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

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

Department of Statistics and Operations Research
University of Santiago de Compostela

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

Air pollution studies

- Investigation of the relation between pollutant concentrations from monitoring sites and the emission sources.

Air pollution studies

- Investigation of the relation between pollutant concentrations from monitoring sites and the emission sources.
- Circular variables (wind direction) play a relevant role.

Air pollution studies

- Investigation of the relation between pollutant concentrations from monitoring sites and the emission sources.
- Circular variables (wind direction) play a relevant role.
- Some previous works:

Air pollution studies

- Investigation of the relation between pollutant concentrations from monitoring sites and the emission sources.
- Circular variables (wind direction) play a relevant role.
- Some previous works:
  - Somerville *et al* (1996): Estimation of the wind direction of maximum air pollutant concentration and identification of emission sources.

Air pollution studies

- Investigation of the relation between pollutant concentrations from monitoring sites and the emission sources.
- Circular variables (wind direction) play a relevant role.
- Some previous works:
  - Somerville *et al* (1996): Estimation of the wind direction of maximum air pollutant concentration and identification of emission sources.
  - Jammalamadaka and Lund (2006), Fernández–Durán (2007): Wind direction and ozone levels.

Air pollution studies

- Investigation of the relation between pollutant concentrations from monitoring sites and the emission sources.
- Circular variables (wind direction) play a relevant role.
- Some previous works:
    - Somerville *et al* (1996): Estimation of the wind direction of maximum air pollutant concentration and identification of emission sources.
    - Jammalamadaka and Lund (2006), Fernández–Durán (2007): Wind direction and ozone levels.
- We focus on sulphur dioxide (SO_2) pollutants.

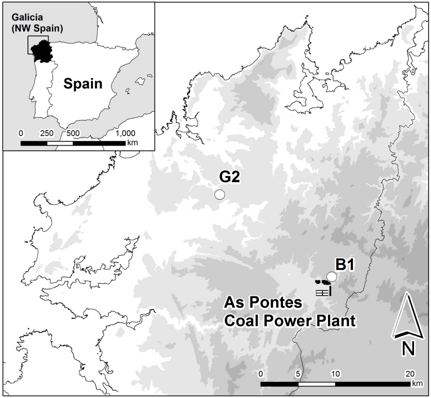


Figure: Locations of monitoring stations and power plant.

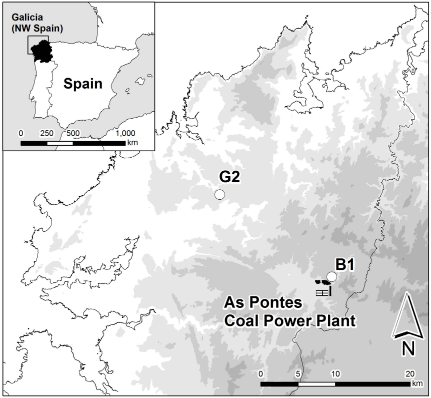


Figure: Locations of monitoring stations and power plant.

Distances to power plant

- B1: 0.9 km
- G2: 18.6 km

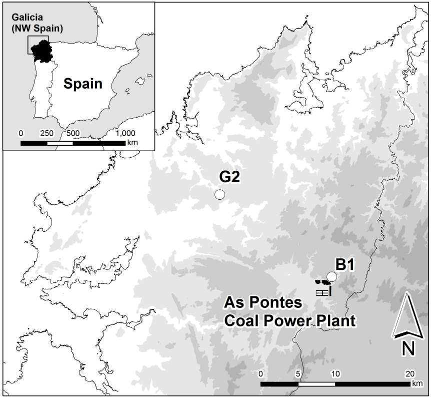


Figure: Locations of monitoring stations and power plant.

Distances to power plant

- B1: 0.9 km
- G2: 18.6 km

Goal of the work

Study wind direction and SO_2 concentration relation.

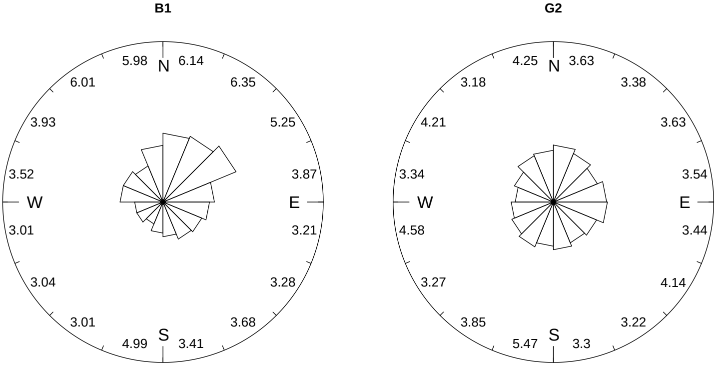


Figure: Rose diagrams for wind direction stations B1 and G2, with average SO_2 concentrations for August 2009.

Definition (Mardia and Jupp, 2000)

A circular random variable Θ *has its support in* $\mathbb{S}^1$*.*

Definition (Mardia and Jupp, 2000)

A circular random variable Θ has its support in $\mathbb{S}^1$. In the a.c. case, its density f_Θ must satisfy:

1. $f_\Theta(\theta) \geq 0,\ \forall \theta \in \mathbb{R}$.
2. $\int_r^{2\pi+r} f_\Theta(\theta) d\theta = 1,\ \forall r \in \mathbb{R}$.
3. $f_\Theta(\theta) = f_\Theta(\theta + 2\pi k),\ \forall \theta \in \mathbb{R},\ \forall k \in \mathbb{Z}$.

