Atmospheric Environment 367 (2026) 121741

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Functional data analysis of air quality time series in Madrid Using FPCA and splines

Joaquín Sancho Val *, Carlos Cajal Hernando, Lourdes Martínez de Baños

Centro Universitario de la Defensa, Academia General Militar, Carretera de Huesca s/n, 50090 Zaragoza, Spain

HIGHLIGHTS

- Functional data analysis applied to air quality in Madrid using FPCA.
- The first three PCs explain over 80% of the variability for most pollutants.
- Station clustering reveals spatial pollution patterns linked to traffic intensity.
- Screeplots and score plots aid interpretation of temporal pollution dynamics.
- Robust methodology supports air quality monitoring and urban planning.

ARTICLE INFO

Keywords:
Air quality
Functional data analysis (FDA)
FPCA
Splines
Madrid
Atmospheric environment

ABSTRACT

Urban air pollution requires analytical tools that capture complex temporal dynamics while remaining interpretable for policy purposes. This study applies Functional Principal Component Analysis (FPCA) to hourly concentrations of NO_2, PM_{10}, and $PM_{2.5}$ measured in 2024 at eight monitoring stations in Madrid. By transforming high-dimensional time series into smooth functional data, FPCA identifies dominant modes of variability and reveals structural differences among sites.

Across all pollutants, the first three components explained more than 85% of total variance. PC1 represented the main seasonal cycle, PC2 reflected short-term fluctuations driven by traffic and meteorology, and PC3 captured episodic peaks. The score plots highlighted consistent clustering: background stations (Casa de Campo, Sanchinarro) showed low levels and weak seasonality; traffic-oriented stations (Plaza Elíptica, Cuatro Caminos, Escuelas Aguirre) displayed high concentrations and stronger cycles; and intermediate sites (Castellana, Plaza Castilla, Méndez Álvaro) exhibited mixed patterns.

These findings confirm FPCA as an efficient diagnostic framework for air quality assessment, offering both dimensionality reduction and interpretability. Beyond methodological value, the approach enables classification of monitoring sites and supports targeted mitigation strategies. FPCA thus provides a transferable tool for urban air quality management and a solid basis for functional data analysis in environmental policy contexts.

1. Introduction

Air pollution remains one of the most pressing environmental and public health challenges worldwide. According to the World Health Organization (WHO), exposure to fine particulate matter ($PM_{2.5}$) and nitrogen dioxide (NO_2) is linked to millions of premature deaths annually, largely due to cardiovascular and respiratory diseases (World Health Organization, 2021; Cohen et al., 2017). In Europe, the European Environment Agency (EEA) has repeatedly identified urban traffic emissions as a dominant contributor to exceedance of air quality standards, with NO_2, PM_{10} and $PM_{2.5}$ being key pollutants of concern (European Environment Agency, 2023). These findings align with global epidemiological evidence showing robust associations between long-term exposure to air pollution and adverse health outcomes (Pope et al., 2009; Dominici et al., 2014; Burnett et al., 2018).

Madrid, like other European capitals, faces recurrent air quality exceedance, particularly in relation to NO_2 from vehicular emissions and particulate matter influenced by both local traffic and regional transport (Querol et al., 2014; Tobías et al., 2020). The complex interplay of emission sources, meteorological conditions, and urban morphology makes the interpretation of air quality dynamics challenging. While conventional statistical methods such as regression models and time series decomposition have been widely applied (Gryparis et al., 2004;

* Corresponding author.
E-mail address: jsanchov@unizar.es (J. Sancho Val).

https://doi.org/10.1016/j.atmosenv.2025.121741
Received 28 August 2025; Received in revised form 26 November 2025; Accepted 10 December 2025
Available online 12 December 2025

Lang et al., 2019), they often fail to fully capture the functional nature of continuous monitoring data and their temporal correlations.

Functional Data Analysis (FDA) has emerged as a powerful statistical tool for analyzing high-resolution environmental datasets (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012). Within FDA, Functional Principal Component Analysis (FPCA) offers a parsimonious representation of pollutant variability by extracting dominant modes of variation from functional time series. These components facilitate interpretation in terms of seasonal cycles, short-term fluctuations, and episodic events. Several works have demonstrated the applicability of FDA for environmental monitoring: outlier detection in air quality data in Spain (Martínez et al., 2014), detection of anomalous pollution episodes in Dublin (Martínez Torres et al., 2020), and the proposal of methodological advances for runs-rules applied to functional data in urban air quality (Sancho et al., 2014). Beyond atmospheric studies, FDA has also been applied to quality control problems, such as harmonic variability in power systems (Sancho et al., 2013), river water quality monitoring (Sancho et al., 2015), and Shewhart-type control charts for environmental indicators (Iglesias et al., 2016).

Recent contributions have highlighted the growing relevance of functional and machine learning methods for urban air pollution analysis. For example, a study (Hasnain et al., 2022) applied forecasting and advanced time-series models to characterize pollutant dynamics across multiple Chinese mega-cities, while another one (Rosca et al., 2025) employed data-driven approaches to predict and forecast air quality in urban areas. Similarly, studies by Querol and colleagues (Querol et al., 2017) have emphasized the need for integrative approaches that combine statistical innovation with actionable policy guidance.

The aim of this study is to apply FPCA to hourly concentrations of NO_2, PM_{10}, and $PM_{2.5}$ recorded in 2024 across eight monitoring stations in Madrid. Specifically, we (i) construct functional representations of pollutant time series, (ii) quantify the proportion of variance explained by leading components, (iii) interpret their temporal meaning in terms of seasonal, episodic, and background variability, and (iv) identify clustering patterns among stations that reflect the influence of traffic and background sources. By combining methodological innovation with policy-relevant interpretation, this work contributes both to the advancement of functional methods in air quality research and to the design of evidence-based strategies for urban air quality management.

In this context, the present study contributes to the existing FPCA-based literature on air quality in several ways. First, we work with multi-year, high-frequency pollutant time series from a dense urban monitoring network in Madrid, which allows us to jointly characterize intra-daily and seasonal variability at the city scale, rather than focusing only on daily or temporally aggregated indicators. Second, we implement a penalized spline representation specifically tailored to urban pollution dynamics, using a smoothing strategy that preserves sharp traffic-related peaks and episodic pollution events while attenuating short-term measurement noise. Third, we combine the resulting functional principal component scores with clustering techniques to derive data-driven typologies of monitoring stations — distinguishing, for example, traffic-oriented and background sites — that can be directly linked to emission sources and regulatory targets. Taken together, these features go beyond previous FPCA applications to air quality in other cities and provide a flexible template that can be extended to multi-city comparisons and to the inclusion of meteorological and emission covariates in future functional models.

2. Data and methods

2.1. Dataset

The dataset used in this study corresponds to the hourly concentrations of three atmospheric pollutants — nitrogen dioxide (NO_2), particulate matter (PM_{10}), and fine particulate matter ($PM_{2.5}$) — obtained from the official *Air Quality Portal of Madrid City Council*. The monitoring period spans the twelve months of the year 2024, thus covering the complete set of seasonal variations.

Table 1
List of monitoring stations in Madrid selected for the analysis (see Fig. 1).

