Journal of the Royal Statistical Society Series C: Applied Statistics, 2025, **74**, 229–248
https://doi.org/10.1093/jrsssc/qlae053
Advance access publication 14 October 2024
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Modelling particle number size distribution: a continuous approach

Israel Martínez-Hernández[1], Carolina Euán[1], Wesley S. Burr[2], Melanie Meis[3], Marta Blangiardo[4] and Monica Pirani[4]

[1]School of Mathematical Sciences, Lancaster University, Lancaster, UK
[2]Department of Mathematics & Statistics, Trent University, Peterborough, Canada
[3]Departamento de Ciencias de la Atmósfera y los Océanos, Facultad de Ciencias Exactas y Naturales, CONICET, Universidad de Buenos Aires, Buenos Aires, Argentina
[4]Department of Epidemiology and Biostatistics, MRC Centre for Environment and Health, Imperial College London, London, UK

Address for correspondence: Israel Martínez-Hernández, School of Mathematical Sciences, Lancaster University, Lancaster LA1 4YW, UK. Email: i.martinezhernandez@lancaster.ac.uk

Abstract

Particulate matter (PM) is well known to be detrimental to health, and it is crucial to apportion PM into the underlying sources to target policies. Particle number size distribution (PNSD) is the most accessible data to identify these sources, which provides information on the PM sizes. Here, we propose a new functional factor model for PNSD, which allows to disentangle PM into sources and contributions while considering the complex dependencies of the data across different sizes and periods. Through a simulation study, we show that this method is able to identify sources correctly, and we use it to analyse hourly PNSD data collected in London for 7 years, finding 6 well-defined sources. Our proposed methodology is fast, accurate, and reproducible.

Keywords: air pollution, functional data, functional factor model, particle number size distribution, source apportionment, ultrafine particles

1 Introduction

Air pollution is well known as a public health emergency, being responsible for approximately 9 million deaths per year (Fuller et al., 2022). One of the most harmful components of air pollution is particulate matter (PM), which is a mixture of solids and liquid particles varying in composition, size, and origin. There is strong evidence on the link between PM and health (see Hime et al., 2018; Mukherjee & Agrawal, 2017; Wu et al., 2018, among others). In particular, the association between cardiovascular and respiratory diseases, cancer, and genetic mutation and PM has been well documented (Moreno-Ríos et al., 2022), and evidence has been judged sufficient to classify outdoor air particles as carcinogenic to humans (Loomis et al., 2013). The relationship between PM and health conditions is particularly strong in large urban environments, which are densely populated (Al-Kindi et al., 2020) and typically suffer from the highest air pollution levels. The knowledge of PM sources is therefore vital for supporting policymakers in identifying actions to improve urban air quality.

Fine and ultrafine particles are those with a diameter smaller than 2,500 and 100 nanometres (nm), respectively. Ultrafine particles are more detrimental to health, as they cause more severe pulmonary inflammation and are retained longer in the lungs (Marval & Tronville, 2022; Murray et al., 2020; Ohlwein et al., 2019). Furthermore, mixtures of particles from different sources can have different toxicity and health effects (see, e.g. Park et al., 2014; Pirani et al., 2015).

Received: November 29, 2023. **Revised:** September 5, 2024. **Accepted:** September 16, 2024
© The Royal Statistical Society 2024.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

Thus, identifying the possible different sources that may be involved in generating fine and ultrafine particles is key to inform public health policies aiming to reduce their emissions (Zeng et al., 2019). Numerous investigations have shown that public policy intervention positively impacts PM emission reduction (Bishop & Bornioli, 2022; Gao et al., 2016).

The process of deriving information about pollution sources and the amount they emit is known as *source apportionment* (SA). In this paper, we consider SA for particle number size distribution (PNSD) measurements (see, e.g. Hopke et al., 2022). PNSD involves considering particle sizes that vary from 10 to 2,500 nm, covering the fine and ultrafine domain. Typically, PNSD data are high-dimensional, highly correlated, and collected at a fine time resolution (e.g. hours) for several years.

In the context of SA, several methods have been used and proposed, spanning from enrichment factors (EF) (Wang et al., 2023), eigenvector analysis Principal Component Analysis (PCA), Factor Analysis (FA) (Thurston et al., 2011) and neural networks among others (Watson et al., 2002). The most widely applied method in the literature has been the positive matrix factorization (PMF); see Paatero and Tapper (1994), Dutton et al. (2010) and Norris et al. (2014) or the review paper by Sun et al. (2020). PMF is a multivariate factor model with the additional constraint that the loadings and scores need to be nonnegative, although, differently from PCA, it has no orthogonality requirement (Chan & Mozurkewich, 2007), and is capable of identifying the relative contributions of each source. While PMF is easy to implement and is used by environmental agencies across the world, it has some drawbacks: (1) it does not guarantee to find the optimal parameters; (2) it will give different results due to the positiveness constraints and the optimization algorithm; (3) as the number of sources needs to be fixed a priori, the results might change for different numbers (Paatero & Tapper, 1994). Additionally, as recent datasets have become larger due to the increase of temporal granularity (e.g. hourly data), PMF in its original form has become less effective, as the one currently used can only handle 500,000 points. While advances have been made in this direction, the available methods take considerable time to run (Hopke et al., 2023). Alternatively, to try to overcome the high dimensionality, one approach that has been used is to aggregate the different sizes and time points into a small number of bins (see, e.g. Baerenbold et al., 2023).

Here, we propose a novel method for carrying out SA, framed into the functional data analysis approach (FDA) (Ramsay & Silverman, 2005). In other words, we assume that the PNSD data are a sequence of positive continuous functions (see Figure 1). FDA deals with data that are defined in an intrinsically infinite-dimensional space and gains computational advantages by representing the continuous functions with finite basis functions. With the FDA approach, the PNSD data are seen as a (functional) time series, where at each time point, an observation is a continuous function. Since we deal with positive functions, we must consider the positiveness constraint. For this, we apply a centred log-ratio transformation that links unrestricted continuous functions and density

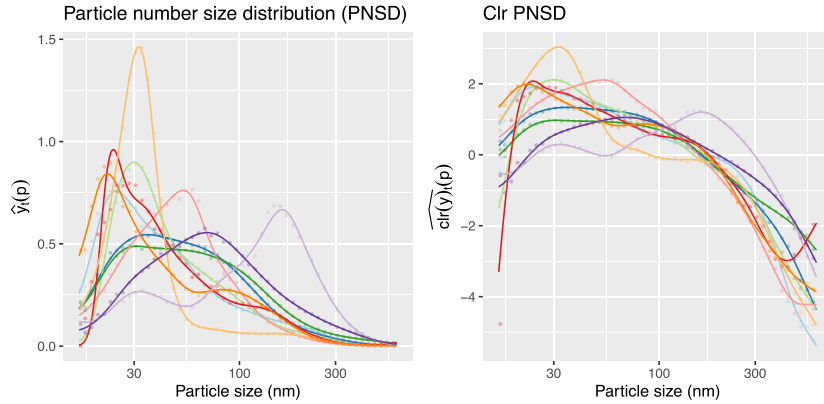

Figure 1. Functional representation of a sample of particle number size distributions—each colour is a different hourly sample at 51 discrete particle sizes. The points represent the observed measurements, and the continuous lines represent the functional fit using a B-spline basis. Left: Functional representation in the Bayes space, $\mathcal{B}^2[a, b]$. Right: Functional representation in the Hilbert space, $L^2[a, b]$.

