Environmental Science and Pollution Research (2023) 30:50067–50095
https://doi.org/10.1007/s11356-023-25790-3

RESEARCH ARTICLE

Unveiling air pollution patterns in Yemen: a spatial–temporal functional data analysis

Mohanned Abduljabbar Hael[1,2]

Received: 14 November 2022 / Accepted: 3 February 2023 / Published online: 15 February 2023

Abstract
The application of spatiotemporal functional analysis techniques in environmental pollution research remains limited. As a result, this paper suggests spatiotemporal functional data clustering and visualization tools for identifying temporal dynamic patterns and spatial dependence of multiple air pollutants. The study uses concentrations of four major pollutants, named particulate matter (PM2.5), ground-level ozone (O_3), carbon monoxide (CO), and sulfur oxides (SO_2), measured over 37 cities in Yemen from 1980 to 2022. The proposed tools include Fourier transformation, B-spline functions, and generalized-cross validation for data smoothing, as well as static and dynamic visualization methods. Innovatively, a functional mixture model was used to capture/identify the underlying/hidden dynamic patterns of spatiotemporal air pollutants concentration. According to the results, CO levels increased 25% from 1990 to 1996, peaking in the cities of Taiz, Sana'a, and Ibb before decreasing. Also, PM2.5 pollution reached a peak in 2018, increasing 30% with severe concentrations in Hodeidah, Marib, and Mocha. Moreover, O_3 pollution fluctuated with peaks in 2014–2015, 2% increase and pollution rate of 265 Dobson. Besides, SO_2 pollution rose from 1997 to 2010, reaching a peak before stabilizing. Thus, these findings provide insights into the structure of the spatiotemporal air pollutants cycle and can assist policymakers in identifying sources and suggesting measures to reduce them. As a result, the study's findings are promising and may guide future research on predicting multivariate air pollution statistics over the analyzed area.

Keywords Functional data analysis · Air pollutants · Clustering · Visualization · Multi-sites · Yemen

Introduction

Air pollution can have serious negative impacts on human health, including cardiovascular and respiratory diseases. Monitoring and controlling air pollutants is crucial for protecting public health and the environment. Commonly used strategies for air pollution monitoring include statistical analysis, data visualization, and identifying correlations and trends in pollution levels. These tools can help identify sources of pollution and inform the development of policies and regulations to reduce and control pollution levels. Additionally, monitoring and control efforts may also include the use of specialized equipment and technology, such as air quality sensors and monitoring stations, to measure and track specific pollutants in the air (Manisalidis et al. 2020; Cook et al. 2021; Li et al. 2022c). Modern statistical approaches, such as spatiotemporal modeling and machine learning techniques, can effectively handle high-dimensional data with temporal and spatial characteristics in environmental air pollution monitoring. These methods can capture the underlying variations and dynamics trends of air pollutants over the entire temporal-spatial scale, making them more suitable for this type of data compared to classical statistical methods (Acal et al. 2022; Wang et al. 2022).

Functional data analysis (FDA) is a powerful technique for working with multi-dimensional air pollutants data. It utilizes additional information, such as the smoothness of the data structure, rate of change, acceleration, and dynamic changes over a large-scale domain, to extract more information from the data compared to traditional vectorial approaches (King et al. 2018; Al-Janabi et al.

Responsible Editor: Marcus Schulz

✉ Mohanned Abduljabbar Hael
2014mohanned@gmail.com

[1] School of Statistics, Jiangxi University of Finance and Economics, Nanchang 330013, China

[2] Department of Data Science and Information Technology, Taiz University, 9674 Taiz, Yemen

2021; Reinholdt Jensen et al. 2022). FDA has been well-established in the literature over the past two decades, with a strong methodological and operational framework. There are several advantages of using FDA compared to traditional vectorial approaches: (i) flexibility: FDA can handle data that is not easily represented by vectors, such as data that is curve- or surface-based, or data that varies over a continuous domain, (ii) smoothness: FDA can incorporate information about the smoothness of the data structure, which is not captured by traditional vector-based methods, (iii) dynamics: FDA can capture dynamic changes in the data, such as rates of change or acceleration, which are not possible with traditional vector-based methods, (iv) high-dimensional data: FDA can handle high-dimensional data, which is a challenge for traditional vector-based methods, (v) modeling: FDA can be used to model complex relationships between variables that are not easily represented by simple linear or polynomial models, (vi) visualization: FDA also allows for better visualization of the data, which can aid in understanding and interpreting the results (Al-Janabi et al. 2020b; Betancourt-Odio et al. 2021). The principles and foundations of the FDA methods are found in Ramsay and Silverman (2002, 2005) besides the nonparametric methods of functional data are presented in a monograph study by Ferraty and Vieu (2006).

The use of FDA techniques in analyzing environmental data has received remarkable attention in the past two decades. Escabias et al. (2005) combined functional logistic regression and principal component analysis of environmental data modeling besides the proposed method were used to estimate drought risk in terms of temporal evolution in temperatures. Ground ozone represents one of the most dangerous environmental pollutants; complex chemical and physical processes generate it in the atmosphere and combustion processes in the troposphere. Several FDA approaches have been proposed to analyze the ozone concentration level, for instance, functional principal components analysis (FPCA) to extract manifest features for ground ozone concentration levels (Caligiuri et al. 2005), smooth-spline-based models to study time trends and oscillations in stratospheric ozone (Meiring 2007), mixed functional methods to model trends in the profiles of stratospheric ozone (Park et al. 2013), and the Kendall's Tau functional statistic (KFT) to discover significant correlations between tropospheric ozone levels in urban and rural sites (Betancourt-Odio et al. 2021). In the same context, Gao (2007) and Gao and Niemeier (2008) used FDA techniques to model the dynamic pattern of nitrogen oxide and diurnal ozone cycles; they showed important results about the structure of spatiotemporal variations in diurnal cycles. In the atmosphere, the concentration of particulate matter (PM) is a highly time–space variable, which follows a periodic cycle dominated by meteorological situations as well as anthropogenic activities. The study by (Broomandi et al. 2021) examined the impact of fine PM2.5 on respiratory and heart diseases. They used a data-driven directed graph representation to infer the causal directionality and spatial embeddedness of PM2.5 concentrations in 14 UK cities over the course of one year. They found notable spatial embedding in the summer and spring and stability to disturbances through the network trophic coherence parameter, with winter being the most significant vulnerability. Many studies have employed FDA to analyze PM and its relationship to air quality. FDA can be used to model the temporal and spatial variation of PM levels, and to identify patterns and trends in the data. It can also be used to estimate the relationship between PM levels and other factors such as weather, traffic, and land use. Additionally, FDA can be used to make predictions about future PM levels and to assess the effectiveness of interventions aimed at reducing PM exposure. For instance, Shaadan et al. (2012) used a functional approach to assess the PM10 pollutant behaviour and compare data from two different years. In another related work, Hörmann et al. (2015) proposed a dynamic version of functional principal component analysis (dynamic FPCs), and the advantage of this approach has been illustrated by applying it to PM10 changes. In another related paper, Kosiorowski et al. (2017) adapted a hierarchical functional time series on a micro-model to forecast day and night PM10 air pollution. In another related study, King et al. (2018) applied modern FDA methods to study the spatial and temporal trends and variability of fine PM components across the USA. In recent years, research on the concentrations of multivariate air pollutants has been investigated by FDA techniques. In another related work, Ruggieri et al. (2013) focused on the principal component analysis of functional data (FPCA) to investigate the variability of multivariate air pollutants data, including (CO, NO_2, PM10, and SO_2). More recently, an analysis of variance based on functional data analysis (FANOVA) has been proposed by Acal et al. (2022). This method has been applied to four air pollutant concentrations, namely PM2.5, benzene, NO_2, and PM10, to assess air pollution changes during the COVID-19 lockdown.

In the environmental pollution framework, an unusually high concentration of air pollutants, known formally as anomalies, may bring problems in the air quality index. Martínez et al. (2014), Sancho et al. (2014), and Torres et al. (2020) implemented a model relying on functional analysis to identify outliers samples, with the overall goal of achieving a better air quality monitoring solution. In another related paper, Shaadan et al. (2015) conducted a study to detect anomalies in daily PM10 functional data, investigate behaviour patterns, and identify potential factors determining PM10 abnormalities at three selected air quality monitoring stations. More applications that demonstrate the usefulness and advantages of FDA methods in environmental data analysis are found in Ocana-Peinado et al. (2008), Valderrama et al. (2010), Embling et al. (2012), Escabias et al. (2013), Ignaccolo et al. (2014), Xiao and Hu (2018), Ochoa et al. (2020), Reinholdt Jensen et al. (2022).

Machine learning is widely used to perform in-depth analysis in various fields such as biomedicine, energy, and economics (Saleh et al. 2023). To make our proposed method more comprehensive, we will compare it to recent algorithms in the context of machine learning and deep learning. For example, in biomedicine, Al-Janabi and Alkaim (2022) proposed a novel optimization method called Lion-AYAD to find optimal DNA protein generated through DNA synthesis. Their results showed the method to be robust with dynamic DNA sequence lengths, with increased accuracy and reduced execution times. In a similar context, Kadhuim and Al-Janabi (2023) presented a model that uses Deep Optimal Neurocomputing Technique (DLSTM-DSN-WOA) and Multivariate Analysis to predict Codon-mRNA. Their proposed model is a pragmatic intelligent data analysis model that reduces computation and handling time for large real data. In the field of renewable energy, Al-Janabi et al. (2020a) proposed deep learning techniques (DCapsNet and DCOM), and Mohammed and Al-Janabi (2022) proposed optimization techniques (FDIRE-GSK) for the generation of electrical energy from natural resources such as wind energy. Another approach, called DRFLLS, has been developed to estimate missing values in various datasets (Al-Janabi and Alkaim 2020). Additionally, the use of machine learning algorithms for high-dimensional functional data classification has become increasingly important in environmental air pollution research. The current study specifically focuses on using the FDA approach for the classification and visualization of high-frequency spatiotemporal air pollution data. Researchers have previously attempted to use FDA methods to cluster air pollution levels, which can help identify patterns and trends in the data and better understand the factors that contribute to air pollution. The use of the FDA, in combination with machine learning algorithms, can help to improve the accuracy and robustness of air pollution classification and visualization.

There have been several studies that have used functional data clustering approaches to analyze the network paths of air quality. To show an example, Ignaccolo et al. (2008) proposed an early study on analyzing the network paths of air quality using functional data clustering; they considered the air pollutant variable as a functional data object and classified them using the Partitioning Around Medoids (PAM) algorithm. Similarly, Ranalli et al. (2016) used FDA and PAM clustering approach to analyze high-frequency spatiotemporal data on the size distribution of particulate matter (PM). In another paper, Kosiorowski and Szlachtowska (2017) proposed a novel k–local functional median algorithm applied to the analysis of a real data set concerning air pollution monitoring. More recently, Bouveyron et al. (2022) developed a functional co-clustering approach based on the functional latent block model (funLBM) and illustrated by the analysis of multivariate air pollution data in the South of France. All these studies have made significant progress in the field of clustering functional air pollution data in terms of methodology and practical applications.

Research on clustering spatiotemporal air pollution using FDA is still an active area of interest, and new studies are needed to further advance the field. Therefore, this study has two main contributions: 1) from a methodological perspective, it presents a method based on the FDA approach for clustering and visualizing spatiotemporal functional data, and 2) from a practical aspect, it applies the proposed method to identify, classify and visualize multiple air pollutants, such as sulphur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), and particulate matter (PM2.5) measured over multiple sites in Yemen during the period of January 1980 to April 2022. As far as the authors know, the air pollution problem in Yemen has not been investigated before, and this is the first study to analyze the multivariate air pollution concentrations using the FDA method. The study highlights several steps to achieve its goal: (1) transforming the discretization air pollution data into functional data to work with functional realm; (2) smoothing the functional air pollution data to improve the structured data from step 1; (3) visualizing the spatiotemporal features of functional air pollution data to discover the mechanism of variability; (4) clustering spatiotemporal functional air pollution data to group similar patterns found in step 3 for both spatial and temporal profiles.

Data and methods

Study area

Yemen, officially the “Republic of Yemen,” is a west Asian country located in the Middle East in the southern part of the Arabian Peninsula. It is bordered by the Kingdom of Saudi Arabia to the north, the Sultanate of Oman to the east, the Arabian Sea to the south, and the Red Sea to the west, and it shares maritime borders with Djibouti, Eritrea, and Somalia. It is the second-largest Arab sovereign state on the peninsula, occupying

555,000 km^2 (214 thousand square miles). Yemen's total coastline extends over a length of approximately 2000 square kilometers (1200 miles). Sana'a is the constitutionally stipulated capital and largest city of Yemen. As of 2021, Yemen has an estimated population of 30,491,000. Yemen lies within latitude and longitude 15° 0′ North and 48° 0′ East which includes an area mostly desert. It also consists of a narrow coastal plain surrounded by rugged mountains.

Yemen's climate is a mixture of temperate, humid, and hot. The western part is exposed to the influence of the monsoon monsoons. Towards the inland eastern region of Yemen, the climate becomes unbearably hot. During the summer, the temperature can reach 54 °C, and the winters are much colder, with frost in some parts. The average annual temperature of the capital, Sanaa, is 18° C. Yemen is also exposed to natural hazards in the form of dust and sand storms. Yemen's climate can be described as a dry subtropical, hot desert climate with low annual rainfall, very high summer temperatures, and a large difference between the maximum and minimum temperatures, especially in the interior regions.

Figure 1 depicts a geopolitical map of Yemen with detailed legends for its major cities, road networks, airports, railways/railroads, and waterways.

Data and variables

In this study, the selection of the 37 major cities for analysis was based on the criteria of high population density and wide geographical coverage across the entire country. The population density of a city is an important factor in determining the level of air pollution, as a higher population density typically leads to higher levels of industrial and vehicular emissions. The geographical location of the cities was also considered to ensure that a diverse range of regions was represented in the analysis. Additionally, the availability of historical air pollution data was taken into account to ensure that an accurate and comprehensive analysis could be performed. The geographical location of the selected cities is illustrated in Fig. 2, which provides a visual representation of the distribution of the cities across the country. This information can be useful in understanding the regional variations in air pollution levels and trends. The name of selected stations and their geographical

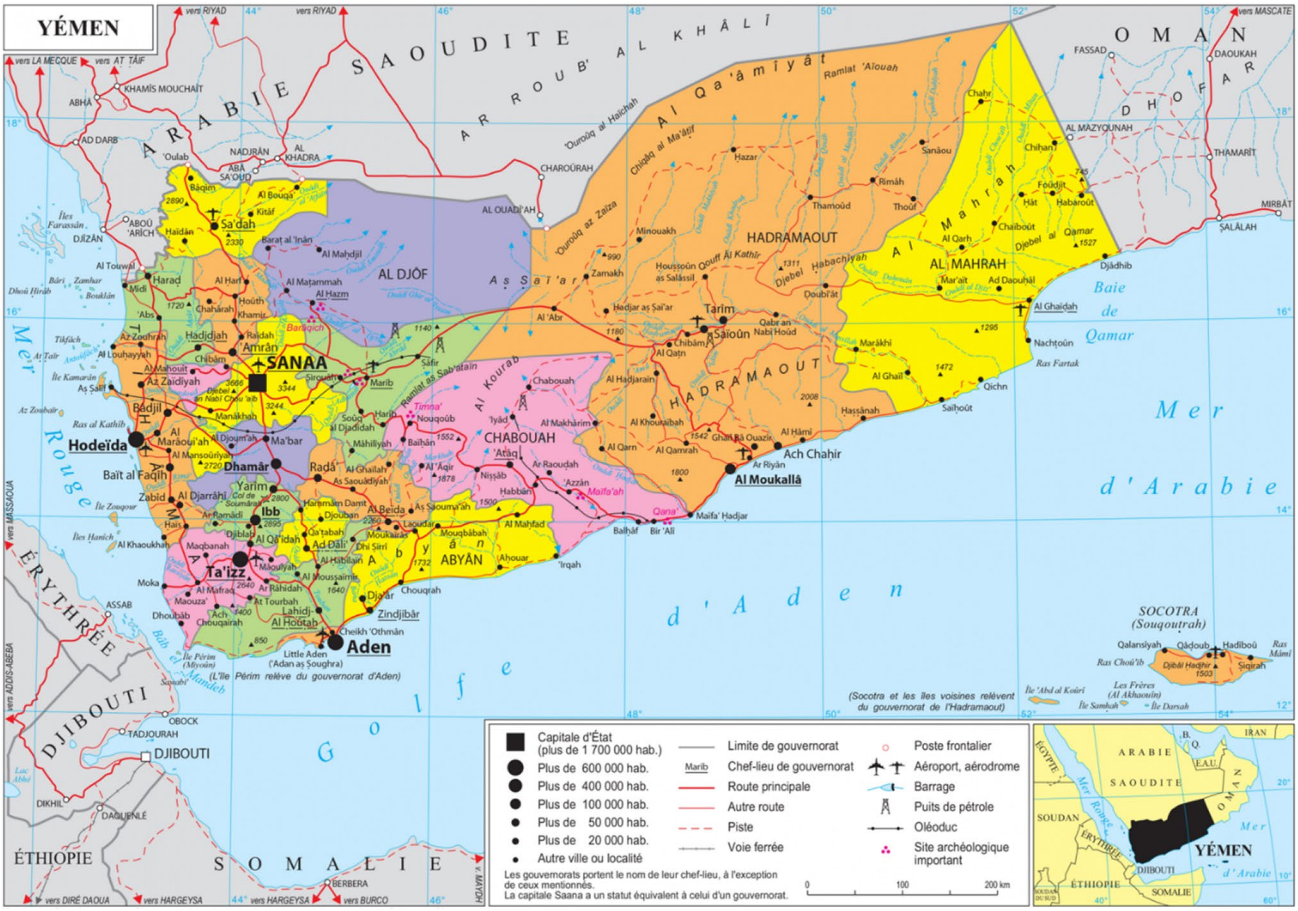


Fig. 1 The map of Yemen with major geographical features (Worldmaps 2023)

