Applied Mathematical Sciences, Vol. 20, 2026, no. 4, 203 - 212
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ams.2026.919338

Circular Classification of Air Quality Regimes from Imperfect Wind-Direction Measurements

Marco Di Marzio

University of Chieti-Pescara, Italy

Stefania Fensore

University of Chieti-Pescara, Italy

Chiara Passamonti

University of Luxembourg, Luxembourg

Abstract

Air pollution is a very critical issue that directly impacts our quality of life. It occurs when air pollutants are present in the atmosphere at high concentrations, surpassing normal levels and significantly degrading our living conditions. Considering the impact of air pollution on human and environmental health, it is crucial to investigate about the levels of air pollutants and their dispersion in the different sites. This phenomenon is strictly related to wind direction because wind is responsible of the dispersion, concentration, and transportation of pollutants in the atmosphere. In this context we pursue a classification task of circular data observed with measurement error and apply a nonparametric statistical discrimination rule to a real dataset involving wind direction and two specific air pollutants. We distinguish two levels of air quality according to the World Health Organization's Air quality guidelines and take into account the amount of particulate matter and nitrogen dioxide pollutants.

Mathematics Subject Classification: 62G05

Keywords: Air Quality Guidelines, Classification task, Directional data, Measurement error

1 Introduction

Air pollution is a complex mix of solid particles, liquid droplets, and gases, stemming from various sources such as household fuel burning, industrial emissions, traffic exhaust, and agricultural practices. They include $PM_{2.5}$ and PM_{10} (particles with an aerodynamic diameter of equal or less than 2.5, and 10 micrometer respectively), ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO) and sulfur dioxide (SO_2). The higher concentrations of these pollutants significantly impact the health of urban populations. Collecting data on wind speed and direction is crucial to comprehensively understand air quality dynamics in any region. This topic has been extensively studied in the literature. Among others, [1] analyzed the relation between wind speed/direction and the concentration/dispersion of particulate matter PM_{10}. To investigate the relationships between meteorological variables and air pollutants, the World Health Organization (WHO) established some Air Quality Guidelines (AQG) which play a crucial role in setting global targets for governments to work towards reducing air pollution and, consequently, improving public health. We link this topic to a circular classification task. Discrimination problems can be tackled through various strategies. The classical, nonparametric one based on kernel density estimation consists on estimating the difference between two densities to get a likelihood decision rule, see for example [13] and [4].

Circular data are observations defined on the circumference of a unit circle, i.e. *angles*. They arise in many scientific fields, for example, biology, geology, meteorology, social sciences, etc. Since they are measured using a periodic scale, ad-hoc methodologies are required to deal with this particular kind of data. Recent and detailed overviews of advances in directional statistics are [16] and [14]. Clearly, discrimination problems could arise in the context of circular data. Indeed, several contributions are available in the literature about this topic, see, among others, [17], [15], [11], [6] and [7].

We discuss this task presenting the case of two groups of angles contaminated by measurement errors. Many reasons can determine the presence of a measurement error, for example, the deterioration or mis-calibration of measurement devices or difficulties in measuring the variables of interest. This is the well-known *errors-in-variables* case. It occurs in various scientific fields, therefore it has been widely studied in the literature, see [2], [3], [8] and [9].

Let Θ and ε be two circular random variables, respectively, referring to the characteristic of interest (which is not observable) and the random measurement error (whose density f_ε is assumed to be symmetric around 0). The available random sample is described by the model

$$\Phi_i = (\Theta_i + \varepsilon_i)\mathsf{mod}(2\pi), \tag{1}$$

where, for $i = 1, \ldots, n$, the ε_is are independent copies of a random angle ε, independent of Θ. The classical approach used to deal with data affected by measurement errors appearing in model (1) is the deconvolution one.

The paper is organized as follows. In Section 2 we discuss the classification rule to discriminate between two groups of data. In Section 3 we present the estimation process for circular densities both in the classical and errors-in-variables cases. Finally, in Section 4 we include a real data case study concerning air pollution and wind direction.