Downloaded from https://academic.oup.com/jrssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

functions (Van den Boogaart et al., 2014). Then, in the space of density functions, we propose a functional factor model. The centred log-ratio function allows us to use FDA tools to define the model. Our approach is extremely fast, taking seconds to run on hourly data over multiple years. Additionally, this approach can take into account temporal dependence in the data, which is not considered in the standard PMF model. Using a bootstrap method, we are able to quantify the uncertainty of the source profiles. Overall, our method can deal with large-scale datasets with a very low computational cost, and the results are consistent in the sense that it does not depend on the optimization algorithm or the number of sources.

The remainder of our paper is organized as follows. In Section 2, we describe the data and the FDA approach, focusing on the estimation of the continuous functions using a B-spline basis function. In Section 3, we describe the functional factor model for positive functions. The inference procedure is presented in Section 4. In Section 5, we conduct a simulation study to evaluate the performance of the proposed methodology under different settings. In Section 6, we analyse an hourly PNSD dataset collected at an urban background site in central London (UK). We conclude in Section 7 by discussing the main contributions of this work and future directions.

2 PNSD as functional data

The illustrative data we will use in this paper to apply the proposed FDA approach are averaged hourly PNSD collected at a single urban background monitoring station located in a residential area of London (North Kensington, 51.52N, 0.21W; 27 m high above sea level). The station belongs to the London Air Quality Network as well as to the national Automatic Urban and Rural Network and is sited in the grounds of Sion Manning School in St Charles Square. PNSD measurements from 1 January 2011 to 31 October 2018 were obtained with SMPS TSI 3080 + CPC TSI 3,775 with long DMA instrument fitted with a diffusion dryer, following the recommendation from EUSAAR/ACTRIS protocol (Wiedensohler et al., 2012), then PNSD were corrected for diffusional losses.

The data $\mathbf{w}_t = (w_t(p_1), \ldots, w_t(p_m))$ are a vector with $w_t(p_j)$ the number of particle size concentration of size p_j at time t (hour). These hourly data were collected for $m = 51$ different particle sizes, ranging from 16.55 to 604.4 nm. Thus, the raw data, $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T\}$, is an hourly multivariate time series with a sample size of 68,664 and 51-dimensional. Due to the large amount of data and high dimensionality, we will consider a functional data approach. Specifically, we will assume that, at each time point, $\mathbf{w}_t$ is a finite sample of a continuous function y_t (with noise).

Representing the PNSD data as continuous functions is reasonable since the particle size p varies continuously between a and b, where a and b can be, e.g. the smallest and largest particle size that can be measured. Thus, we assume that our data is a sequence of continuous functions $\{y_t : [a, b] \to \mathbb{R}^+, t = 1, \ldots, T = 68{,}664\}$. Furthermore, we normalize each function y_t such that it integrates one. That is, we model the sequence of continuous density functions $\{y_t\}$.

The original data had a moderate percentage of missingness; of the $68{,}664 \times 51$ observations in the original data set, 655,553 were missing (18.3%). Missing data in particle concentration are usually related to downtime of the measuring device and, therefore, occurs simultaneously for all size bins. The PNSD missing duration ranges from 1 hr to 15 days, with 80% of the missing duration less than three consecutive days. To resolve this issue, we applied a time series imputation method before transforming the data to the functional space, using the algorithm discussed in Castel and Burr (2021). These were all imputed for the final set of 51 time series considered.

We should note that although we assume y_t is a continuous density function at each time point, in practice, we observe $\mathbf{w}_t$, a discrete and noisy version at m points, $p_1, \ldots, p_m$ (Ramsay & Silverman, 2005). Thus, we will need to estimate the continuous function at each time point t. A common procedure is to use basis functions, e.g. cubic B-spline, $\psi_1, \ldots, \psi_L$, such that $y_t(p) \approx \sum_{i=1}^{L} c_{i,t}\psi_i(p)$. Basis representation is a well-known problem in FDA, and for instance, the coefficients $c_{i,t}$ can be estimated using a least squares estimator. The main assumption is that y_t is an element of a Hilbert space. However, our data (density functions) do not fit under the common assumption of Hilbert spaces. For instance, given that the set of density functions is not a vector space with the usual $(+, \cdot)$ operation, it is not guaranteed that summing two of these will result in another density function. In addition, our data have the additional constraint

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

that the estimated curve must be nonnegative. For these reasons, we use an alternative approach and follow the method proposed by Talská et al. (2018), briefly described below (also see Hron et al., 2016):

1. We assume that $\{y_t\}$ belongs to a Bayes space, $\mathcal{B}^2[a, b]$ (as explained in Section 3.1). We apply an isometric isomorphism from Bayes space to Hilbert space $L^2[a, b]$, the centred log-ratio (clr) transformation: $\text{clr}(y_t)(p) := \log y_t(p) - \frac{1}{b-a}\int_a^b \log y_t(s)\,\mathrm{d}s$.
2. The transformed density, $\text{clr}(y_t)$, belongs to $L^2[a, b]$, and it can be approximated using FDA methods for basis representation: $\text{clr}(y_t)(p) \approx \sum_{j=1}^{L} c_{j,t}\psi_j(p)\,\mathrm{d}p$.
3. For each t, we then estimate the basis representation by minimizing the cost function

$$\lambda \sum_{i=1}^{m}\left\{\text{clr}(y_t)(p_i) - \sum_{j=1}^{L} c_{j,t}\psi_j(p_i)\right\}^2 + (1-\lambda)\int_a^b \left\{\frac{\partial^2}{\partial p^2}\sum_{j=1}^{L} c_{j,t}\psi_j(p)\right\}^2 \mathrm{d}p,$$

   with respect to $\{c_{1,t}, \ldots, c_{L,t}\}$.[1] The parameter $\lambda \in (0, 1)$ controls the smoothness of the estimated function.
4. Then, the estimated continuous function $\widehat{y}_t$ is obtained via $\widehat{\text{clr}(y_t)}$ in the Hilbert space by applying the inverse clr transformation.

Steps 1–4 guarantee that $\widehat{y}_t$ is a positive function and integrates one. This procedure was applied to each hourly density function to obtain the continuous curves at each hour t, where we set $\lambda = 0.997$ using generalized cross-validation criteria (see Appendix A). The number of basis functions is defined by placing 15 equally-spaced knots in the log scale of the domain $[a, b]$ (different numbers of knots, from 10 to 25, were tested, and results were similar). Figure 1 shows the effect of basis representation for a subsample of our PNSD data (the x-axis is in log-scale for better visualization). Each colour represents a different hourly density $\widehat{y}_t$, the dots are the observed data $\mathbf{w}_t$ (discrete version), and the line describes the continuous representation $\widehat{y}_t$. Overall, the functional representation using B-splines functions on the Hilbert space, $\widehat{\text{clr}(y_t)}(p)$, captures the variety of the different density shapes in a seemingly accurate way. Here, we will analyse the sequence of continuous density functions $\{\widehat{y}_1, \ldots, \widehat{y}_T\}$. Henceforth, we will use the notation y_t instead of $\widehat{y}_t$.