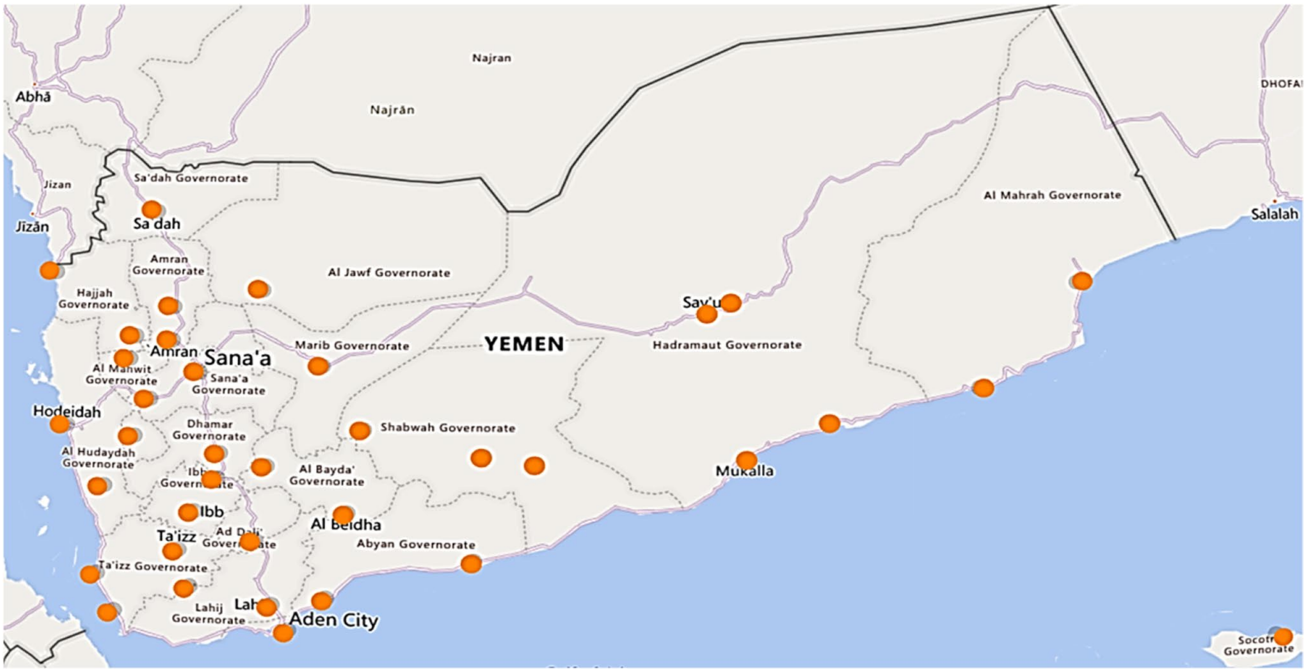


Fig. 2 The spatial distribution of the selected sites in Yemen

characteristics (latitude, longitude, and average elevation) are given in Supplementary Table 1.

The concentrations of a pollutant are typically gauged by using the metric of micrograms per meter cube($\mu g/m^3$). There are two different metrics to measure the particulate matter concentrations: (PM2.5) and (PM10) refer to the particles that are less than $2.5\mu g/m^3$ and $10\mu g/m^3$ in diameter, respectively. This study focuses on four main air pollutants: PM2.5, O_3, SO_2, and CO. The area-averaged monthly records for four primary air pollutant variables for several locations in Yemen During the period 1980–2022 were extracted as satellite data from (NASA 2023). The total sample size for CO, SO2, and PM2.5 pollutants measurements is equal to (n = 500) discrete records discrete records$X_i(t_j), t_j \in [1, 500], i = 1, \ldots, 37,$ the *ith* discrete observation $X_i(t_j), j = 1, \ldots, 500$ indicates pollutants values for the *ith* month. The total sample size for O3 pollutant measurements is equal to (n = 508) discrete records$X_i(t_j), t_j \in [1, 508], i = 1, \ldots, 37,$ the *ith* discrete observation $X_i(t_j), j = 1, \ldots, 508$ indicates pollutants values for the *ith* month.The discrete monthly pollutants dataset $X_i(t_j)$ will be transformed into continuous functions by adapting a suitable basis functions system. More details about the variables and their features are summarized in Table 1.

Model hypothesis and limitations

The aim of this paper is to analyze and categorize the dynamic changes in air pollution concentrations using functional analysis techniques and a functional mixtures clustering model. The method was applied to multivariate high-dimensional air pollution data collected from cities in Yemen from 1980 to 2022. Fourier transformation,

Table 1 Details of variables and data

Variables	Spatial resolution	Units	Temporal resolution	Time Period	Source
SO_2	0.5×0.625^o	kg/m^3	Average monthly	Oct,1980—Apr, 2022	MERRA-2 Model *
PM2.5	0.5×0.625^o	kg/m^3	Average monthly	Oct,1980—Apr, 2022	MERRA-2 Model *
O_3	0.5×0.625^o	Dobson	Average monthly	Jan,1980—Apr, 2022	MERRA-2 Model *
CO	0.5×0.625^o	kg/m^2	Average monthly	Oct,1980—Apr, 2022	MERRA-2 Model *

* Datasets extracted by considering the unified source

B-spline functions, and generalized cross-validation were utilized to smooth and reconstruct the data. Two enhanced 3D visualization tools were used to examine the spatiotemporal variations in air pollutants and a functional mixture model was employed to classify the functional air pollutants data based on their spatiotemporal characteristics. The paper sets forth three hypotheses: (i) that the air pollution data is measured at a set of ordered times and the discrete observations $X_i\left(t_j\right)$ are dense and regular over a specified time interval, (ii) that the air pollutants data follows a Gaussian mixture model-based FDM(Σ_k, β_k) model, and (iii) that the number of clusters (K) and the intrinsic dimension (d) must be predetermined. The first hypothesis is supported by the regular observation of air pollution data from 1980 to 2022. The second hypothesis is supported by the use of Gaussian mixture models in functional data analysis, and the option to use a robust mixture model in the future. The third hypothesis, which assumes a fixed number of clusters, is based on the belief that using selection methods to determine K would lead to unclear results. In this study, the number of clusters was set at 4, as it was determined to provide the best segmentation of the air pollution data.

Model estimation

The structure of the spatiotemporal functional data (STFD) model in a multivariate pollutant's context is a statistical framework for analyzing functional data that varies over both space and time. The model typically consists of several components: (i) spatial component: this captures the spatial variation in the functional data, often represented as a spatial random effect, (ii) temporal component: this captures the temporal variation in the functional data, often represented as a temporal random effect, (iii) functional component: this captures the functional variation in the data, often represented as a functional principal component analysis (FPCA) model, (iv) covariate component: this captures the relationship between the functional data and any additional covariate information, often represented as a linear or nonlinear regression model, and (v) error component: this captures the residual variation in the data not explained by the other components (King et al. 2018; Wang et al. 2020; Hael et al. 2020). Overall, the STFD model is a flexible framework that can be used to analyze multivariate pollutant's objects over space and time and can be extended to include other sources of variation or additional information as needed. Additional information has been included in the supplementary section regarding the theoretical framework, including elements like the structure of spatiotemporal functional data, the basis functions and smoothing techniques used, and the functional mixture model for analyzing STFD. This section will cover the concepts of model estimation, including the Expectation, Discrimination, and Maximization phases. Expectation (E) phase: In the E-step, the model uses the current estimate of the parameters to calculate the probability of each data point belonging to each cluster. Whereas Discrimination (D) phase: In the D-step, the model uses the probabilities calculated in the E-step to re-estimate the parameters of the clusters. While Maximization (M) step: In the M-step, the model uses the re-estimated parameters from the D-step to update the overall estimate of the parameters of the model (Preda 2007; Bouveyron et al. 2015). The Expectation-Discrimination-Maximization (EDM) procedure is a three-step process used to estimate the parameters of the mixture model. It involves alternating between the E-step, the D-step, and the M-step until a specified criterion is met. In this study, the criterion was set as 100 iterations (q = 100). The EDM procedure continues to iterate until this iteration number is reached.

The expectation phase

The expectation is the first step which computes the posterior probabilities $t_{ik}^{(q)}$ under the condition the current value of the parameter $\theta^{(q)}$, at iteration q. The probability $P(z_{ik} = 1)$ points out that the curve brings from the kth component and $P(z_{ik} = 0)$ otherwise. In the functional discriminative model, the posterior probabilities $t_{ik}^{(q)}$, i = $1, 2, \ldots, n$;k = $1, 2, \ldots, K$ that each curve suits the kth component can be given as (Bouveyron et al. 2015):

$$t_{ik}^{(q)} = E\left[z_{ik}\middle|y_i, \theta^{(q)}\right] = \frac{\pi_k^{(q)}\phi(y_i, \theta_k^{(q)})}{\sum_{l=1}^{K} \pi_l^{(q)}\phi(y_i \mid \theta_l^{(q)})} \tag{1}$$

where $\theta_k^{(q)} = (\pi_k^{(q)}, \mu_k^{(q)}, \Sigma_k^{(q)}, \beta^{(q)})$ are the combination of parameters for the kth mixture component and $\phi(.)$ is the Gaussian density. The model parameters will be updated in the Maximization(M) step (mentioned below) and estimated at an optimal point in the last iteration (q).

The discrimination phase

The Discrimination (D) step is aimed to determine the orientation matrix U^q of the discriminative latent space F conditionally on the posterior probabilities $t_{ik}^{(q)}$ through maximizing the standard Fisher's (F) criterion (Preda 2007):

$$\hat{U}^{(q)} = \underset{U}{Max}\ \text{trace}\left(\left(U^t S U\right)^{-1} U^t S_B^{(q)} U\right) \tag{2}$$

The S refers to the whole sample covariance matrix and $S_B^{(q)}$ refers to the soft between-cluster covariance matrix, which is defined as: $S_B^{(q)} = \frac{1}{n}\sum_{k=1}^{K} n_k^{(q)}(m_k^{(q)} - \overline{y})(m_k^{(q)} - \overline{y})^t$. In the functional unsupervised classification framework with an unobserved variable (z), the Fisher criterion optimizes the discriminative function $U \in L_2[0, T]$ by (Bouveyron et al. 2015):

$$\underset{U}{Max}\frac{Var\left(\mathbb{E}\left\{\int_0^T X(t)U(t)dt \mid z\right\}\right)}{Var\left(\int_0^T X(t)U(t)dt\right)} \tag{3}$$

The optimization procedure of (4) is the eigenfunction U associated with the highest eigenvalue λ of the following generalized eigenproblem:

$$\int_0^T B(t,s)U(s)ds = \lambda \int_0^T C(t,s)U(s)ds \tag{4}$$

The estimation of the covariance operator $C(t,s)$ based on the basis function $(\psi_j)_j = 1, \dots, p$, is given as:

$$\hat{C}(t,s) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} y_{ij}\psi_j(t)\right)\left(\sum_{j=1}^{p} y_{ij}\psi_j(t)\right) \tag{5}$$

The $B(t,s)$ here indicates the integral between cluster covariance operators and conditionally on the posterior probabilities $t_{ik}^{(q)}$ obtained from the Expectation (E)-step, the estimator of $B(t,s)$ at iteration (q) is defined as:

$$\hat{B}^{(q)}(t,s) = \frac{1}{n}\sum_{k=1}^{k}\frac{1}{n_k^{(q)}}\left(\sum_{i=1}^{n} t_{ik}^{(q)}v(t)\right)\left(\sum_{i=1}^{n} t_{ik}^{(q)}v(s)\right) \tag{6}$$

The maximization phase

The Maximization step (M) is aimed to estimate the parameters of the functional latent mixture model. In this step, maximizing the conditional expectation of the complete data log-likelihood conditionally to the orientation matrix U^q is computed in the following form (Bouveyron et al. 2015):

$$Q(\theta) = \frac{-1}{2}\sum_{k=1}^{K} n_k^{(q)}\left[-2\log\left(\pi_k\right) + A + \log\left(\left|\Sigma_k\right|\right) + B + D + h\right] \tag{7}$$

The explanation of these notations is as follows: θ indicates the parameters of the mixture model $\theta = (\pi_k, \mu_k, \Sigma_k, \beta_k)$, $A = \text{trace}\left(\Sigma_k^{-1}U^{(q)t}C_k^{(q)}U^{(q)}\right)$, $B = (p-d)\log\left(\beta_k\right)$, $\mathrm{D} = \frac{\text{trace}\left(C_k^{(q)}\right) - \sum_{j=1}^{d} u_j^{(q)t}C_k^{(q)}u_j^{(q)}}{\beta_k}$, $C_k^{(q)} = \frac{1}{n_k^{(q)}}\sum_{i=1}^{n} t_{ik}^{(q)}\left(y_i - \mu_k^{(q)}\right)\left(y_i - \mu_k^{(q)}\right)^t$ presents the empirical covariance matrix of the kth cluster,$u_j^{(q)}$ is the jth column vector of $U^{(q)}$, and $h = plog(2\pi)$ is a constant term. At iteration q, the maximization of $Q(\theta)$ is conditional on $U^{(q)}$ conduces to the estimation of mixture parameters of the FDM(Σ_k, β_k) model according to the following update formulas (Bouveyron et al. 2015):

$$\Rightarrow \text{The proportion parameter}, \pi_k^{(q)} = \frac{n_k^{(q)}}{n} \tag{8}$$

$$\Rightarrow \text{The mean parameter}, \mu_k^{(q)} = \frac{1}{n_k^{(q)}}\sum_{i=1}^{n} t_{ik}^{(q)}U^{(q)t}y_i \tag{9}$$

$$\Rightarrow \text{The covariance parameter}, \Sigma_k^{(q)} = U^{(q)t}C_kU^{(q)} \tag{10}$$

$$\Rightarrow \text{The noise parameter}, \beta_k^{(q)} = \frac{\text{trace}\left(C_k\right) - \sum_{j=1}^{d} u_j^{(q)t}C_ku_j^{(q)}}{p-d} \tag{11}$$

The flowchart illustrating the methods proposed in the current study can be found in Fig. 3. Additionally, this flowchart serves as a framework for understanding the terminology used in the study's methodology and statistics. Besides, the proposed method for the visualization and clustering of the STFD uses a multivariate framework implemented in R programming language with the help of several package environments. Spatial–Temporal Functional Air Pollution Data Analyzer (STFAPDA), which is a useful tool for analysing and understanding the dynamics of air pollution, is the name of the proposed algorithm. The algorithm consists of several steps, which are described in detail as follows:

