FUNCTIONAL DATA ANALYSIS APPROACH FOR IDENTIFYING REDUNDANCY IN AIR QUALITY MONITORING STATIONS

Annalina Sarra [1], Adelia Evangelista[1], Tonio Di Battista[1] and Sergio Palermi[2]

[1] Department of Philosophical, Pedagogical and Economic-Quantitative Sciences , University of Chieti-Pescara, (e-mail: annalina.sarra@unich.it, adelia.evangelista@unich.it, tonio.dibattista@unich.it)

[2] Agency of Environmental Protection of Abruzzo (ARTA), Pescara, Italy, (e-mail: s.palermi@artaabruzzo.it)

ABSTRACT: The assessment of air quality is of great importance for defining measures for pollution reduction and ensuring the public health protection. The monitoring stations are the tools established to measure and manage the compliance with national ambient air quality standards. Because these networks need considerable financial resources, many studies are aimed at identifying possible redundancy in air quality monitoring sites. Following these lines of research, we focus on ascertaining if the spatial distributions of NO_2, PM_{10}, $PM_{2.5}$ and benzene concentrations are homogenously distributed in the urban area of Pescara-Chieti (Central Italy). To this end we adopt a multivariate functional model-based clustering algorithm.

KEYWORDS: air quality, redundancy, meteorological normalization, FDA, model-based clustering.

1 Introduction

In recent decades there has been a growing interest in monitoring air pollution levels, especially in urban areas. Countries all over the world have set up air quality monitoring networks for collecting unbiased, accurate and comparable data on the air quality and supporting policies that lessen the impact on human health and the environment. In order to save money and avoid data duplication, it is preferable to use the fewest number of stations possible to meet monitoring goals. There are numerous studies in the literature that look for potential redundancy in air quality monitoring networks (see Wilson *et al.*, 2005 for a review). The majority of them concentrate on determining whether or not the pollutant is uniformly distributed throughout the area and on the intra-urban

variation of air pollutant concentrations. In this study, we address the problem of identifying possible redundancy in air quality monitoring stations using the FDA (Ramsay & Silverman, 2005) paradigm. FDA has gained considerable interest in the literature over the past two decades, and several benefits of using FDA over conventional vectorial approaches have been emphasized, such as the possibility to extract more information from the data (the smoothness of the data structure, rate of change, acceleration, and dynamic changes over a large-scale domain). In this work, we analyze the multivariate air pollution concentrations using a multivariate functional model-based clustering approach proposed by Schmutz *et al.*, 2020. The data set used is comprised of hourly measurements of air quality and weather data obtained from the automatic reporting platform operated by the Regional Agency for Environmental Protection of the Abruzzo Region (ARTA), in the urban area of Pescara-Chieti (Central Italy). We also implement a meteorological normalization to control for changes in the weather and lower the variability in air quality time series. The remainder of this paper is structured as follows. Section 2 describes the study area and the data used for the analysis, as well as the meteorological normalization procedure conducted. Section 3 provides background information on the functional clustering algorithm employed. Finally, Section 4 conveys the main findings of the analysis.

2 Study area and data

The study focused on the Chieti-Pescara urban area in the Abruzzo Region (Central Italy), which includes the conurbation of the major cities Pescara and Chieti, and the neighboring municipalities of Montesilvano and Francavilla al Mare. It is a nearly flat area located in the terminal stretch (about 15 km long) of the Pescara river valley, which flows into the Adriatic Sea. The valley industrial and vehicular traffic are the main contributors to air pollution, with domestic heating having a sizable impact during the winter. For this study, we take into account NO_2, PM_{10}, $PM_{2.5}$ and benzene measurements obtained from ARTA automatic reporting platform between January 2017 and December 2019 at 5 five monitoring sites divided into two categories: urban background (3 sites: Teatro d'Annunzio, Chieti, and Francavilla) and urban traffic (2 sites: Via Firenze and Montesilvano). The dataset also includes the following meteorological factors: wind speed, wind direction, temperature, relative humidity, solar radiation, air pressure and precipitation, measured on the ground at air quality monitoring stations. Since weather strongly influences pollutants formation and transport, in this paper we consider a meteorologi-

cal/weather normalization. More specifically, in our air quality data analysis over time, we control for changes of meteorology by means of boosted regression trees, as implemented in the R package deweather (Carslaw, 2021).