Finally, we notice that $\{y_1, \ldots, y_T\}$ is a time-ordered functional data with time dependence. It is not difficult to see that PNSD at time t is correlated with PNSD at time $t + 1$. Figure 2 shows the autocorrelation plots for lags $h = 0$, 1, 24, and 168 (hours), corresponding to one hour, day, and week, respectively. Here, we observe two types of correlations: the correlation between different particle sizes and the time lag correlation. For time lags smaller than 24, there is a negative correlation between the smaller particle sizes (<30 nm) and the large particle sizes (>200 nm). For closer particle size counts, there is a strong positive correlation that seems to vanish after a few days.

3 Methodology

In this section, we assume that our data $\{y_t\}$ is a sequence of continuous density functions $y_t : [a, b] \to \mathbb{R}^+$, representing the normalized PNSD for all different particle sizes observed at times t (the continuous functions y_t are obtained using the procedure described in Section 2 and illustrated in Figure 1).

We aim to investigate the different profiles of the sources and the contributions of these sources to the total PNSD, $\{y_t\}$. There are two main aspects we need to take into account. First, we need to define a proper sample space for the density functions; otherwise, the analysis can lead to meaningless results (it is not a linear space). Second, we need to consider that the sequence of density functions is time-dependent, i.e. the data are a positive functional time series. In our modelling, we consider these features of the data. For the first aspect, we consider a proper space of density

[1] By construction, we have $\int \text{clr}(y_t)\,\mathrm{d}p = 0$. Then, the representation $\sum_{j=1}^{L} c_{j,t}\psi_j(p)$ should integrate zero as well. This induces a constraint on the coefficients $c_{j,t}$, and it is considered in the minimization process. For more details, see Talská et al. (2018) and Machalová et al. (2016).

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

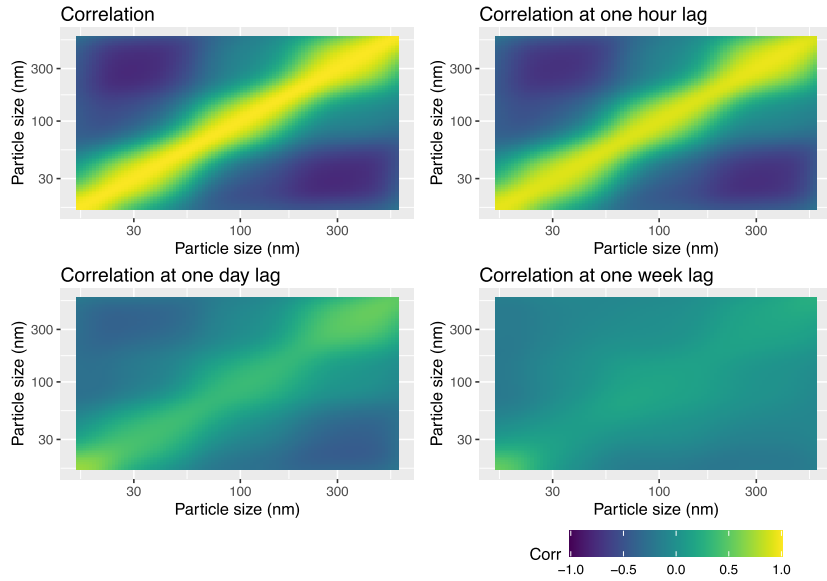

Figure 2. Correlation between functional densities at different time lags.

functions, denoted by $\mathcal{B}^2$ and described in Section 3.1, and for the second aspect, we define a positive dynamic functional factor model in this space. To perform inference, we use an isomorphism, clr, between $\mathcal{B}^2$ and the well-known L^2 space. The estimation procedure to obtain the profiles considers the time dependence of $\{y_t\}$.

3.1 Bayes space

As in the multivariate analysis, it is crucial to define a proper space where the intensity functions can be defined as values of a functional random variable. Let $\mathcal{B}^2[a, b]$ be the space of positive functions f with support $[a, b]$ such that $\int (\log f)^2(p)\,\mathrm{d}p < \infty$. The space $\mathcal{B}^2[a, b]$ contains redundant elements in the following sense. Let f and g be positive functions such that $f = cg$, where c is a positive number. Then, f and g provide the same relative information. That is, the normalized and the nonnormalized PNSD data represent the same proportion of particle numbers. For this reason, a space of equivalent classes of positive functions is preferred. Define the equivalent relation $f \equiv g$ if and only if there exists a positive constant c such that $f = cg$. Then, the Bayes space is defined as the equivalence positive functions in $[a, b]$. By practical convenience, $\mathcal{B}^2[a, b]$ will be the set of density functions, the positive functions that integrate one.

Now, define the sum and multiplication operation in $\mathcal{B}^2[a, b]$ as follows. For $f, g \in \mathcal{B}^2[a, b]$, and $\alpha \in \mathbb{R}$,

$$f \oplus g(t) = \frac{f(t)g(t)}{\int_a^b f(s)g(s)\,\mathrm{d}s}; \quad (\alpha \odot f)(t) = \frac{f^\alpha(t)}{\int_a^b f^\alpha(s)\,\mathrm{d}s}.$$

The operation $\oplus$ represents the addition of two elements in $\mathcal{B}^2[a, b]$, and it is called *perturbation*. The operation $\odot$ represents the multiplication by a real scalar α and it is called *powering*. With these two operations, $\mathcal{B}^2[a, b]$ is a vector space where the constant function $f(t) = \frac{1}{b-a}$ (uniform density on $[a, b]$) is the neutral element of perturbation.

Now, define an inner product $\langle f, g\rangle_{\mathcal{B}}$ in $\mathcal{B}^2[a, b]$ as

$$\langle f, g\rangle_{\mathcal{B}} = \frac{1}{b-a}\iint \{\log f(t) - \log f(s)\}\{\log g(t) - \log g(s)\}\,\mathrm{d}t\,\mathrm{d}s.$$

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

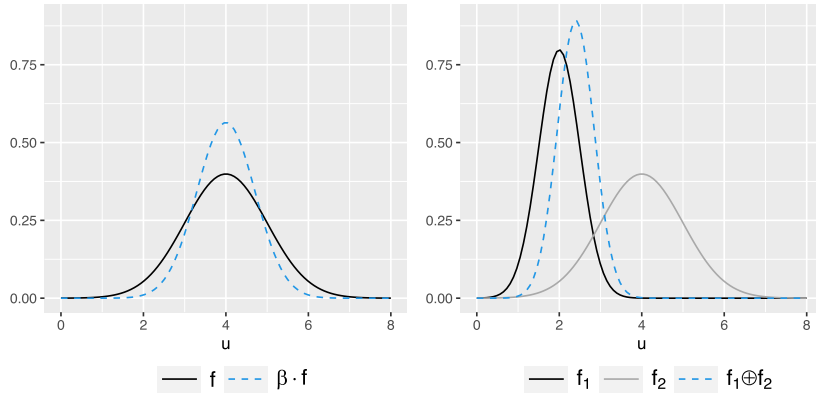

Figure 3. Bayes operations example with truncated Normal densities (μ, σ^2). Left: Powering (multiplication) operation with $\beta = 2$ and $f \sim (4, 1)$. Right: Perturbation (addition) operation with $f_1 \sim (2, 1/2)$ and $f_2 \sim (4, 1)$.