Station code	Station name	Type
08	Escuelas Aguirre	Urban background
24	Casa de Campo	Peri-urban background
38	Cuatro Caminos	Urban traffic
47	Méndez Álvaro	Urban traffic
48	Castellana	Urban traffic
50	Plaza Castilla	Urban traffic
56	Plaza Elíptica	Urban traffic
57	Sanchinarro	Urban background

The pollutants were selected due to their recognized impact on urban air quality and human health, and because they are consistently measured across a significant subset of the monitoring network. From the full set of stations, we selected eight monitoring sites that jointly report the three pollutants. These stations (Table 1) were chosen because they not only provide simultaneous data for NO_2, PM_{10}, and $PM_{2.5}$, but also represent different urban typologies and geographical areas of Madrid (e.g., traffic-oriented sites, background stations, and suburban locations).

Fig. 1 shows the geographical distribution of the selected eight stations within the city of Madrid. This spatial selection ensures that the analysis captures heterogeneous patterns of pollution dynamics across the metropolitan area.

2.1.1. Exploratory data quality analysis

To ensure the robustness of the functional data analysis, we performed an exploratory assessment of data quality for each monitoring station and pollutant. Table 2 summarizes the total number of available values, missing values (and their percentage), the observed range (minimum and maximum), mean concentration, and the number of statistical outliers.

Missing values were identified as non-numeric entries or records explicitly flagged as invalid. Outliers were defined using the interquartile range (IQR) method, identifying any daily concentration below $Q1-1.5 \times IQR$ or above $Q3+1.5 \times IQR$. These diagnostics inform the reliability of the functional representations and confirm the absence of systematic data loss.

2.1.2. Treatment of outliers in functional data construction

The presence of statistical outliers in environmental data is a known challenge due to episodic events (e.g., dust intrusions, traffic surges) or sensor anomalies. In our dataset, outliers were identified using the standard interquartile range (IQR) criterion: values falling below $Q1-1.5 \times IQR$ or above $Q3+1.5 \times IQR$ were flagged as potential anomalies.

As shown in Table 2, all stations and pollutants contained a small to moderate number of outliers, particularly in $PM_{2.5}$ and PM_{10} time series. Rather than excluding these observations, we retained them during functional data construction and relied on the smoothing properties of spline-based FDA to attenuate their influence. Specifically, smoothing splines with automatic penalty selection (via Generalized Cross-Validation) were applied to each time series, ensuring a balance between data fidelity and robustness to local anomalies.

This approach allows the model to preserve meaningful episodic peaks -relevant for environmental interpretation—while minimizing the impact of transient noise or measurement errors. Visual inspection of the resulting functional curves confirmed that the smoothing process effectively reduced the distortion caused by extreme values without suppressing genuine signal variability.

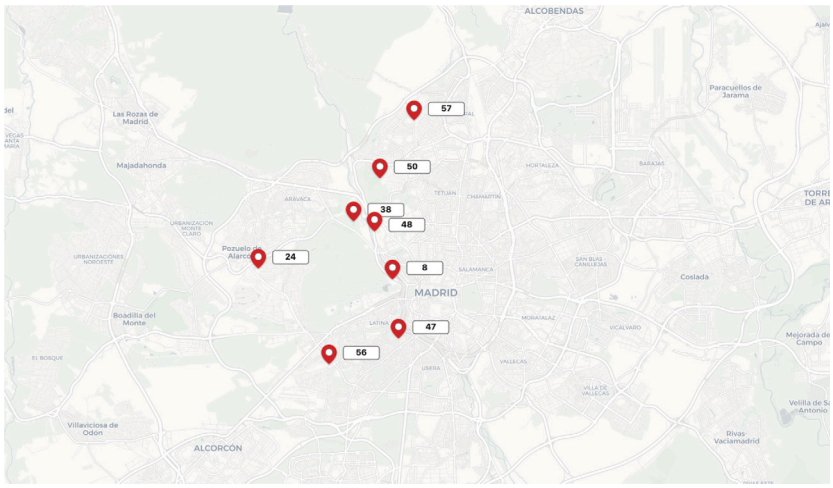

Fig. 1. Location of the eight monitoring stations in Madrid selected for the analysis. These sites were chosen for simultaneously reporting NO_2, PM_{10}, and $PM_{2.5}$ throughout 2024, and for representing different urban environments.

Table 2
Summary of data completeness and quality per station and pollutant.

Station–Pollutant	Total	Missing (%)	Mean	Min	Max	Outliers
Escuelas Aguirre – NO_2	341	0.0	25.43	0	62	1
Escuelas Aguirre – PM_{10}	341	0.0	9.83	−1	36	20
Escuelas Aguirre – $PM_{2.5}$	341	0.0	21.23	0	131	11
Casa de Campo – NO_2	341	0.0	13.65	0	46	17
Casa de Campo – PM_{10}	341	0.0	5.75	0	25	9
Casa de Campo – $PM_{2.5}$	341	0.0	13.62	0	122	11
Cuatro Caminos – NO_2	341	0.0	23.97	0	89	16
Cuatro Caminos – PM_{10}	341	0.0	8.83	0	38	6
Cuatro Caminos – $PM_{2.5}$	341	0.0	18.33	0	113	11
Méndez Álvaro – NO_2	341	0.0	18.13	0	58	1
Méndez Álvaro – PM_{10}	341	0.0	7.78	0	26	3
Méndez Álvaro – $PM_{2.5}$	341	0.0	16.43	0	96	10
Castellana – NO_2	341	0.0	20.56	0	58	12
Castellana – PM_{10}	341	0.0	9.32	0	36	9
Castellana – $PM_{2.5}$	341	0.0	19.42	0	118	15
Plaza Castilla – NO_2	341	0.0	24.09	0	60	1
Plaza Castilla – PM_{10}	341	0.0	9.47	0	278	7
Plaza Castilla – $PM_{2.5}$	341	0.0	19.93	0	294	13
Plaza Elíptica – NO_2	341	0.0	28.47	0	58	0
Plaza Elíptica – PM_{10}	341	0.0	9.35	0	37	6
Plaza Elíptica – $PM_{2.5}$	341	0.0	17.53	0	75	9
Sanchinarro – NO_2	341	0.0	18.56	0	68	15
Sanchinarro – PM_{10}	341	0.0	7.34	0	26	6
Sanchinarro – $PM_{2.5}$	341	0.0	14.5	0	108	11

2.2. Mathematical background

2.2.1. Smoothing

Let $x_i(t)$ denote the observed pollutant concentration at monitoring station i over time $t \in [0, T]$. Since the original dataset consists of hourly measurements, we construct smooth functions $x_i(t)$ via spline smoothing:

$$x_i(t) \approx \sum_{k=1}^{K} c_{ik}\phi_k(t) \tag{1}$$

where $\{\phi_k(t)\}$ is a B-spline basis of dimension K, and c_{ik} are the corresponding coefficients for station i.

The optimal level of smoothness is determined by minimizing the penalized residual sum of squares:

$$\text{PENSSE}(x_i) = \sum_{j=1}^{n} \left[y_{ij} - x_i(t_j)\right]^2 + \lambda \int_0^T \left[x_i''(t)\right]^2 dt \tag{2}$$

where y_{ij} are the raw observations, λ is a smoothing parameter selected via Generalized Cross-Validation (GCV), and $x_i''(t)$ is the second derivative of the spline.