2 Statistical discrimination for labeling air polluted regions

Typically, the regional air quality is evaluated using the Air Quality Index (AQI), a numerical scale that indicates the current level of air pollution. The AQI scale provides six classes of air quality. In this paper we treat the air pollution as a binary variable. Indeed, according to the AQG, we distinguish only two levels of air quality (*low risk* and *high risk*) and link them to wind direction formulating a classification problem.

Considering that wind direction could be measured with error, we discuss the discrimination rule taking into account the possible presence of measurement error (for more details see [8]). Nonparametric classification is typically performed by involving kernel density estimation. We aim to discriminate between two populations described by the unknown circular densities f_1 and f_2. In the error-free case two random samples respectively drawn from these densities are used, while in the errors-in-variables context data from these populations are not available, therefore two random samples defined according to model (1) are employed. A general discrimination rule is based on a classifier given by the difference from the kernel density estimators of f_1 and f_2, which allocates the observation θ to the population described by f_1 if it is greater than or equal to 0 at $\theta \in (0, 2\pi]$.

Formally, letting $\dot{g}(\theta; \kappa_1, \kappa_2)$ be a generic kernel estimator of $g(\theta) = f_1(\theta) - f_2(\theta)$, and assuming n_1 (contaminated) observations from f_1 and $n_2 = n - n_1$ ones from f_2, a discriminant rule can be defined by assigning θ to the first population if

$$\dot{g}(\theta; \kappa_1, \kappa_2) := \dot{f}_1(\theta; \kappa_1) - \dot{f}_2(\theta; \kappa_2) \geq 0. \tag{2}$$

This topic has been studied in the Euclidean error-free case by [4] and in the circular errors-in-variables case by [10]. In the following we discuss the way to estimate the densities f_1 and f_2 appearing in formula (2).

3 Wind direction populations estimation

Wind direction is a circular random variable because the results of the measurements are angles. Therefore, the wind direction populations estimation task requires a statistical methodology tailored for circular data.

Let $\Theta_1, \ldots, \Theta_n$ be a random sample of angles from an unknown circular probability density function f_Θ. The kernel estimator of f_Θ at $\theta \in [0, 2\pi)$ (see [12] and [5]) is defined as

$$\hat{f}_\Theta(\theta; \kappa) = \frac{1}{n} \sum_{i=1}^{n} K_\kappa(\Theta_i - \theta), \tag{3}$$

where K_κ is a circular kernel, i.e. a periodic, unimodal, symmetric density function with concentration parameter $\kappa > 0$, which admits a convergent Fourier series representation as follows

$$K_\kappa(\theta) = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{\ell=1}^{\infty} \gamma_\ell(\kappa) \cos(\ell\theta) \right\}. \tag{4}$$

The role of the kernel function in the estimation process is to give more weight to the observations which are in a neighbourhood of the estimation point θ. The width of that neighbourhood is determined by the smoothing parameter κ, which plays the role of the inverse of the *Euclidean* bandwidth, in the sense that smaller values of κ give wider neighbourhoods.

Letting $\eta_j(K_\kappa) := \int_0^{2\pi} K_\kappa(u) \sin^j(u) du$, a circular kernel K_κ is said to be a second sin-order kernel if $\eta_0(K_\kappa) = 1$, $\eta_1(K_\kappa) = 0$ and $\eta_2(K_\kappa) = \frac{1}{2}(1 - \gamma_2(\kappa)) \neq 0$. Classical examples of second sin-order kernels include the von Mises density, the wrapped Normal and the wrapped Cauchy densities.

Now, assume that we deal with an errors-in-variables problem, where the goal is to estimate the density f_Θ of a random angle Θ, when only i.i.d. observations $\Phi_1, \ldots, \Phi_n$ defined according to model (1) are available. In this classical measurement errors problem the density f_Φ of Φ is the circular convolution of f_Θ and f_ε. Then, estimating f_Θ amounts to a *circular density deconvolution*, and has been addressed by [8] who defined the following kernel estimator for f_Θ at $\theta \in [0, 2\pi)$

$$\tilde{f}_\Theta(\theta; \kappa) := \frac{1}{n} \sum_{i=1}^{n} \widetilde{K}_\kappa(\Phi_i - \theta), \tag{5}$$