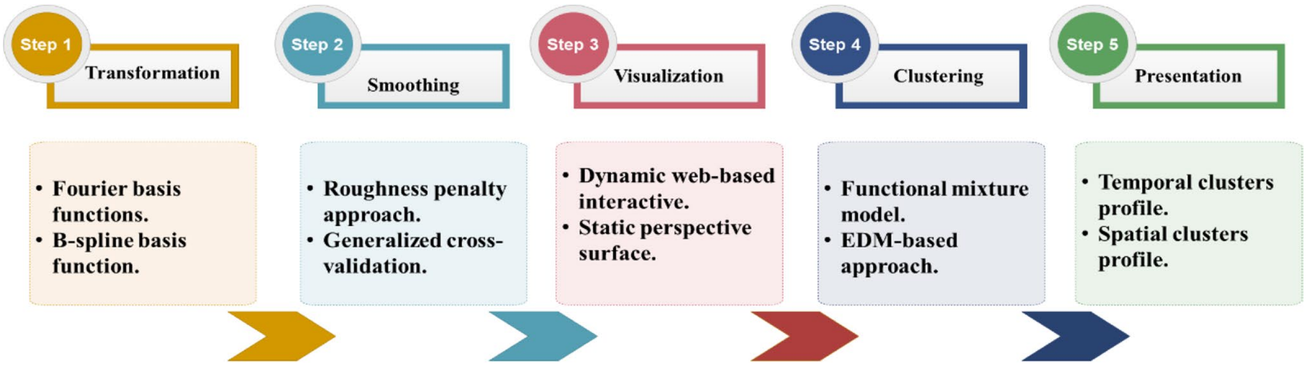
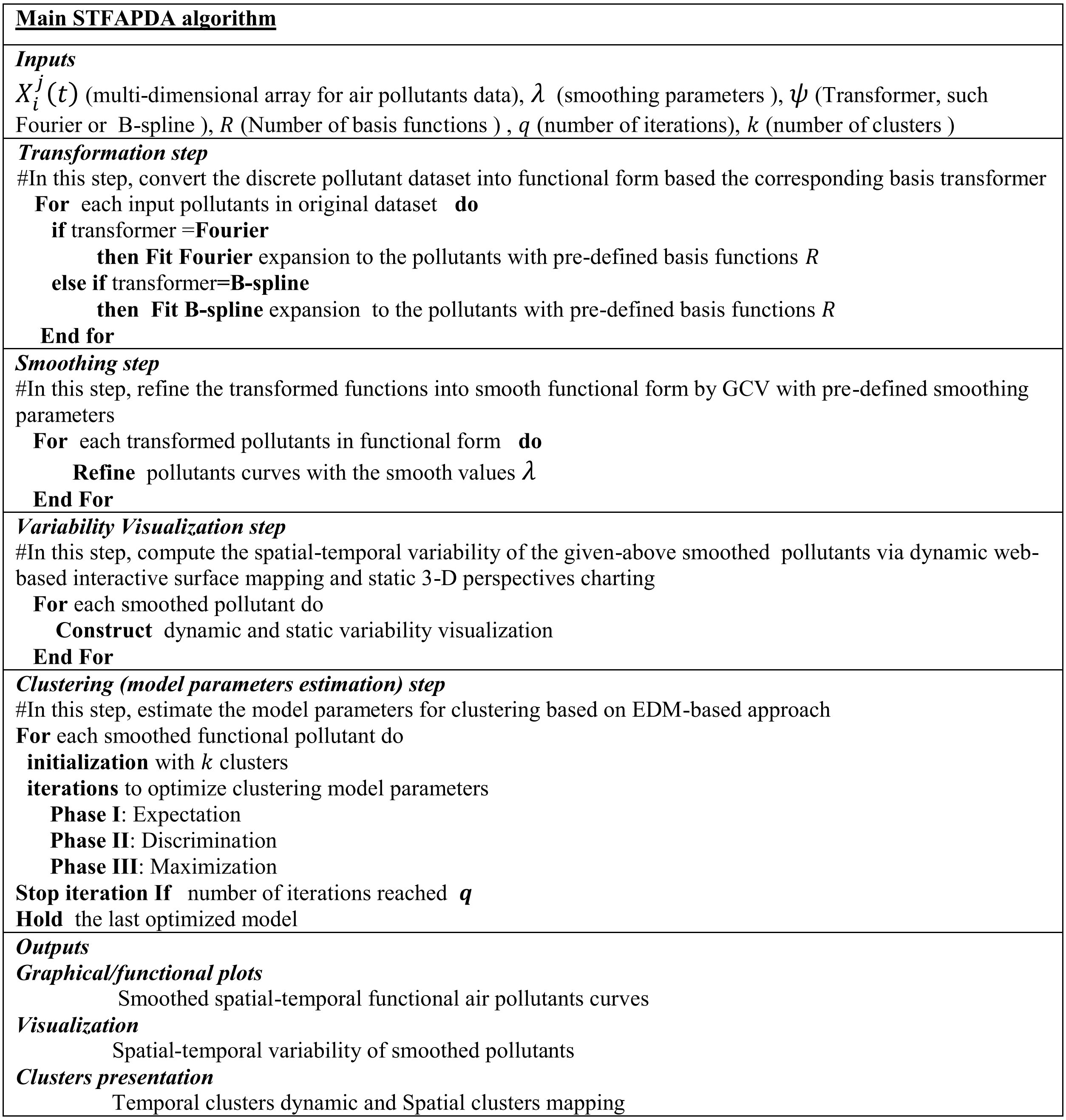

Step 1
Transformation
Fourier basis functions.
B-spline basis function.
Step 2
Smoothing
Roughness penalty approach.
Generalized cross-validation.
Step 3
Visualization
Dynamic web-based interactive.
Static perspective surface.
Step 4
Clustering
Functional mixture model.
EDM-based approach.
Step 5
Presentation
Temporal clusters profile.
Spatial clusters profile.

Fig. 3 The flowchart of the proposed methods in the current study



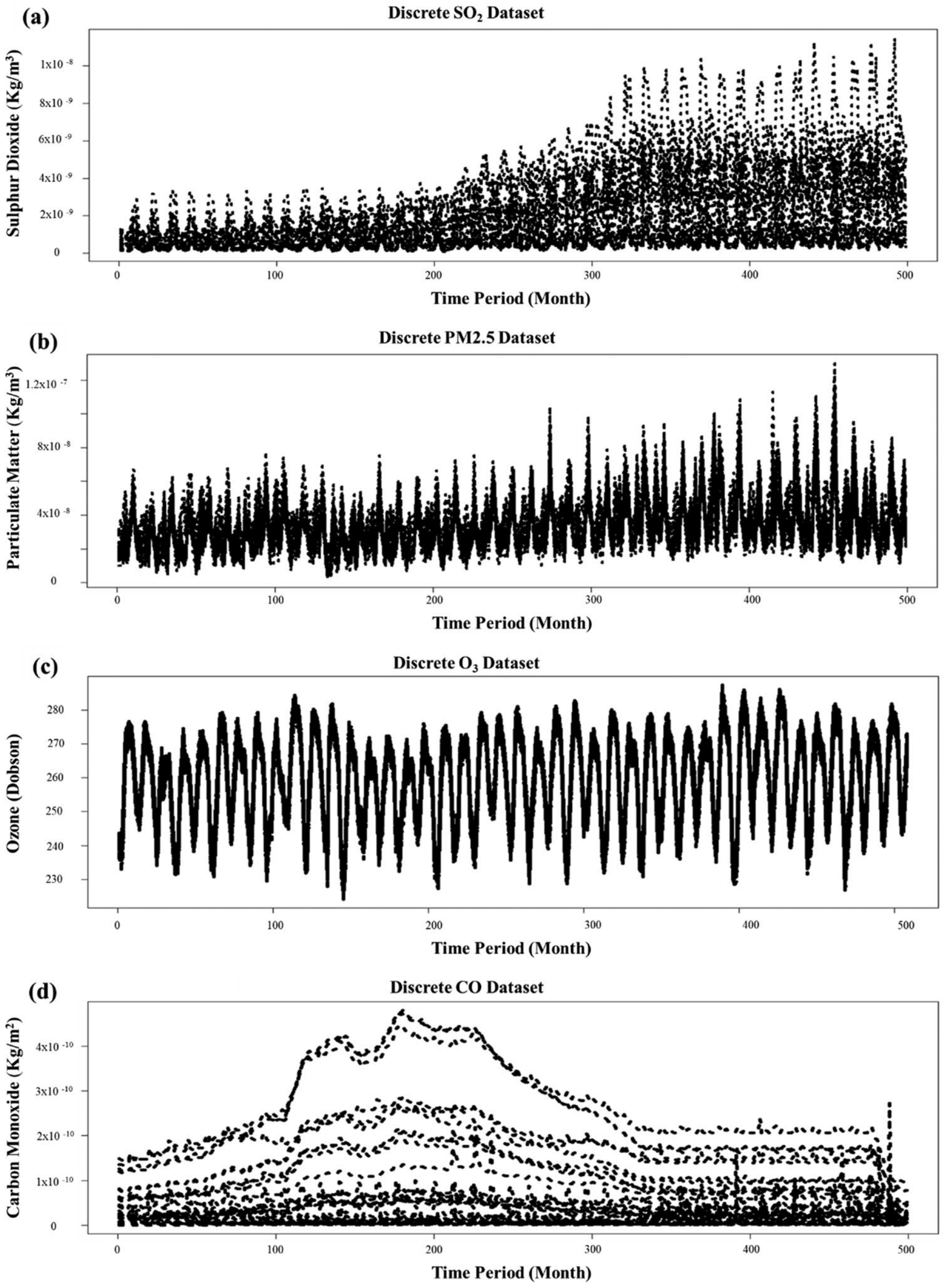
```
Main STFAPDA algorithm
Inputs
$X_i^j(t)$ (multi-dimensional array for air pollutants data), $\lambda$ (smoothing parameters ), $\psi$ (Transformer, such
Fourier or B-spline ), $R$ (Number of basis functions ) , $q$ (number of iterations), $k$ (number of clusters )
Transformation step
#In this step, convert the discrete pollutant dataset into functional form based the corresponding basis transformer
  For each input pollutants in original dataset do
    if transformer =Fourier
        then Fit Fourier expansion to the pollutants with pre-defined basis functions $R$
    else if transformer=B-spline
        then Fit B-spline expansion to the pollutants with pre-defined basis functions $R$
  End for
Smoothing step
#In this step, refine the transformed functions into smooth functional form by GCV with pre-defined smoothing
parameters
  For each transformed pollutants in functional form do
      Refine pollutants curves with the smooth values $\lambda$
  End For
Variability Visualization step
#In this step, compute the spatial-temporal variability of the given-above smoothed pollutants via dynamic web-
based interactive surface mapping and static 3-D perspectives charting
  For each smoothed pollutant do
     Construct dynamic and static variability visualization
  End For
Clustering (model parameters estimation) step
#In this step, estimate the model parameters for clustering based on EDM-based approach
For each smoothed functional pollutant do
 initialization with $k$ clusters
 iterations to optimize clustering model parameters
    Phase I: Expectation
    Phase II: Discrimination
    Phase III: Maximization
Stop iteration If number of iterations reached $q$
Hold the last optimized model
Outputs
Graphical/functional plots
            Smoothed spatial-temporal functional air pollutants curves
Visualization
            Spatial-temporal variability of smoothed pollutants
Clusters presentation
            Temporal clusters dynamic and Spatial clusters mapping
```

Results and discussion

Transforming and smoothing data

The discretization datasets of four primary air pollutants over 37 major cities in Yemen during the period from 1980 to 2022 are given in Fig. 4. Discretization is the process of dividing continuous data into discrete intervals or bins. In the context of air pollution data, discretization is often used to convert continuous measurements of pollutants (e.g., in micrograms per cubic meter) into categorical levels (e.g., low, medium, high). The discretization points chosen are determined by the nature of the data and the research question. As a result, it is critical to remember that the discretization points used can have a significant impact on the analysis's outcome, and it may be necessary to experiment with different options in order to find the best representation of the data. In Fig. 5, discrete air pollution data (CO, SO_2, PM2.5, and O_3) with the spatial elements are presented. The initial step is converting the discretely observed air pollutants curves (CO, SO_2, PM2.5, and O_3) into continuous functional objects for reconstructing the data framework. In

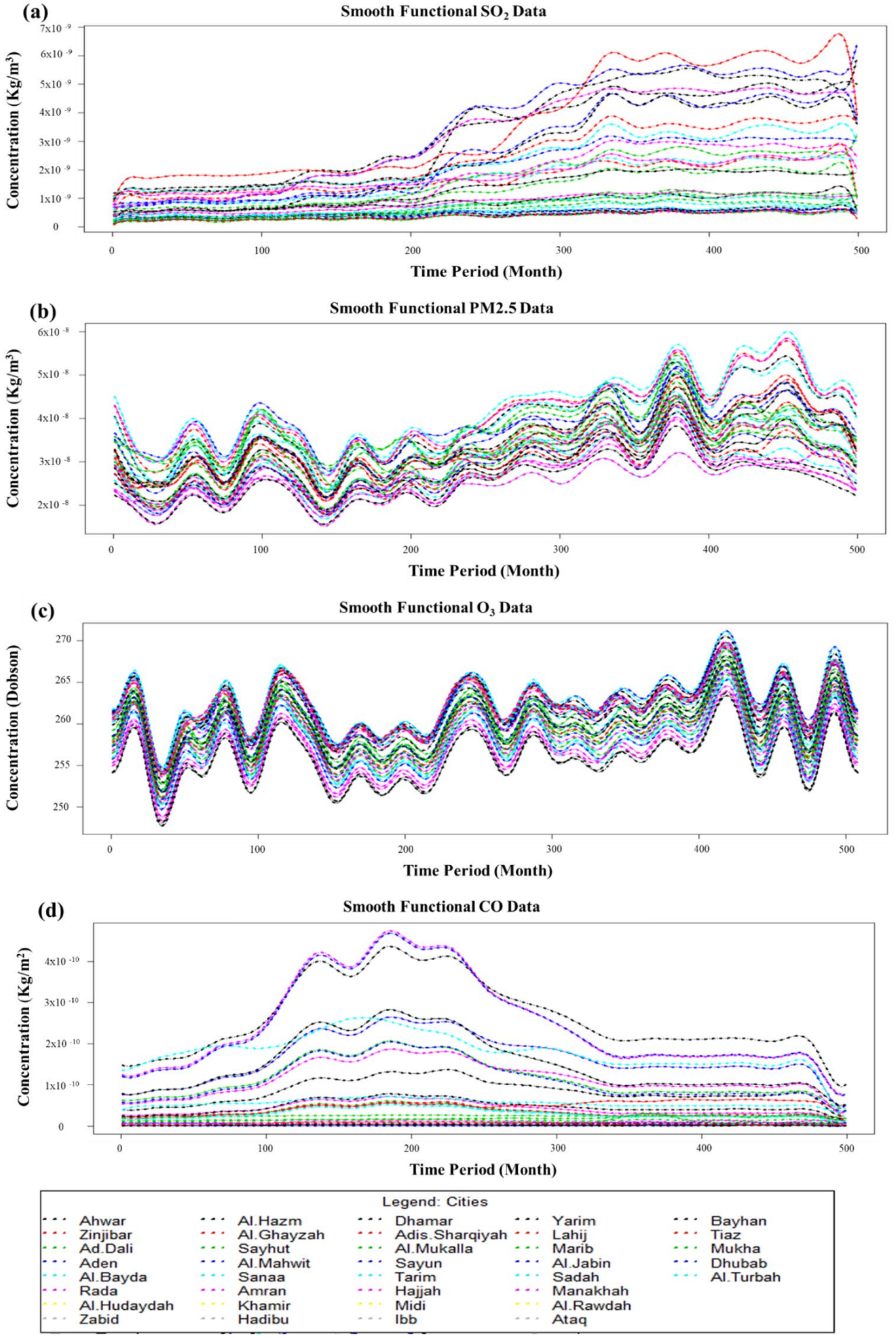
Fig. 4 Discretization points for **a** SO_2, **b** PM2.5, **c** O_3, and **d** CO datasets over 37 cities in Yemen



particular, Fourier basis functions are applied to the O_3 and PM2.5 air pollutants curves, which exhibit seasonal variability throughout the entire data domain. It is observed that some functional O_3 and PM2.5 data objects have a high level of fluctuations, which is unusual.

Spatial–temporal functional data analysis is a technique that can be used to transform and smooth air pollution data. It involves analyzing data over both space and time and modeling the data as a collection of functions rather than a set of discrete points. This allows for a more accurate representation of the data, as well as the ability to smooth and interpolate missing values. This technique utilizes various methods such as functional principal component analysis and functional regression to analyze the data and gain insights from it. Additionally, it allows modeling the temporal and spatial correlations in the data, which can help to understand the underlying patterns and trends in the air pollution data. Transforming and smoothing air pollution data involves several steps: data preprocessing, data transformation, data smoothing, data visualization, and data modeling. Data preprocessing involves cleaning and preparing the data for analysis. Data transformation

Fig. 5 Transformed and smoothed functional data of multivariate air pollutants over multiple cities **a** SO_2, **b** PM2.5, **c** O_3, and **d** CO (legend below)



involves converting the data into a more suitable format for analysis. Data smoothing involves removing noise or random variations from the data to make patterns and trends more visible. Data visualization involves creating visual representations of the data to gain insights and identify patterns and trends. Data modeling involves developing statistical or machine learning models to better understand the data and make predictions about future air pollution patterns. These steps are not always applied in a linear fashion and the analyst may have to iterate over the process to reach a final solution. Figure 5 depicts the charts of transforming and smoothing multivariate functional air pollutants data for multiple locations. The data of transforming and smoothing from air pollutants (CO, SO_2, PM2.5, and O_3) will be discussed in the following subsections.

