

# Functional data analysis approach for identifying redundancy in air quality monitoring stations amid Covid-19

Annalina Sarra \*, Adelia Evangelista\*\*  
Tonio Di Battista\*\*\* Sergio Palmeri \*\*\*\*

\*University of Chieti-Pescara (Italy)  
annalina.sarra@unich.it,

\*\*University of Chieti-Pescara (Italy)  
adelia.evangelista@unich.it

\*\*\*University of Chieti-Pescara (Italy)  
tonio.dibattista@unich.it

\*\*\*\*Regional Agency for the Environmental Protection of Abruzzo (Italy)  
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  $\text{NO}_2$ ,  $\text{PM}_{10}$ ,  $\text{PM}_{2.5}$  and benzene concentrations are homogeneously distributed in the urban area of Pescara-Chieti (Central Italy). Air pollution data are collected from five monitoring stations of regional network. To perform the statistical analysis, we considered two timeframes amid Covid pandemic: before and during lockdown. By applying a functional model-based clustering, we investigated whether the effect of lockdown varies across different types of monitoring sites, with the ultimate aim to find out redundancy.

## 1 Introduction

In recent decades there has been a growing interest in monitoring air pollution levels, especially in urban areas, and in determining if ambient air quality standards are exceeded. Air pollution is generated when particles, biological molecules or other harmful materials are introduced into the earth's atmosphere. The evaluation of air quality status is one of the most serious environmental issues due the consequences to the acute exposure to air pollutants, that may cause serious health concerns (Kelishadi and Poursafa, 2010). The monitoring and the analysis of pollution data is of paramount importance because they could help environmental agencies in designing their policies against pollution. Around the world, countries have established air quality monitoring networks to obtain objective, accurate and comparable air

quality information of a specific area, and to support measures to reduce impact on human health and the natural environment. Despite the air quality monitoring stations are expensive, very often their number on certain territories is redundant. On the other hand, it is desirable to use as few stations as possible to meet monitoring objectives, in order to reduce cost and avoid overlapping data. Potentially, the elimination of redundant monitoring stations will provide better support for managers to formulate a more adequate air pollution control strategy. For all air quality monitoring networks the identification of the redundant measurements is, in fact, an important task, not only for determining the cost of pollution monitoring, but also for determining the integrity of pollution information monitoring and the accuracy of air quality assessment. In literature, there are many studies aimed at identifying possible redundancy in air quality monitoring networks (see, for a review, Wilson et al. (2005)). Most of them focuses on examining the intra-urban variation of air pollutant concentrations and whether or not the pollutant is homogeneously distributed across the area. In describing and quantifying pollutant concentrations uniformity at the intra-urban and other scales and determine whether pollution information captured by a monitoring station was correlated with that of other stations, commonly used techniques are correlation coefficients (Lin, 1989; Pinto et al., 2004), absolute differences, coefficients of variation, and coefficients of divergence (Pinto et al., 2004). Other alternatives are based on some multivariate statistical techniques, mainly principal component analysis and cluster analysis, employed for the site selection of air quality monitoring stations (Nunez-Alonso et al., 2019). Actually, air quality pollutants represent a typical example of data in which functional data analysis (FDA), can be a useful approach to be followed. In fact, these data are continuous over time even if they are collected at a daily, monthly or annual frequency (Wang et al., 2019). Thus, a suitable framework for working with the entire time spectrum is offered by the FDA paradigm. Over the years there has been an increasing interest in adopting FDA, with motivating examples retrieved from different application areas. One can refer, among others, to Fortuna et al. (2020), Torres et al. (2010), Viviani and Gron (2005). However, to our knowledge, only few works formulate functional models that explicitly exploit the functional form of the air quality data which gives new ways to gather information more than a single value or matrix obtained in the traditional univariate and multivariate context (see, for instance, Wang et al. (2019)). Within the field of FDA, in this paper, we consider the issue of functional data clustering. Specifically, we approach the problem of identifying possible redundancy in air quality monitoring stations, by a multivariate functional model-based clustering. Our interest in adopting clustering algorithms for air quality functional data relies on the fact they represent an effective method for finding representative curves, corresponding to different modes of variation, of pollutant concentrations, measured at the monitoring stations designed as urban background stations, representative of the population average exposure, as well as, at urban traffic ones. So, in this way, we are able to extract additional information contained in the mean curves of each group, useful for the problem at hand. Air pollution data for this study consist of measures of four pollutants: Nitrogen Dioxide ( $\text{NO}_2$ ), atmospheric particulate matter ( $\text{PM}_{10}$  and  $\text{PM}_{2.5}$ ) and benzene, obtained from the hourly air quality reporting platform of Chieti-Pescara urban area (Central Italy), run by Regional Agency for the Environmental protection of Abruzzo Region (ARTA). The analysed period is from February to April 2020, and in part coincides with the restrictions of human activities adopted during the quarantine policies, decided by the Italian Government on 10 March 2020, to contrast COVID-19 pandemic. The lockdown measures had included a shutdown of cross-area travel and the requirement that