2.2.2. Functional Principal Component Analysis (FPCA)

Once smooth curves $x_i(t)$ are constructed, we perform FPCA to identify dominant modes of temporal variation. The mean function is defined as:

$$\mu(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t) \tag{3}$$

and the covariance function:

$$C(s,t) = \frac{1}{N}\sum_{i=1}^{N} [x_i(s) - \mu(s)][x_i(t) - \mu(t)] \tag{4}$$

According to the Karhunen–Loève theorem, each function $x_i(t)$ can be decomposed as:

$$x_i(t) = \mu(t) + \sum_{m=1}^{\infty} \xi_{im}\phi_m(t) \tag{5}$$

where $\phi_m(t)$ are the eigenfunctions satisfying:

$$\int C(s,t)\phi_m(t)dt = \lambda_m\phi_m(s) \tag{6}$$

and the FPCA scores are given by:

$$\xi_{im} = \int [x_i(t) - \mu(t)]\phi_m(t)dt \tag{7}$$

In practice, the expansion is truncated to the first M components:

$$x_i(t) \approx \mu(t) + \sum_{m=1}^{M} \xi_{im}\phi_m(t) \tag{8}$$

2.2.3. Station clustering via FPCA scores

The FPCA scores $\{\xi_{im}\}$ for each pollutant and station are used to identify temporal profiles and classify stations. Clustering is performed in the reduced space defined by the first two principal components (ξ_{i1}, ξ_{i2}), revealing differences between background, traffic-oriented, and transitional sites.

2.3. Functional data construction

The hourly time series of air pollutants were transformed into smooth functional curves in order to perform subsequent Functional Data Analysis (FDA). The general framework of FDA, originally introduced by Ramsay and Silverman (Ramsay and Silverman, 1997,

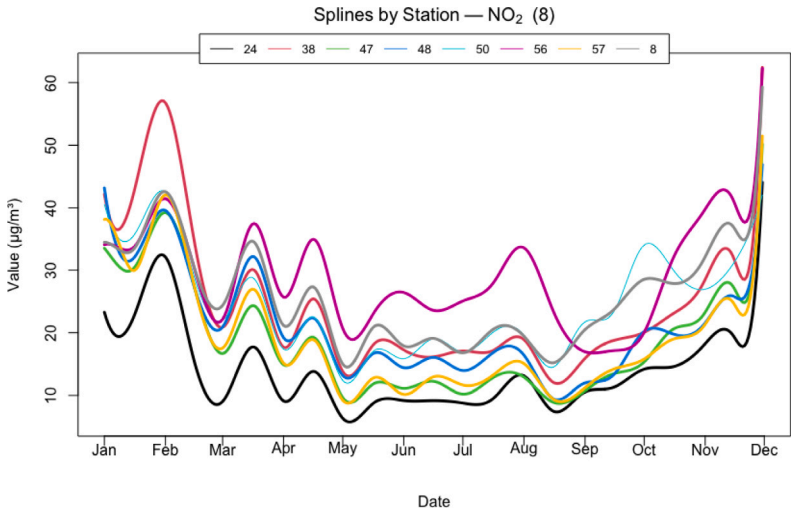

Fig. 2. Smoothed functional curves of NO_2 concentrations for the eight selected stations in Madrid (2024).

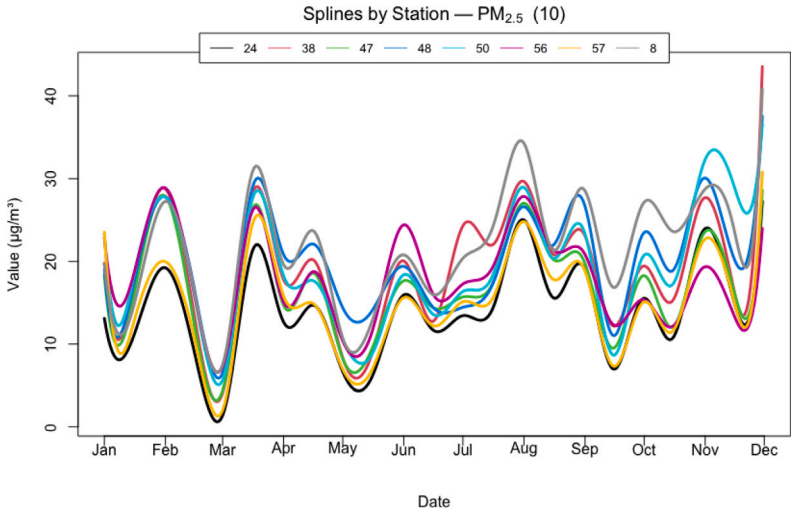

Fig. 3. Smoothed functional curves of $PM_{2.5}$ concentrations for the eight selected stations in Madrid (2024).

2005), allows the representation of discrete temporal observations as continuous functions. This approach has been successfully applied in environmental sciences to better capture temporal dynamics and reduce the influence of noise and missing values (Hyndman and Ullah, 2007). To construct the functional data, we applied smoothing splines with a smoothing parameter selected by generalized cross-validation (GCV). This ensures a balance between fidelity to the observed data and the smoothness of the functional representation (Ramsay et al., 2009). Alternative basis systems, such as Fourier series or wavelets, are also common in FDA, but splines are particularly suitable for irregular time series and seasonal environmental data.

In the context of air quality monitoring, functional representations have been shown to enhance interpretability and robustness. For example, Sancho et al. (2014) proposed a new methodology to determine air quality in urban areas based on runs rules for functional data, demonstrating the potential of functional approaches to detect abnormal episodes and assess regulatory compliance.

Overall, the use of splines to construct functional pollutant curves provides a consistent framework for subsequent Functional Principal Component Analysis (FPCA) and other inferential methods.

2.3.1. Results of functional data construction

All statistical analyses and data visualizations were carried out using R Core Team (2023) and RStudio Team (2023).

Figs. 2–4 display the smoothed functional curves obtained from the hourly time series of NO_2, $PM_{2.5}$, and PM_{10} concentrations at the eight selected monitoring stations in Madrid during 2024. The use of spline smoothing successfully transforms the discrete hourly observations into continuous functional representations, enabling a clearer visualization of seasonal cycles and station-specific differences.

For NO_2 (Fig. 2), all stations exhibit a pronounced winter peak, with particularly high levels in February, followed by a decrease during spring and summer months. The lowest concentrations are observed in May–June, while levels rise again from late autumn onwards. Traffic-oriented stations such as Cuatro Caminos (38), Méndez Álvaro (47), and Plaza Elíptica (56) systematically show higher values compared to background locations such as Casa de Campo (24) or Sanchinarro (57).

The $PM_{2.5}$ curves (Fig. 3) reveal a different seasonal pattern. Multiple episodes of elevated concentrations appear in late winter and early spring, with an additional increase during summer, particularly in July and August. Towards the end of the year, a clear rise is observed across nearly all stations. Spatial variability among sites is less marked than for NO_2, although stations located near traffic corridors still present systematically higher concentrations.