Table 2 Details on type and number of basis functions, smooth method, and parameters for all functional variables

Functional variable	Type of basis	No. of basis functions	Smoothing technique	Smoothing parameter
SO_2	Cubic B-spline basis	30	GCV	0.06
PM2.5	Fourier transformation	30	GCV	0.04
O_3	Fourier transformation	35	GCV	0.03
CO	Cubic B-spline basis	25	GCV	0.05

Analysis of SO_2 concentrations

The process of converting discrete data to a continuous functional form is known as functional data analysis. B-spline functions are commonly used in this process because they are flexible and can be used to approximate a wide range of shapes. B-splines are piecewise polynomial functions that are defined over a set of control points, called knots. When applying B-spline functions to discrete SO_2 pollution data, the goal is to find a smooth curve that accurately represents the underlying trend in the data. However, in some cases, the resulting curve may have a rough level that can be improved using a smoothing method. The GCV standard is a commonly used method for smoothing functional data. It is a computational method that allows for the estimation of the optimal smoothing parameter, which controls the degree of roughness in the final curve. The present smoothing aims to improve the structure of the functional SO_2 data and facilitate its subsequent interpretation and analysis. This can be useful for identifying patterns and trends in the data that may be difficult to discern when working with discrete data. Additionally, the continuous functional form of the data can be used in further statistical analysis, such as regression or model fitting. Figure 5a describes the use of B-spline bases function with cubic degree and a smoothing parameter of 0.06 to transform and smooth SO_2 air pollutant data. This process makes the data more informative and allows for an easy visual representation of the shape of the SO_2 pollutants over the entire domain. From Fig. 5a, it is clear that the SO_2 shape has two significant periods of long-term changes, with higher variations in the middle and end of the domain. The highest peak of SO_2 is located around the year 2021 with an average pollutant level of 0.0000007 kg/m^3. Table 2 gives more details on the type and number of basis functions, smooth method, and parameters for all functional variables. Previous research has shown that the functional form has not been used to study SO_2 in the past, but there have been studies using the classical form. The study by Al-Janabi et al. (2020b) used a different analysis method and included other pollutants, whereas the current study focuses specifically on SO_2 and processes the data in a discontinuous/discretization form without any data transformation or smoothing, which sets it apart from the aforementioned study.

Analysis of PM2.5 concentrations

Fourier basis functions are used to adapt to the seasonal variability of PM2.5 pollution in the entire data domain. The structure of functional PM2.5 air pollutant data is also improved using a penalized roughness method based on generalized cross-validation criteria for easier interpretation. Thirty Fourier bases functions and a smoothing parameter of 0.04 are used for converting and smoothing PM2.5 air pollutant data. The smoothing parameters determine the strength of the smoothing applied to the data. A smaller smoothing parameter will result in a smoother function with less noise but may also smooth out important features or outliers in the data. A larger smoothing parameter will result in a less smooth function with more noise but will also be more responsive to changes in the data, including outliers. The use of Fourier basis functions in this context is to transform the data into a different domain where it can be more easily smoothed. Smoothing functions can be configured to handle outliers in various ways such as excluding or down-weighting them or treating them as valid data points. Additionally, outliers can be handled before the smoothing step through preprocessing techniques such as removing or transforming the data. Table 2 and Fig. 5b provide additional information and a graphical representation of PM2.5 pollutant data. The proposed procedures make the data more informative and easier to understand. Figure 4b shows that the PM2.5 shape exhibits significant variations at the end of the time domain, which indicates a change in the levels of the pollutant compared to the rest of the time domain. This could potentially be due to changes in industrial activity, transportation patterns, or other factors that contribute to PM2.5 emissions. It is important to further investigate the cause of these variations to understand the impact on air quality and take appropriate measures to reduce emissions. Additionally, Fig. 5b also highlights that some functional PM2.5 data objects have a high level of fluctuations. These fluctuations could be caused by various factors such as weather conditions, changes in population density, and industrial activity. These fluctuations can have a significant impact on air quality and public health, so it is important to analyze them more deeply in later sections of the research. This could include identifying specific sources of emissions, evaluating the effectiveness of current air quality regulations, and developing strategies to reduce emissions and improve air quality. The highest peak of PM2.5 is located at the time domain labeled around 450 to 470, corresponding to the years 2017–2018 with an approximate pollutant average of 0.000006 kg/m^3. Previous studies have examined the PM2.5 pollutant variable using a functional form, including King et al. (2018) and more recently Acal et al. (2022). King et al. (2018) used both B-spline and restricted maximum likelihood (REML) techniques to transform and smooth the PM2.5 data, while Acal et al. (2022) only used B-spline smoothing. In contrast, recent studies have also analyzed PM2.5 pollutant data without any transformation or smoothing, such as Al-Janabi et al. (2021) and Wang et al. (2022).

Analysis of O_3 concentrations

There are several methods for transforming and smoothing O_3 concentrations. One common method for transforming O_3 concentrations is to convert them from a raw measurement to a pollutant index, such as the Air Quality Index (AQI), which provides a more easily understandable and comparable value. Another method is to apply a smoothing technique, such as moving average or lowest smoothing, to reduce the impact of random measurement errors and to better identify trends and patterns in the data. It's important to note that before applying any transformation or smoothing technique, it is necessary to ensure that the data is of good quality and that any outliers or missing values have been properly handled.

The use of Fourier basis functions for adapting to O_3 pollutant data with seasonal variability is a common approach in functional data analysis. The Fourier basis functions are able to capture the periodic nature of the data and can be used to model and smooth the data (Hael 2021). The use of the penalized roughness method with the GCV criteria is a method for selecting the optimal smoothing parameter (Guo et al. 2022). The GCV criteria is a way of evaluating the performance of the model by comparing the observed data to the predicted data. The penalized roughness method is a way of controlling the smoothness of the model by adding a penalty term to the objective function. By using the GCV criteria, the optimal smoothing parameter is chosen to minimize the difference between the observed and predicted data while also controlling the smoothness of the model. In this specific case, thirty-five Fourier basis functions and a smoothing parameter of 0.03 were used for converting and smoothing the O_3 pollutant data. This approach aims at reducing the high fluctuations level present in some functional O_3 data objects. The overall goal is to improve the structure of functional O_3 air pollutant data and make it with enhanced sights. Table 2 and Fig. 5c provide additional information and a graphical representation of the O_3 pollutant data. The transformation and smoothness procedures make the data more informative and easier to understand. Figure 5c shows that the functional shape of the O_3 pollutant has several peaks, indicating significant dynamic changes over time. The highest peaks occur in the time domain labeled around 410 to 430, corresponding to the years 2014–2015, with an average pollutant level of 265 Dobson. The functional analysis framework used in this study is similar to that of a study by (Bouveyron et al. 2022) in that it uses the Fourier basis to reconstruct the functions because the O_3 pollutant exhibits clear periodicity. Additionally, several recent studies have examined O_3 pollutant data without any alteration or smoothing, such as those by Yang et al. (2020), Liu et al. (2022), and Shams et al. (2022). These studies differ from the current one in terms of the way they analyze and process O_3 pollutant data.

Analysis of CO concentrations

In this study, B-spline functions are being applied to discrete CO pollution data to convert it to a continuous functional form. The rough level of some curves can be improved using a smoothing method, and the GCV standard is used to handle the transformed functional data. This smoothing aims to improve the structure and interpretability of the functional CO data. Twenty-five B-spline bases functions with a cubic (fourth-order) degree and a smoothing parameter of 0.05 are being used for this transformation and smoothing process. The use of the GCV criteria to penalize the functional CO data suggests that the data with high noise levels are being treated as outliers. Outliers in this context are likely to be data points that are significantly different from the rest of the data and do not fit the underlying structure of the functional data. By penalizing these data points, the noise is removed, and the structure of the functional CO data is improved. However, it is important to note that it's not always correct to consider that these data are outliers, they might be real events, so it is necessary to examine the data carefully and consider other factors before removing any data. One possible way to avoid removing real events and to smooth the data is to use moving averages. Table 2 and Fig. 5d provide detailed information about the process of CO data and its graphical representation. The proposed procedures make the CO pollution data more informative and easier to understand. The table displays the numerical values of the CO pollution data, while Fig. 5d provides a visual representation of the data. From Fig. 5d, it is clear that the shape of the CO pollutants has high significant variations over a long period of time. The CO concentrations started in the time domain labeled around 100, corresponding to the year 1990. The highest peak is located in the time domain labeled around 195, corresponding to the year 1996 with an approximate pollutant average of 0.000004 kg/m^2. Additionally, a more accurate analysis for CO curve clustering will be explained in "Clustering of functional CO data"

Section of the paper. This section will provide a deeper understanding of the data and possible patterns in pollution levels over time. This will assist in identifying areas that need to be addressed for pollution control and management. CO had not been studied before using the functional form, according to earlier research. However, this pollutant has received some traditional research. Al-Janabi et al. (2021) and Li et al. (2022d) currently studied the CO pollutant using different analysis methods in a recent study. As a result, the current study differs from their study in that the data was processed in a discontinuous/discretization form without any transformation or smoothing.

Visualizing the spatiotemporal variability

Visualizing the spatiotemporal variability of data for major pollutants is important for several reasons: (i) understanding patterns and trends: visualizing the data can help identify patterns and trends in pollutant levels over time and in different locations. this information can inform decision-making and management strategies for reducing pollution, (ii) identifying high-risk areas: by visualizing the data, it is possible to identify areas or times where pollutant levels are particularly high or variable. this can help target interventions and resources to areas that need them the most, (iii)

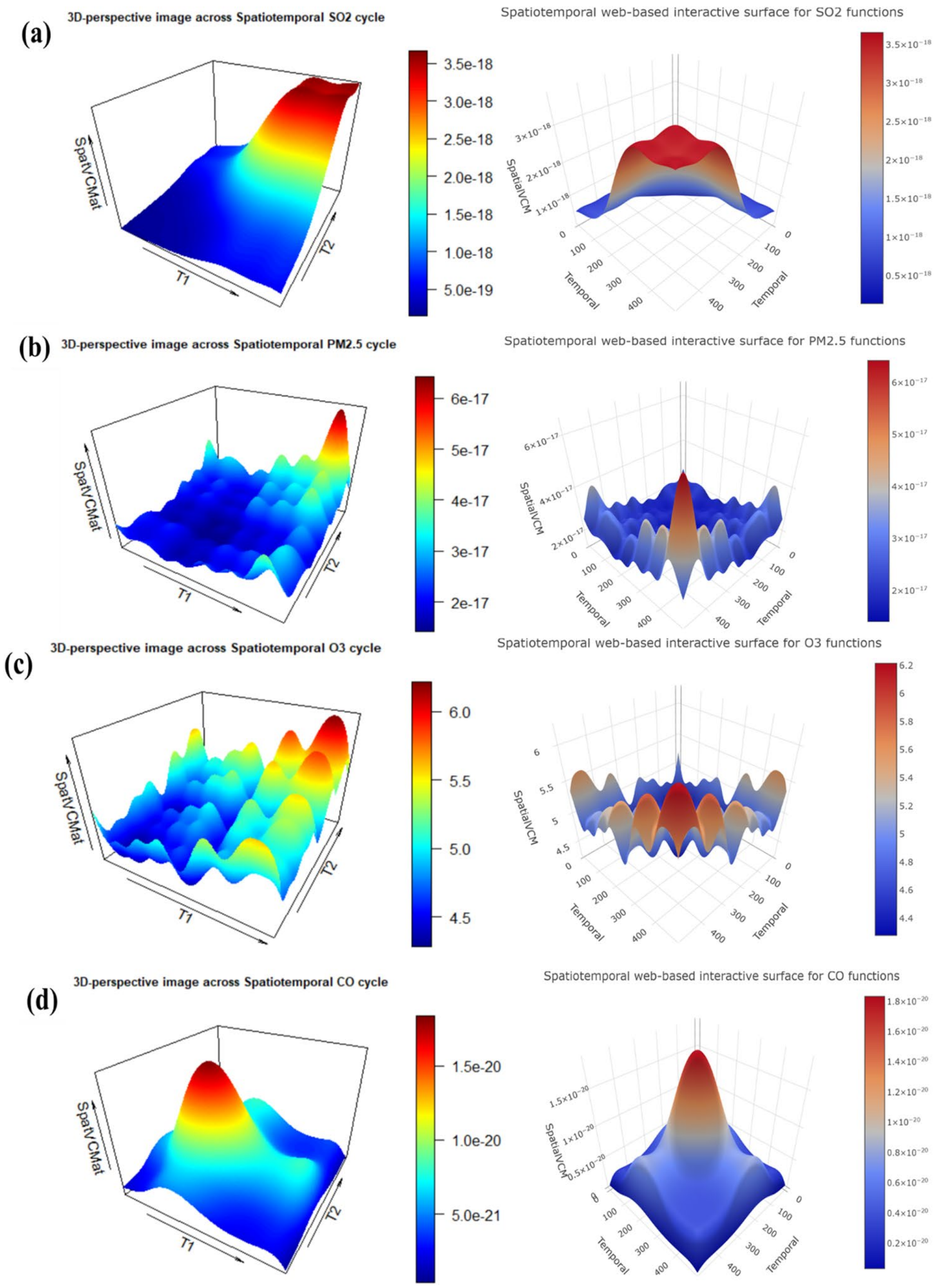


Fig. 6 3D-visualization of Spatiotemporal functional data for multivariate air pollutants using dynamic web-based interactive method (right penal), and static perspective surface (left panel) **a** SO_2, **b** PM2.5, **c** O_3, and **d** CO

communication and education: visualizing the data can be an effective way to communicate and educate the public about air pollution and its impacts. by providing clear and easy-to-understand information, it can help to raise awareness and engage the community in efforts to reduce pollution, (iv) monitoring progress: visualizing the data over time can help to track the progress of pollution reduction efforts and evaluate the effectiveness of different management strategies, (v) decision-making and policy formulation: the spatiotemporal variability data when visualized, it can be used to evaluate the distribution, intensity, and frequency of pollutants, which can be used as an input for decision-making and policy formulation (Li et al. 2022a; Nikolaou et al. 2023). Overall, visualizing the spatiotemporal variability of data for major pollutants can provide valuable insights that can help to improve air quality and protect public health.

The use of static 3-D perspectives charting and dynamic web-based interactive surface mapping as visualization tools for high-dimensional, voluminous, and spatial–temporal pollutants data is a promising approach. Static 3-D perspectives charting allows for the visualization of the data in a three-dimensional space, providing a clear and intuitive representation of the distribution of pollutants over time and in different locations. This can be particularly useful for identifying patterns and trends in the data, and for identifying areas or times where pollutant levels are particularly high or variable. Dynamic web-based interactive surface mapping, on the other hand, allows for the creation of interactive maps that can be viewed on the web. This can be useful for visualizing the data in a geographic context, and for exploring the data in more detail. The interactivity allows users to navigate and zoom in and out of the data, and to view the data at different levels of detail. Both techniques can be used to analyze the spatial–temporal data for four major pollutants variables (SO_2, PM2,5, O_3, and CO) and provide a comprehensive understanding of the patterns and trends in the data, as shown in Fig. 6. The use of these visualization tools can be very effective for data analysis, decision-making, and policy formulation, and for communicating and educating the public about air pollution and its impacts. The next subsections analyze the spatial–temporal data for four major pollution variables (SO_2, PM2,5, O_3, and CO).

Visualizing of SO_2 data

The visualization in Fig. 6a shows two different tools for displaying spatiotemporal variation in SO_2 pollutant data. The first tool is a static 3D chart with a gradient heatmap (left penal), while the second tool is a dynamic, web-based interactive surface map (right penal). The proposed two tools for visualizing SO_2 pollutants can aid users in understanding the spatiotemporal dynamics of low and high air pollution over a long-term period. The dynamic web-based interactive surface mapping tool, in particular, has interactive features that allow users to click buttons to view the concentration of SO_2 pollution. The heatmap metrics in this tool provide detailed information on the structure of heterogeneity and fluctuations in SO_2 levels. Specifically, a red color on the gradient heatmap indicates intense high variations across spatial and temporal SO_2 cycles, while a blue color on the gradient heatmap indicates low variations across these cycles. It is clear that the highest variability of SO_2 occurs from 2007 to 2021, with insignificant variations observed in the rest of the time interval, particularly at the beginning of the interval. Using the same technique, (Ranaarif and Yuwono 2021) recently observed variations in the concentration and spatial distribution patterns of SO_2 pollutants in Bali Island from 2011 to 2020. The main difference between current SO_2 visualization tools and the study by (Ranaarif and Yuwono 2021) is that the latter used traditional visualization approaches to analyze changes in the concentration and spatial distribution patterns of SO_2 pollutants, while current tools may use more advanced techniques such as 3D modeling or interactive maps.