local people stay home, which minimized industrial, transportation, and commercial activities. Many reports give evidence that the nationwide lockdowns, imposed in numerous countries to stop the spread of the Covid-19 infection, have had a positive effect on air quality all over the world, causing a reduction in the level of pollutants (Surech and Sharma, 2020), for the limited transportation and economic activities. In some sense, the mitigation measures against COVID-19 could be regarded as a naturally controlled experiment, in which there were much lower emissions of air pollutants than is typical. For the purposes of our work, by comparing the concentrations of  $\text{NO}_2$ ,  $\text{PM}_{10}$ ,  $\text{PM}_{2.5}$  and benzene before and during lockdown, we can assess if the restrictive measures adopted during March and April 2020 brought about a significant reduction among all the air monitoring stations of network of Chieti-Pescara urban area and if some possible redundancies can be noted. The rest of the document is structured as follows: Section 2 introduces the methodology used whereas Section 3 describes the study area and the data used for the analysis. Results are presented in section 4. Finally, the section 5 concludes the paper.

## 2 Methodology

Functional data analysis has received increased attention over the past years to express discrete observations arising from time series, waveforms or surfaces in the form of a function. The core idea of FDA is to treat the data not as multivariate observations but as (discretized) values of possibly smooth functions. FDA extends the classical techniques to data transformed in functions or curves, with the advantage of reducing thousands of observations to a few coefficients but conserving information about the functional form. During the last decade, it has emerged an important literature on FDA methods. A comprehensive introduction to the foundations and applications of FDA can be found in Ramsay and Silverman (2002, 2005), whereas nonparametric functional methods are summarised in a monograph by Ferraty and Vieu (2006). In the following subsection, we provide the basic technical details of the model-based clustering algorithm we are considering in this paper.

### 2.1 Multivariate functional model-based clustering

Cluster analysis is a set of popular data analysis techniques aimed at the selection and grouping of homogeneous elements into a set of data that are similar to one another without using any prior knowledge on the groups labels (Jain and Dubes, 1988).

One of the major approach to clustering analysis is represented by model-based clustering. By model-based algorithms, the data are viewed as coming from a mixture of probability distributions in which each component constitutes a different cluster (Hastie et al., 2009). Such models have a well-established theoretical backgrounds and dedicated literature, dealing with estimation algorithms for such models, can be found in Dempster et al. (1977), Hunter and Lange (2004) and Nguyen (2017). In recent times, a growing area of investigation has regarded the application of these methods within the FDA framework (e.g., Bouveyron and Jacques (2011); Jacques and Preda (2014), Nguyen and Chamroukhi (2018)). In our work, we rely on a novel clustering technique for multivariate functional data proposed by Schumtz et al. (2020). Our choice is essentially based on the flexibility of the selected method in handling multivariate data. Using a multivariate functional principal component analysis, the considered