Definition (Mardia and Jupp, 2000)

A circular random variable Θ has its support in $\mathbb{S}^1$. In the a.c. case, its density f_Θ must satisfy:

1. $f_\Theta(\theta) \geq 0, \forall \theta \in \mathbb{R}$.
2. $\int_r^{2\pi+r} f_\Theta(\theta) d\theta = 1, \forall r \in \mathbb{R}$.
3. $f_\Theta(\theta) = f_\Theta(\theta + 2\pi k), \forall \theta \in \mathbb{R}, \forall k \in \mathbb{Z}$.

Example (von Mises distribution $vM(\mu, \kappa)$)

The von Mises distribution has density

$$\varphi_{vM}(\theta; \mu, \kappa) = (2\pi I_0(\kappa))^{-1} \exp\left[\kappa \cos(\theta - \mu)\right],$$

where $\mu \in [0, 2\pi)$ is the circular mean, $\kappa \geq 0$ is the concentration in μ direction. Its distribution is denoted by Ψ_{vM}.

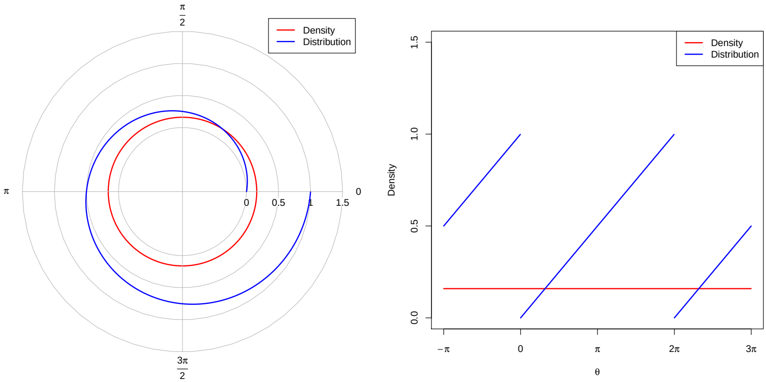


Figure: Circular and linear representations of the density and distribution of a von Mises with $\kappa = 0$ (**circular uniform distribution**).

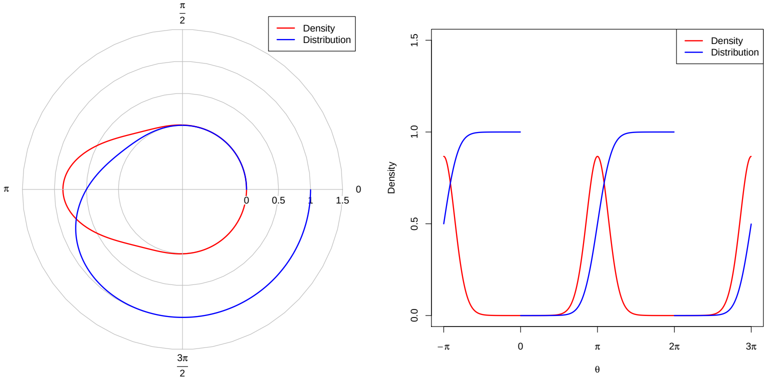
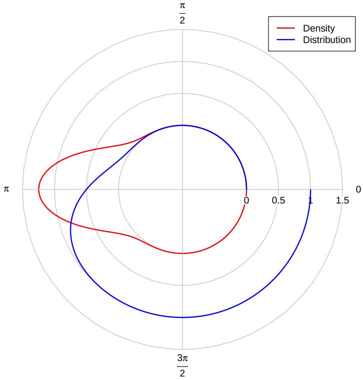


Figure: Circular and linear representations of the density and distribution of a von Mises with $\mu = \pi$ and $\kappa = 5$.



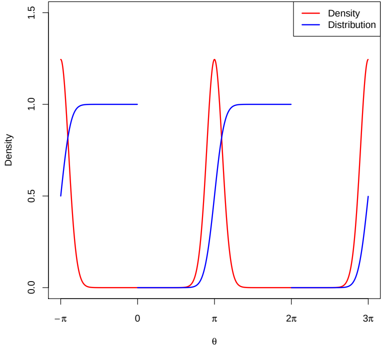


Figure: Circular and linear representations of the density and distribution of a von Mises with $\mu = \pi$ and $\kappa = 10$.

Denote by

- Θ a circular variable with density φ and distribution Ψ.
- X a linear variable with density f and distribution F.

Denote by

- Θ a circular variable with density φ and distribution Ψ.
- X a linear variable with density f and distribution F.

Joint distribution of (Θ, X) ?

Denote by

- Θ a circular variable with density φ and distribution Ψ.
- X a linear variable with density f and distribution F.

Joint distribution of (Θ, X) ?

Theorem (Johnson and Wehrly, 1978)

Let g be a circular density. Then

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

is a circular–linear density with marginal densities φ and f.

Denote by

- Θ a circular variable with density φ and distribution Ψ.
- X a linear variable with density f and distribution F.

Joint distribution of (Θ, X) ?

Theorem (Johnson and Wehrly, 1978)

Let g be a circular density. Then

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

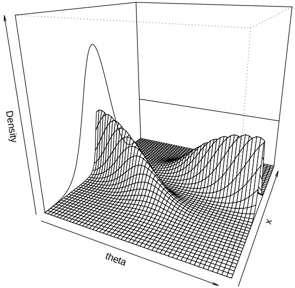
is a circular–linear density with marginal densities φ and f.

- It is a construction of p from φ and f, not a characterization.
- Θ and X independent $\Leftrightarrow g(\omega) = (2\pi)^{-1}$, $\forall\omega \in [0, 2\pi)$

Example (Circular uniform and Normal marginal densities)

Densities: $\varphi = (2\pi)^{-1}$, $f = \phi$ and $g = \varphi_{vM}(\mu, \kappa)$.