Let $\|\cdot\|_{\mathcal{B}}$ denote the induced norm and $d_{\mathcal{B}}$ denote the induced distance, i.e. $\|f\|_{\mathcal{B}} = \sqrt{\langle f, f\rangle_{\mathcal{B}}}$, and $d_{\mathcal{B}}(f, g) = \|f \ominus g\|_{\mathcal{B}}$, where $\ominus$ is defined as $(f \ominus g)(t) = f \oplus (-1 \odot g)(t)$. Then, we have that $(\mathcal{B}^2[a, b], \oplus, \odot, \langle\cdot,\cdot\rangle_{\mathcal{B}})$ forms a separable Hilbert space (see Talská et al., 2020).

We assume that our functional time series $\{y_t\}$ are elements of the space $\mathcal{B}^2[a, b]$. In the context of the PNSD data, y_t will be assumed to be the result of the perturbations of latent profiles of K sources. For example, if there are only two sources with latent profiles f_1 and f_2, then y_t is the result of $[\beta_1 \odot f_1] \oplus [\beta_2 \odot f_2]$, where β_1 and β_2 are real numbers indicating the contribution level of f_1 and f_2 to the total PNSD, respectively. To illustrate this composition, assume that the profiles f_1 and f_2 are truncated Normal density functions to the range $[0, 8]$ with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively.[2] First, let us consider the operation $\beta_1 \odot f_1$. If $\beta_1 > 0$, then this can be interpreted as a precision parameter of f_1 since $\beta_1 \odot f_1(t) \equiv f_1^{\beta}(t) \propto \exp\{-\frac{1}{2\sigma_1^2/\beta}(t - \mu_1)^2\}$. This is, $\beta_1 \odot f_1(t)$ is a truncated Normal density function with mean μ_1 and variance σ_1/β_1. Thus, for a fixed σ_1, β_1 indicates the precision contribution of source one with profile f_1. If $\beta_1 \leq 0$, this is interpreted as a null contribution of the corresponding source. Specifically, if $\beta_1 = 0, \beta_1 \odot f_1$ is the uniform distribution. If $\beta_1 < 0$, then instead of concentrating the mode of f_1, it replaces it with its reciprocal, bringing the mode of f_1 to zero and small values of f_1 become large numbers, i.e. $\beta_1 \odot f_1(t)$ will be a 'U' wide shape. Now, let us focus on $[\beta_1 \odot f_1] \oplus [\beta_2 \odot f_2]$. Without loss of generality, assume that $\beta_1 = \beta_2 = 1$, otherwise, we can always rewrite σ_1 and σ_2. We have that $f_1 \oplus f_2$ is a truncated Normal density function with mean parameter $(\mu_1\sigma_2 + \mu_2\sigma_1)/(\sigma_1^2 + \sigma_2^2)$ and variance parameter $\sigma_1^2\sigma_2^2/(\sigma_1^2 + \sigma_2^2)$. This means that the PNSD observed from the combinations of two sources will be a truncated Normal density function with the highest peak at $(\mu_1\sigma_2 + \mu_2\sigma_1)/(\sigma_1^2 + \sigma_2^2)$, which is an average of the two means weighted by the variances. Figure 3 illustrates different scenarios of two source contributions with profiles f_1 and f_2. The left plot shows f_1 and $\beta_1 \odot f_1$, with $\beta_1 = 2$. In this case, we see that $\beta_1 \odot f_1$ is sharper than f_1, meaning that the particle sizes close to the mean have a strong contribution to the total PNSD. While the right plot shows f_1, f_2, and $f_1 \oplus f_2$. In this case, we see that $f_1 \oplus f_2$ is a new density in between f_1 and f_2 but closer to f_1 since f_1 presents less uncertainty than f_2.

Given the above motivation, we formally define the functional factor model for density functions.

3.2 Functional dynamic factor model for density functions

In this section, we describe the dynamic functional factor model in the space $\mathcal{B}^2[a, b]$. Let K denote the total number of sources, and let f_k, $k = 1, \ldots, K$, be the latent densities representing the kth source profiles density. Let $\boldsymbol{\beta}_t = (\beta_{1,t}, \ldots, \beta_{K,t})^{\top}$ be a multivariate time series, with $\beta_{k,t}$ representing

[2] $f_i(t) \equiv \phi(t; \mu_i, \sigma_i^2)\mathbb{1}_{[a,b]}(t)$, where $\phi(t; \mu, \sigma^2)$ denotes the density function of a Normal distribution with mean μ and variance σ^2, and f_i integrates one.

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

the contribution of the kth source at time t. Let $\{\epsilon_t\}$ be a sequence of zero-mean functional error in $\mathcal{B}^2[a, b]$. We assume that the sequence of density functions, $\{y_1, \ldots, y_T\}$, follows the model:

$$y_t(p) = \bigoplus_{k=1}^{K} \left[\beta_{k,t} \odot f_k\right](p) \oplus \epsilon_t(p), \quad p \in [a, b]. \tag{1}$$

We refer to model (1) as a positive functional factor (PFF) model.

As in the usual functional factor model definition, we assume that the set of functions $\{f_k\}$ are orthogonal:

$$\langle f_i, f_j \rangle_{\mathcal{B}} = \mathbb{1}(i = j), \tag{2}$$

where $\mathbb{1}(i = j)$ takes value 1 if $i = j$ and zero otherwise. Thus, the data y_t are driven by K latent components. The sequence of functional white noise $\{\epsilon_t\}$ is such that $\mathbb{E}(\varepsilon_t) = 0_{\mathcal{B}}$, i.e. the mean $\mathbb{E}(\varepsilon_t)$ is the uniform distribution on $[a, b]$ and $\{\epsilon_t\}$ represents the rest of the variability not accounted by the profiles.

It is important to mention that, if $\{\boldsymbol{\beta}_t\}$ is time dependent, then $\{y_t\}$ is time dependent. This property will be easier to see after introducing the isomorphism clr in Section 4. In our context, the density functions are time dependent, so we will assume that $\{\boldsymbol{\beta}_t\}$ is time dependent.

Given the model (1), our next step is to perform inference (the learning process), and we will discuss this in the next section.

4 Inference

Here, we describe how the latent components of the PFF model (1) are estimated. For this, we first establish a connection between $\mathcal{B}^2[a, b]$ and the $L^2[a, b]$ space, let us define a functional as

$$\begin{aligned} \operatorname{clr} : \mathcal{B}^2[a, b] &\to L^2[a, b] \\ f &\mapsto \operatorname{clr}(f), \end{aligned}$$

where $\operatorname{clr}(f)(t) := \log f(t) - \frac{1}{b-a}\int_a^b \log f(s)\, \mathrm{d}s$. This clr functional is an isometric isomorphism, and we have that

$$\operatorname{clr}(f \oplus g) = f + g, \quad \operatorname{clr}(\alpha \odot f) = c \cdot f, \quad \langle f, g \rangle_{\mathcal{B}} = \langle \operatorname{clr}(f), \operatorname{clr}(g) \rangle_{L^2},$$

where $\langle \cdot, \cdot \rangle_{L^2}$ is the usual inner product in $L^2[a, b]$. The transformation clr plays a key role in our modelling providing a tool to go back and forth between spaces without losing information. Particularly, it allows us to perform inferences in $\mathcal{B}^2[a, b]$ by using the usual methods in $L^2[a, b]$. That is, we will perform the inference in the $L^2[a, b]$ space and then transform the results back to the $\mathcal{B}^2[a, b]$ space.