where the weight function

$$\widetilde{K}_\kappa(\phi) := \frac{1}{2\pi} \left\{ 1 + 2 \sum_{\ell=1}^{\infty} \frac{\gamma_\ell(\kappa)}{\lambda_\ell(\kappa_\varepsilon)} \cos(\ell\phi) \right\} \tag{6}$$

is a circular deconvolution kernel, with $\gamma_\ell(\kappa)$ and $\lambda_\ell(\kappa_\epsilon) \neq 0$ (for each ℓ) respectively being the ℓ-th Fourier coefficient of a circular kernel K_κ and the ℓ-th Fourier coefficient of the error density f_ε (with concentration κ_ε).

Note that $\tilde{f}_\Theta$ has the form of a classical kernel density estimator $\hat{f}_\Theta$, and its generic Fourier coefficient is given by the ratio between the Fourier coefficient of the kernel K_κ and the error one (see formulae (3) and (5) to compare the estimators, and (4) and (6) to compare the kernels). As for the order of the kernel, we have that $\widetilde{K}_\kappa$ retains the same sin-order of K_κ.

The asymptotic bias and variance of the circular density estimators (3) and (5), generically denoted by $\dot{f}$, are collected as follows. Consider estimator $\dot{f}_\Theta(\theta;\kappa)$ employing a second sin-order kernel K_κ with Fourier coefficients $\gamma_\ell(\kappa)$, $\ell \in \mathbb{Z}^+$, aimed to estimate f_Θ. If *i*) f_Θ is twice continuously differentiable in a neighbourhood of $\theta \in [0, 2\pi)$, and *ii*) κ increases with n in such a way that, for any $\ell \in \mathbb{Z}^+$, $\lim_{n\to\infty} \gamma_\ell(\kappa) = 1$, and $\lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{\infty} \gamma_\ell^2(\kappa) = 0$, then

$$\mathsf{B}[\dot{f}_\Theta(\theta;\kappa)] = \frac{(1-\gamma_2(\kappa))}{2}\frac{f_\Theta^{(2)}(\theta)}{2} + o(1-\gamma_2(\kappa)),$$

$$\mathsf{VAR}[\dot{f}_\Theta(\theta;\kappa,\kappa_\varepsilon)] = \frac{(1 + 2\sum_{\ell=1}^{\infty} \alpha_\ell^2(\kappa;\kappa_\varepsilon))\, f_\Theta(\theta)}{2\pi n} + o\left(\frac{\sum_{\ell=1}^{\infty} \alpha_\ell^2(\kappa;\kappa_\varepsilon)}{n}\right),$$

where the coefficient $\alpha_\ell(\kappa;\kappa_\varepsilon)$ appearing in the variance formulation is $\gamma_\ell(\kappa)$ for estimator (3) and $\gamma_\ell(\kappa)/\lambda_\ell(\kappa_\varepsilon)$ for estimator (5). Note that the bias of the deconvolution estimator is the same as the bias of the error-free case estimator.

4 A real data case study

Starting from the Air Quality Guidelines provided by the World Health Organization, that are reported in Table 1, we aim to classify wind direction according to the level of air pollutant. Specifically, we use both the standard kernel estimator (which does not take into account the presence of the measurement error) and the deconvolution one, that are discussed in Section 3 and employed in classifier (2). We consider data from New South Wales Air Qual-

Table 1: WHO recommended 2021 AQG levels of pollutants.

Pollutant	2021 AQG level
$PM_{2.5}$, $\mu g/m^3$	5
PM_{10}, $\mu g/m^3$	15
O_3, $\mu g/m^3$	60
NO_2, $\mu g/m^3$	10
SO_2, $\mu g/m^3$	40
CO, mg/m^3	4

ity Monitoring Network (`https://www.airquality.nsw.gov.au`). We choose a site located at Western Sydney, Earlwood, and select hourly data from July to December 2021, and, to reduce serial correlation, we retain alternate days. Figure 1 shows the wind directions according to the level of pollutant.