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



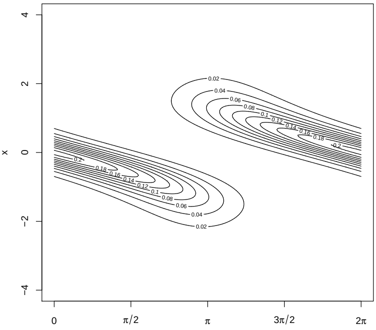


Figure: Joint density p_1 with $\mu = \pi$ and $\kappa = 2$.

Example (Circular uniform and Normal marginal densities)

Densities: $\varphi = (2\pi)^{-1}$, $f = \phi$ and $g = \varphi_{vM}(\mu, \kappa)$.

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

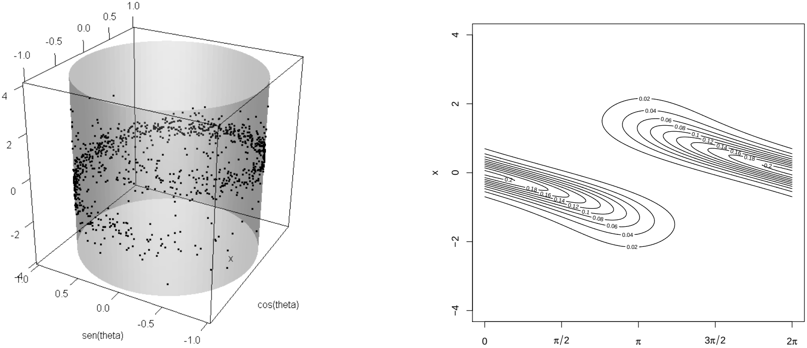
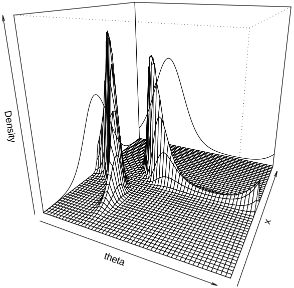


Figure: Joint density p_1 with $\mu = \pi$ and $\kappa = 2$.

Example (von Mises and Normal marginal densities)

Densities: $\varphi = \varphi_{vM}(\mu_1, \kappa_1)$, $f = \phi$ and $g = \varphi_{vM}(\mu, \kappa)$.

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



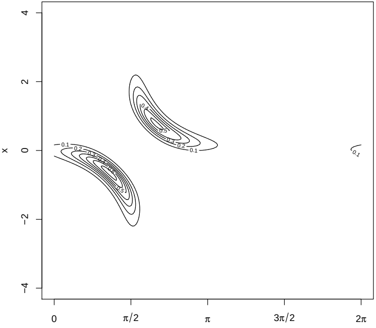


Figure: Joint density p_2 with $\mu_1 = \frac{\pi}{2}$, $\kappa_1 = 2$, $\mu = \pi$ and $\kappa = 5$.

Example (von Mises and Normal marginal densities)

Densities: $\varphi = \varphi_{vM}(\mu_1, \kappa_1)$, $f = \phi$ and $g = \varphi_{vM}(\mu, \kappa)$.

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

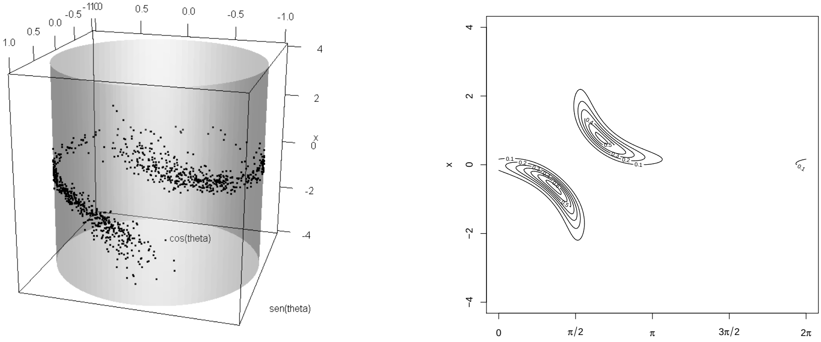


Figure: Joint density p_2 with $\mu_1 = \frac{\pi}{2}$, $\kappa_1 = 2$, $\mu = \pi$ and $\kappa = 5$.

Our model

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

Our model

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

Estimation algorithm

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

Our model

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

Estimation algorithm

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

Our model

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

Estimation algorithm

1. Obtain estimators for the marginal densities $\hat{\varphi}$, $\hat{f}$ and the corresponding marginal distributions $\hat{\Psi}$, $\hat{F}$.
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 $\hat{g}$.
3. Obtain the circular–linear density estimator as

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

Estimation approaches

- **Parametric**. Estimate parametrically φ, f and g, for example by ML. Fernández–Durán (2007) estimates the model using ML for the linear density and Nonnegative Trigonometric Sums for the circular densities.

Estimation approaches

- **Parametric**. Estimate parametrically φ, f and g, for example by ML. Fernández–Durán (2007) estimates the model using ML for the linear density and Nonnegative Trigonometric Sums for the circular densities.
- **Mixed**. Estimate φ and f parametrically (some intuition) and g nonparametrically (no intuition) or other possible combinations.

Estimation approaches

- **Parametric**. Estimate parametrically φ, f and g, for example by ML. Fernández–Durán (2007) estimates the model using ML for the linear density and Nonnegative Trigonometric Sums for the circular densities.
- **Mixed**. Estimate φ and f parametrically (some intuition) and g nonparametrically (no intuition) or other possible combinations.
- **Nonparametric**. Estimate nonparametrically both marginals φ and f and the joining density g by kernel smoothing.