Notice that if we apply the functional clr on both sides of equation (1), we obtain the corresponding model in L^2 space as

$$\operatorname{clr}(y_t)(p) = \sum_{k=1}^{K} \beta_{k,t} \operatorname{clr}(f_k)(p) + \operatorname{clr}(\epsilon_t)(p), \quad p \in [a, b], \tag{3}$$

which is the usual functional factor model [see model (4)]. However, one should notice that since y_t is a density function, i.e. $\int y_t(p)\, \mathrm{d}p = 1$, the transformed function $\operatorname{clr}(y_t)$ has a zero integral constraint; $\int \operatorname{clr}(y_t)(p)\, \mathrm{d}p = 0$. This constraint is considered when estimating the continuous function in terms of the basis function (see step 3 in Section 2).

Let $X_t = \operatorname{clr}(y_t)$, $F_t = \operatorname{clr}(f_t)$, and $\varepsilon_t = \operatorname{clr}(\epsilon_t)$. Then, $\{X_t\}$ is a real-valued functional time series that follows the model

$$X_t(p) = \sum_{k=1}^{K} \beta_{t,k} F_k(p) + \varepsilon_t(p), \quad p \in [a, b] \tag{4}$$

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

where F_k are the factor curves, $\{\beta_{t,k}\}_{t\geq 1}$ is the scalar factor time series, and $\varepsilon_t(p)$ is functional white noise. Notice that $\{\beta_{t,k}\}_{t\geq 1}$ is the same time series in both spaces.

We use the model (4) to estimate the profiles. Since $\{X_t\}$ is time-dependent, one should carefully estimate the latent components. In Martínez-Hernández et al. (2022), a nonparametric estimation method for this model was proposed, assuming the functional data $\{X_t\}$ is time-dependent. The factor curves are modelled with B-spline basis functions. Specifically, F_k are assumed to have the form $F_k(p) = \sum_{i=1}^{L} a_{k,i}\psi_i(p)$. Then, F_k are estimated by taking into account the time dependence of the functional data $\{X_t\}$. To that end, F_k is estimated by considering the linear space generated by all factor curves $\mathcal{H}_F := \text{span}\{F_1, \ldots, F_K\}$. The space $\mathcal{H}_F$ is uniquely defined. Then, the estimator $\{\widehat{F}_1, \ldots, \widehat{F}_K\}$ are defined as a basis function for the linear space $\mathcal{H}_F$, which turns out to be the eigenfunctions of the long-run covariance operator of the functional time series $\{X_t\}$. Here, we follow this idea, and we only describe the overall idea of the estimator; the details can be found in Martínez-Hernández et al. (2022).

Let $c_h(u, v) = \text{Cov}(X_0(u), X_h(v))$ be the auto-covariance function at lag h of a functional data X_t. Using c_h as a kernel, the covariance operator Γ_h at lag h of $\{X_t\}$ is defined as

$$\Gamma_h(z)(u) = \int c_h(u, v)z(v)\,dv, \quad z \in L^2.$$

Under the model (4), we have that, the covariance operator at lag h of X_t is

$$\Gamma_h = \sum_{i,j=1}^{K} \text{Cov}(\beta_{i,t}, \beta_{j,t+h})F_i \otimes F_j + \mathbb{1}(h = 0)\Gamma_\varepsilon, \quad h \in \mathbf{Z}, \tag{5}$$

where $\otimes$ denotes the tensor product, and Γ_ε is the corresponding covariance operator of ε_t. Then, it can be shown that the operator Λ, defined as

$$\Lambda := \sum_{h\in\mathbb{Z}, h\neq 0} \Gamma_h$$

characterizes the linear space $\mathcal{H}_F$. In fact, the first K eigenfunctions of Λ corresponding to nonzero eigenvalues can be defined as the basis functions of $\mathcal{H}_F$. Notice that Λ contains all past dependence information.

Let $\widehat{\Lambda}$ be a consistent estimator of Λ, and let $\zeta_1, \ldots, \zeta_K$ be the first K eigenfunctions of $\widehat{\Lambda}$ with nonzero eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_K > 0$. That is, for each $k \in \{1, \ldots, K\}$, $\widehat{\Lambda}(\zeta_k) = \lambda_k\zeta_k$. Now, we define $\widehat{F}_k = \zeta_k$. Given the estimator $\widehat{F}_k$ in $L^2[a, b]$, we recover the estimator of kth source profiles density as $\widehat{f}_k = \exp\{\widehat{F}_k\}$.

Regarding the estimator of the trajectory of $\{\boldsymbol{\beta}_t\}$, we notice that the time series $\{\boldsymbol{\beta}_t\}$ is the same in both spaces $L^2[a, b]$ and $\mathcal{B}^2[a, b]$. Also, under a mild condition on the cross-covariance of $\text{clr}(\varepsilon_t)$ and the latent components, and using assumption (2), we have that $\beta_{k,t}$ is recovered by projecting the data in $L^2[a, b]$ onto $\widehat{F}_k$. Therefore, the estimated trajectories are obtained in the $L^2[a, b]$ space as follows

$$\{\widehat{\beta}_{k,t}\}_{t=1}^{T} = \{\langle \text{clr}(\widehat{f}_k), X_t\rangle\}_{t=1}^{T}, \quad k = 1, \ldots, K.$$

Finally, we describe how to quantify the uncertainty of the estimated profiles $\widehat{f}_k$. Since our method is nonparametric, we use a bootstrap method by resampling the model residuals in the L^2 space. Specifically, let $\widehat{\varepsilon}_t := X_t - \widehat{X}_t$, where $\widehat{X}_t = \sum_{k=1}^{K} \widehat{\beta}_{k,t}\widehat{F}_k$. Let $\widehat{\boldsymbol{\varepsilon}} = \{\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_T\}$ be the set of the estimated residuals. Then, to obtain a simulation of a functional time series, we first obtain T_{boot} residuals $\{\varepsilon_1^{\text{boot}}, \ldots, \varepsilon_{T_{\text{boot}}}^{\text{boot}}\}$ by sampling with repetition from $\widehat{\boldsymbol{\varepsilon}}$, and then obtain the simulated data (bootstrap sample) as

$$X_t^{\text{boot}} = \sum_{k=1}^{K} \widehat{\beta}_{k,t}\widehat{F}_k + \varepsilon_t^{\text{boot}},$$

with sample size T_{boot}. This process is repeated ℓ times and thus obtaining ℓ functional time series; $\{X_t^{\text{boot},1}\}_t, \ldots, \{X_t^{\text{boot},\ell}\}_t$. For each bootstrap sample, $\{X_t^{\text{boot},i}\}_t$, we fit the model (4) to obtain

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

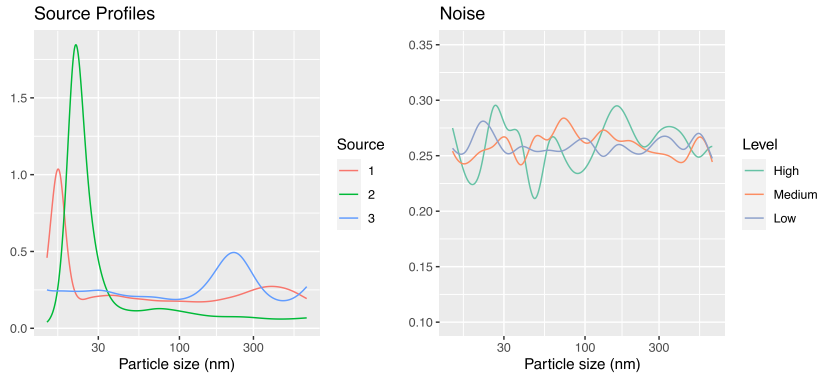

Figure 4. Simulation study setting. Left: the three profiles used in Model (1). Right: three examples of functional noise ϵ_t used in Model 1 to simulate data.