3 Model based clustering algorithm

The main steps involved by the model-based clustering algorithm for high-dimensional data (fun-HDDC) introduced by Schmutz *et al.*, 2020 can be summarized as follows. Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are the observed multivariate curves, representing in our case the air quality data. The goal is to group them into K homogenous clusters, where K is fixed a priori. The core idea is to transform the high-dimensional data into group-specific subspaces. For each group k ($k = 1, \dots K$), let $d_k < R$ denote the intrinsic dimension of a low dimensional latent subspace in which the curve of each cluster could be described. Through a principal component analysis for multivariate functional data, curves are expressed into a group-specific basis

$$\varphi_r^k(t) = \sum_{l=1}^{R} q_{krl}\phi_l(t), 1 \leq r \leq R \tag{1}$$

obtained through a linear transformation from the matrix of principal factors $\left\{\phi_r^j\right\}_{1\leq j\leq p, 1\leq r\leq R}$ where q_{krl} are the basis expansion coefficients of the eigenfunctions, contained in an orthogonal matrix $R \times R$. Thus, each multivariate curve n_k, of cluster k, can be represented by its score $(\delta_i^k)_{1\leq i\leq n_k}$.
The scores are assumed to follow a Gaussian distribution $\delta_i^k \sim N(\mu_k, \Delta_K)$ with $\mu_k \in R^R$ the mean function and Δ_K the corresponding covariance matrix. Actually, the novel approach (fun-HDDC) is an extension of the work of Jacques & Preda, 2014, and it is advantageous from two perspectives: modeling all principal component scores with estimated variances that are not zero, and proposing a criterion for choosing the number of clusters using the expectation-maximization (EM) algorithm.

4 Results

In this section, we present the results obtained through the use of the algorithm illustrated in Section 3. All the analyses were performed using the R packages *fda* and *funHDDC* (R Core Team, 2022). The observed pollutant time series and the meteorological normalized pollutant time series have been

transformed into functional data with a process of smoothing, with 30 basis and cubic B-spline. In either instances, the model-based clustering algorithm applied here is the [AkjBQkDk] model (see, Schmutz *et al.*, 2020 for more details) and provided the partition of monitoring stations into two groups. We find out that the composition of the identified groups does not change after performing a meteorological normalization: the first cluster contains the monitoring stations of Chieti and Francavilla whereas the remaining monitoring sites of Via Firenze, Montesilvano and Teatro d'Annunzio are grouped in the second cluster. For NO_2, PM_{10} and $PM_{2.5}$, we observe that cluster 2 exhibits higher values throughout the period considered than cluster 1; conversely for benzene an opposite behaviour is recorded.

Interestingly, the functional multivariate clustering algorithm reveals a potential misclassification since the Pescara urban background station of "Teatro d'Annunzio" is grouped with two traffic stations. This result highlights the peculiarity of the municipality of the Pescara, characterized by a considerable population density and a capillary road network, with high volumes of traffic that insists on an area little extended. In this context, urban traffic emissions represent the dominant source of atmospheric pollution and make background stations similar to traffic ones.

References

CARSLAW, D.C. 2021. Deweather-An R Package to Remove Meteorological Variation from Air Quality Data. *Available online:https://github.com/davidcarslaw/deweather*.

JACQUES, J., & PREDA, C. 2014. Model based clustering for multivariate functional data. *Computational Statistics and Data Analysis.*, **71**, 92–106.

R CORE TEAM. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAMSAY, J.O, & SILVERMAN, B.W. 2005. *Functional data analysis, 2nd edn.* New York: Springer-Verlag.

SCHMUTZ, A., JACQUES, J., BOUVEYRON, C., CHÉZE, L., & MARTIN, P. 2020. Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics.*, **35**, 1101–1131.

WILSON, G., KINGHAM, S., PEARCE, J., & STURMAN, A. 2005. A review of intraurban variationsin particulate air pollution: implications for epidemiological research. *Atmospheric Environment.*, **34**, 6444–6462.