Kernel estimation

- Let f be a linear density and $X_1, \ldots, X_n$ a sample from $X \sim f$. The kernel estimator of f is

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

Kernel estimation

- Let f be a linear density and $X_1, \ldots, X_n$ a sample from $X \sim f$. The kernel estimator of f is

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

- For a circular density φ and a sample $\Theta_1, \ldots, \Theta_n$, the kernel estimator defined by Hall, Watson and Cabrera (1987) is

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

- In the linear case,

$$h_{\mathsf{AMISE}} = \mathcal{O}\left(n^{-\frac{1}{5}}\right).$$

- In the linear case,

$$h_{\mathsf{AMISE}} = \mathcal{O}\left(n^{-\frac{1}{5}}\right).$$

- For the circular case, Taylor (2008) shows that for the von Mises distribution,

$$\nu_{\mathsf{AMISE}} = \mathcal{O}\left(n^{\frac{2}{5}}\right).$$

- In the linear case,

$$h_{\mathsf{AMISE}} = \mathcal{O}\left(n^{-\frac{1}{5}}\right).$$

- For the circular case, Taylor (2008) shows that for the von Mises distribution,

$$\nu_{\mathsf{AMISE}} = \mathcal{O}\left(n^{\frac{2}{5}}\right).$$

- The circular bandwidth parameter ν behaves as $1/h^2$.

- In the linear case,

$$h_{\mathsf{AMISE}} = \mathcal{O}\left(n^{-\frac{1}{5}}\right).$$

- For the circular case, Taylor (2008) shows that for the von Mises distribution,

$$\nu_{\mathsf{AMISE}} = \mathcal{O}\left(n^{\frac{2}{5}}\right).$$

- The circular bandwidth parameter ν behaves as $1/h^2$.
- Large values of ν undersmooth and small ones oversmooth (inverse behaviour of h).

- In the linear case,

$$h_{\mathsf{AMISE}} = \mathcal{O}\left(n^{-\frac{1}{5}}\right).$$

- For the circular case, Taylor (2008) shows that for the von Mises distribution,

$$\nu_{\mathsf{AMISE}} = \mathcal{O}\left(n^{\frac{2}{5}}\right).$$

- The circular bandwidth parameter ν behaves as $1/h^2$.
- Large values of ν undersmooth and small ones oversmooth (inverse behaviour of h).
- A possible choice of ν is by LSCV

$$\nu_{\mathsf{LSCV}} = \arg\min_{\kappa\geq 0} \int \hat{f}_\kappa(\omega)^2 d\omega - \frac{2}{n}\sum_{i=1}^n \hat{f}_\kappa^{-i}(\Theta_i).$$

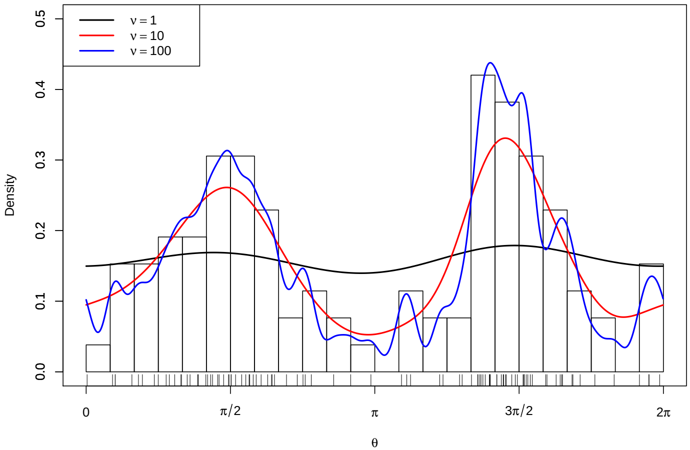


Figure: Effects of the circular bandwidth in the density estimator. Sample of size $n = 100$ from an equal mixture of $vM(\frac{\pi}{2}, 2)$ and $vM(\frac{3\pi}{2}, 5)$.

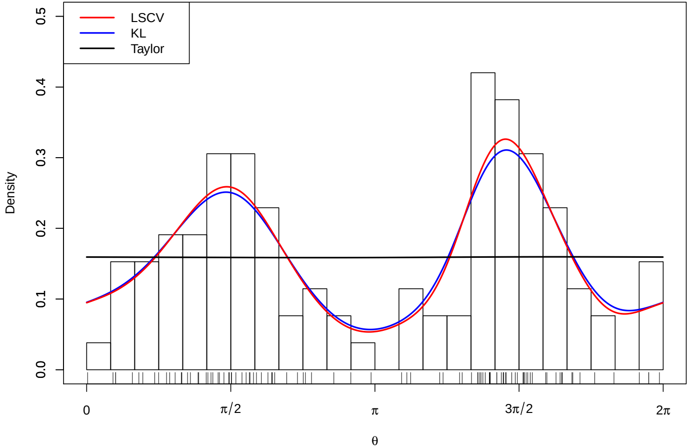
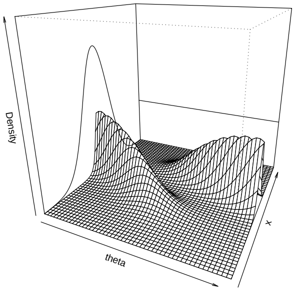


Figure: LSCV, KL and Taylor bandwidths. Sample of size $n = 100$ from an equal mixture of $vM(\frac{\pi}{2}, 2)$ and $vM(\frac{3\pi}{2}, 5)$.

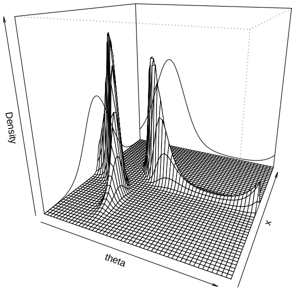
Example 1

- φ a circular uniform.
- $f \sim \mathcal{N}(0,1)$.
- $g \sim vM(\pi, 2)$.