Visualizing of PM2.5 data

The visualization of variance–covariance 3D surfaces for spatiotemporal functional PM2.5 pollutant data is an important tool for understanding and identifying patterns in air quality. Two different methods have been implemented for the visualization of spatiotemporal PM2.5 variation. The first is a static 3-D perspective chart with a gradient heatmap, as shown in the left panel of Fig. 6b. The second is a dynamic web-based interactive surface mapping that provides enhanced animation, as shown in the right panel of Fig. 6b. The implementation of these visualization tools is crucial for assisting users in developing an awareness of the air quality in a specific area. It allows users to identify the spatiotemporal dynamics of low and high air pollution over a long-term period, as well as provide detailed information about the structure of heterogeneity and fluctuations in PM2.5 concentrations. For example, the red color on the gradient heatmap indicates an intense high variation across spatial and temporal PM2.5 cycles, while the blue color on the gradient heatmap indicates a low variation across spatial and temporal PM2.5 cycles. The PM2.5 visualization results have also revealed different fluctuations throughout the whole domain, with the highest fluctuation period of the PM2.5 pollutants observed between 2016 and 2020. Additionally, the dynamic web-based interactive surface mapping offers interactive features such as buttons that allow users to view the PM2.5 pollution concentration in an easy-to-understand format, making it a powerful tool for understanding air quality patterns. In addition to the visualization methods described earlier, it is important to note that some existing visualization approaches are limited in their ability to process large PM2.5 data sets and support dynamic visualizations. To compare the current PM2.5 visualization tools with recent previous studies, the current study can be compared to the approach used by Li et al. (2016) who employed a visualization approach to analyze the air quality index based on PM2.5 pollutant in Beijing, China. Similar studies include those by Tang et al. (2021) and Medhi and Gogoi (2021) used visualization tools to analyze the impact of COVID-19 on PM2.5 in China and India, respectively. All of these previous studies have used traditional visualization and interpolation methods which differ from the current study which used spatial–temporal dynamic visualization for PM2.5 functions. This current study presents a novel approach that addresses the limitations of traditional visualization methods and provides a more comprehensive understanding of PM2.5 data.

Visualizing of O_3 data

The visualization of variance–covariance 3D surfaces for spatiotemporal functional O_3 pollutant data is shown in Fig. 6c. Two different tools have been implemented for the visualization of spatiotemporal O_3 variation. The first one is a static 3-D perspective charting with a gradient heatmap drawn in Fig. 6c (left panel). The second one is a dynamic web-based interactive surface mapping which provides enhanced animation drawn in Fig. 6c (right panel). The visualization of O_3 pollutants could help the users to identify the spatiotemporal dynamic of low and high air pollution over a long-term period. Moreover, in the dynamic web-based interactive surface mapping, interactive features are designed where users are able to click buttons to view the O_3 pollution concentration. It is easy to understand the O_3 variability by heatmap metrics which provide detailed information regarding the structure of heterogeneity and fluctuations. Specifically, the red color on the gradient heatmap indicates an intensely high variation across spatial and temporal O_3 cycles, and the blue color on the gradient heatmap indicates a low variation across spatial and temporal O_3 cycles. The O_3 visualization results discovered different fluctuations throughout the whole domain, with the highest fluctuation period of the O3 pollutant being during the period from 2013 to 2017. Additionally, recent studies (Nurgazy et al. 2019; Gagliardi and Andenna 2020; Ahmad et al. 2022) have used machine learning approaches for the visualization of O_3 pollutants, such as the presentation of CAVisAP, a context-aware system for outdoor air pollution visualization by using internet of thing (IoT) platforms, and the exploration of surface ozone behavior by using machine learning approaches. However, different from these studies, the current analysis considers the dynamic visualization of O_3 pollutants over spatial–temporal.

Visualizing of CO data

The visualization of variance–covariance 3D surfaces for spatiotemporal functional CO pollutant data is shown in Fig. 6d. Two different tools have been implemented for the visualization of spatiotemporal CO variation. The first one is a static 3-D perspective charting with a gradient heatmap, as depicted in the left panel of Fig. 6d. The second tool is a dynamic web-based interactive surface mapping, which provides enhanced animation and is shown in the right panel of Fig. 6d. It is observed that the significant central part of the high CO variability occurs in the periods from January 1990 to 2000 and it is insignificant elsewhere. In recent literature, the CO pollutant concentrations have been analyzed with other pollutants using traditional and machine learning methods. For example, Grace et al. (2020) used a traditional visualization approach to analyze CO pollutant concentrations with other pollutants. Jain and Kaur (2021) proposed machine learning and visualization techniques for the analysis of air pollution concentrations during the COVID-19 pandemic. The clustering method used in the study aims to identify and differentiate different layers within the temporal and spatial cycle of air pollution. The results of this analysis, including any functional clustering of temporal and spatial pollution variables, will be discussed in a subsequent section of the study.

Clustering the spatiotemporal functional data

Clustering is a technique used in machine learning to group similar data points together. In the context of spatiotemporal functional data of air pollutants, clustering can be used to group locations with similar air pollution patterns over time. This can be useful for identifying hotspots of pollution and for understanding the factors that contribute to air pollution in different areas. There are several different clustering algorithms that can be used, such as k-means, hierarchical clustering, and density-based clustering (Schmutz et al. 2020). The choice of algorithm will depend on the specific characteristics of the data and the research question being addressed. In order to cluster spatiotemporal functional data, it is necessary to first extract relevant features from the data that capture the patterns of interest (Hael et al. 2021). This could include measures of the overall level of pollution, the variability of pollution over time, or the similarity of pollution patterns between different locations. These features can then be used as inputs to the clustering algorithm (Shi et al. 2022). Once the clusters have been identified, various visualization and analysis techniques can be used to explore the results. This can include mapping the clusters to examine their spatial distribution, plotting the time series data for each cluster to examine their temporal patterns, and comparing the pollution levels and patterns across different clusters

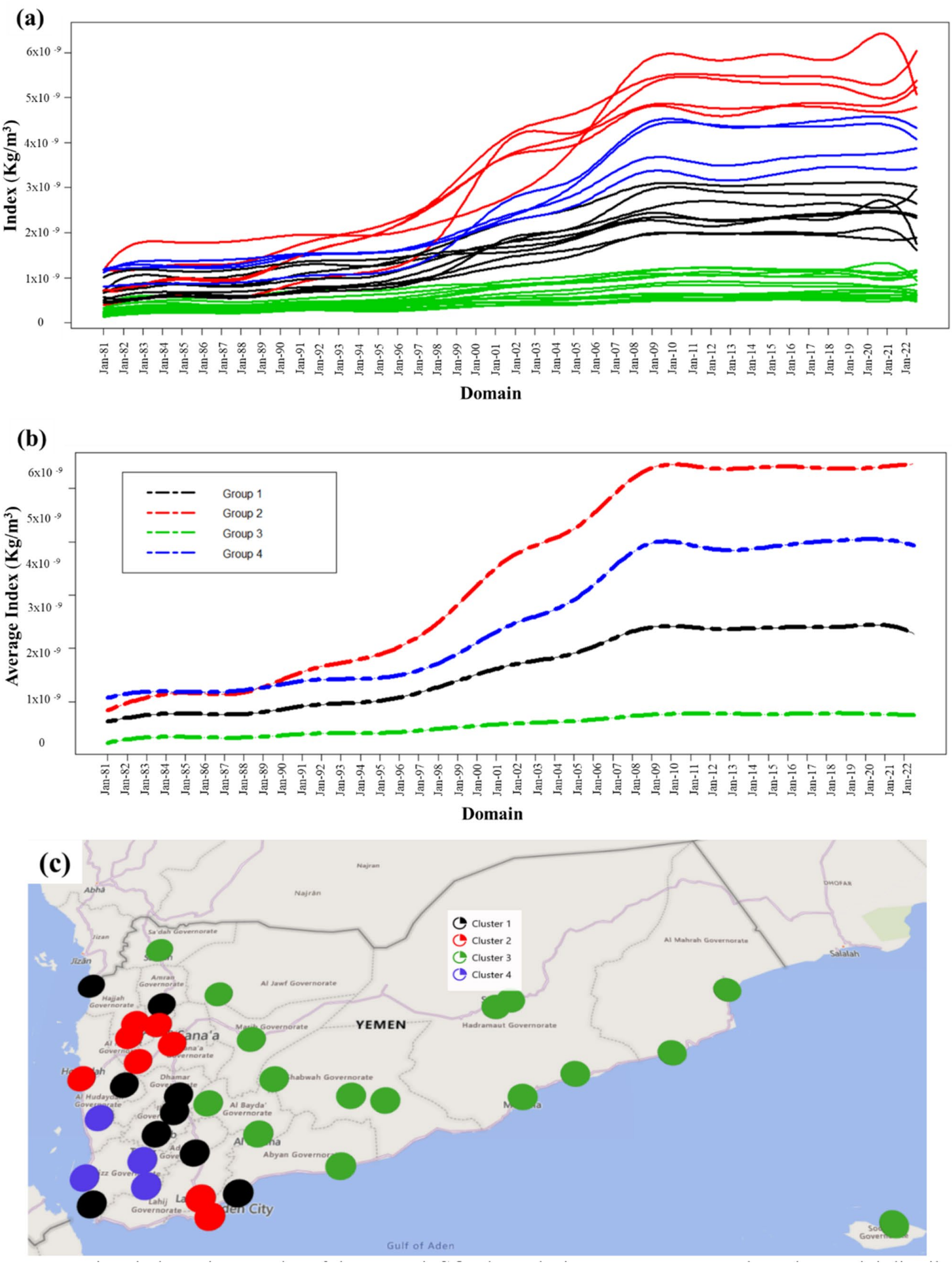


Fig. 7 **a** Functional clustering results of the smooth SO_2 data, **b** its group means, and **c** the spatial distribution map of the obtained clusters

Table 3 Summary of the cluster characteristics for spatial–temporal SO_2 patterns in cities of Yemen

Pollutant	Classifier	Color	Cities/locations	Ratio	Features (shape/size)	Pollution degree
SO_2	Class A	Black	Zinjibar, Ad Dali, Khamir, Dhamar, Midi,Ibb,Yarim, Al-Jabin, Dhubab	24.32%	Snaky-line shape after 1997	Moderate pollution
	Class B	Red	Aden, Al-Hudaydah, Al-Mahwit, Sanaa, Amran, Hajjah, Lahij, Manakhah	21.62%	Three-profiles shape, constant before 1990, then high increase after 1991, stable after 2011	Extremally pollution
	Class C	Green	Ahwar, Al-Bayda,Rada, Al-Hazm, Al-Ghayzah, Sayhut Hadibu, Adis-Sharqiyah, Al-Mukalla, Sayun, Tarim, Marib, Sadah, Al-Rawdah, Ataq, Bayhan	43.24%	No changes with straight-line shape through all domain	Clean
	Class D	Blue	Zabid,Tiaz, Mukha, Al-Turbah	10.81%	Three-profiles shape, constant before 1990, then increase after 1991, stable after 2011	High pollution

to identify key differences. The study presented an advanced clustering approach based on functional mixture models to effectively deal with high-dimensional, large-scale, and spatial–temporal air pollutants data. The results of this functional clustering approach provided meaningful insights into the spatial–temporal dynamics of pollutants and informative visualizations for both temporal cluster profiles and spatial cluster mapping. In this section, clustering the spatiotemporal functional data of air pollutants (SO_2, PM2.5, O_3, and CO) will be discussed.

Clustering of functional SO_2 data

Sulfur dioxide (SO_2) is a gaseous air pollutant composed of sulfur and oxygen. It can have a significant impact on human health, animal health, and the environment. In this study, the functional clustering of smooth functional SO_2 pollutant levels was conducted over multiple cities in Yemen and the results were used to identify groups of cities with similar levels of pollution. The spatial distribution of these clusters is presented in Fig. 7. Cluster group 3, indicated by the color green, is identified as having the least polluted cities in terms of SO_2. These cities are primarily located in southern Yemen and represent approximately half of the total cities studied, including the Al-Mahrah Governorate, Hadramout Governorate, Abyan Governorate, and Shabwah Governorate. The largest source of sulfur dioxide emissions is the burning of high-sulfur fossil fuels by heavy equipment power plants and other industrial facilities. Other sources of sulfur dioxide emissions include natural sources such as volcanoes and industrial processes such as extracting minerals from ore, as well as ships, locomotives, and other vehicles. Table 3 provides more detailed information about the memberships, spatiotemporal features, and degree of pollution for each cluster group. It is worth noting that short-term exposure to sulfur dioxide can damage the respiratory system, especially lung function, and irritate the eyes. It causes coughing and mucus secretion and exacerbates chronic bronchitis and asthma conditions. Generally, sulfur dioxide emissions led to high sulfur dioxide concentrations in the air, which can form other sulfur oxides (SOx) that can be harmful as well. Based on a thorough review of recent literature, it appears that the clustering of spatial–temporal dynamics of SO_2 data using a functional data framework has not been studied or investigated before. However, a few studies have analyzed SO_2 pollutant data using vectorized-based methods. For example, the study conducted by Al-Janabi et al. (2021) employed the intelligent prediction method called IFCsAP to handle SO_2 pollutant data along with other pollutant variables. Another study by Kujawska et al. (2022) used an artificial neural network model to forecast sulfur dioxide levels in the air. Both of these studies were focused on forecasting SO_2 pollutant levels using artificial data analytics, while the proposed functional model in this current study aims to cluster the hidden features of spatial–temporal SO_2 dynamics.

Clustering of functional PM2.5 data

In this study, we aim to investigate the dynamic behavior of PM2.5 air pollutants and identify potential spatiotemporal functional clusters over various locations in Yemen. Our proposed approach will provide a meaningful result and offer a graphical interpretation of the spatiotemporal variations of PM2.5. As shown in Fig. 8, our findings indicate that the highest PM2.5 variability is concentrated in the cities located on the western side of the Red Sea, as represented by cluster 2 (red color) which is characterized by two unique peaks: a low-volume peak and large-volume peak. Clusters 4 and 1 also show considerable fluctuations throughout the entire domain. It is worth noting that the PM2.5 air pollution concentration levels significantly decreased during the COVID-19 outbreak period. Table 4 provides detailed information on the cluster memberships, spatiotemporal features, and degree of PM2.5 pollution.