## FDA approach for identifying redundancy in air quality monitoring stations

algorithm fits the data into group-specific functional subspaces. Actually, this method represents an improvement of the first Gaussian model-based clustering method based on a principal component analysis for multivariate functional data, Funclust, proposed by Jacques and Preda (2014). Even if the above mentioned authors proposed a more flexible method compared with the previous ones, it suffers from some limitations. Indeed, by using an approximation of the notion of the density of distribution for functional data, Jacques and Preda (2014) modelled only a given proportion of principal components, ignoring significant part of the available information. Conversely, the method we are considering exploits all available information, by modeling all non-null variance principal components. Following Schumtz et al. (2020), let us first assume that the observed curves are independent realization of a  $L_2$ -continuous multivariate stochastic process  $\mathbf{X}$ , where

$$\mathbf{X} = \{\mathbf{X}(t), t \in [0, T]\} = \{(X^1(t), \dots, X^p(t))\}_{t \in [0, T]} \quad (1)$$

for which the sample paths, i.e. the observed curves  $\mathbf{X}_i = (X_i^1, \dots, X_i^p)$  belongs to  $L_2[0, T]$ . Our goal is to explore a functional data set in order to automatically group the observed multivariate curves  $\mathbf{X}_i$  into  $K$  homogenous clusters. In problem of clustering, it is important to find out the number of clusters which should be a compromise between fit to the available data and over-fitting. In this respect, we first assume that  $K$  is fixed a priori. We denote with  $Z_{ik}$  a latent variable such that  $Z_{ik} = 1$  if the multivariate curve  $\mathbf{X}_i$  belongs to cluster  $k$  and is equal to 0, otherwise. The number of the curves within cluster  $k$  is defined with  $n_k = \sum_{i=1}^n z_{ik}$ . For each group  $k$ , ( $k = 1, 2, \dots, K$ ), let  $d_k < R$  denote the intrinsic dimension of a low-dimensional functional 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:

$$\varphi_r^k(t) = \sum_{l=1}^R q_{krl} \phi_l(t), 1 \leq r \leq R \quad (2)$$

obtained through a linear transformation from the matrix of principal factors  $\{\phi_r^j\}_{1 \leq j \leq p, 1 \leq r \leq R}$  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 the covariance matrix defined as follow:



$$\Delta_k = \left( \begin{array}{cc|cc} a_{k1} & 0 & & \\ & \ddots & & \\ 0 & a_{kd_k} & & \\ \hline & & b_k & 0 \\ 0 & & & \ddots \\ & & 0 & b_k \end{array} \right) \begin{matrix} d_k \\ \\ \\ R - d_k \end{matrix} \quad (3)$$

Thanks to the assumptions made on  $\Delta_k$ , it is possible to finely model the variance of the first  $d_k$  principal components, while just a unique parameter  $b_k$  models the remaining ones considered as noise components.

This is deemed the most general model (indicated as  $[a_{kj}b_kQ_kd_k]$ ), from which five sub-models are derived, determined by the constrained applied on model parameters and labelled as follows:  $([a_kb_kQ_kd_k], [a_{kj}b_kQ_kd_k], [a_kb_kQ_kd_k], [ab_kQ_kd_k], [abQ_kd_k])$ . The maximum likelihood estimates of parameters in the functional mixture model is facilitated by adopting the EM algorithm (Dempster, 1977). As for the choice of the hyperparameter  $K$ , classical model selection tools include AIC (Akaike information criterion, Akaike (1974)), BIC (Bayesian information criterion, Schwarz (1978)), and ICL (Integrated completed likelihood criterion, Biernacki et al. (2000)). For mixture models the selection, generally, takes place through the Bayesian information criterion. A comprehensive discussion on model inference and on the choice of the number of clusters is provided in Schumtz et al. (2020).

### 3 Study area and data

#### 3.1 Study area

For the purposes of assessing air quality, an area of greatest criticality was identified in the Abruzzo Region, consisting of the conurbation of the major centers Pescara and Chieti, which also includes the neighboring municipalities of Montesilvano and Francavilla al Mare (Fig.1), for a total population of 283602 inhabitants at 01/01/2018.