Example 2

- $\varphi \sim vM\left(\frac{\pi}{2}, 2\right)$.
- $f \sim \mathcal{N}(0,1)$.
- $g \sim vM(\pi, 5)$.



Our model

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

Our model

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

- In terms of copulas, can be expressed as

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

Our model

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

- In terms of copulas, can be expressed as

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

- Copula formulation helps for random simulation from p_1 and p_2 (our examples).

Definition

A copula C is a bivariate distribution function with uniform marginals. It allows to express joint distributions in terms of marginal distributions.

Definition

A copula C is a bivariate distribution function with uniform marginals. It allows to express joint distributions in terms of marginal distributions.

Sklar's Theorem

Let X, Y two random variables with joint distribution F and marginals F_1 and F_2. There exists a copula C such that

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

Definition

A copula C is a bivariate distribution function with uniform marginals. It allows to express joint distributions in terms of marginal distributions.

Sklar's Theorem

Let X, Y two random variables with joint distribution F and marginals F_1 and F_2. There exists a copula C such that

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

Sklar's Theorem in terms of densities:

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

Definition

A copula C is a bivariate distribution function with uniform marginals. It allows to express joint distributions in terms of marginal distributions.

Sklar's Theorem

Let X, Y two random variables with joint distribution F and marginals F_1 and F_2. There exists a copula C such that

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

Sklar's Theorem in terms of densities:

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

Our model:

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

Copula simulation (for our examples)

Copula simulation (for our examples)

Consider a circular–linear variable (Θ, X) with joint distribution $P = C_{\Theta,X}(\Psi, F)$.

Copula simulation (for our examples)

Consider a circular–linear variable (Θ, X) with joint distribution $P = C_{\Theta,X}(\Psi, F)$.

Simulation from (Θ, X)

1. Simulate $(U, V) \sim C_{\Theta,X}$ (U and V are uniforms).
2. Compute $\Theta = \Psi^{-1}(U)$ and $X = F^{-1}(V)$.
3. $(\Theta, X) \sim P$.

Copula simulation (for our examples)

Consider a circular–linear variable (Θ, X) with joint distribution $P = C_{\Theta,X}(\Psi, F)$.

Simulation from (Θ, X)

1. Simulate $(U, V) \sim C_{\Theta,X}$ (U and V are uniforms).
2. Compute $\Theta = \Psi^{-1}(U)$ and $X = F^{-1}(V)$.
3. $(\Theta, X) \sim P$.

- Simulation by copulas is easier due to the structure of p.
- The copula density is $c(u, v) = 2\pi g(2\pi(u + v))$.

Simulation setting

- **Parametric**: Maximum Likelihood.
- **Mixed**: φ and f by ML and g by kernel estimation with LSCV bandwidth.
- **Nonparametric**: Linear and circular kernel estimation with linear BCV bandwidth and circular LSCV bandwidths.

Simulation setting

- **Parametric**: Maximum Likelihood.
- **Mixed**: φ and f by ML and g by kernel estimation with LSCV bandwidth.
- **Nonparametric**: Linear and circular kernel estimation with linear BCV bandwidth and circular LSCV bandwidths.

- Other bandwidth selectors: Seather & Jones (linear); Kullblack–Leibler and Taylor (circular). Similar results.

Simulation setting

- **Parametric**: Maximum Likelihood.
- **Mixed**: φ and f by ML and g by kernel estimation with LSCV bandwidth.
- **Nonparametric**: Linear and circular kernel estimation with linear BCV bandwidth and circular LSCV bandwidths.

- Other bandwidth selectors: Seather & Jones (linear); Kullblack–Leibler and Taylor (circular). Similar results.
- Sample sizes: $n = 50, 200, 500, 1000$. Samples generated using copula simulation.

Error criterion

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

Error criterion

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

- MISE aproximated by Monte Carlo with $M = 1000$ replicates.

Error criterion

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

- MISE aproximated by Monte Carlo with $M = 1000$ replicates.
- Benchmark: Parametric model.

Error criterion

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

- MISE aproximated by Monte Carlo with $M = 1000$ replicates.
- Benchmark: Parametric model.
- Relative MISE efficiencies for Mixed and Nonparametric approaches.

	n	Estimation method: Param.	Estimation method: Mixed	Estimation method: Nonpar.	Relative efficiency: Mixed	Relative efficiency: Nonpar.
Example 1	50	0.0054	0.0095	0.0168	0.5674	0.3208
	200	0.0013	0.0029	0.0052	0.4802	0.2483
	500	0.0005	0.0012	0.0025	0.4267	0.2078
	1000	0.0003	0.0007	0.0014	0.3897	0.1840
Example 2	50	0.0402	0.0483	0.0977	0.8331	0.4115
	200	0.0104	0.0137	0.0363	0.7595	0.2862
	500	0.0043	0.0060	0.0185	0.7140	0.2296
	1000	0.0021	0.0032	0.0107	0.6783	0.2006

Table: MISE for estimating the circular–linear density in Example 1 and 2.

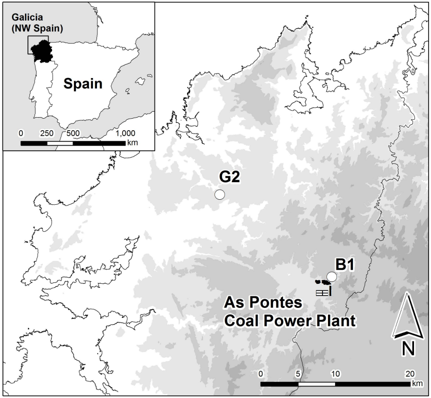


Figure: Locations of monitoring stations and power plant.