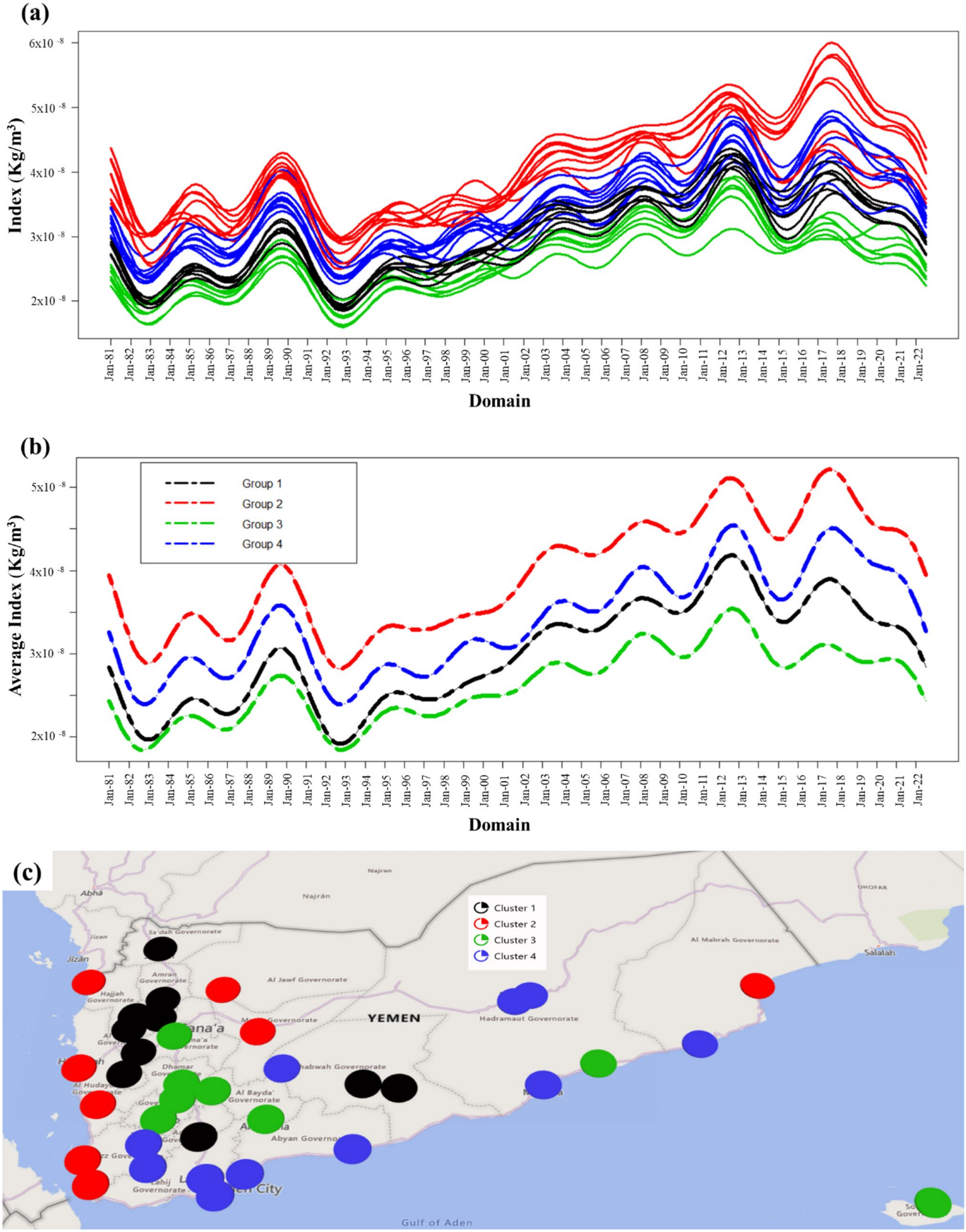


Fig. 8 **a** Functional clustering results of the smooth PM2.5 data, **b** its group means, and **c** the spatial distribution map of the obtained clusters

Table 4 Summary of the cluster characteristics for spatial–temporal PM2.5 patterns in cities of Yemen

Pollutant	Classifier	Color	Cities/locations	Ratio	Features (shape/size)	Pollution degree
PM2.5	Class A	Black	Ad Dali, Al-Mahwit, Amran,Khamir, Hajjah, Al-Jabin, Sadah, Manakhah, Al-Rawdah, Ataq	27.03%	Multiple peaks with medium volumes	Moderate pollution
	Class B	Red	Al-Hudaydah,Zabid,Al-Hazm, Al-Ghayzah, Midi, Marib, Mukha, Dhubab	21.62%	Large-size peak (2017–2018), low-size peak (1989–1990)	Strongly pollution
	Class C	Green	Al-Bayda,Rada, Sanaa, Hadibu, Dhamar, Adis-Sharqiyah,Ibb, Yarim	21.62%	Fluctuations with low-size peaks through domain	Slight pollution
	Class D	Blue	Ahwar,Zinjibar,Aden, Sayhut, Al-Mukalla, Sayun,Tarim, Lahij, Bayhan, Tiaz, Al-Turbah	29.73%	Multiple peaks with medium volumes	Moderate pollution

It is concluded that the levels of PM2.5 in Yemen follow a periodic cycle that is controlled by meteorological factors such as temperature and solar radiation. This conclusion aligns with the findings of Abdul-Rahim et al. (2022) who found that statistical analysis revealed a positive correlation between PM2.5 concentrations and temperature for both fall and summer samples. However, the analysis also revealed a positive correlation between PM2.5 concentrations and relative humidity for fall samples and a negative correlation for summer samples. While their study only focused on a small area in Sanna city, the results may be applicable to many other cities in Yemen. Additionally, previous studies have shown that the levels of PM2.5 in Yemen are influenced by the rhythm of human activities that modulate anthropogenic emission rates. The change in PM2.5 air pollution is tied to population-weighted exposure levels (PWEL). As per the research (Li et al. 2022b), areas with high PWEL and rapid increases in PM2.5 concentrations were primarily found in developing countries such as India, Bangladesh, Nepal, and Pakistan, as well as in the developed country of Saudi Arabia, and the least developed countries of Yemen and Myanmar. Moreover, the study by Fang et al. (2020) found that the regions with the highest levels of pollution are primarily located in China, Southeast Asia, South Asia, West Asia, and North Africa, particularly in the Arabian Gulf region. The study also identified energy intensity and per capita electricity consumption as the primary drivers of PM2.5 concentrations, whereas an expanding forest area was found to significantly decrease PM2.5 concentrations. In recent years, there has been a growing body of research that has focused on studying the PM2.5 pollutant using functional data analysis. For example, a study by Wang et al. (2019) adapted the framework of functional data analysis to compare the fluctuation patterns of PM2.5 concentration between provinces in China from 1998 to 2016, both spatially and temporally. Another study by Liang et al. (2021) used a spatial-functional mixture method to model and cluster PM2.5 concentrations across China. The current study is similar to these two studies in that it also uses the same functional framework. However, there are also several recent studies that have used a different approach to cluster PM2.5 pollutant data, such as the studies by Jorquera and Villalobos (2020), Liu et al. (2020), Su et al. (2020), and Park et al. (2022). These studies differ from the current study in that they use classical analysis frameworks to process and cluster PM2.5 pollutant data.

Clustering of functional O_3 data

The ozone (O_3) is formed when sunlight and heat cause chemical reactions between volatile organic compounds (VOC) and nitrogen oxides (NOX), also known as hydrocarbons. These reactions can occur both near the ground, in the troposphere, and high in the stratosphere. In the stratosphere, O_3 forms a protective layer that shields the Earth from harmful ultraviolet radiation from the sun, but at ground level, O_3 is a harmful air pollutant (Wang et al. 2020). In Fig. 9, the functional clustering of smooth functional ground-level O_3 levels in multiple cities in Yemen is presented, along with the cluster mean and the spatial distribution of the obtained clusters. The data indicates that the highest ozone concentrations are found in coastal cities and islands located on the western side of the Red Sea, the southern side of the Arabian Sea, and the Gulf of Aden. This may be due to the higher levels of pollutants, such as volatile organic compounds and nitrogen oxides, present in these areas, which contribute to ozone formation. Additionally, the unique meteorological conditions in these regions may also make them more susceptible to ozone formation. Table 5 provides a summary of the obtained cluster profiles for ozone air population levels and the degree of air pollution in cities of Yemen.

High ozone concentrations near ground level can have serious consequences for human health, as well as for crops, animals, and other substances. O_3 is a powerful oxidant that can irritate the respiratory system, causing symptoms such as coughing, sore throat, and chest discomfort. People with

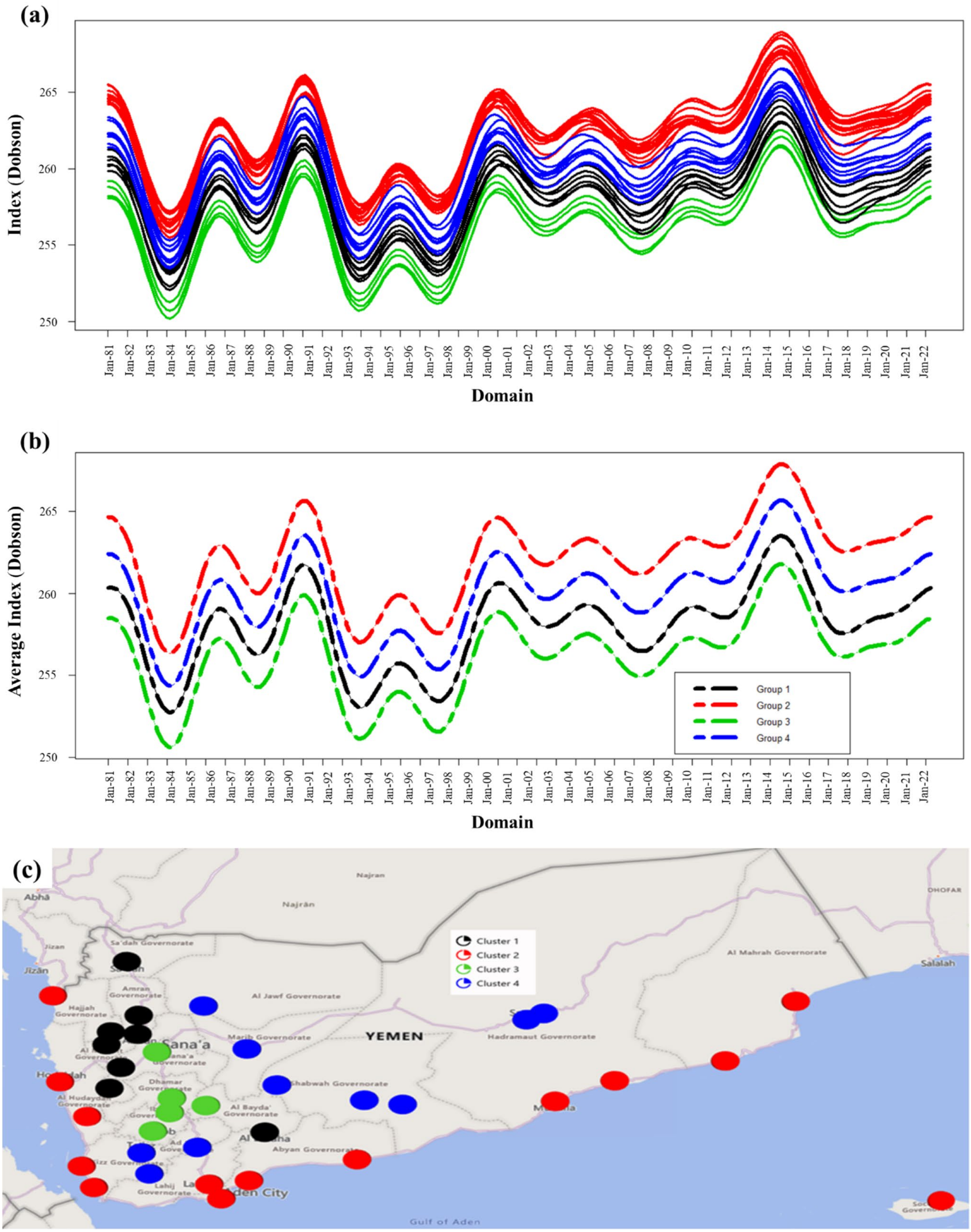


Fig. 9 **a** Functional clustering results of the smooth O_3 data, **b** its group means, and **c** the spatial distribution map of the obtained clusters

Table 5 Summary of the cluster characteristics for spatial–temporal ozone patterns in cities of Yemen

Pollutant	Classifier	Color	Cities/locations	Ratio	Features (shape/size)	Pollution degree
O_3	Class A	Black	Al-Bayda, Al-Mahwit, Amran, Khamir, Hajjah, Al-Jabin, Sadah, Manakhah,	21.62%	Two peaks with moderate volume in 1990 and 2015, multiple peaks with low volume in other years	Moderate pollution
	Class B	Red	Ahwar,Zinjibar, Aden, Al-Hudaydah, Zabid, Al-Ghayzah, Sayhut, Hadibu, Adis-Sharqiyah, Al-Mukalla, Midi, Lahij, Mukha, Dhubab	37.84%	Two peaks with large volume in 1990 and 2015, multiple peaks with medium volume in other years	Severe pollution
	Class C	Green	Rada, Sanaa, Dhamar, Ibb, Yarim,	13.51%	Several peaks with low volume in 1990, 2000, and 2015	Slight pollution
	Class D	Blue	Ad Dali, Al-Hazm, Sayun, Tarim, Marib, Al-Rawdah, Ataq, Bayhan, Tiaz, Al-Turbah	27.03%	Two peaks with large volume in 1990 & 2015, multiple peaks with medium/low volume in other years	High pollution

asthma and other lung conditions are particularly vulnerable to the effects of ozone pollution, as it can worsen their symptoms and increase the risk of respiratory infections. Long-term exposure to O_3 can also lead to inflammation and damage to the cells lining the lungs, which can increase the risk of chronic lung diseases such as bronchitis and emphysema. Additionally, high O_3 concentrations can weaken the immune system's ability to fight off bacterial infections in the respiratory tract. There are many factors that influence the development of ground-level O_3, including wind direction and speed, temperature, timing cycles, and vehicle driving patterns. O_3 is formed when pollutants from cars, power plants, and other sources react with sunlight, so weather conditions play a key role in determining O_3 levels. O_3 is typically a pollutant in the summer, when temperatures are high and sunlight is abundant, and it is a major component of smog in many urban areas during the summer months. Due to its relation to climate conditions, ground-level ozone is also known as "summer smog." It is important to note that Ozone, though it is harmful at ground level, is beneficial in the upper atmosphere where it protects the earth from harmful UV rays.

The analysis in this study builds upon the work of Schmutz et al. (2020) by utilizing a functional clustering framework to analyze O_3 pollutant curves. However, it also diverges from previous studies, such as those conducted by Pineda Rojas et al. (2019) and Saeipourdizaj et al. (2022). Pineda Rojas et al. (2019) employed traditional clustering techniques to examine the spatial patterns that lead to peak ozone hourly concentrations, using Monte Carlo outcomes as the basis for their analysis. On the other hand, Saeipourdizaj et al. (2022) utilized a classical spatiotemporal mixture model-based clustering framework to cluster days of the year 2017, based on hourly O_3 amounts collected from four stations in Tabriz. This study takes a different approach, utilizing the functional clustering framework to analyze O_3 pollutant curves, which sets it apart from these previous studies.

Clustering of functional CO data

In this sub-section, we will present and discuss the main results of spatiotemporal functional clustering that have been adapted to the transformed air pollination data structure. Specifically, we will examine the functional clustering findings of the smooth functional CO pollutant over multiple cities in Yemen, including the group average and the spatial distribution of the obtained clusters, as depicted in Fig. 10. Overall, the spatiotemporal functional dynamic pattern of the CO air pollutant can be divided into three distinct phases. The first phase, which spans from January 1991 to December 2001, is characterized by a prominent polluting peak with a high volume of pollution. The second phase, which begins in January 2009 and ends in December 2019, is characterized by a stable and constant polluting pattern. The final phase, which is related to the COVID-19 pandemic, is representative of the COVID-19 lockdown period, during which pollution levels decreased dramatically. It has been observed that there has been a significant decrease in carbon monoxide (CO) air pollution in Yemen starting from January 2020 to April 2022. This has led to an overall improvement in air quality in the country. The objective of the study is to identify and classify the spatiotemporal patterns of functional CO data across multiple locations in Yemen. Our proposed method has been able to provide the best partition of potential clusters, which have been divided into four main groups. Specifically, cluster group 2 (colored red) comprises three major cities in Yemen—Sanaa, Taiz, and Ibb—which are considered to be more polluted compared to other cities in the country. Following group 2, group 4 comprises moderate polluting cities located on the western sides of Yemen. Group 3 (colored green) includes cities with zero CO pollution throughout the whole domain, owing to low population density and fewer human activities. The detailed profile, characteristics, and degree of pollution for CO concentration for each cluster are listed in Table 6. The study by Grace et al. (2020) employed the commonly used method of Fuzzy

