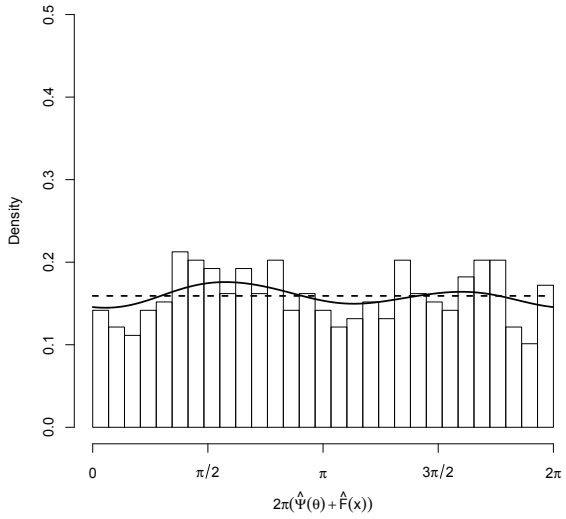

Figure 8: Monitoring station G2. Solid line: joining density estimator using circular kernel method with cross–validatory bandwidth. Dashed line: circular uniform density. Histogram for artificial sample $\{2\pi(\hat{\Psi}(\theta_i) + \hat{F}(x_i))\}_{i=1}^{n}$.

Similarly to the analysis carried out for B1, we have computed different correlation coefficients obtaining low values: $\rho_{CL} = 0.0103$, for the circular–linear correlation coefficient and $D_n = 0.0124$, for the rank correlation coefficient. Based on this correlation coefficient, the p–value for the test for independence is 0.0622. Tests for circular uniformity on the joining density have been also run, all of them showing no evidences to reject the null hypothesis of circular uniformity. Hence, there is no evidence of relation between SO_2 concentrations and wind direction in G2.

The contour plot for the estimation of the circular–linear density is shown in Figure 9 jointly with the bivariate density contour under independence. The plots are quite similar, which is not surprising. Note that the modes in this figure correspond with western winds (see Figure 2, right plot). The station is located NW from the power plant and considering it as the main emission source in the area, low levels of SO_2 concentrations are expected.

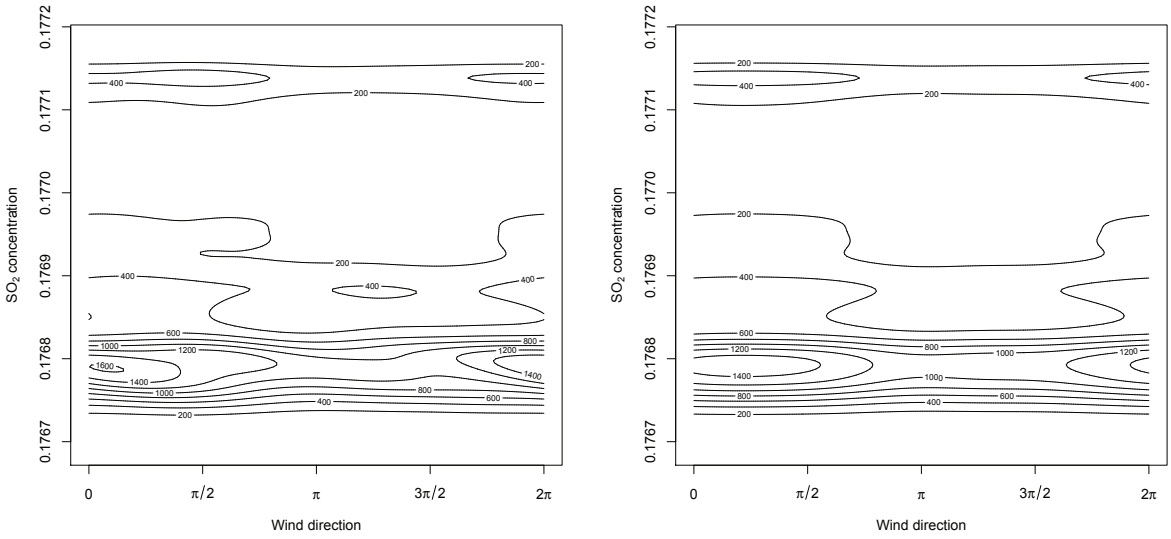

Figure 9: Monitoring station G2. Contour plot for joint density estimator for wind direction and SO_2 concentrations (left plot). Joint density estimation under independence (right plot).