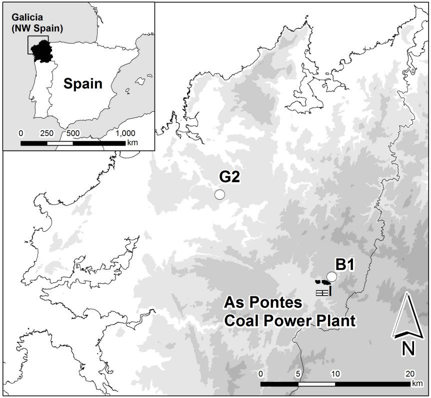


Figure: Locations of monitoring stations and power plant.

Raw data

- SO_2 measured in $\mu g/m^3$. Detection limit: $> 3\mu g/m^3$.
- Wind direction.

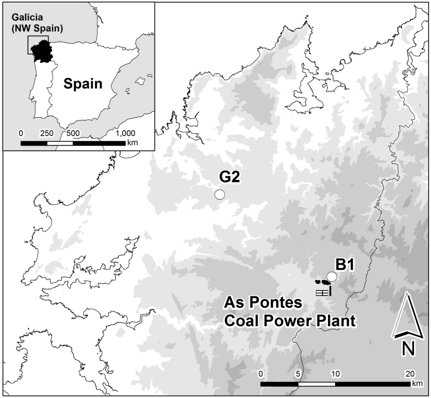


Figure: Locations of monitoring stations and power plant.

Raw data

- SO_2 measured in $\mu g/m^3$. Detection limit: $> 3\mu g/m^3$.
- Wind direction.

Our data

- Hourly averaged SO_2 and wind direction (circular mean).
- Perturbation to avoid repeated data.
- Box–Cox in SO_2.

How to proceed with repeated data?

- Linear case: Azzalini (1981) proposes a perturbation that allows a consistent estimation of the distribution:

$$\widetilde{X}_i = X_i + b\varepsilon_i,\ \varepsilon_i \sim \text{Epanech}\left(-\sqrt{5}, \sqrt{5}\right),$$

How to proceed with repeated data?

- Linear case: Azzalini (1981) proposes a perturbation that allows a consistent estimation of the distribution:

$$\widetilde{X}_i = X_i + b\varepsilon_i,\ \varepsilon_i \sim \text{Epanech}\left(-\sqrt{5}, \sqrt{5}\right),$$

where $b = C^* n^{-\delta}$. Optimum choice of δ is $\frac{1}{3}$, derived from $b_{\text{AMISE}} = \mathcal{O}(n^{-\frac{1}{3}})$. C^* is chosen as $1.3\hat{\sigma}$.

How to proceed with repeated data?

- Linear case: Azzalini (1981) proposes a perturbation that allows a consistent estimation of the distribution:

$$\widetilde{X}_i = X_i + b\varepsilon_i,\ \varepsilon_i \sim \text{Epanech}\left(-\sqrt{5}, \sqrt{5}\right),$$

where $b = C^* n^{-\delta}$. Optimum choice of δ is $\frac{1}{3}$, derived from $b_{\text{AMISE}} = \mathcal{O}(n^{-\frac{1}{3}})$. C^* is chosen as $1.3\hat{\sigma}$.

- Circular case: **open problem**. Our perturbation:

$$\widetilde{\theta}_i = \theta_i + d\varepsilon_i,\ \varepsilon_i \sim vM(0,1),$$

How to proceed with repeated data?

- Linear case: Azzalini (1981) proposes a perturbation that allows a consistent estimation of the distribution:

$$\widetilde{X}_i = X_i + b\varepsilon_i,\ \varepsilon_i \sim \text{Epanech}\left(-\sqrt{5}, \sqrt{5}\right),$$

where $b = C^* n^{-\delta}$. Optimum choice of δ is $\frac{1}{3}$, derived from $b_{\text{AMISE}} = \mathcal{O}(n^{-\frac{1}{3}})$. C^* is chosen as $1.3\hat{\sigma}$.

- Circular case: **open problem**. Our perturbation:

$$\widetilde{\theta}_i = \theta_i + d\varepsilon_i,\ \varepsilon_i \sim vM(0,1),$$

with $d = n^{-\frac{1}{5}}$. Analogy with the bidimensional ($\mathbb{S}^1$) linear $b_{\text{AMISE}} = \mathcal{O}(n^{-\frac{1}{5}})$ (Liu and Yang, 2008).

Are wind direction and SO_2 independent?

Are wind direction and SO_2 independent?

Circular–linear correlation coefficients (Mardia, 1976):

- ρ_{CL}: R^2 for $X \sim \cos(\Theta) + \sin(\Theta)$.
- D_n: ranks correlation. Test for $H_0 : D_n = 0$.

Are wind direction and SO_2 independent?

Circular–linear correlation coefficients (Mardia, 1976):

- ρ_{CL}: R^2 for $X \sim \cos(\Theta) + \sin(\Theta)$.
- D_n: ranks correlation. Test for $H_0 : D_n = 0$.

Our model

Θ and X independent $\Leftrightarrow g(\omega) = (2\pi)^{-1}, \forall \omega \in [0, 2\pi)$

Are wind direction and SO_2 independent?

Circular–linear correlation coefficients (Mardia, 1976):

- ρ_{CL}: R^2 for $X \sim \cos(\Theta) + \sin(\Theta)$.
- D_n: ranks correlation. Test for $H_0 : D_n = 0$.

Our model

Θ and X independent $\Leftrightarrow g(\omega) = (2\pi)^{-1}, \forall \omega \in [0, 2\pi)$

Uniformity tests (Mardia and Jupp, 2000):

- Kuiper: Kolmogorov–type test.
- Watson: Cramer–von Mises test.
- Rayleigh: Alternative hypothesis is a unimodal distribution.
- Rao's Spacing test.