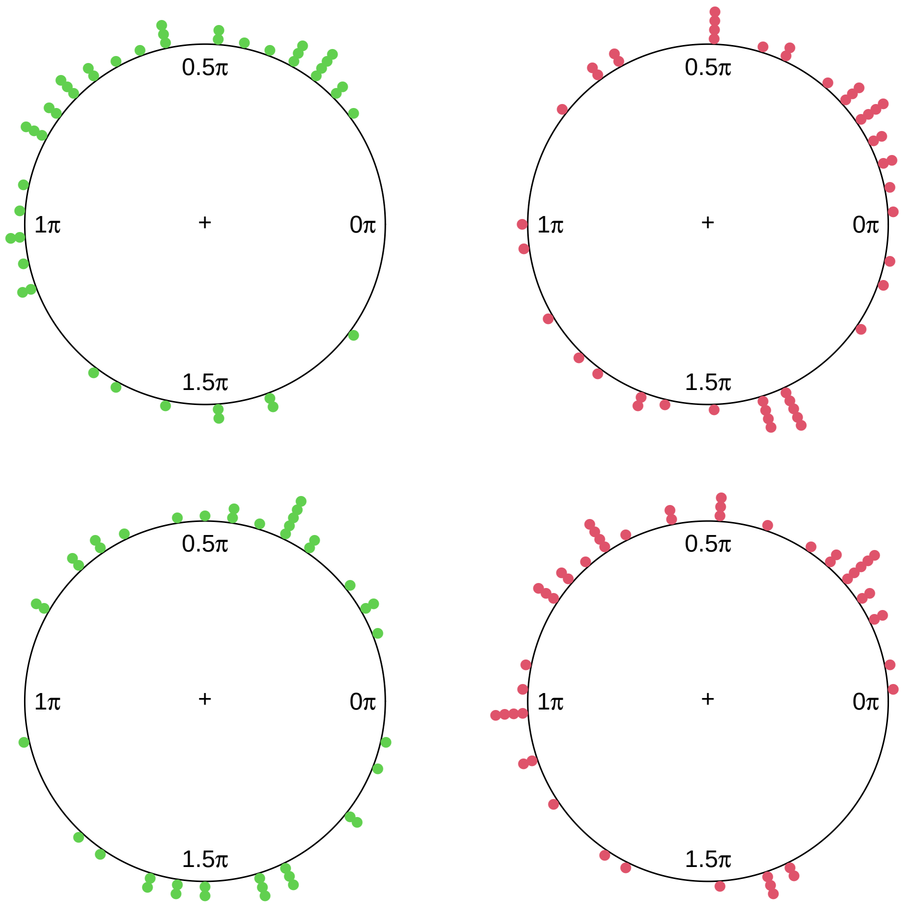

Figure 1: Top: wind direction groups of size $n_1 = 44$ and $n_2 = 47$ in correspondence of *low* (green) and *high* (red) levels of NO_2 pollutant. Bottom: wind direction groups of size $n_1 = 42$ and $n_2 = 48$ in correspondence of *low* (green) and *high* (red) levels of $PM_{2.5}$ pollutant.

Due to the fact that air pollution in a region strongly depends on wind direction, in our analysis, we consider a bivariate dataset including wind direction and a specific pollutant.

We make two groups of wind directions, respectively, of sizes n_1 and n_2 according to the related level of air pollutant.

In order to distinguish between *high* and *low* levels of air pollution, we refer to the thresholds reported by the World Health Organization's Air quality guidelines.

We perform two separated classifications where we take into account the amount of pollutant of particulate matter ($PM_{2.5}$), first, and nitrogen dioxide (NO_2), then.

We are allowed to use a deconvolution classifier to discriminate between the two groups of observations because wind direction is typically affected by measurement errors.

As for the error density, we hypothesize different scenarios assuming two possible error models, i.e. the wrapped Laplace (wL) and the von Mises (vM) ones.

The concentration parameters of the error densities are chosen in such a way that the variability of the measurement error is comparable between the two considered scenarios.