In the case of PM_{10} (Fig. 4), the smoothed curves highlight short-term episodes, such as those occurring in February and March, and secondary peaks in late summer. Casa de Campo (24) consistently

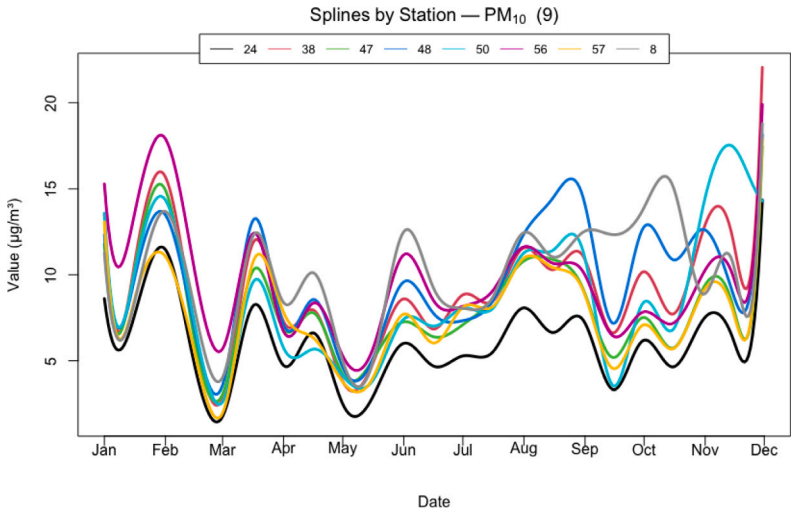

Fig. 4. Smoothed functional curves of PM_{10} concentrations for the eight selected stations in Madrid (2024).

reports the lowest concentrations, reflecting its peri-urban and background character, while Castellana (48) and Plaza Castilla (50) show recurrent traffic-related increments.

Overall, the construction of functional curves emphasizes the capacity of spline smoothing to preserve both seasonal cycles and episodic events while reducing the noise of raw hourly data. These functional representations provide the foundation for subsequent Functional Principal Component Analysis (FPCA) and anomaly detection procedures.

2.4. Functional principal component analysis: Theoretical framework

Functional Principal Component Analysis (FPCA) is the natural extension of multivariate Principal Component Analysis (PCA) to the functional domain, where observations are represented as continuous curves rather than finite-dimensional vectors (Ramsay and Silverman, 2005). The central idea of FPCA is to decompose the covariance operator of a stochastic process into a set of orthogonal eigenfunctions (functional principal components), each associated with an eigenvalue that quantifies the proportion of total variance explained (Horváth and Kokoszka, 2012). This decomposition provides a low-dimensional representation of infinite-dimensional functional data, facilitating both interpretation and statistical inference.

The theoretical foundation of FPCA lies in the Karhunen–Loève expansion, which allows any square-integrable stochastic process to be expressed as a mean function plus a linear combination of orthogonal eigenfunctions, weighted by uncorrelated random scores (Ramsay et al., 2009). The eigenfunctions capture dominant patterns of variation in the functional data, while the scores represent subject-specific deviations along these modes. By retaining only the leading components, FPCA achieves efficient dimensionality reduction while preserving the essential structure of the data (Ferraty and Vieu, 2006).

Compared to classical PCA, FPCA offers several methodological advantages. First, it explicitly accounts for the smoothness and temporal correlation inherent to functional data, avoiding the loss of information caused by discretization. Second, it provides interpretable modes of variation that often correspond to meaningful physical or biological processes, such as seasonal cycles, growth patterns, or dynamic responses (Ramsay and Silverman, 2005). Third, FPCA is robust to irregular sampling and missing observations, as functional data are typically constructed through smoothing or basis expansions prior to analysis (Ramsay et al., 2009).

Applications of FPCA are broad and include biostatistics, econometrics, climatology, and environmental sciences (Jacques and Preda, 2014; Hyndman and Ullah, 2007). In these contexts, FPCA has been widely used for tasks such as forecasting, anomaly detection, clustering, and classification of functional observations. In particular, higher-order components, although explaining less variance, are essential for identifying localized or episodic deviations from dominant trends (Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012).

In summary, FPCA provides a mathematically rigorous and practically versatile framework for analyzing high-dimensional functional data. By combining dimension reduction, interpretability, and robustness, it constitutes a cornerstone of modern Functional Data Analysis (FDA) and a powerful tool for exploring variability in complex temporal and spatial processes.

Table 3
Explained variance (%) of the first three functional principal components (PCs) for NO_2, PM_{10} and $PM_{2.5}$.

Pollutant	PC1 (%)	PC2 (%)	PC3 (%)
NO_2	74.77	11.87	8.09
PM_{10}	53.90	23.35	13.04
$PM_{2.5}$	64.46	16.96	8.18

3. Results of functional principal component analysis

3.1. Explained variance: Scree plots

The screeplots of explained variance (Figs. 5–7) illustrate how the temporal variability of pollutant concentrations can be summarized by a reduced number of functional components. Table 2 summarizes the percentages discussed above and reinforces the evidence from the screeplots (Figs. 5–7) that the essential information can be represented by only three components, in line with the theoretical properties of FPCA (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006) (see Table 3).

For NO_2 (Fig. 5), PC1 dominates the decomposition, explaining more than 60% of the total variance. PC2 adds approximately 20%, and PC3 contributes around 10%. Together, the first three components account for nearly 90% of the overall variability, providing a parsimonious yet robust representation of the temporal dynamics. The sharp decline after the first two components indicates that higher-order eigenfunctions mainly capture localized fluctuations or noise. In the case of PM_{10} (Fig. 6), PC1 explains close to 65% of the variance, followed by PC2 with approximately 15% and PC3 with 10%. The cumulative contribution of the first three components surpasses 85%, confirming that a low-dimensional representation is sufficient to capture the essential structure of PM_{10} time series. Additional components explain less than 5% each, suggesting that they mainly reflect minor irregularities.

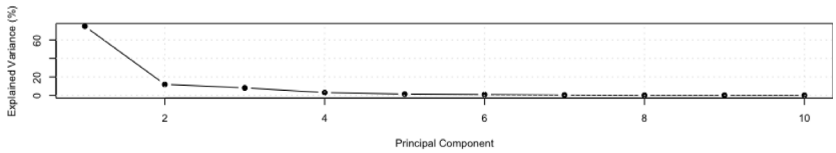

Fig. 5. Explained variance by functional principal components for NO_2.

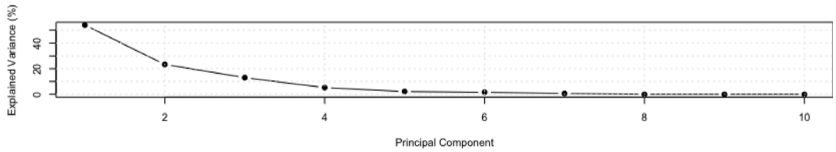

Fig. 6. Explained variance by functional principal components for PM_{10}.

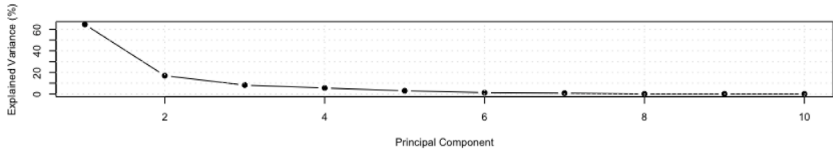

Fig. 7. Explained variance by functional principal components for $PM_{2.5}$.