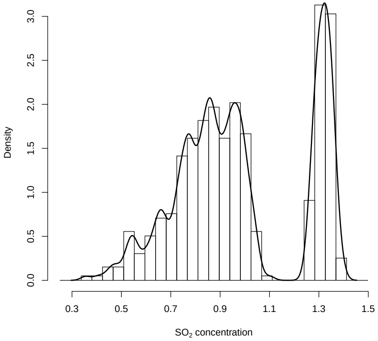


Figure: SO_2 concentration (Box–Cox) in B1.

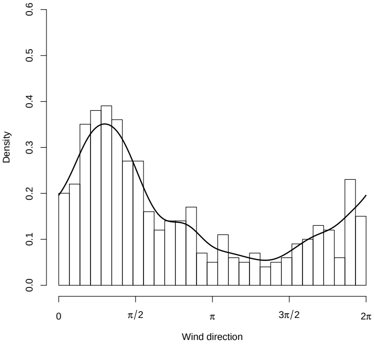


Figure: Wind direction in B1.

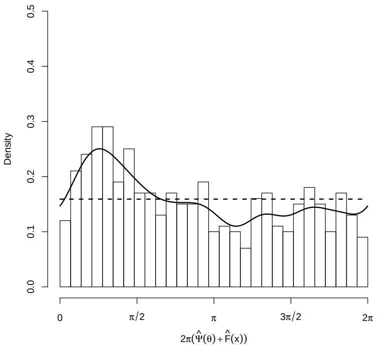


Figure: Estimation of g in B1.

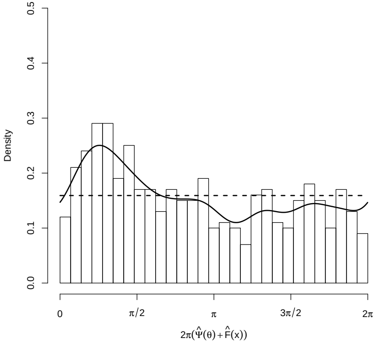


Figure: Estimation of g in B1.

Test	Statistic	p–value
Kuiper	2.8196	< 0.01
Watson	0.6425	< 0.01
Rayleigh	0.1552	< 0.01
Rao	140.8554	< 0.05

Table: Uniformity tests for g.

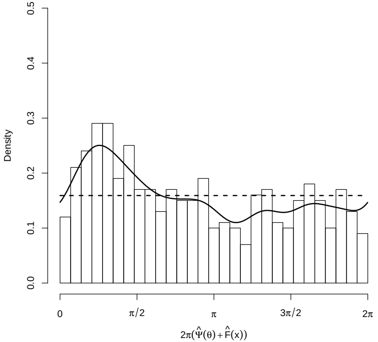


Figure: Estimation of g in B1.

Test	Statistic	p–value
Kuiper	2.8196	< 0.01
Watson	0.6425	< 0.01
Rayleigh	0.1552	< 0.01
Rao	140.8554	< 0.05

Table: Uniformity tests for g.

Circular–linear correlation

- $\rho_{CL} = 0.1515$.
- $D_n = 0.1422$ with p–value=0.

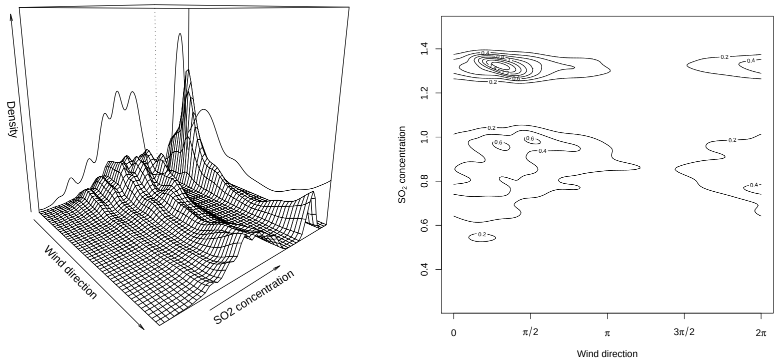


Figure: Surface and contourplot of the estimated circular–linear density in B1.

Another exploratory tool: circular regression.

- Consider a circular regression model:

$$Y = m(\Theta) + \varepsilon, \quad m(\theta) = \mathbb{E}\left(Y|\Theta = \theta\right)$$

with ε a zero–mean variable independent from Θ.

Another exploratory tool: circular regression.

- Consider a circular regression model:

$$Y = m(\Theta) + \varepsilon, \quad m(\theta) = \mathbb{E}\left(Y | \Theta = \theta\right)$$

with ε a zero–mean variable independent from Θ.

- The circular Nadaraya–Watson, with von Mises kernel, is

$$\hat{m}(\theta; \nu) = \frac{\sum_{i=1}^{n} y_i \cdot \varphi_{vM}(\theta - \theta_i; 0, \nu)}{\sum_{i=1}^{n} \varphi_{vM}(\theta - \theta_i; 0, \nu)}$$

Another exploratory tool: circular regression.

- Consider a circular regression model:

$$Y = m(\Theta) + \varepsilon, \quad m(\theta) = \mathbb{E}\left(Y|\Theta = \theta\right)$$

  with ε a zero–mean variable independent from Θ.