It is an almost entirely flat area, located in the terminal stretch (about 15 km long) of the Pescara river valley, which flows into the Adriatic Sea, right at Pescara, the most populous town (about 120,000 inhabitants). The climate is temperate warm, with an average annual temperature between 15 and 16 C; rainfall settles at 650 mm per year, while the prevailing ventilation follows the direction of the valley axis (prevalence of wind directions from the SW in the winter and from the NE in the summer, mainly governed by sea and land breezes). Especially in the winter semester, situations of atmospheric stability are frequent, with the characteristic phenomenon of thermal inversion that favors the accumulation of pollutants in the lower layers of the atmosphere. The sources of air pollution consist mainly of vehicular traffic and industrial activities scattered along the valley, with a further contribution, in the winter season, due to domestic heating.

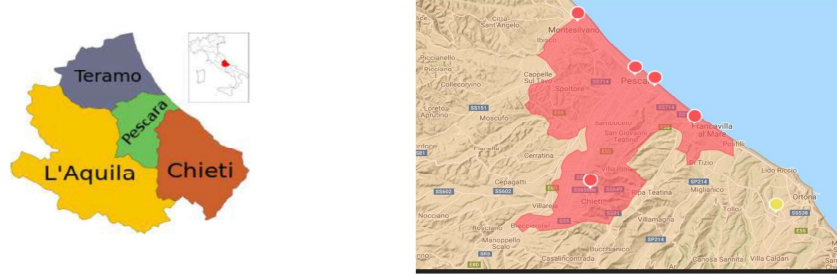


FIG. 1 – Abruzzo region- Central Italy (left panel)- Map of Pescara-Chieti conurbation (right panel) with the position of monitoring stations.

### 3.2 Data

Air pollution data for this study consist of measurements of  $\text{NO}_2$ ,  $\text{PM}_{10}$ ,  $\text{PM}_{2.5}$  and benzene ( $\text{C}_6\text{H}_6$ ) obtained from the automatic reporting platform, run by Regional Agency for the Environmental protection of Abruzzo Region (ARTA). These variables are measured in micrograms per cubic meter ( $\mu\text{g}/\text{m}^3$ ) and information are obtained from five monitoring sites. The air monitoring stations of Pescara (Via Firenze) and Montesilvano are designed as *traffic type* and are located where the pollution level is most influenced by traffic emissions from neighboring roads with medium-high traffic intensity; conversely, air quality data collected from the monitoring stations of Pescara (Teatro d'Annunzio), Chieti and Francavilla are deemed *Background measuring stations*, located where the pollution level is not influenced mostly by emissions from specific sources and are representative of the population average exposure. Daily measurements of pollutants have been collected from February to April 2020. The comparison was made for the time period of 1<sup>st</sup> February to 10<sup>th</sup> March 2020 (before lockdown period) versus 11<sup>st</sup> March 2020 to 20<sup>th</sup> April 2020 (during lockdown period). Since weather strongly influences pollutants formation and transport, the analyzed data set also includes daily weather variables, such as wind, rain, temperature. Because changes in wind (both intensity and direction), coupled with the often non linear relationships between changes in emissions and changes in meteorological conditions, have a remarkable impact on the transport and diffusion of pollutants over the area of interest, in this paper we consider a meteorological/weather normalisation. 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, 2020).

### 3.3 Mean pollutants concentrations

In Table 1, we display the mean concentrations of each pollutant, in each monitoring site, before and during the lockdown phases, after removing meteorological variation in air pollution trends. The effect of the lockdown period was most obvious on  $\text{NO}_2$  influence: it is a pollutant strongly linked to road traffic and the post lockdown decrease is clearly visible and

TAB. 1 – Mean concentrations of pollutants: pre and during lockdown

Pollutant	Before Lockdown					During Lockdown				
	Traffic		Background			Traffic		Background		
	Fi	Mo	Ch	Fr	Th	Fi	Mo	Ch	Fr	Th
NO <sub>2</sub>	23.6	24.8	20.0	15.9	34.7	10.5	9.5	8.5	7.9	10.5
PM <sub>10</sub>	24.3	21.8	22.0	17.8	21.8	29.3	24.1	26.3	24.9	30.4
PM <sub>2.5</sub>	14.9	14.0	16.9	11.6	15.1	17.5	16.0	20.5	15.1	18.0
Benzene	1.01	0.57	0.91	0.88	0.55	0.65	0.38	1.15	0.94	0.77

Legend of monitoring stations: Fi (Pescara, Via Firenze)- Mo (Montesilvano)-Ch (Chieti)- Fr (Francavilla)-Th (Pescara, Teatro d’Annunzio)

marked in all monitoring sites, even if differences emissions in magnitude exist depending on the stations.