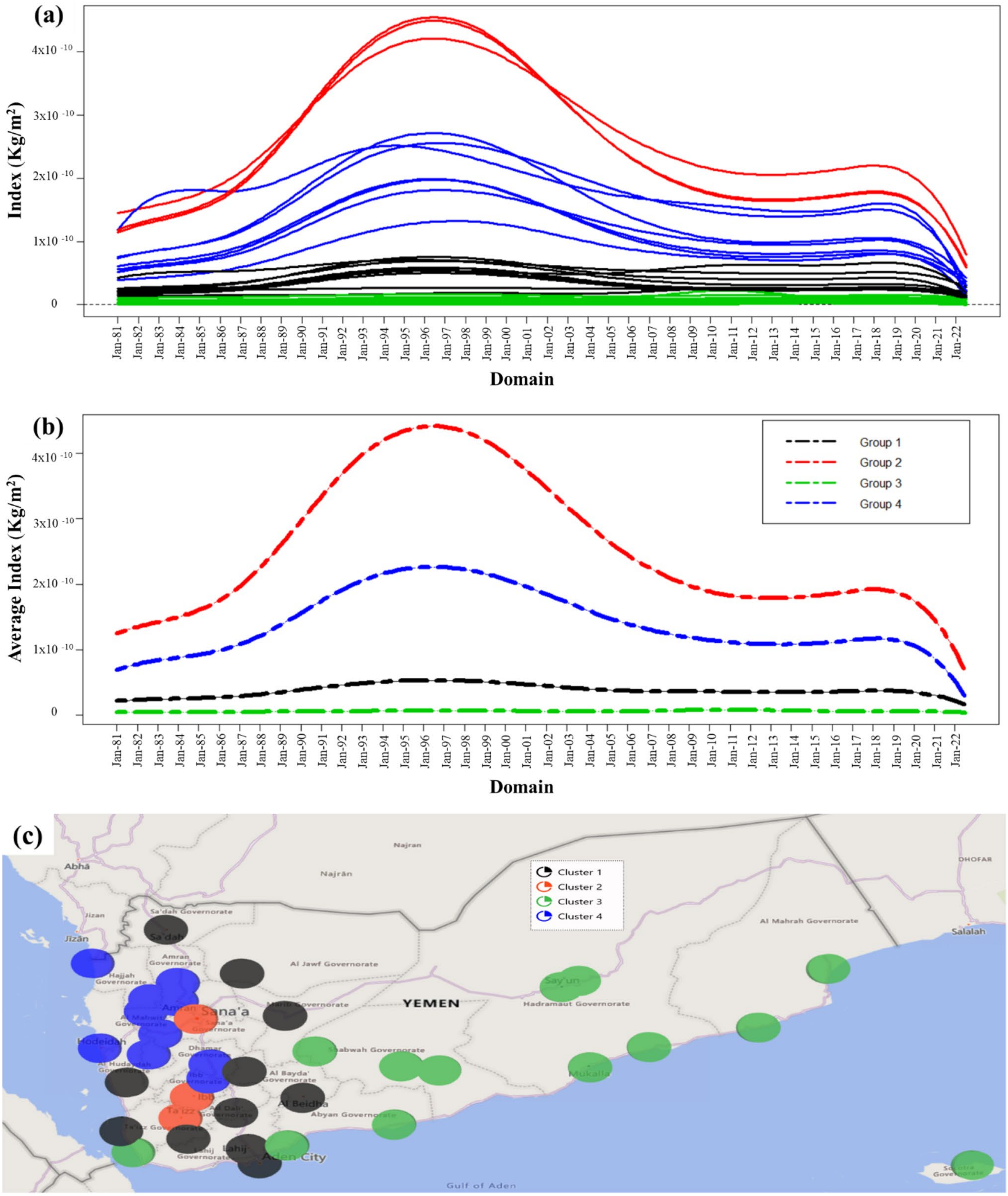


Fig. 10 **a** Functional clustering results of the smooth CO data, **b** its group means, and **c** the spatial distribution map of the obtained clusters

Table 6 Summary of the cluster characteristics for spatial–temporal CO patterns in cities of Yemen

Pollutant	Classifier	Color	Cities/locations	Ratio	Features (shape/size)	Pollution degree
CO	Class A	Black	Ad Dali,Aden,Al-Bayda,Rada,Zabid,Al-Hazm,Lahij,Marib,Sadah,Mukha,Al-Turbah	29.73%	Snaky-line shape during 1992–2000, straight-line shape before 1990 and after 2005	Light pollution
	Class B	Red	Sanaa, Ibb, Taiz	8.12%	Large-volume normal/Gaussian peak during 1991–2001, constant change after 2002, decline from 2020 to 2022	Extremely pollution
	Class C	Green	Ahwar,Zinjibar,Al-Ghayzah,Sayhut,Hadibu,Adis-Sharqiyah,Al-Mukalla,Sayun,Tarim,Al-Rawdah,Ataq,Bayhan,Dhubab	35.13%	No changes with straight-line shape through all domain	Clean
	Class D	Blue	Al-Hudaydah,Al-Mahwit,Amran,Khamir,Dhamar,Hajjah,Midi,Yarim,Al-Jabin,Manakhah	27.03%	Low-volume kurtosis peak during 1991–2001, decline from 2020 to 2022	Middle pollution

c-Means clustering to examine and present the data on CO pollutants alongside other pollutants, using real-time sensor data. On the other hand, the research conducted by Jain and Kaur (2021) introduced the use of machine learning and visualization techniques for forecasting and analyzing the air quality in major cities in India, taking into account six major pollutants, including CO.

Finally, the proposed method in this study can be compared to recent approaches that utilize Big Data and intelligent computation, such as those presented by Al-Janabi et al. (2021), Al-Janabi et al. (2020b), and Al-Janabi et al. (2019). These methods aimed to predict multiple air pollution concentrations using the Intelligent Forecaster of Concentrations caused air pollution (IFCsAP) (Al-Janabi et al. 2021), a pragmatic method based on intelligent big data analytics (Al-Janabi et al. 2019), and intelligent computation (Al-Janabi et al. 2020b). However, the main difference between our method and these previous studies is that our focus is on clustering and visualizing spatial–temporal air pollutant curves through functional data approaches, while their focus was on predicting discrete air pollutant data through intelligent big data analytics. As previously stated, it is more efficient to use statistical methods that can analyze the temporal and spatial variations of pollutants over time. The functional data analysis approach enables the examination of the entire time spectrum of pollutant variables. The table below compares the current study to recent studies that employed similar methods within the functional framework. However, the functional techniques used differ based on the purpose of the study. For example, Acal et al. (2022) focused on investigating the potential impact of the COVID-19 lockdown on air quality in the Pescara-Chieti urban area in Italy, which is known for high air pollution levels. Betancourt-Odio et al. (2021) used functional data and Kendall's functional Tao (KFT) to study the relationship between O_3 pollution levels in rural and urban areas in the Spanish Community of Madrid. Their findings indicate a complex, non-linear relationship between urban and rural areas. Torres et al. (2020) compared the effectiveness of three different analytical methods in identifying pollution episodes and outliers. More information about these studies is summarized in Table 7. A comparison of the current proposed method with other approaches applied to environmental pollution data is shown in Table 7. The comparison focuses on both functional and non-functional frameworks, univariate and multivariate settings, and lists the advantages and disadvantages of each approach. The disadvantages are based on the author's opinion, but other drawbacks may also exist. This information can be useful for understanding how the current work builds upon or differs from previous research in the field and can provide insights into the strengths and limitations of different methods.

Conclusions and recommendations

The study aimed to visualize and cluster the dynamic behavior of multiple air pollution concentrations using functional analysis techniques and functional mixtures clustering model. The method was applied to multivariate high dimensional air pollution data from cities in Yemen from January 1980 to April 2022. Fourie transformation, B-spline functions, and generalized-cross validation were used to reconstruct and smooth data. The study used two enhanced 3D visualization tools to explore the spatiotemporal variations in the functional air pollutants cycle and a functional mixture model was used to identify and classify the spatiotemporal functional air pollutants data. The study found four substantial clusters for all functional air pollutants variables and demonstrated the ability to identify, visualize, and classify the continuous functional dynamic patterns of air pollutants

Table 7 Comparison of the currently proposed method with other approaches applied to environmental pollution data

Ref	Study area/ country	Pollutants/ variables	Study period/ data interval	Framework (non-/semi-/ functional)	Type of analysis (univariate / or multivariate)	Domain (temporal, spatial–temporal)	Methods/ approaches	Advantage/significance	Disadvantage / limitation
Current	Yemen	CO, SO_2, PM2.5, O_3	Oct,1980–Apr, 2022	Functional	Multivariate	Spatial–temporal	Functional mixture clustering model and enhanced visualization tools	Able to Visualize the temporal-spatial dynamic structure of smoothed air pollutant curves Able to Cluster the underlying features of air pollutants curves	Data need to be smoothed properly/carefullyThe number of clusters is predefined by user
Saeipourdizaj et al. (2022)	Tabriz (Iran)	PM2.5, O3	2017	Conventional	Univariate	Spatial–temporal	spatiotemporal mixture model–based clustering (STMC)	The approach used to cluster the days of 2017 according to the hourly concentrations of O3 and PM10 collected from four stations of Tabriz using the mixture model–based clustering for spatiotemporal data	It depends on discrete-based points It is impossible for functional data framework The sample size is small, and data was insufficiency
Wang et al. (2022)	Inner Mongolia (China)	PM2.5	2016–2018	Non-functional(conventional)	Univariate	Spatial–temporal	Geographically weighted regression	The established model can better estimate the spatial distribution of PM2.5 concentration in the study area	The sample size is small It depends on discretization points The model only deals with one level of pollutant
Bouveyron et al. (2022)	South of France	PM10, NO_2, O_3	2013–2018	Functional	Multivariate	Spatial–temporal	A functional latent block model for co-clustering multivariate functional data	The proposed co-clustering approach is a powerful resource to analyze multivariate functional data. it is useful to identify intrinsic data structure and summarize variables profiles over long periods of time	The model need high computation The model arguments should be defined correctly with some specifications It deals with multi-dimensional variables that make it complex

Table 7 (continued)

Ref	Study area/ country	Pollutants/ variables	Study period/ data interval	Framework (non-/semi-/ functional)	Type of analysis (univariate / or multivariate)	Domain (temporal, spatial–temporal)	Methods/ approaches	Advantage/significance	Disadvantage / limitation
Reinholdt Jensen et al. (2022)	Chassieu (France)	Pollutant flush signals	2004 to 2011	Functional	Univariate	Temporal	Functional Principal Components Analysis and functional clustering	Able to identify different flush patterns emerging from the analyzed data The capacity to analyze Total Suspended Solids (TSS) and Mass Volume (MV) curves	limited to univariate functional case Data not properly smoothed due to applied local linear smoothing Defined on equally spaced domain; sample sizes must be the same
Acal et al. (2022)	Chieti-Pescara (Central Italy)	NO_2, PM10, PM2.5, benzene	Feb. 1 – Mar. 10, 11st Mar.–18th Apr. 2020	Functional	Multivariate	Spatial–temporal	Functional analysis of variance	The approach has proven to be beneficial to monitoring the evolution of air quality in both different periods of time	Defined on equally spaced domain, sample sizes must be the same limited to specific measurements and is not generalized
Betancourt-Odio et al. (2021)	Madrid (Spain)	Ozone	2011–2018	Semi-Functional	Univariate	Temporal	Kendall's Tau functional statistic (KFT)	The approach used to identify significant correlations between studied areas	Data is not smoothed The functional structure is not clear It dealt with one pollutant and ignore the others It needs to be improved
Al-Janabi et al. (2021)	Iraq	PM2.5, PM10, NO_2, CO, O_3, SO_2	2021	Non-Functional	Univariate	real-time	Intelligent forecaster of concentrations caused air pollution (IFCsAP)	The approach is efficient, cost-effective and easy to use It can be deployed in all places (environment with concentrations of air pollution)	It depends on discretization point It doesn't work with curves data and functional framework
Torres et al. (2020)	Dublin (Ireland)	NOx, NO_2	2013	Semi-Functional	Univariate	Temporal	Functional control process	The approach helps to identify the sources of pollution more effectively. It enhances the value of monitoring data and exceedances of limit values	It is mixed with conventional methods It is limited to specific measurements and is not generalized

Table 7 (continued)

Ref	Study area/ country	Pollutants/ variables	Study period/ data interval	Framework (non-/semi-/ functional)	Type of analysis (univariate / or multivariate)	Domain (temporal, spatial–temporal)	Methods/ approaches	Advantage/significance	Disadvantage / limitation
Al-Janabi et al. (2020b)	Beijing (China) and London(UK)	PM2.5, PM10, NO_2,CO, ozone O_3, SO_2	KDD Cup 2018	Non-Functional	Multivariate	Discrete	Smart air quality prediction model (SAQPM)	The approach is working with intelligent computation, and deep learning techniques It is able to make an intelligent predictor for the concentrations of air pollutants over the next 2 days	It doesn't work with curves data and functional framework It depends on the discretization point It Increases the computational complexity
Wang et al. (2019)	China	PM2.5	1998 to 2016	Functional	Univariate	Spatial–temporal	Functional data analysis tools such as Functional Principal Component and Adaptive Clustering Analysis	The study employed FDA techniques to analyze PM2.5 concentration data for each province. Beside it used several methods to decompose, cluster, and testing the differences	limited to the univariate functional case it does not work with missing values Sample size is small because data was yearly not monthly
King et al. (2018)	United States	PM2.5	2003 to 2015	Functional	Univariate	Spatial–temporal	Functional Principal Components Analysis	The approach can predict yearly profiles for locations at which data are not available It offers dimension reduction for easier data visualization It allows to study changes of pollutant levels annually	It dealt with one pollutant and ignore the others There are lacks pre-processing analysis Defined on the equally spaced domain; sample sizes must be the same

(SO_2, PM2.5, O_3, and CO) over multi-sites in Yemen. Some main results have been concluded as follows:

Yemen has experienced substantial dynamic patterns of air pollution concentrations over different spatial locations from the period 1980–2022.

The obtained results have also provided evidence that vehicular emission is the primary source of air pollution in Yemen besides industrial activity and mixing factors are also shown to be the secondary contributing factors towards air pollution variation.

The functional clustering findings showed a noteworthy decline in CO emissions during the COVID-19 pandemic; additionally, the cities of Sanaa, Ibb, and Tiaz were classified as the more polluted cities in Yemen.

Regarding the Ground-level O_3 pollutant, the results showed great fluctuations with increase and decrease during the entire domain; however, there was no effect on ozone level concentrations due to the COVID-19 pandemic period.

Although PM2.5 concentrations have witnessed an extremally significant increase before the COVID-19 pandemic period, they have shown a noticeable decrease during the COVID-19 pandemic period.

In general, the results showed that there was stability and no significant changes in SO_2 levels, particularly during the last two decades.

Overall, ambient air pollution can be controlled and reduced with the implementation of strategic measures, led by sound leadership and development efforts to help emerging economies recover from past losses. Successful pollution control methods, that are technically, politically, and economically feasible for a specific country, can be shared globally to minimize air pollution. Recommendations to control and reduce air pollutants include the development and implementation of new environmental standards, the use of intervention techniques to decrease concentration, the prohibition of polluting materials and fuels in urban and rural areas, regulation of private vehicles, and an increase in public transportation. Additionally, promoting the use of clean fuels and implementing effective policies to ensure standard operating protocols in workplaces, industries, and hospitals can help control the spread of pathogenic microbes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-023-25790-3.

Data availability Data are accessible from the corresponding author upon request.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication The author consents to publish this manuscript.

Competing interests The author declares no competing interests.

References

Abdul-Rahim AK, Al-Sowaidi NA, Eadan ZA (2022) Concentration of fine particulate matter (PM2.5) and black carbon (BC) in aerosol samples in Al-Zubairy Area In Sana'a, Yemen. Electron J Univ Aden Basic Appl Sci 3(204–213):7. https://doi.org/10.47372/ejua-ba.2022.3.187

Acal C, Aguilera AM, Sarra A et al (2022) Functional ANOVA approaches for detecting changes in air pollution during the COVID-19 pandemic. Stoch Environ Res Risk Assess 36:1083–1101. https://doi.org/10.1007/s00477-021-02071-4

Ahmad M, Rappenglück B, Osibanjo OO, Retama A (2022) A machine learning approach to investigate the build-up of surface ozone in Mexico-City. J Clean Prod 379:134638. https://doi.org/10.1016/j.jclepro.2022.134638

Al-Janabi S, Alkaim A (2022) A novel optimization algorithm (Lion-AYAD) to find optimal DNA protein synthesis. Egypt Informatics J 23:271–290. https://doi.org/10.1016/j.eij.2022.01.004

Al-Janabi S, Alkaim A, Al-Janabi E et al (2021) Intelligent forecaster of concentrations (PM2.5, PM10, NO2, CO, O3, SO2) caused air pollution (IFCsAP). Neural Comput Applic 33:14199–14229. https://doi.org/10.1007/s00521-021-06067-7

Al-Janabi S, Alkaim AF (2020) A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation. Soft Comput 24:555–569. https://doi.org/10.1007/s00500-019-03972-x

Al-Janabi S, Alkaim AF, Adel Z (2020a) An Innovative synthesis of deep learning techniques (DCapsNet & DCOM) for generation electrical renewable energy from wind energy. Soft Comput 24:10943–10962. https://doi.org/10.1007/s00500-020-04905-9

Al-Janabi S, Mohammad M, Al-Sultan A (2020b) A new method for prediction of air pollution based on intelligent computation. Soft Comput 24:661–680. https://doi.org/10.1007/s00500-019-04495-1

Al-Janabi S, Yaqoob A, Mohammad M (2019) Pragmatic method based on intelligent big data analytics to prediction air pollution. In: Farhaoui Y (ed) Big Data and Networks Technologies. Springer International Publishing, Cham, pp 84–109

Betancourt-Odio A, Valencia D, Soffritti M, Budría S (2021) An analysis of ozone pollution by using functional data: rural and urban areas of the Community of Madrid. Environ Monit Assess 193. https://doi.org/10.1007/s10661-021-09180-1

Bouveyron C, Côme E, Jacques J (2015) The discriminative functional mixture model for a comparative analysis of bike sharing systems. Ann Appl Stat 9:1726–1760. https://doi.org/10.1214/15-AOAS861

Bouveyron C, Jacques J, Schmutz A et al (2022) Co-clustering of multivariate functional data for the analysis of air pollution in the south of France. Ann Appl Stat 16(3):1400–1422. https://doi.org/10.1214/21-AOAS1547

Broomandi P, Geng X, Guo W et al (2021) Dynamic complex network analysis of PM2.5 concentrations in the UK, using hierarchical directed graphs (V1.0.0). Sustain 13:1–14. https://doi.org/10.3390/su13042201

Caligiuri LM, Costanzo GD, Reda A (2005) The study of ground ozone concentration levels : a functional analysis approach based on principal components analysis. WIT Trans Ecol Environ 82:59–67. https://doi.org/10.2495/AIR050071

Cook Q, Argenio K, Lovinsky-Desir S (2021) The impact of environmental injustice and social determinants of health on the role of air pollution in asthma and allergic disease in the United States.