The kernel used is the von Mises density. The smoothing degrees are selected by using a cross validation criterion minimizing the Kullback–Leibler loss ($\hat{\kappa}_1 = 5.8$ and $\hat{\kappa}_2 = 26.7$ for NO_2; $\hat{\kappa}_1 = 4.1$ and $\hat{\kappa}_2 = 21.9$ for $PM_{2.5}$).

For the implementation of the deconvolution classifier, the number of Fourier coefficients is equal to 20. Results, reported in terms of misclassification rates, are collected in Table 2.

We can note that, regardless of the measurement error model, the deconvolution method provides an advantage ranging from 10% to 25% with respect to the standard circular kernel density classifier in terms of misclassification rate, for both the $PM_{2.5}$ and NO_2 air pollutants.

Table 2: Misclassification rates (%) for the two pollutants, obtained using the standard circular kernel classifier $\hat{g}$ and the deconvolution one $\tilde{g}$.

Pollutant	f_ε	$\hat{g}$	$\tilde{g}$
$PM_{2.5}$	wL	32.2	28.9
	vM	32.2	25.6
NO_2	wL	29.7	25.3
	vM	29.7	22.0

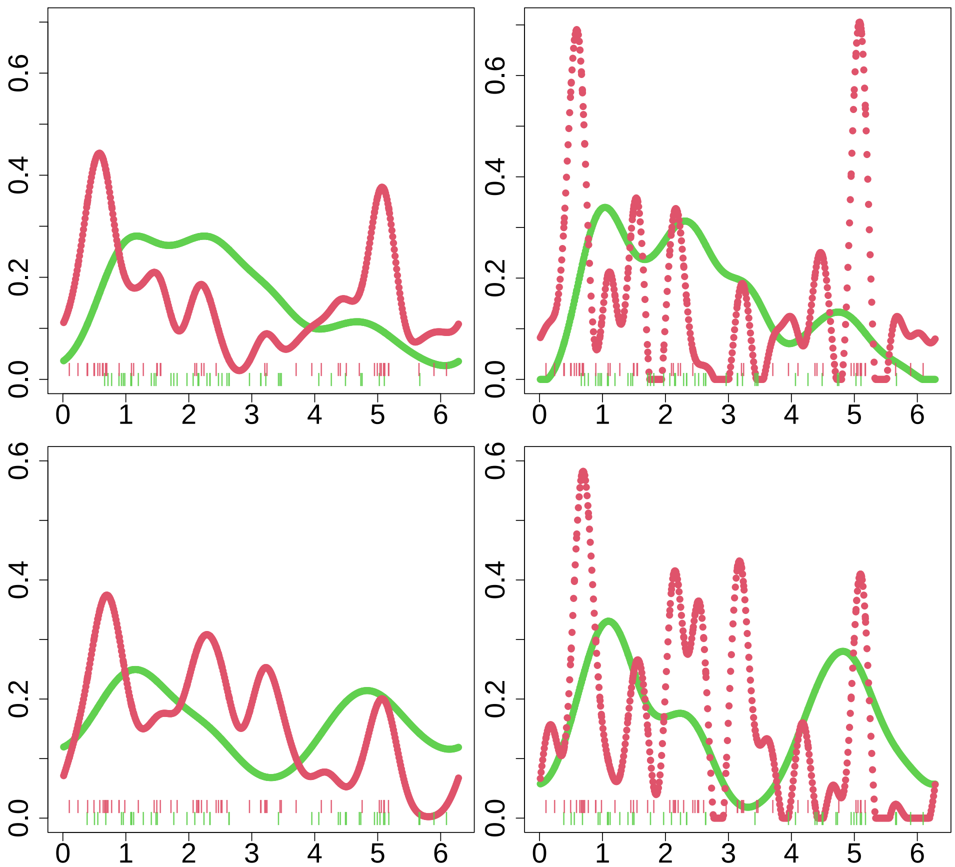

Figure 2: Kernel density estimates using classical (left) and deconvoluted (right) estimators of two groups, *low* (green) and *high* (red) levels of NO_2 (top) and $PM_{2.5}$ (bottom) pollutants. Vertical bars denote the observations.