Conversely, the impact of lockdown on PM<sub>10</sub> and PM<sub>2.5</sub> is the most complex of the four pollutants studied: background and traffic stations undergo an increase during the period of lockdown directives, implying that the monitoring sites might be under the effect of multiple non-transportation related emission sources, mainly domestic heating, which presumably increased during lockdown. Some differences can be found in the variations of benzene between pre and during lockdown phases: benzene levels basically dropped in the traffic-measuring stations whereas an opposite trend can be found in the background measuring sites. A possible explanation for this phenomenon could be found in domestic heating systems based on biomass combustion, which could affect more intensely the residential areas (where the background monitoring stations are located), far away from the main traffic routes.

## 4 Results

In this section we present a functional descriptive analysis and the results obtained through the use of the algorithm illustrated in Section 2. All the analyses were performed using the R packages *fda*, *funFEM* and *funHDDC* (R Development Core Team, 2020). The observed pollutant time series have been transformed into functional data with a process of smoothing, with 20 basis and cubic B-spline. In order to dynamically analyze the behaviour of the four pollutants concentrations, we extract information from the first order derivatives of the reconstructed curves, to study the velocity, and from the second derivatives to approximate the acceleration (Ramsay and Silverman, 2002, 2005). The investigation of NO<sub>2</sub> derivatives reveals that in all monitoring stations the speed and the relative acceleration decreased, especially in the first 20 days, as a result of the lockdown imposed by the authorities, while in the remaining period it tends to be stable. On the contrary, the PM<sub>10</sub> trend is very unstable and in the last 20 days it shows a strong deceleration. We also detected a marked decrease of speed and acceleration for PM<sub>2.5</sub> in the first period and in the last days. The same trend is also recorded for benzene, with the exception of the Montesilvano monitoring station which recorded a decrease only in the first days of the pre-lockdown period. The model-based clustering algorithm was applied here with the  $[a_k b_k Q_k d_k]$  model, using a basis of 20 natural cubic splines and provided the partition of monitoring stations into two groups. In what follows, we display the clustering of

pollutants time series and the estimated mean functions of the groups, differentiated according to the time period considered (Fig.2 and Fig.3).

Interestingly, we find out that the composition of identified groups varies according to the period taken into account. Looking at the pre-lockdown phase, it appears that cluster 1 (red line) contains the traffic measuring stations of Pescara-Via Firenze (Fi), Montesilvano (Mo) and the background station of Pescara-Teatro d'Annunzio (Th). In the right panels of Fig.2 we display the mean profile of each cluster. For the  $\text{NO}_2$  and  $\text{PM}_{10}$ , we observe that cluster 1 experiences higher values throughout this period than cluster 2 (green line); conversely, for the benzene an opposite behaviour is recorded whereas for  $\text{PM}_{2.5}$  the mean profiles values of two groups appear very similar. In addition, for the benzene, the mean profiles of two clusters are easy distinguishable. The analysis conducted during the lockdown phase and visualized in Fig.3, leads to the definition of two groups with similar and very close patterns; however, the two monitoring stations of Firenze (Fi) and Montesilvano (Mo), classified as traffic-measuring stations are not placed in the same cluster.