J Allergy Clin Immunol 148:1089-1101.e5. https://doi.org/10.1016/j.jaci.2021.09.018

Embling CB, Illian J, Armstrong E et al (2012) Investigating fine-scale spatio-temporal predator-prey patterns in dynamic marine ecosystems: A functional data analysis approach. J Appl Ecol 49:481–492. https://doi.org/10.1111/j.1365-2664.2012.02114.x

Escabias M, Aguilera AM, Valderrama MJ (2005) Modeling environmental data by functional principal component logistic regression. Environmetrics 16(1):95–107. https://doi.org/10.1002/env.696

Escabias M, Valderrama MJ, Aguilera AM et al (2013) Stepwise selection of functional covariates in forecasting peak levels of olive pollen. Stoch Environ Res Risk Assess 27:367–376. https://doi.org/10.1007/s00477-012-0655-0

Fang K, Wang T, He J et al (2020) The distribution and drivers of PM2.5 in a rapidly urbanizing region: The Belt and Road Initiative in focus. Sci Total Environ 716:137010. https://doi.org/10.1016/j.scitotenv.2020.137010

Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer Series in Statistics, New York, p 76. https://doi.org/10.1007/0-387-36620-2

Gagliardi RV, Andenna C (2020) A machine learning approach to investigate the surface ozone behavior. Atmosphere (basel) 11:1–16. https://doi.org/10.1016/j.jclepro.2022.134638

Gao HO (2007) Day of week effects on diurnal ozone / NOx cycles and transportation emissions in Southern California. Transp Res Part D Transp Environ 12:292–305. https://doi.org/10.1016/j.trd.2007.03.004

Gao HO, Niemeier DA (2008) Using functional data analysis of diurnal ozone and NOx cycles to inform transportation emissions control. Transp Res Part D Transp Environ 13:221–238. https://doi.org/10.1016/j.trd.2008.02.003

Grace RK, S KA, Monisha B, Kaarthik A (2020) Analysis and visualization of air quality using real time pollutant data. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp 34–38. https://doi.org/10.1109/ICACCS48705.2020.9074283

Guo B, Wu H, Pei L et al (2022) Study on the spatiotemporal dynamic of ground-level ozone concentrations on multiple scales across China during the blue sky protection campaign. Environ Int 170:107606. https://doi.org/10.1016/j.envint.2022.107606

Hael MA (2021) Modeling of rainfall variability using functional principal component method: a case study of Taiz region, Yemen. Model Earth Syst Environ 7:17–27. https://doi.org/10.1007/s40808-020-00876-w

Hael MA, Ma H, AL-kuhali HA (2021) Unsupervised classification of wind speed directions based on functional discriminative latent mixture model. In: 2021 12th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), pp 110–118. https://doi.org/10.1109/PAAP54281.2021.9720313

Hael MA, Yongsheng Y, Saleh BI (2020) Visualization of rainfall data using functional data analysis. SN Appl Sci 2:461. https://doi.org/10.1007/s42452-020-2238-x

Hörmann S, Kidzi L, Hallin M (2015) Dynamic functional principal components. J R Stat Soc Ser B Stat Methodol 77:319–348. https://doi.org/10.1111/rssb.12076

Ignaccolo R, Ghigo S, Giovenali E (2008) Analysis of air quality monitoring networks by functional clustering. Environmetrics 19:672–686. https://doi.org/10.1002/env.946

Ignaccolo R, Mateu J, Giraldo R (2014) Kriging with external drift for functional data for air quality monitoring. Stoch Environ Res Risk Assess 28:1171–1186. https://doi.org/10.1007/s00477-013-0806-y

Jain P, Kaur G (2021) A learning-based visualization application for air quality evaluation during COVID'19 pandemic in open data centric services. Machine Learning Algorithms and Applications 1–22. https://doi.org/10.1002/9781119769262.ch1

Jorquera H, Villalobos AM (2020) Combining cluster analysis of air pollution and meteorological data with receptor model results for ambient pm2.5 and pm10. Int J Environ Res Public Health 17:1–25. https://doi.org/10.3390/ijerph17228455

Kadhuim ZA, Al-Janabi S (2023) Codon-mRNA prediction using deep optimal neurocomputing technique (DLSTM-DSN-WOA) and multivariate analysis. Results Eng 17:100847. https://doi.org/10.1016/j.rineng.2022.100847

King MC, Staicu AM, Davis JM et al (2018) A functional data analysis of spatiotemporal trends and variation in fine particulate matter. Atmos Environ 184:233–243. https://doi.org/10.1016/j.atmosenv.2018.04.001

Kosiorowski D, Mielczarek D, Rydlewski JP (2017) Forecasting of a hierarchical functional time series on example of macromodel for the day and night air pollution in silesia region - a critical overview. Cent Eur J Econ Model Econom 2018:53–73

Kosiorowski D, Szlachtowska E (2017) K-local median algorithm for functional data in empirical analysis of air pollution. Proceedings from the 11th ProfessorA. Zelias International Conference, pp 153–162. http://pliki.konferencjazakopianska.pl/proceedings_2017/index.html

Kujawska J, Kulisz M, Aubakirova Z (2022) Application of artificial neural networks model to predict the levels of sulfur dioxides in the air of Zamość, Poland. J Phys Conf Ser 2412:1–10. https://doi.org/10.1088/1742-6596/2412/1/012005

Li G, Zhang S, Shi X et al (2022) Spatiotemporal variability and diffusive emissions of greenhouse gas in a shallow eutrophic lake in Inner Mongolia. China. Ecol Indic 145:109578. https://doi.org/10.1016/j.ecolind.2022.109578

Li H, Fan H, Mao F (2016) A visualization approach to air pollution data exploration-a case study of air quality index (PM2.5) in Beijing. China Atmosphere (basel) 7:1–20. https://doi.org/10.3390/atmos7030035

Li S, Shafi S, Zou B, et al (2022b) PM(2.5) Concentration exposure over the belt and road region from 2000 to 2020. Int J Environ Res Public Health 19. https://doi.org/10.3390/ijerph19052852

Li X, Hu Z, Cao J, Xu X (2022c) The impact of environmental accountability on air pollution: a public attention perspective. Energy Policy 161. https://doi.org/10.1016/j.enpol.2021.112733

Li Y, Ma Z, Han T et al (2022d) Long-term declining in carbon monoxide (CO) at a rural site of Beijing during 2006–2018 implies the improved combustion efficiency and effective emission control. J Environ Sci 115:432–442. https://doi.org/10.1016/j.jes.2020.11.011

Liang D, Zhang H, Chang X, Huang H (2021) Modeling and regionalization of China's PM2.5 using spatial-functional mixture models. J Am Stat Assoc 116:116–132. https://doi.org/10.1080/01621459.2020.1764363

Liu H, Long Z, Duan Z, Shi H (2020) A new model using multiple feature clustering and neural networks for forecasting hourly PM2.5 concentrations, and its applications in China. Engineering 6:944–956. https://doi.org/10.1016/j.eng.2020.05.009

Liu Y, Tang G, Liu B et al (2022) Decadal changes in ozone in the lower boundary layer over Beijing, China. Atmos Environ 275:119018. https://doi.org/10.1016/j.atmosenv.2022.119018

Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E (2020) Environmental and health impacts of air pollution: a review. Front Public Heal 8:1–13. https://doi.org/10.3389/fpubh.2020.00014

Martínez J, Saavedra Á, García-Nieto PJ et al (2014) Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). Appl Math Comput 241:1–10. https://doi.org/10.1016/j.amc.2014.05.004

Medhi S, Gogoi M (2021) Visualization and Analysis of COVID-19 Impact on PM2.5 Concentration in Guwahati city. In: 2021 International

Conference on Computational Performance Evaluation (ComPE), pp 12–16. https://doi.org/10.1109/ComPE53109.2021.9752244

Meiring W (2007) Oscillations and time trends in stratospheric ozone levels: a functional data analysis approach. J Am Stat Assoc 102(479):788–802. https://doi.org/10.1198/016214506000000825

Mohammed GS, Al-Janabi S (2022) An innovative synthesis of optmization techniques (FDIRE-GSK) for generation electrical renewable energy from natural resources. Results Eng 16:100637. https://doi.org/10.1016/j.rineng.2022.100637

NASA (2023) Giovanni (The Bridge Between Data and Science). https://giovanni.gsfc.nasa.gov/giovanni/. Accessed 22 Jan 2023

Nikolaou N, Dallavalle M, Stafoggia M et al (2023) High-resolution spatiotemporal modeling of daily near-surface air temperature in Germany over the period 2000–2020. Environ Res 219:115062. https://doi.org/10.1016/j.envres.2022.115062

Nurgazy M, Zaslavsky A, Jayaraman PP, et al (2019) CAVisAP: context-aware visualization of outdoor air pollution with IoT platforms. In: 2019 International Conference on High Performance Computing & Simulation (HPCS), pp 84–91. https://doi.org/10.1109/HPCS48598.2019.9188167

Ocana-Peinado F, Valderrama MJ, Aguilera AM (2008) A dynamic regression model for air pollen concentration. Stoch Environ Res Risk Assess 22:59–63. https://doi.org/10.1007/s00477-007-0153-y

Ochoa JO, Muñoz DPO, León CLU (2020) Functional analysis of variance of air pollution caused by fine particles. Univ Sci 25:1–16. https://doi.org/10.11144/JAVERIANA.SC25-1.FAOV

Park A, Guillas S, Petropavlovskikh I (2013) Trends in stratospheric ozone profiles using functional mixed models. Atmos Chem Phys 11473–11501. https://doi.org/10.5194/acp-13-11473-2013

Park J-M, Lee T-J, Kim D-S (2022) Improving PMF source reconciliation with cluster analysis for PM2.5 hourly data from Seoul. Korea. Atmos Pollut Res 13:101398. https://doi.org/10.1016/j.apr.2022.101398

Pineda Rojas AL, Leloup JA, Kropff E (2019) Spatial patterns of conditions leading to peak O3 concentrations revealed by clustering analysis of modeled data. Air Qual Atmos Heal 12:743–754. https://doi.org/10.1007/s11869-019-00694-9

Preda C (2007) Regression models for functional data by reproducing kernel Hilbert spaces methods. J Stat Plan Inference 137:829–840. https://doi.org/10.1016/j.jspi.2006.06.011

Ramsay JO, Silverman BW (2002) Applied functional data analysis: methods and case studies. Springer Series in Statistics, New York

Ramsay JO, Silverman BW (2005) Functional data analysis. Springer Series in Statistics, New York. https://doi.org/10.1007/b98888

Ranaarif S, Yuwono AS (2021) Analysis of the distribution of sulfur dioxide (SO2) pollutant in Bali Island for the 2011–2020 period. IOP Conf Ser Earth Environ Sci 871:1–15. https://doi.org/10.1088/1755-1315/871/1/012031

Ranalli MG, Rocco G, JonaLasinio G et al (2016) Functional exploratory data analysis for high-resolution measurements of urban particulate matter. Biometrical J 58:1229–1247. https://doi.org/10.1002/bimj.201400251

Reinholdt Jensen DM, Sandoval S, Aubin JB et al (2022) Classifying pollutant flush signals in stormwater using functional data analysis on TSS MV curves. Water Res 217:118394. https://doi.org/10.1016/j.watres.2022.118394

Ruggieri M, Plaia A, Di Salvo F, Agró G (2013) Functional principal component analysis for the explorative analysis of multisite-multivariate air pollution time series with long gaps. J Appl Stat 40:795–807. https://doi.org/10.1080/02664763.2012.754852

Saeipourdizaj P, Musavi S, Gholampour A, Sarbakhsh P (2022) Clustering the concentrations of PM10 and O3: application of spatiotemporal model–based clustering. Environ Model Assess 27:45–54. https://doi.org/10.1007/s10666-021-09802-5

Saleh B, Ma A, Fathi R et al (2023) Optimized mechanical properties of magnesium matrix composites using RSM and ANN. Mater Sci Eng B 290:1–21. https://doi.org/10.1016/j.mseb.2023.116303

Sancho J, Martínez J, Pastor JJ et al (2014) New methodology to determine air quality in urban areas based on runs rules for functional data. Atmos Environ 83:185–192. https://doi.org/10.1016/j.atmosenv.2013.11.010

Schmutz A, Jacques J, Bouveyron C et al (2020) Clustering multivariate functional data in group-specific functional subspaces. Comput Stat 35:1101–1131. https://doi.org/10.1007/s00180-020-00958-4

Shaadan N, Deni SM, Jemain AA (2012) Assessing and comparing PM10 pollutant behaviour using functional data approach. Sains Malaysiana 41:1335–1344

Shaadan N, Jemain AA, Latif MT, Deni SM (2015) Anomaly detection and assessment of PM10 functional data at several locations in the Klang Valley, Malaysia. Atmos Pollut Res 6:365–375. https://doi.org/10.5094/APR.2015.040

Shams SB, Walden VP, Hannigan JW et al (2022) Analyzing ozone variations and uncertainties at high latitudes during sudden stratospheric warming events using MERRA-2. Atmos Chem Phys 22:5435–5458. https://doi.org/10.5194/acp-22-5435-2022

Shi G, Liu J, Zhong X (2022) Spatial and temporal variations of PM2.5 concentrations in Chinese cities during 2015–2019. Int J Environ Health Res 32:2695–2707. https://doi.org/10.1080/09603123.2021.1987394

Su SH, Chang CW, Chen WT (2020) The temporal evolution of PM2.5 pollution events in taiwan: clustering and the association with synoptic weather. Atmosphere (basel) 11:1–14. https://doi.org/10.3390/atmos11111265

Tang Z, Mei Z, Sui G, Zou J (2021) Visualizing the impact of COVID-19 on PM2.5 concentrations in China. Reg Stud Reg Sci 8:51–53. https://doi.org/10.1080/21681376.2021.1872414

Torres JM, Pérez JP, Val JS, et al (2020) A functional data analysis approach for the detection of air pollution episodes and outliers: a case study in Dublin, Ireland. Mathematics 8. https://doi.org/10.3390/math8020225

Valderrama MJ, Ocaña FA, Aguilera AM, Ocaña-Peinado FM (2010) Forecasting pollen concentration by a two-step functional model. Biometrics 66:578–585. https://doi.org/10.1111/j.1541-0420.2009.01293.x

Wang D, Zhong Z, Bai K, He L (2019) Spatial and temporal variabilities of PM2.5 concentrations in China using functional data analysis. Sustain 11. https://doi.org/10.3390/su11061620

Wang J, Liu D, Xu X et al (2022) Analysis of the temporal and spatial pattern of air pollution and the heterogeneity of its influencing factors in central Inner Mongolia from 2016 to 2018. Front Environ Sci 10:1–15. https://doi.org/10.3389/fenvs.2022.893437

Wang Y, Xu K, Li S (2020) The functional spatio-temporal statistical model with application to O3 pollution in Beijing, China. Int J Environ Res Public Health 17. https://doi.org/10.3390/ijerph17093172

Worldmaps (2023) Geopolitical map of Yemen from World and country maps. https://www.worldmaps.info/Yemen/. Accessed 22 Jan 2023

Xiao W, Hu Y (2018) Functional data analysis of air pollution in six major cities. J Phys Conf Ser 1053. https://doi.org/10.1088/1742-6596/1053/1/012131

Yang G, Liu Y, Li X (2020) Spatiotemporal distribution of ground-level ozone in China at a city level. Sci Rep 10:1–12. https://doi.org/10.1038/s41598-020-64111-3

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.