5 Final comments

The circular–linear density estimation algorithm based on Johnson and Wehrly (1978) proposal allows for the introduction of flexible estimators in the marginal components. Although we have consider kernel density estimators for the marginal components, other flexible density estimators could be used. In addition, the interpretation of the circular–linear density model in terms of copulas, enables the performance of simulation studies.

A natural question that may arise is the adequacy of model (3) for a certain circular–linear bivariate variable. In the simulation study presented in this paper, both examples satisfy this condition. However, for the data analysis, we have implicitly assumed that the underlying density admits such a

representation. To the best of our knowledge, there are no suitable tests for assessing if a circular–linear density can be expressed in this way.

In real data application, the precision of measurement devices may pose some extra problems in the data analysis. In our case, the lack of precision results in the appearance of repeated values, and a data perturbation procedure was needed in order to apply the algorithm. Data perturbation for circular data needs further investigation, although it is not in the scope of this work. However, we have checked by simulations that the applied perturbation did not affected the distribution of the data. This perturbation is based on Liu and Yang (2008) results, who derive the optimal bandwidth for multivariate kernel density estimation. Another possible problem that may be encountered in practice, for linear variables, is censoring, due to detection limits or other phenomena. Under censoring, the observation values are only partially known, and suitable estimation procedures for density estimation with censored data should be applied.

The simulation study and real data analysis has been carried out in R 2.11.1 (R Development Core Team (2010)), using self–programmed code and packages circular and CircStats. The computational cost of the method is not high, and makes its application feasible in practice. For the real data analysis, the average time for B1 is 131.57 seconds, taking the computation of the joining density 75.78 seconds. For G2, the average running time for the algorithm is 127.31 seconds. Running times were measured in a regular laptop.

Acknowledgements

The authors acknowledge the support of Project MTM2008–03010, from the Spanish Ministry of Science and Project 10MDS207015PR from Dirección Xeral de I+D, Xunta de Galicia.

References

Azzalini, A. (1981) A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, **68**, 326-328.

Di Marzio, M., Panzera, A. and Taylor, C.C. (2009) Local polynomial regression for circular predictors. *Statistics and Probability Letters*, **79**, 2066-2075.

Fernández–Durán, J.J. (2004) Circular distributions based on nonnegative trigonometric sums. *Biometrics*, **60**, 499-503.

Fernández–Durán, J.J. (2007) Models for circular–linear and circular–circular data constructed from circular distributions based on nonnegative trigonometric sums. *Biometrics*, **63**, 579-585.

Hall, P., Watson, G.S. and Cabrera, J. (1987) Kernel density estimation with spherical data. *Biometrika*, **74**, 751-762.

Jammalamadaka, S.R. and Lund, U.J. (2006) The effect of wind direction on ozone levels: a case study. *Environmental and Ecological Statistics*, **13**, 287-298.

Jammalamadaka, S.R. and SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific Press. Singapore.

Johnson, M.E. (1987). *Multivariate Statistical Simulation*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons. New York.

Johnson, R. A. and Wehrly, T. (1978). Some angular–linear distributions and related regression models. *Journal of the American Statistical Association*, **73**, 602-606.

Liu, R. and Yang, L. (2008) Kernel estimation of multivariate cumulative distribution function. *Journal of Nonparametric Statistics*, **20**, 661-677.

Mardia, K.V. (1976) Linear–circular correlation and rhythmometry. *Biometrika*, **63**, 403-405.

Mardia, K.V. and Jupp, P.E: (2000) *Directional Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons. Chichester.