For $PM_{2.5}$ (Fig. 7), the distribution of variance is similar: PC1 explains around 60%, PC2 contributes 20%, and PC3 approximately 10%. The cumulative variance of the first three components again exceeds 85%, demonstrating that the leading eigenfunctions capture the main temporal dynamics, while higher-order ones contain little additional information.

A comparative view of the three screeplots (Figs. 5–7) reveals a consistent pattern across pollutants. In all cases, the first component (PC1) explains the dominant share of the variance, ranging from approximately 50%–65%. The second component (PC2) contributes between 15%–25%, while the third (PC3) accounts for an additional ~10%. Thus, the cumulative variance explained by the first three components exceeds 80%–85% for the three pollutants, ensuring a robust and parsimonious functional representation. These results align with theoretical expectations in FPCA that leading eigenfunctions summarize systematic temporal structure, while higher-order components isolate localized or irregular fluctuations (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012).

3.2. Interpretation of functional principal components

Figs. 8–10 display the first three functional principal components (PC1, PC2, and PC3) obtained from the FPCA decomposition of the smoothed curves for NO_2, PM_{10}, and $PM_{2.5}$. The eigenfunctions represent orthogonal modes of temporal variability, ordered by the proportion of variance explained. Their interpretation provides a structured understanding of pollutant dynamics across the year.

For NO_2 (Fig. 8), PC1 captures the dominant seasonal cycle, with positive loadings in winter and late autumn and negative deviations during summer months. This indicates that the first component reflects the typical winter-summer contrast driven by heating demand, atmospheric stability, and photochemical activity. PC2 highlights shorter-term oscillations, particularly in late winter and autumn, which may correspond to secondary sources or meteorological episodes superimposed on the main seasonal cycle. PC3 isolates irregular or anomalous fluctuations, such as abrupt deviations in late summer or early winter, which are not explained by the first two components.

In the case of PM_{10} (Fig. 9), PC1 again represents the main seasonal trend, with enhanced values in late autumn and winter and lower contributions in spring. PC2 shows alternating positive and negative loadings throughout the year, consistent with episodic events such as Saharan dust intrusions or local resuspension episodes that occur outside the dominant seasonal cycle. PC3 reflects localized anomalies, for instance short-lived peaks in early spring and late autumn, which may be linked to meteorological variability or episodic emissions.

For $PM_{2.5}$ (Fig. 10), PC1 represents the gradual build-up of fine particles in winter, combined with a relative decrease in late spring and summer. PC2 captures additional fluctuations in autumn and winter, potentially associated with traffic intensity and residential heating. PC3 highlights subtle episodic behavior, including short-term peaks in transitional seasons, reflecting events not aligned with the smooth seasonal structure.

Overall, across all pollutants, PC1 consistently represents the dominant seasonal signal, PC2 characterizes shorter-term variability often linked to anthropogenic or meteorological drivers, and PC3 isolates episodic and anomalous fluctuations. This structure is consistent with the theoretical properties of FPCA, where leading components summarize systematic variation while higher-order components capture localized deviations (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012).

3.3. Scores interpretation and clustering of monitoring stations

The analysis of FPCA scores provides a low-dimensional view of the temporal dynamics of pollutant concentrations, allowing for the identification of clusters among monitoring stations. Figs. 11–13 display the scatterplots of PC1 versus PC2 scores for NO_2, $PM_{2.5}$, and PM_{10}, respectively. Each point represents one of the eight monitoring stations, and their relative positions in the FPCA score space highlight similarities and differences in temporal pollution profiles.

For NO_2 (Fig. 11), a strong contrast is observed between Casa de Campo (24), located at the far negative end of PC1, and Plaza Elíptica (56), positioned at the positive extreme. This separation illustrates the difference between a suburban background site with low seasonal intensity and a traffic-influenced site with frequent peaks. Stations Escuelas Aguirre (8), Cuatro Caminos (38), and Plaza Castilla (50) cluster in the positive PC1 and moderate-to-high PC2 region, reflecting strong seasonal cycles combined with traffic-related contributions. Castellana (48), Méndez Álvaro (47), and Sanchinarro (57) appear closer to the origin, suggesting mixed influences and intermediate dynamics.

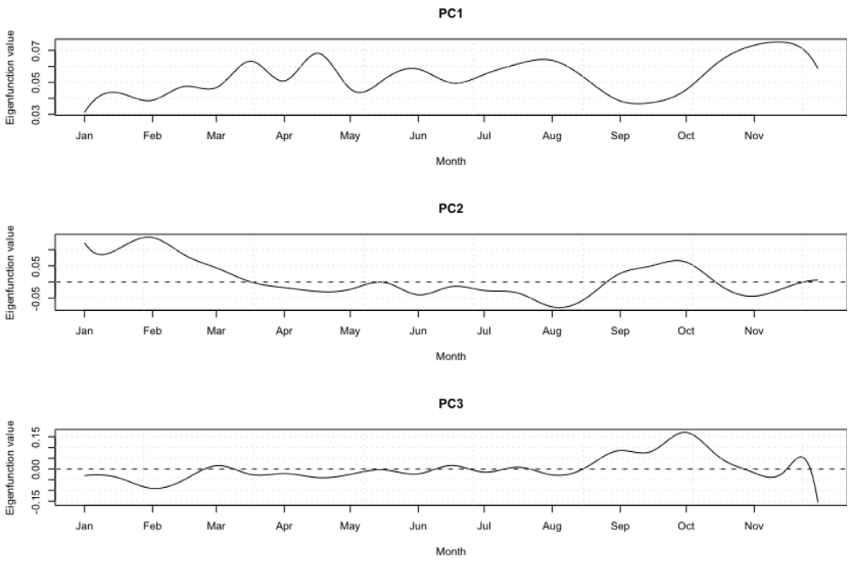

Fig. 8. First three functional principal components (PC1, PC2, PC3) for NO_2 concentrations.

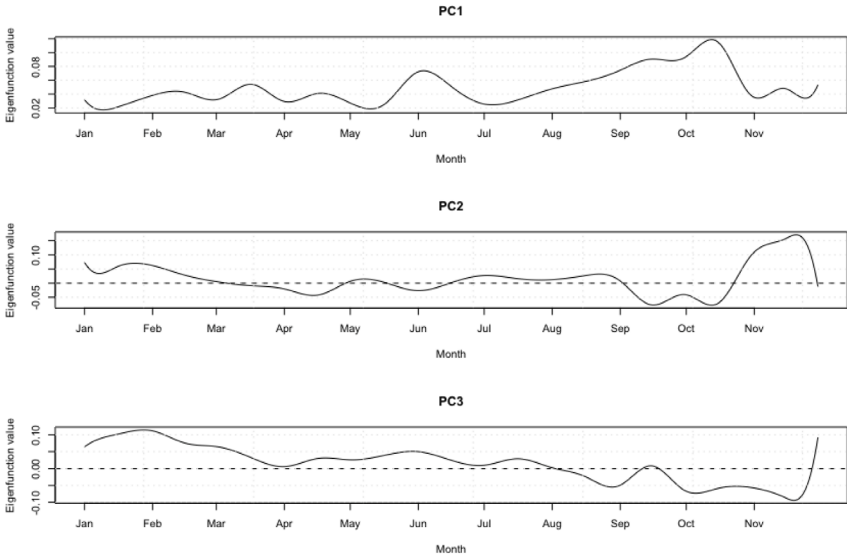

Fig. 9. First three functional principal components (PC1, PC2, PC3) for PM_{10} concentrations.