$\{\widehat{F}_k^{*,i},\ k = 1, \ldots, K\}$. Finally, we transform these profiles back to obtain a bootstrap sample of the source profiles, $\{\widehat{f}_k^{*,i},\ k = 1, \ldots, K\}$. Thus, to obtain and visualize the uncertainty of an estimated profile $\widehat{f}_k$, we use the collection of densities from the bootstrap output, $\{\widehat{f}_k^{*,1}, \ldots, \widehat{f}_k^{*,I}\}$. Specifically, the confident region for $\widehat{f}_k$ is $\widehat{f}_k(p) \pm 2\widehat{\sigma}(p)$, where, for each $p \in [a, b]$, $\widehat{\sigma}(p)$ is the standard deviation of the values $\{\widehat{f}_k^{*,1}(p), \ldots, \widehat{f}_k^{*,I}(p)\}$ (see Figure C1).

5 Simulation study

In this section, we investigate the performance of our proposed methodology to recover the truth sources in the PFF model (1). We consider three ($K = 3$) sources that mimic the shape of PNSD sources found in previous studies (Baerenbold et al., 2023). Figure 4 (left) shows these profiles (they integrate one): sources one and two are mainly concentrated in the smallest particle sizes, while source three represents concentrations of larger particle sizes. We add a noise term $\epsilon_t(p)$, specified through uniform densities in $[a, b]$ (see Talská et al., 2018). In other words, for a given t, we draw a sample of M uniform data points in $[a, b]$, then compute the histogram counts, where the breakpoints are set equal to the values of the different particle sizes. Finally, we use a continuous basis fit to this histogram to obtain a sample of the error term. Figure 4 (right) shows three simulated noise curves using this procedure for different M values. We refer to these cases as noise levels: High ($M = 2{,}000$), Medium ($M = 5{,}000$), and Low ($M = 10{,}000$).

We consider the following setting for the time series contributions, $\{\beta_{k,t}\}$, for the three sources:

$$\beta_{k,t} = \mu_k + w_{k,t},$$

where $w_{k,t}$ is an AR(1) process with autoregressive parameter $\phi_k = 0.8$ and $\sigma_k^2 = 0.05$ for $k = 1, 2, 3$. The mean level was set as $\mu_1 = 1$, $\mu_2 = 1.5$, and $\mu_3 = 2$. Notice that we select a small value for σ_k^2 to control the noise unrelated to the model noise component ($\epsilon_t(p)$).

We generate a sample of normalized PNSD functions for each simulation setting, $y_t(p)$, using the PFF model for $t = 1, \ldots, T$. Then, we applied our inference methodology described in Section 4 to obtain the estimated profiles, $\{\widehat{f}_1, \widehat{f}_2, \widehat{f}_3\}$. To measure how accurately we can estimate the sources, we compute the total variation distance (TVD) between the three sources, $d_{TV}(f_k, \widehat{f}_k) = 1 - \int_a^b \min\{f_k(p), \widehat{f}_k(p)\}\mathrm{d}p$ for $k = 1, 2, 3$. The TVD between two densities has a graphical interpretation as $1 - \delta$, where δ represents the common area below the two densities (Euán et al., 2018). The TVD is bounded, $0 \leq d_{TV} \leq 1$, and the closer to zero, the better. We also compute the integrated mean squared error (ISE) between $\mathrm{clr}(\widehat{f}_t)$ and $\mathrm{clr}(f_t)$, as $\mathrm{ISE}(\mathrm{clr}(f_k), \mathrm{clr}(\widehat{f}_k)) = \int_a^b (\mathrm{clr}(f_k(p)) - \mathrm{clr}(\widehat{f}_k(p)))^2\mathrm{d}p$. The ISE will correspond to the squared distance in the Bayes space.

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

Table 1. Simulation results

		TVD			ISE		
T	Noise	Source 1	Source 2	Source 3	Source 1	Source 2	Source 3
240	High	0.134 (0.046)	0.284 (0.005)	0.129 (0.014)	0.5 (0.276)	0.913 (0.101)	0.867 (0.081)
240	Medium	0.134 (0.046)	0.284 (0.006)	0.13 (0.014)	0.495 (0.278)	0.897 (0.092)	0.88 (0.073)
240	Low	0.132 (0.047)	0.283 (0.007)	0.13 (0.015)	0.482 (0.288)	0.888 (0.099)	0.878 (0.082)
720	High	0.129 (0.045)	0.284 (0.006)	0.129 (0.012)	0.473 (0.261)	0.895 (0.103)	0.877 (0.062)
720	Medium	0.137 (0.049)	0.284 (0.006)	0.13 (0.01)	0.516 (0.287)	0.902 (0.105)	0.883 (0.053)
720	Low	0.131 (0.045)	0.284 (0.005)	0.129 (0.012)	0.477 (0.265)	0.891 (0.09)	0.884 (0.056)

Note. The mean value of the total variation distance (TVD) and the integrated mean squared error (ISE) across 100 replicates is reported. Observed standard deviations are indicated between brackets. The recovery of the sources is reasonably good.

Table 1 reports the simulation results, mean and standard deviation values computed across 100 replicates. Overall, there is evidence of robustness to the noise level. In recovering the actual sources, we can observe from the TVD values that source 2 was recovered less accurately. We can also observe from the ISE that source 1 was the best-identified source. We looked into the estimated profiles for each source for the medium noise level (see Appendix B), and we confirm that the shape is overall well recovered, but sources with lower dispersion in the Bayes space seem harder to identify. However, looking at the ISE and TVD for all sources jointly, the recovery is reasonably good using the proposed methodology.

6 Data analysis

In this section, we analyse the PNSD data obtained from the urban background site in London North Kensington introduced in Section 2. Overall, we identified six sources whose profiles are presented in Figure 5. Before going into details of our findings, we first describe how we select the different parameters involved in the inference procedure, such as the number of sources, K. Then, we present the results.

6.1 Model specifications

Using the sequence $\{y_t\}$, we fit the PFF model (1), with $[a, b] = [16.55, 604.4]$, which are the smallest and largest particle sizes in the data. An important component of this model is the estimation of Λ, the long-run covariance of $\{\text{clr}(y_t)\}_t$. For this, we use a smooth periodogram-type estimator with lag parameter $h = 120$ (5 days) (see Rice & Shang, 2017). Notice that this estimation of Λ is independent of the number of factors K. To select the number of factors (sources), we adopt a ratio-based approach. The ratio-based approach is as follows. Assuming model (4), then K is the number of nonzero eigenvalues of Λ. However, the nonzero eigenvalues of $\widehat{\Lambda}$ are not exactly K (due to the uncertainty of the estimation and the zero-eigenvalues of $\widehat{\Lambda}$ are unlikely to be zero exactly). Thus, the ratio-based estimator is defined as $\widehat{K} = \arg\min_{1\le k\le k_u} \widehat{\lambda}_{k+1}/\widehat{\lambda}_k$, where $\widehat{\lambda}_k$ are the eigenvalues of $\widehat{\Lambda}$ and k_u is a number that is large enough such that $K < k_u$. For our data application, we observed that the first eigenvalue was much bigger than the rest. Therefore, we adapted and selected the second minimum of the ratio-based criteria. We used $k_u = 10$ as the upper bound of the maximum factors.