Nelsen, R.B. (2006) *An Introduction to Copulas*. Springer Series in Statistics. Springer–Verlag.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Austria.

Somerville, M. C., Mukerjee, S., Fox, D. L. and Stevens, R. K. (1994). Statistical approaches in wind sector analyses for assessing local source impacts. *Atmospheric Environment*, **28**, 3483-3493.

Somerville, M.C., Mukerjee, S. and Fox, D.L. (1996) Estimating the wind direction of maximum air pollutant concentration. *Environmetrics*, **7**, 231-243.

Taylor, C.C. (2008) Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis*, **52**, 3493-3500.

Wand, M.P. and Jones, M.C. (1995) *Kernel smoothing*. Chapman and Hall Ltd. London.

Wehrly, T. and Johnson, R.A. (1980) Bivariate models for dependence of angular observations and a related Markov process. *Biometrika*, **67**, 255-256.

Reports in Statistics and Operations Research

2005

05-01 SiZer Map for Evaluating a Bootstrap Local Bandwidth Selector in Nonparametric Additive Models. M. D. Martínez-Miranda, R. Raya-Miranda, W. González-Manteiga and A. González-Carmona.

05-02 The Role of Commitment in Repeated Games. I. García Jurado, Julio González Díaz.

05-03 Project Games. A. Estévez Fernández, P. Borm, H. Hamers.

05-04 Semiparametric Inference in Generalized Mixed Effects Models. M. J. Lombardía, S. Sperlich.

2006

06-01 A unifying model for contests: effort-prize games. J. González Díaz.

06-02 The Harsanyi paradox and the "right to talk" in bargaining among coalitions. J. J. Vidal Puga.

06-03 A functional analysis of NOx levels: location and scale estimation and outlier detection. M. Febrero, P. Galeano, W. González-Manteiga.

06-04 Comparing spatial dependence structures. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.

06-05 On the spectral simulation of spatial dependence structures. R. M. Crujeiras, R. Fernández-Casal.

06-06 An L_2-test for comparing spatial spectral densities. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.

2007

07-01 Goodness-of-fit tests for the spatial spectral density. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.

07-02 Presmothed estimation with left truncated and right censores data. M. A. Jácome, M. C. Iglesias-Pérez.

07-03 Robust nonparametric estimation with missing data. G. Boente, W. González-Manteiga, A. Pérez-González.

07-04 k-Sample test based on the common area of kernel density estimators. P. Martínez-Camblor, J. de Uña Álvarez, N. Corral-Blanco.

07-05 A bootstrap based model checking for selection-biased data. J. L. Ojeda, W. González-Manteiga, J . A. Cristobal.

07-06 The Gaussian mixture dynamic conditional correlation model: Bayesian estimation, value at risk calculation and portfolio selection. P. Galeano, M. C. Ausín.

2008

08-01 ROC curves in nonparametric location-scale regression models. W. González-Manteiga, J. C. Pardo Fernández, I. Van Keilegom.

08-02 On the estimation of α-convex sets. B. Pateiro-López, A. Rodríguez-Casal.

2009

09-01 Lasso Logistic Regression, GSoft and the Cyclyc Coordinate Descent Algorithm. Application to Gene Expression Data. M. García-Magariños, A. Antoniadis, R. Cao, W. González-Manteiga.

2010

10-01 Asymptotic behaviour of robust estimators in partially linear models with missing responses: The effect of estimating the missing probability on simplified marginal estimators. A. Bianco, G. Boente, W. González-Manteiga, A. Pérez-González.

10-02 First-Price Winner-Takes-All Contents. J. González-Díaz.

10-03 Goodness of Fit Test for Interest Rate Models: an approach based on Empirical Process. A. E. Monsalve-Cobis, W. González-Manteiga, M. Febrero-Bande.

2011

11-01 Exploring wind direction and SO_2 concentration by circular–linear density estimation. E. García–Portugués, R.M. Crujeiras, W. González–Manteiga.

Previous issues (2001 – 2003):
http://eio.usc.es/reports.php