## 5 Conclusions

A functional model-based clustering approach for comparing air quality monitoring sites of the urban area of Chieti-Pescara in two specific time periods, was proposed here. Clustering was shown to depend on the timeframe considered. The revised methodology allows to draw the following conclusions, relevant for formulating adequate air pollution control strategy. In pre-lockdown period, the functional clustering assigns the Pescara urban background station (Teatro "d'Annunzio") to a group composed of the two traffic stations (Montesilvano and Firenze). This potential misclassification actually highlights, within the Chieti-Pescara urban conurbation under analysis, the peculiarity of the municipality of 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. Furthermore, the inspection of the curves relating to individual pollutants highlights the leading role of  $\text{NO}_2$  (pollutant specifically linked to road traffic) in determining clustering. The lockdown resulted in a heavy reduction in traffic volumes in the entire analysis area, and, therefore, in a net reduction in the concentration of  $\text{NO}_2$ , detectable in the background stations as well as in the traffic ones. As a result,  $\text{NO}_2$  loses its role as a "guide" variable in determining the outcomes of clustering, which is linked to the evolution of other pollutants, in particular, particulate matter airborne ( $\text{PM}_{2.5}$  and  $\text{PM}_{10}$ ), whose response to lockdown is less clear and unambiguous than that of  $\text{NO}_2$ . The belonging to the two clusters seems to be determined more by the geographical position of the stations than by their type. The cluster characterized by the highest concentrations includes stations, both traffic and background, belonging to the municipalities of Pescara and Chieti, where most of the population of the entire area is concentrated, while the other cluster includes the Montesilvano station (traffic) and that of Francavilla (background), located on the edge of the area under study, where the urban fabric becomes less dense. Therefore, it is possible to identify the population density as the main parameter that explains the belonging to one of the two clusters. This is confirmed by the fact that, if on the one hand the lockdown has significantly reduced road traffic, on the other it has forced people home, producing an increase in emissions from domestic heating systems, which are not a sec-

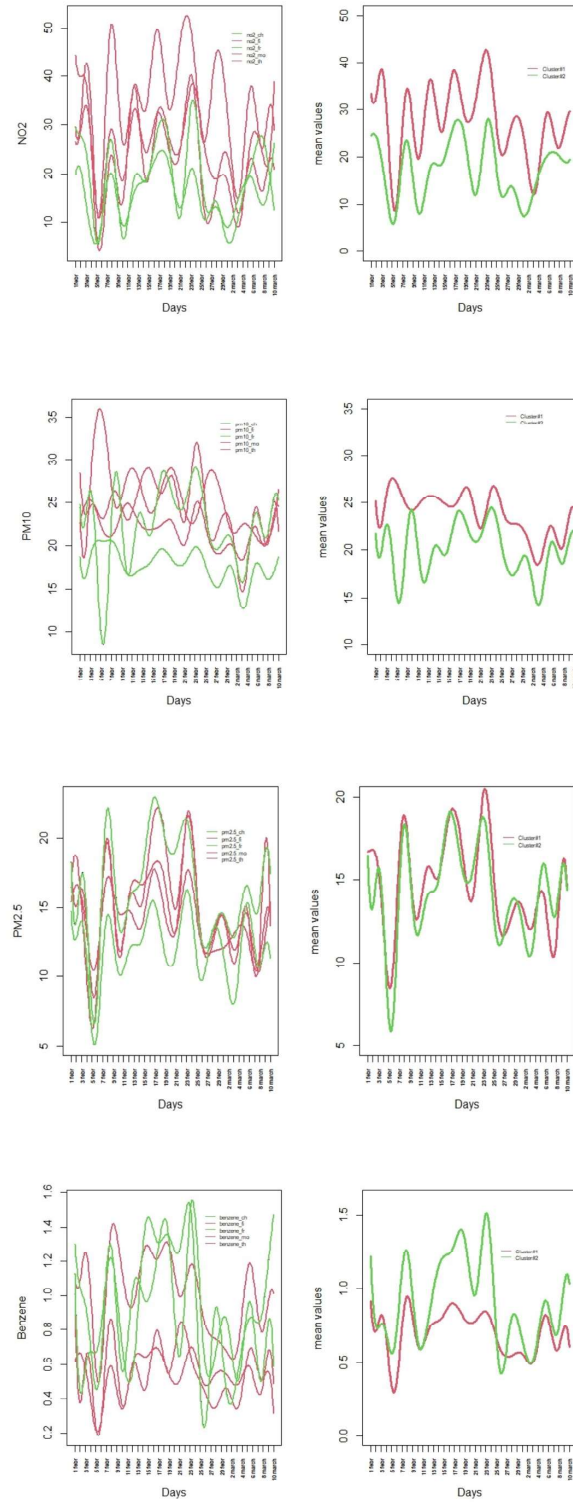


FIG. 2 – Clustering of pollutant time series and estimated mean functions in the pre-lockdown period: 1st February-10th March 2020