- The circular Nadaraya–Watson, with von Mises kernel, is

$$\hat{m}(\theta;\nu) = \frac{\sum_{i=1}^n y_i \cdot \varphi_{vM}(\theta - \theta_i; 0, \nu)}{\sum_{i=1}^n \varphi_{vM}(\theta - \theta_i; 0, \nu)}$$

- Possible selection of ν:

$$\nu_{\text{LSCV}} = \arg\min_{\kappa \geq 0} \sum_{i=1}^n \left(y_i - \hat{m}^{-i}(\theta_i;\kappa)\right)^2$$

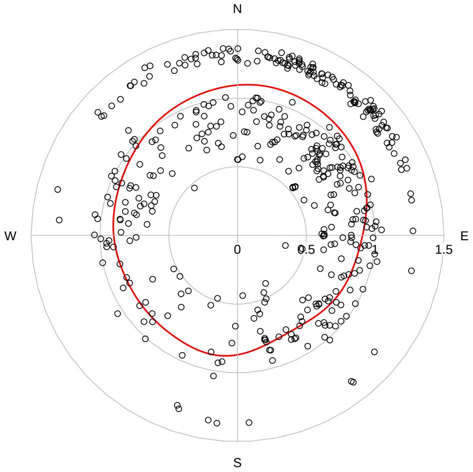


Figure: Circular regression of SO_2 (Box–Cox) in wind direction for B1.

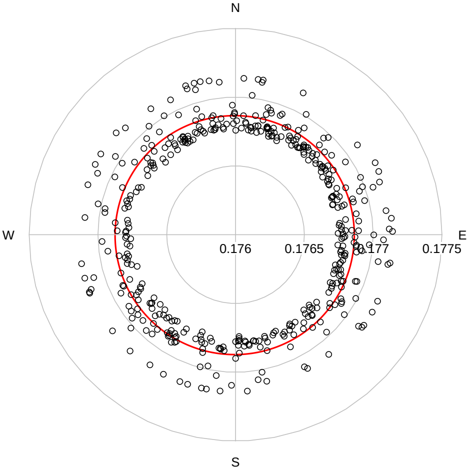


Figure: Circular regression of SO_2 (Box–Cox) in wind direction for G2.

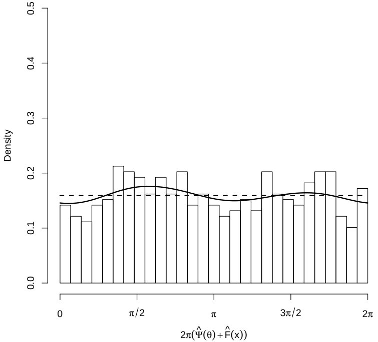


Figure: Estimation of g in G2.

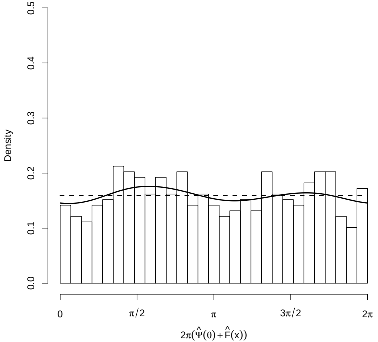


Figure: Estimation of g in G2.

Test	Statistic	p–value
Kuiper	1.2042	> 0.15
Watson	0.0748	> 0.10
Rayleigh	0.0259	0.737
Rao	130.7370	> 0.10

Table: Uniformity tests for g.

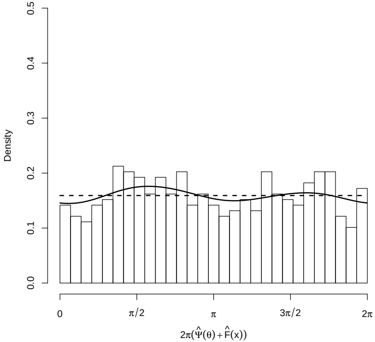


Figure: Estimation of g in G2.

Test	Statistic	p–value
Kuiper	1.2042	> 0.15
Watson	0.0748	> 0.10
Rayleigh	0.0259	0.737
Rao	130.7370	> 0.10

Table: Uniformity tests for g.

Circular–linear correlation

- $\rho_{CL} = 0.0103$.
- $D_n = 0.0124$ with p–value$=0.0622$.

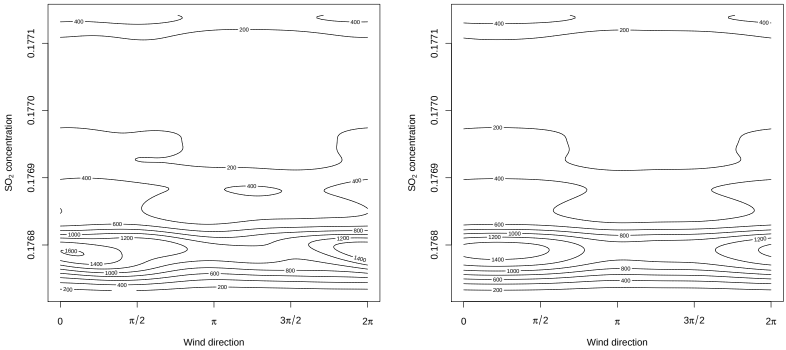


Figure: Right: contourplot of the estimated density in G2. Left: contourplot under independence.

Conclusions

- B1:
  - Moderate dependence between wind direction and SO_2.
  - Higher SO_2 concentrations linked to the NE and N wind, opposite direction to the power plant.
- G2: independence between wind direction and SO_2.

Open problems

1. Circular data perturbation.
2. Goodness–of–fit test for the Johnson and Wehrly family of circular–linear distributions.

References

A. Azzalini. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68(1):326–328, 1981.

P. Hall, G. S. Watson, and J. Cabrera. Kernel density estimation with spherical data. *Biometrika*, 74(4):751–762, 1987.

R. A. Johnson and T. E. Wehrly. Some angular-linear distributions and related regression models. *J. Amer. Statist. Assoc.*, 73(363):602–606, 1978.

K. V. Mardia and P. E. Jupp. *Directional statistics*. John Wiley & Sons Ltd., Chichester, 2000.

R. B. Nelsen. *An introduction to copulas*. Springer, New York, 2006.

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

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

Department of Statistics and Operations Research
University of Santiago de Compostela

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGACIÓN OPERATIVA