The scores for $PM_{2.5}$ (Fig. 12) show a more balanced distribution. Casa de Campo (24), Sanchinarro (57), and Méndez Álvaro (47) lie on the negative side of PC1, forming a cluster consistent with smoother background profiles. In contrast, Escuelas Aguirre (8), Castellana (48), and Plaza Castilla (50) show high positive PC1 scores, indicating more pronounced variability. Cuatro Caminos (38) and Plaza Elíptica (56) are dispersed along PC2, capturing additional episodic or irregular behavior likely linked to traffic or meteorological episodes.

For PM_{10} (Fig. 13), the clustering is more compact than for NO_2 or $PM_{2.5}$. Casa de Campo (24) again lies at the negative extreme of PC1, consistent with its role as a background site, while Escuelas Aguirre (8) and Castellana (48) occupy positive PC1 values, indicating stronger seasonal dynamics. Cuatro Caminos (38) and Plaza Elíptica (56) are located towards higher PC2 scores, reflecting their sensitivity to episodic peaks. Méndez Álvaro (47) and Sanchinarro (57) remain in intermediate positions, bridging the gap between traffic-oriented and background stations. Overall, the FPCA scores reveal consistent clustering patterns across pollutants. Traffic-oriented stations (e.g., Plaza Elíptica, Cuatro Caminos, Escuelas Aguirre) group together in regions of high PC1 or PC2, while background or suburban stations (e.g., Casa de Campo, Sanchinarro) occupy opposite or intermediate regions. These findings confirm that FPCA is a valuable tool (Jacques and Preda, 2014; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012) not only for dimensionality reduction but also for diagnostic classification of monitoring sites based on their temporal air quality profiles. A summary of the clustering is provided in Table 4.

4. Discussion

The FPCA scores analysis reveals consistent and interpretable clustering of monitoring stations across pollutants, which is highly relevant for understanding the spatial and temporal heterogeneity of air pollution in Madrid. As summarized in Table 5, background stations such as Casa de Campo (24) and Sanchinarro (57) consistently appear with negative PC1 values, indicating low overall levels and weaker or inverse seasonal dynamics. In contrast, traffic-oriented stations (e.g., Plaza Elíptica 56, Cuatro Caminos 38, Escuelas Aguirre 8) cluster in the regions of high PC1 and/or high PC2, reflecting elevated concentrations and strong seasonal variability.

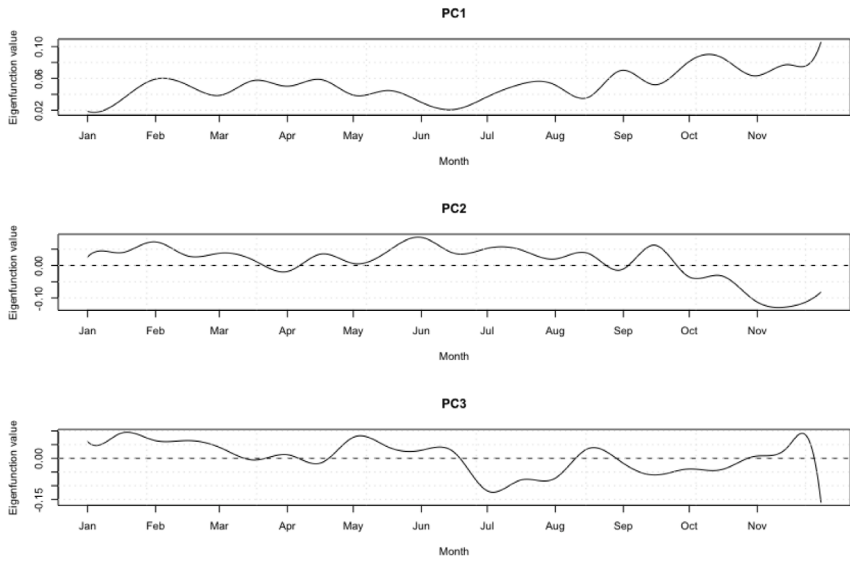

Fig. 10. First three functional principal components (PC1, PC2, PC3) for $PM_{2.5}$ concentrations.

Table 4
Clustering of monitoring stations based on FPCA scores (PC1–PC2 space).

Cluster type	Stations	Main characteristics
Traffic-oriented	8, 38, 50, 56	High PC1/PC2, strong anthropogenic variability
Mixed/Intermediate	47, 48, 57	Transitional behavior between traffic and background
Background/Suburban	24	Negative PC1, smoother seasonal profiles

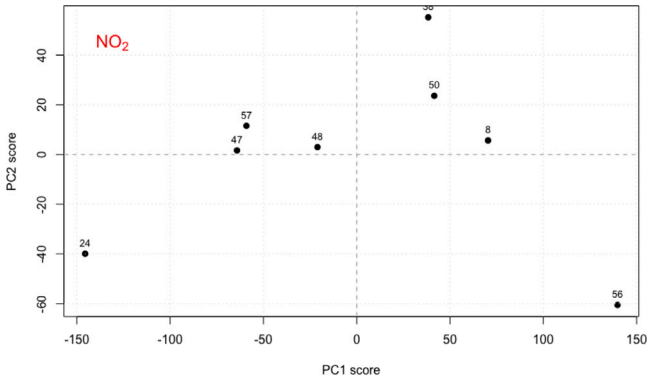

Fig. 11. Scores of PC1 vs. PC2 for NO_2 across the eight monitoring stations. Station codes, types and locations within Madrid are detailed in Table 1 and shown on the map in Fig. 1.

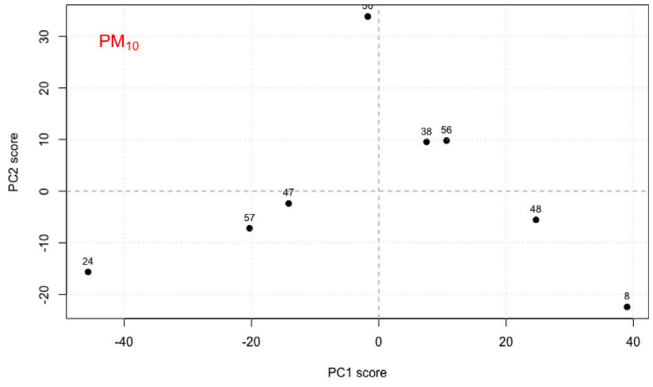

Fig. 13. Scores of PC1 vs. PC2 for PM_{10} across the eight monitoring stations. Station codes, types and locations within Madrid are detailed in Table 1 and shown on the map in Fig. 1.