# FDA approach for identifying redundancy in air quality monitoring stations

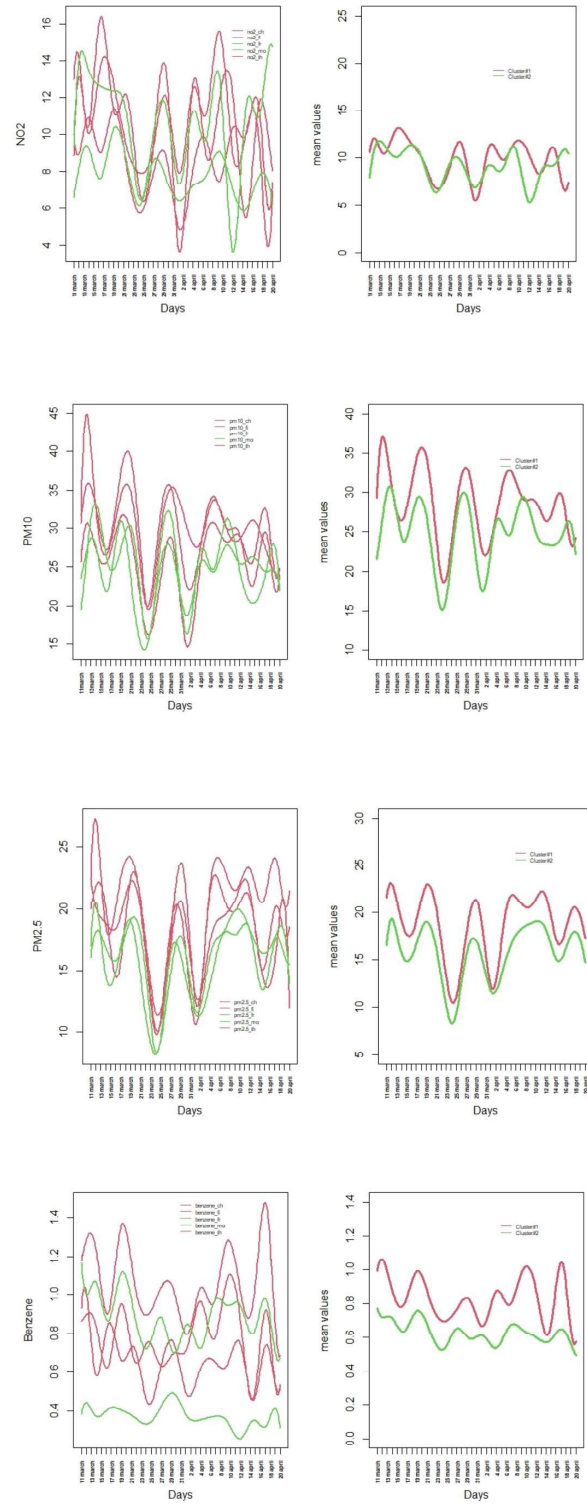


FIG. 3 – Clustering of pollutant time series and estimated mean functions during the lockdown period: 11st March-20th April 2020

ondary source of  $PM_{10}$  and  $PM_{2.5}$  emissions. It is worth noting that domestic heating, limited to biomass systems, whose diffusion has grown considerably in the last decade, also impacts on benzene levels. The "geographical", rather than functional, connotation of clustering is also explained, in our opinion, by considerations inherent to the circulatory regime prevailing in the river valley which constitutes a large part of the territory of analysis, with alternating winds from the SW (land breeze) and from NE (sea breeze) which produce transport of pollutants on the SW-NE axis of the valley and make the concentrations of the same homogeneous between the Chieti and Pescara stations, even regardless of their type, while the stations on the edge of the area itself (Montesilvano and Francavilla), less affected by this circulatory regime, are characterized by lower levels of  $PM_{10}$  /  $PM_{2.5}$  and benzene pollution. In the area under investigation, the presence of various background stations is undoubtedly appropriate for capturing local peculiarities related to both the type of predominant emission sources, the settlement context and the transport of pollutants. The risk of redundancy is minimal if the design of the monitoring network is preceded by a careful examination of these characteristics. The experiment offered by the lockdown has allowed, by creating an unprecedented scenario in which the source of "road traffic" has been drastically reduced, to highlight the importance of the various factors that contribute to determining the levels of pollution, in particular those related to population density and to the dominant regimes of atmospheric circulation.