Given the number of factors $\widehat{K}$, we then only consider the first $\widehat{K}$ eigenfunctions of $\widehat{\Lambda}$: $\widehat{\zeta}_1, \ldots, \widehat{\zeta}_{\widehat{K}}$. Then, we transform back these $\widehat{K}$ functions to the Bayes space with clr^{-1} to obtain $\widehat{\xi}_1, \ldots, \widehat{\xi}_{\widehat{K}}$, which are the candidates for the profiles estimators. Other possible candidates are $-1 \odot \widehat{\xi}_k$, $k = 1, \ldots, \widehat{K}$, this is because the solution of the eigenfunctions is up to sign. We choose signs following the next criteria: (1) the estimated profile should be a quasiconcave function, and (2) the corresponding coefficients $\{\beta_{k,t}\}$ are such that $\bar{\beta}_{k,\cdot} > 0$, for $k = 1, \ldots, K$. These criteria are based

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

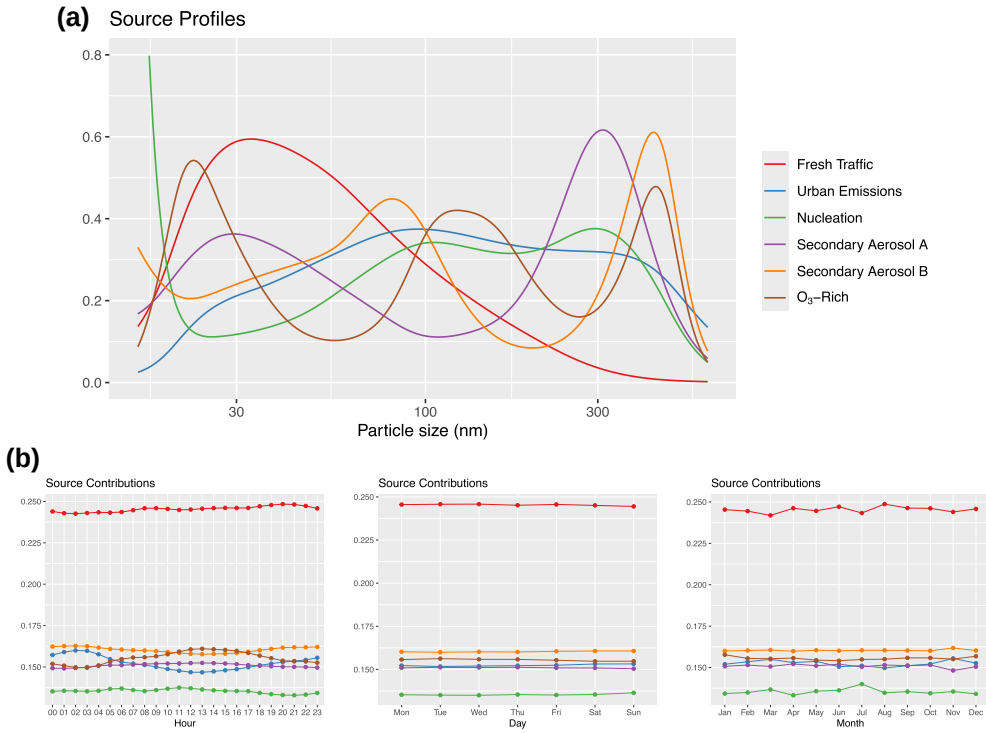

Figure 5. (a) Estimated sources profiles. (b) Average contributions per hour, day and month. Sources ID have been relabelled according to the correlation with the main pollutants, shape of profiles and wind plots.

Table 2. Correlations between estimated sources (contribution series $\tilde{\beta}_{k,t}$) and air pollutants and temperature (hourly time series) from 1 January 2011 to 31 October 2018

	Source 1	Source 2	Source 3	Source 4	Source 5	Source 6
CO	−0.24	0.32	0.13	−0.35	0.27	−0.14
NO_x	−0.17	0.28	0.07	−0.31	0.24	−0.13
NO_2	−0.13	0.30	0.01	−0.33	0.31	−0.17
NO	−0.17	0.23	0.09	−0.26	0.18	−0.09
O_3	0.04	−0.19	0.04	0.17	−0.26	0.19
SO_2	−0.12	0.14	0.10	−0.18	0.13	−0.06
PM_{10}	−0.45	0.48	0.24	−0.39	0.33	−0.19
$PM_{2.5}$	−0.48	0.50	0.28	−0.41	0.34	−0.19
Air temperature	0.01	−0.16	0.10	0.06	−0.14	0.12

on interpretability reasons, e.g. having a 'U' shape of a profile is not expected as a potential source profile. Once the sign is chosen, we adjust the profiles as follows. We rescale $\beta_{k,t}$ such that $\mathbb{E}(\beta_{k,t})$ is one. The corresponding rescaling value is added to the profile. After these steps, we obtain the final estimators: $\widehat{f}_1, \ldots, \widehat{f}_K$, where $\widehat{f}_k$ is either $c_k \odot \widehat{\xi}_k$ or $-c_k \odot \widehat{\xi}_k$, with c_k being the rescaled value.

The interpretation of $\{\widehat{\beta}_{k,t}\}$ might be challenging since the estimation is performed in the Bayes space. In fact, $\widehat{\beta}_{k,t} > 1$ means a strong contribution of the peak particle size, $0 < \widehat{\beta}_{k,t} < 1$ means a weak contribution of $\widehat{f}_k$ (increase in dispersion), and $\widehat{\beta}_{k,t} \leq 0$ means null contribution. The latter can be confusing to environmental scientists because this differs from commonly used approaches such as PMF, where all contributions are positive. To facilitate interpretation, we adopt a

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

projection approach. Specifically, for each source k, we project the data y_t on the profile $\widehat{f}_k$, that is, $\tilde{\beta}_{k,t} := \langle y_t, \widehat{f}_k \rangle$. Then, the time series $\{\tilde{\beta}_{k,t}\}_t$ is defined as the contribution of the source k. This definition is intuitive since this represents the PNSD in the subspace generated by f_k and then the contribution of f_k on the total PNSD.

Finally, to help with the source interpretation, we compute correlations $\text{Cor}(\tilde{\beta}_{k,t}, z_t)$, $k = 1, \ldots, \widehat{K}$, between the source contributions and (i) air temperature; (ii) the following pollutants: Carbon monoxide (CO), Nitrogen oxide (NO_x), Nitrogen dioxide (NO_2), Nitric oxide (NO), Ozone (O_3), Sulphur dioxide (SO_2), and PM concentrations (PM_{10} and $PM_{2.5}$) (see Table 2). Data on air pollutants and temperature were obtained using the R package `openair` (Carslaw & Ropkins, 2012).