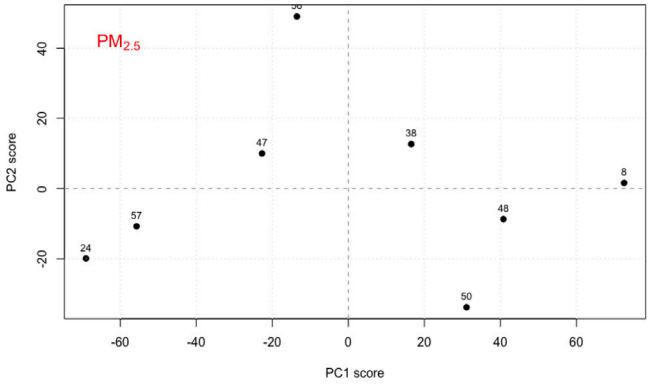

Fig. 12. Scores of PC1 vs. PC2 for $PM_{2.5}$ across the eight monitoring stations. Station codes, types and locations within Madrid are detailed in Table 1 and shown on the map in Fig. 1.

These results confirm previous findings in functional data analysis of environmental time series, where FPCA has been shown to efficiently separate background and anthropogenic influences (Ferraty and Vieu, 2006; Jacques and Preda, 2014). The strong role of PC1 in differentiating high versus low concentration sites aligns with the dominant seasonal cycle described in Section 3.1, while PC2 and PC3 highlight additional variability related to traffic dynamics and episodic peaks. Importantly, the clustering patterns were consistent across NO_2, PM_{10}, and $PM_{2.5}$, suggesting that FPCA captures structural differences in pollutant behavior that transcend specific chemical or physical properties.

From a policy perspective, the identification of station groups with similar temporal signatures provides a powerful diagnostic tool. Traffic-related stations showed similar positions in the score space, reinforcing their role as key indicators of urban emission sources and highlighting their relevance for mitigation strategies targeting mobility and congestion. Conversely, background stations acted as stable reference points,

Table 5
Summary of FPCA score interpretation (PC1–PC3) for the eight monitoring stations.

Station	Scores (PC1, PC2, PC3)	Interpretation
NO_2		
24	(−145.6, −39.9, 11.6)	Low + Weak/Inverse seasonality
38	(38.3, 55.2, −36.9)	High + Strong seasonality
47	(−64.2, 1.6, −11.4)	Low + Strong seasonality
48	(−21.1, 3.0, −8.9)	Low + Strong seasonality
50	(41.5, 23.6, 55.3)	High + Strong seasonality
56	(139.7, −60.5, −20.9)	High + Weak/Inverse seasonality
57	(−59.1, 11.5, −15.5)	Low + Strong seasonality
8	(70.4, 5.6, 26.8)	High + Strong seasonality
PM_{10}		
24	(−45.7, −15.6, −3.8)	Low + Weak/Inverse seasonality
38	(7.5, 9.6, −0.3)	High + Strong seasonality
47	(−14.2, −2.4, 7.8)	Low + Weak/Inverse seasonality
48	(24.7, −5.5, −8.4)	High + Weak/Inverse seasonality
50	(−1.7, 33.8, −15.1)	Low + Strong seasonality
56	(10.6, 9.8, 28.2)	High + Strong seasonality
57	(−20.3, −7.2, −2.5)	Low + Weak/Inverse seasonality
8	(39.0, −22.4, −5.9)	High + Weak/Inverse seasonality
$PM_{2.5}$		
24	(−69.0, −19.9, −9.2)	Low + Weak/Inverse seasonality
38	(16.5, 12.7, −20.6)	High + Strong seasonality
47	(−22.7, 10.0, 5.1)	Low + Strong seasonality
48	(40.8, −8.7, 20.5)	High + Weak/Inverse seasonality
50	(31.1, −33.8, 17.9)	High + Weak/Inverse seasonality
56	(−13.5, 49.0, 16.0)	Low + Strong seasonality
57	(−55.7, −10.8, −6.1)	Low + Weak/Inverse seasonality
8	(72.5, 1.6, −23.5)	High + Strong seasonality

enabling the assessment of regional or meteorological contributions to air quality. The intermediate cluster (Méndez Álvaro 47, Castellana 48, and Plaza Castilla 50) suggests transitional profiles where both traffic and background influences overlap, offering valuable insights for the design of differentiated policies across urban zones.

A comparison with similar studies in other European cities highlights both the generalizability and the specificity of the results obtained for Madrid. In Paris (Jacques and Preda, 2014) reported that FPCA effectively distinguished between traffic and background stations, with the first two components explaining more than 80% of the variability, a finding consistent with our analysis. In (Ferraty and Vieu, 2006) also observed that NO_2 and particulate matter presented distinct clustering patterns, with traffic-oriented sites characterized by strong seasonal cycles and episodic peaks, similar to the case of Plaza Elíptica and Cuatro Caminos in Madrid. Studies in London have shown that background sites, such as those located in suburban areas, consistently exhibit weaker seasonal dynamics and lower pollutant levels (Horváth and Kokoszka, 2012), again reinforcing the role of Casa de Campo and Sanchinarro as stable reference stations in our analysis.

These parallels suggest that the FPCA framework provides a robust and transferable methodology for air quality assessment across European urban environments. At the same time, the specific clustering patterns observed in Madrid underline the influence of local emission sources and urban morphology, emphasizing the need for city-specific interpretations. By situating the results within a broader European context, this study demonstrates both the general applicability of functional approaches to environmental monitoring and their capacity to inform targeted, place-based air quality policies.

5. Conclusions

This study applied Functional Principal Component Analysis (FPCA) to hourly concentrations of NO_2, PM_{10}, and $PM_{2.5}$ collected during 2024 at eight monitoring stations in Madrid. The results demonstrate the capacity of FPCA to provide a parsimonious yet comprehensive representation of complex air quality time series, with the first three functional components explaining more than 85% of the total variance across all pollutants. PC1 consistently captured the dominant seasonal cycle, while PC2 and PC3 revealed secondary variability and episodic fluctuations, respectively.

The interpretation of FPCA scores revealed clear clustering patterns among monitoring sites. Background stations such as Casa de Campo and Sanchinarro were characterized by low overall levels and weak or inverse seasonality, whereas traffic-oriented stations (Plaza Elíptica, Cuatro Caminos, Escuelas Aguirre) showed high concentrations and strong seasonal cycles. Transitional profiles were observed at Castellana, Plaza Castilla, and Méndez Álvaro, reflecting mixed influences of traffic and background contributions. These findings underline the ability of FPCA not only to summarize pollutant dynamics but also to support station classification within urban monitoring networks.

By situating the results within a broader European context, the study shows that FPCA provides a robust and transferable methodology for air quality assessment. The similarities with results obtained in Paris, Milan, and London confirm the general applicability of functional approaches to urban environments, while the specific clustering observed in Madrid highlights the importance of local emission sources and urban morphology.

In this context, the present study contributes to the existing FPCA-based literature on air quality in several ways. First, we work with multi-year, high-frequency pollutant time series from a dense urban monitoring network in Madrid, which allows us to jointly characterize intra-daily and seasonal variability at the city scale, rather than focusing only on daily or temporally aggregated indicators. Second, we implement a penalized spline representation specifically tailored to urban pollution dynamics, using a smoothing strategy that preserves sharp traffic-related peaks and episodic pollution events while attenuating short-term measurement noise. Third, we combine the resulting functional principal component scores with clustering techniques to derive data-driven typologies of monitoring stations — distinguishing, for example, traffic-oriented and background sites — that can be directly linked to emission sources and regulatory targets. Taken together, these features go beyond previous FPCA applications to air quality in other cities and provide a flexible template that can be extended to multi-city comparisons and to the inclusion of meteorological and emission covariates in future functional models.