For illustrative purposes, we report in Figure 2 the density estimates obtained for the two groups of wind directions using both the estimators in the scenario where the wrapped Laplace error is assumed. We can see that when the deconvolution classifier is employed the estimates are less smooth, and, as the consequence, more observations are correctly classified.

References

[1] Cichowicz, R., Wielgosinski, G., Fetter, W., Effect of wind speed on the level of particulate matter PM10 concentration in atmospheric air during winter season in vicinity of large combustion plant, *J. Atmos. Chem.*, **77** (1) (2020), 35–48. https://doi.org/10.1007/s10874-020-09401-w

[2] Delaigle, A., An alternative view of the deconvolution problem, *Stat. Sin.*, (2008), 1025–1045.

[3] Delaigle, A., Nonparametric kernel methods with errors-in-variables: constructing estimators, computing them, and avoiding common mistakes, *Aust. N. Z. J. Stat.*, **56** (2) (2014), 105–124. https://doi.org/10.1111/anzs.12066

[4] Di Marzio, M., Taylor, C. C., Kernel density classification and boosting: an L_2 analysis, *Stat. Comput.*, **15** (2005), 113–123. https://doi.org/10.1007/s11222-005-6203-8

[5] Di Marzio, M., Panzera, A., Taylor, C. C., Kernel density estimation on the torus, *J. Stat. Plan. Inference*, **141** (6) (2011), 2156–2173. https://doi.org/10.1016/j.jspi.2011.01.002

[6] Di Marzio, M., Fensore, S., Panzera, A., Taylor, C. C., Kernel density classification for spherical data, *Stat. Probab. Lett.* **144** (2019), 23–29. https://doi.org/10.1016/j.spl.2018.07.018

[7] Di Marzio, M., Fensore, S., Panzera, A., Taylor, C. C., Local binary regression with spherical predictors, *Stat. Probab. Lett.*, **144** (2019), 30–36. https://doi.org/10.1016/j.spl.2018.07.019

[8] Di Marzio, M., Fensore, S., Panzera, A., Taylor, C. C., Density estimation for circular data observed with errors, *Biometrics* **78** (1) (2022), 248–260. https://doi.org/10.1111/biom.13431

[9] Di Marzio, M., Fensore, S., Taylor, C. C., Kernel regression for errors-in-variables problems in the circular domain, *Stat. Methods Appl.*, **32** (4) (2023), 1217–1237. https://doi.org/10.1007/s10260-023-00687-0

[10] Di Marzio, M., Fensore, S., Panzera, A., & Passamonti, C., Low-bias discrimination of circular data with measurement errors, *Advances in Data Analysis and Classification*, (2025), 1-20. https://doi.org/10.1007/s11634-025-00655-4

[11] Fernandes, K., Cardoso, J. S., Discriminative directional classifiers, *Neurocomputing*, **207** (2016), 141–149. https://doi.org/10.1016/j.neucom.2016.03.076

[12] Hall, P., Watson, G. S., Cabrera, J., Kernel density estimation with spherical data, *Biometrika*, **74** (4) (1987), 751–762. https://doi.org/10.1093/biomet/74.4.751

[13] Hall, P., Wand, M. P., On nonparametric discrimination using density differences, *Biometrika*, **75** (3) (1988), 541–547. https://doi.org/10.1093/biomet/75.3.541

[14] Ley, C., Verdebout, T. (eds.), *Applied Directional Statistics*, Modern Methods and Case Studies, CRC Press, 2018.
https://doi.org/10.1201/9781315228570

[15] Lopez-Cruz, P. L., Bielza, C., Larranaga, P., Directional naive Bayes classifiers, *Pattern Anal. Appl.*, **18** (2015), 225–246.
https://doi.org/10.1007/s10044-013-0340-z

[16] Mardia, K. V., Jupp, P. E., *Directional Statistics*, Wiley, 2009.
https://doi.org/10.1002/9780470316979

[17] SenGupta, A., Roy, S., A simple classification rule for directional data, *Advances in Ranking and Selection, Multiple Comparisons, and Reliability: Methodology and Applications*, (2005), 81–90. https://doi.org/10.1007/0-8176-4422-9_5

Received: June 15, 2026; Published: July 14, 2026