## References

- Akaike, H. (1974). A new look at the statistical model identification . *IEEE Transactions on Automatic Control* 9, 716–723.
- Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood . *IEEE Trans PAMI* 22, 719–725.
- Bouveyron, C. and J. Jacques (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification* 5, 281–300.
- Carslaw, C. (2020). deweather: Remove the influence of weather on air quality data (R package version 0.5). Technical report, <https://github.com/davidcarslaw/deweather>.
- Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39, 1–38.
- Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis*. New York: Springer Series in Statistics. Springer.
- Fortuna, F., S. Gattone, and T. Di Battista (2020). Functional estimation of diversity profiles. *Environmetrics* 31, 67–68.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, second edition.
- Hunter, D. R. and K. Lange (2004). A tutorial on EM algorithms. *The American Statistician* 58, 30–37.
- Jacques, J. and C. Preda (2014). Model based clustering for multivariate functional data. *Computational Statistics and Data Analysis* 71, 92–106.

## FDA approach for identifying redundancy in air quality monitoring stations

- Jain, A. K. and R. C. Dubes (1988). *Algorithms for Clustering Data*. New York: Upper Saddle River, NJ: Prentice-Hall, Inc.
- Kelishadi, R. and P. Poursafa (2010). Air pollution and non-respiratory health hazards for children. *Archives of Medical Science* 6, 483–495.
- Lin, L. I.-K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45, 255–268.
- Nguyen, H. D. (2017). An introduction to Majorization-Minimization algorithms for machine learning and statistical estimation. *WIREs Data Mining and Knowledge Discovery* 7, e1198.
- Nguyen, H. D. and F. Chamroukhi (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *WIREs Data Mining and Knowledge Discovery* 8, 92–106.
- Nunez-Alonso, D., L. Perez-Arribas, S. Manzoor, and J. Caceres (2019). Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies in the Madrid Region. *Journal of Analytical Methods in Chemistry* 2019, 109–129.
- Pinto, J., A. Lefohn, and D. Shadwick (2004). Spatial variability of PM<sub>2.5</sub> in urban areas in the United States. *Journal of Air Waste Management* 54, 440–449.
- R Development Core Team, R. (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Technical report.
- Ramsay, J. and B. Silverman (2002). *Applied functional data analysis*. New York: Springer Series in Statistics. Springer-Verlag.
- Ramsay, J. and B. Silverman (2005). *Functional data analysis*. New York, second edition: Springer Series in Statistics.
- Schumtz, A., J. Jacques, C. Bouveyron, L. Cheze, and P. Martin (2020). Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics* 5, 281–300.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6 (2), 461–464.
- Surech, J. and T. Sharma (2020). Social and travel lockdown impact considering coronavirus disease (COVID-19) on air quality megacities of India: Present benefits, future challenges and way forward. *Aerosol and Air Quality Research* 20, 1222–1236.
- Torres, J., P. Garcia Nieto, L. Alejano, and R. A.N. (2010). Detection of outliers in gas emissions from urban areas using functional data analysis. *Journal of Hazourds Materials* 186, 144–149.
- Viviani, R. and M. Gron, G. and Spitzer (2005). Functional principal component analysis of fMRI data. *Human Brain Mapping* 24, 109–129.
- Wang, D., Z. Zhangqi, K. Bai, and L. He (2019). Spatial and Temporal Variabilities of PM<sub>2.5</sub> Concentrations in China Using Functional Data Analysis. *Sustainability* 14, 679–696.
- Wilson, G., S. Kingham, J. Pearce, and A. Sturman (2005). A review of intraurban variations in particulate air pollution: implications for epidemiological research. *Atmospheric Environment* 34, 6444–6462.