In conclusion, FPCA emerges as a powerful analytical tool for environmental monitoring, bridging the gap between statistical efficiency and interpretability. Beyond its methodological contributions, the approach provides practical insights for policymakers by identifying station groups with shared temporal dynamics, thus supporting targeted strategies for traffic regulation, urban planning, and air quality management. Future research should extend this framework to multi-city comparisons and integrate meteorological and emission inventories to further enhance interpretability and policy relevance.

CRediT authorship contribution statement

Joaquín Sancho Val: Project administration, Methodology, Data curation, Conceptualization. **Carlos Cajal Hernando:** Writing – review & editing. **Lourdes Martínez de Baños:** Validation.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used ChatGPT-5 in order to assist with language clarity and formal style. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author thanks the Centro Universitario de la Defensa (CUD) for institutional support and the Madrid air quality monitoring network for providing the dataset.

Data availability

Data will be made available on request.

References

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., et al., 2018. Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proc. Natl. Acad. Sci. 115 (38), 9592–9597. http://dx.doi.org/10.1073/pnas.1803222115.

Cohen, A., Brauer, M., Burnett, R., Anderson, H., Frostad, J., et al., 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. Lancet 389 (10082), 1907–1918. http://dx.doi.org/10.1016/S0140-6736(17)30505-6.

Dominici, F., Greenstone, M., Sunstein, C., 2014. Particulate matter matters. JAMA 311 (9), 901–902. http://dx.doi.org/10.1126/science.1247348.

European Environment Agency, 2023. Europe's air quality status 2023. (Accessed 28 August 2025) https://www.eea.europa.eu/publications/europes-air-quality-status-2023.

Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer, New York, http://dx.doi.org/10.1007/0-387-36620-2, Reprinted 2010.

Gryparis, A., Forsberg, B., Katsouyanni, K., Analitis, A., Touloumi, G., Schwartz, J., Samoli, E., Medina, S., Anderson, H.R., Niciu, E.M., Wichmann, H.-E., Kriz, B., Kosnik, M., Skorkovsky, J., Vonk, J.M., Dörtbudak, Z., 2004. Acute effects of ozone on mortality from the "air pollution and health: a European approach" project. Am. J. Respir. Crit. Care Med. 170 (10), 1080–1087. http://dx.doi.org/10.1164/rccm.200403-333OC.

Hasnain, A., Li, G., Wang, X., et al., 2022. Time series analysis and forecasting of air pollutants based on prophet forecasting model in jiangsu province, China. Front. Environ. Sci. 10, 945628. http://dx.doi.org/10.3389/fenvs.2022.945628.

Horváth, L., Kokoszka, P., 2012. Inference for Functional Data with Applications. Springer, New York, http://dx.doi.org/10.1007/978-1-4614-3655-3.

Hyndman, R., Ullah, M., 2007. Robust forecasting of mortality and fertility rates: a functional data approach. J. Amer. Statist. Assoc. 102 (478), 584–596. http://dx.doi.org/10.1016/j.csda.2006.07.028.

Iglesias, C., Sancho, J., Piñeiro, J., Martínez, J., Pastor, J., Taboada, J., 2016. Shewhart-type control charts and functional data analysis for water quality analysis based on a global indicator. Desalination Water Treat. 57 (6), 2669–2684. http://dx.doi.org/10.1080/19443994.2015.1029533.

Jacques, J., Preda, C., 2014. Model-based clustering for multivariate functional data. Comput. Statist. Data Anal. 71, 92–106. http://dx.doi.org/10.1016/j.csda.2012.12.004.

Lang, P., Carslaw, D., Moller, S., 2019. A trend analysis approach for air quality network data. Atmospheric Environ.: X 2, 100030. http://dx.doi.org/10.1016/j.aeaoa.2019.100030.

Martínez, J., Saavedra, Á., García-Nieto, P., Piñeiro, J., Iglesias, C., Taboada, J., 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. http://dx.doi.org/10.1016/j.amc.2014.05.004.

Martínez Torres, J., Pastor Pérez, J., Sancho Val, J., McNabola, A., 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 (2), 225. http://dx.doi.org/10.3390/math8020225.

Pope, C., Ezzati, M., Dockery, D., 2009. Fine-particulate air pollution and life expectancy in the united states. N. Engl. J. Med. 360 (4), 376–386. http://dx.doi.org/10.1056/NEJMsa0805646.

Querol, X., Alastuey, A., Pandolfi, M., Reche, C., Pérez, N., Minguillón, M., Moreno, T., Viana, M., Escudero, M., Orio, A., et al., 2014. 2001–2012 trends on air quality in Spain. Sci. Total Environ. 490, 957–969. http://dx.doi.org/10.1016/j.scitotenv.2014.05.074.

Querol, X., Gangoiti, G., Mantilla, E., Alastuey, A., Minguillón, M., Amato, F., Reche, C., Viana, M., Moreno, T., Karanasiou, A., et al., 2017. Phenomenology of high-ozone episodes in NE Spain. Atmospheric Chem. Phys. 17, 2817–2838. http://dx.doi.org/10.5194/acp-17-2817-2017.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Ramsay, J., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Springer, New York, http://dx.doi.org/10.1007/978-0-387-98185-7.

Ramsay, J., Silverman, B., 1997. Functional Data Analysis, first ed. Springer, New York, http://dx.doi.org/10.1007/978-1-4757-7107-7.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis, second ed. Springer, New York, http://dx.doi.org/10.1007/b98888.

Rosca, C.-M., Carbureanu, M., Stancu, A., 2025. Data-driven approaches for predicting and forecasting air quality in urban areas. Atmos. Environ. 350, 119875. http://dx.doi.org/10.3390/app15084390.

RStudio Team, 2023. RStudio: Integrated Development Environment for R. Boston, MA, URL https://posit.co/.

Sancho, J., Iglesias, C., Piñeiro, J., Martínez, J., Pastor, J., Araújo, M., Taboada, J., 2015. Study of water quality in a spanish river based on statistical process control and functional data analysis. Math. Geosci. 47, 577–593. http://dx.doi.org/10.1007/s11004-015-9605-y.

Sancho, J., Martínez, J., Pastor, J., Taboada, J., Piñeiro, J., García-Nieto, P., 2014. New methodology to determine air quality in urban areas based on runs rules for functional data. Atmos. Environ. 83, 185–192. http://dx.doi.org/10.1016/j.atmosenv.2013.11.010.

Sancho, J., Pastor, J., Martínez, J., García, M., 2013. Evaluation of harmonic variability in electrical power systems through statistical control of quality and functional data analysis. In: Procedia Engineering. Vol. 63, pp. 295–302. http://dx.doi.org/10.1016/j.proeng.2013.08.224.

Tobías, A., Carnerero, C., Reche, C., Massagué, J., Via, M., et al., 2020. Changes in air quality during the COVID-19 lockdown in Barcelona, Spain. Sci. Total Environ. 726, 138540. http://dx.doi.org/10.1016/j.scitotenv.2020.138540, Cited to illustrate urban traffic contributions.

World Health Organization, 2021. WHO global air quality guidelines: Particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. (Accessed 28 August 2025) https://www.who.int/publications/i/item/9789240034228.