Finally, once the model is defined, we obtain the uncertainty of the estimated profiles using the bootstrap method described in Section 4. For this, we set $l = 100$ bootstrap replicates and bootstrap series of $T_{\text{boot}} = 10{,}000$ length. The plots for the confidence bootstrap regions are shown in Figure C1.

6.2 Results

Here, we present the results for the data described in Section 2, which covers the period 2011–2018. The number of sources was selected following the criteria selection described in Section 6.1. We identified $\widehat{K} = 6$ (see Figure A2 to visualize the ratio values). Figure 5 shows the profiles and their contributions. PM sources are also associated with meteorological conditions, particularly with wind. Hence, to help with the interpretation, we downloaded hourly wind speed (m/s) and direction (angle) at the London Heathrow site from the NOAA Integrated Surface Database accessed through the R package `worldmet` (Carslaw, 2023). We then produced polar plots (Figure 6), showing how the contribution of each source varies by wind direction (i.e. the direction the wind is coming from) and wind speed (represented by the radial axis such that zero wind speed is at the centre of the plot and increases away from the centre to 20 ms^{-1}). The colour scale represents the mean concentration that corresponds to a specific wind speed and wind direction interval (Carslaw & Ropkins, 2012; Uria-Tellaetxe & Carslaw, 2014).

We can see that these six sources are well defined and cover the entire domain of bin sizes (Figure 5). Comparing these profiles with Rivas et al. (2020), who ran the commonly used PMF approach for PNSD collected at the same monitoring station, we identified two additional sources, showing the capability of our proposed method to disentangle the complex pollution mix in an urban setting. Below, we present the six-source solution yielded by our analysis.

The first source contributions increase slightly in late spring-summer (see Figure 5b, red line), with a peak around 35 nm (Figure 5a). It is the sources showing the highest contribution on average per hour with relative minimums during early morning and early afternoon and relative maximums in the morning and at night. Moreover, the average daily contribution decreases

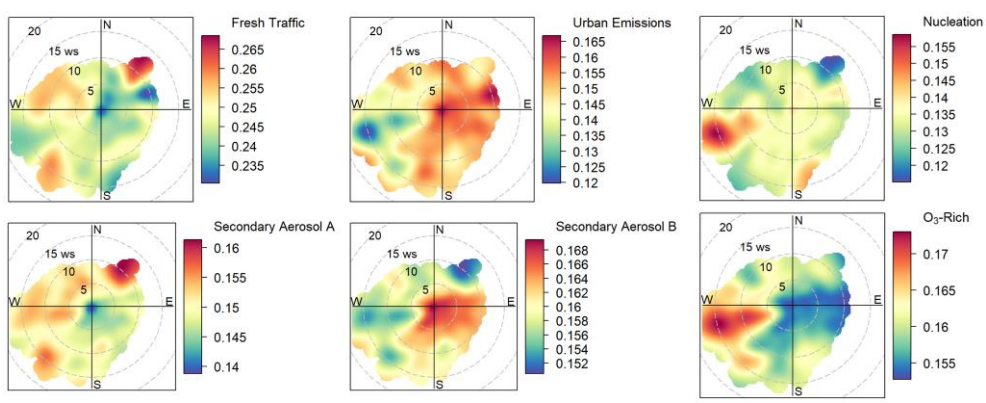

Figure 6. Polar plots showing how the hourly PNSD contributions to the sources are affected by the hourly wind direction and wind velocity during the years 2011–2018.

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

monotonically from Friday to Sunday. It does not show correlations with any pollutant but a moderate negative correlation with PM_{10} and $PM_{2.5}$. The polar plots displayed in Figure 6 show that this source is characterized by strong wind speeds from the East, which is reasonable since North Kensington station is located West of the city centre of London, where traffic is expected to be most dense. Even more, on the south-west, there is the Westway Roundabout, which contributes to high concentrations due to the braking and accelerating manoeuvres of cars. Similarly to Rivas et al. (2020) and Beddows et al. (2015), we labelled this source as **fresh traffic**.

The second source is highly correlated with all gases and PM, except O_3 (see Table 2). Particularly, we distinguish the highest correlation values for PM, where the generation of these particles is associated with industrial combustion, production process, road transport, and waste processing plants, among others (Holman, 1999). From Figure 5b, we can see that this source (blue line) has contribution values slightly larger at the beginning of the day and a relative minimum during the noon hours. In addition, from Figure 5a we could notice a peak at 99 nm, while from Figure 6, we can appreciate a stronger association with easterly winds, which are typically affected by European mainland (Beddows et al., 2015). In this context, this source could represent **urban emissions**, which is consistent with Rivas et al. (2020). Moreover, we can easily distinguish lower contribution values for the summer season.

The third source has a moderate correlation with CO, PM, and air temperature (see Table 2). The contribution values indicate that there is no visible variation throughout the day (Figure 5b, green line), except during midday, where solar radiation is maximum, along with a maximum peak in July, which is consistent with the month of the greatest contribution of solar radiation. The profile presents a peak for the smallest bin (16 nm), so it could be labelled as **nucleation**. Consistent with earlier studies (Beddows et al., 2015; Harrison et al., 2019; Rivas et al., 2020), we found that this source is associated primarily with winds from the west and the west-south-westerly sector (Figure 6).

The fourth source can be labelled as **secondary aerosol A**, which is typically characterized by a bimodal shape, with a first peak at small bin sizes, around 30 nm, and a second peak at very large bin sizes, around 292 nm, in accordance with Beddows et al. (2015) [see Figure 5a, purple line]. Also, we can distinguish some large contribution values for the middle of the day (Figure 5b, purple line) and predominant easterly wind (Figure 6). In addition, we have a positive correlation with O_3 (Table 2).

The fifth source is correlated with all gases and PM (except O_3). The corresponding profile is bimodal (Figure 5a, orange line), with a first peak around 80 nm and a second peak around 360 nm. In Figure 5b (orange line), we can see that its maximum contributions are at early and late hours, with a minimum around mid-day. In Figure 6, we can notice a predominant association with south-easterly winds. Using the shape profile and the contributions behaviour, we can say that this source corresponds to **secondary aerosol B** pollution. This is in accordance with Tremper et al. (2022).

Finally, for source 6, we highlight the correlation with O_3. The corresponding profile presents a multimodal shape, with peaks around 15, 80, and 340. Also, we can see that contribution values are higher around 7 a.m. and 2 p.m., with a dominance of westerly winds (Figure 6). Based on these observations, we label this source as O_3-Rich. The exact nature of this source is unclear, as pointed out in Hopke et al. (2022). The correlation with O_3 and the higher contribution in the early to mid-afternoon suggests that this source represents particles that have grown through the condensation of secondary material, including both organic and inorganic species.

7 Discussion

We proposed a novel functional data analysis approach to SA for PNSD data. Our proposal is a functional factor model in the Bayes space, i.e. a functional factor model for density functions. We evaluated the performance of our method on simulated data and on a time series dataset from a monitoring site in London. Our findings suggest that our method provides a reliable tool to estimate and recover the different sources of the PNSD data. As described in Section 6, we found six different sources, which are consistent with previous studies: fresh traffic, urban emissions, nucleation, secondary aerosol A, secondary aerosol B, and O_3-Rich. In addition,

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

we provided strategies to select the different tuning parameters and a bootstrap method to quantify uncertainty in the estimated source profiles. Moreover, our approach gains computational advantages by using basis representations for the continuous functions, and it can handle high-dimensional and high-frequency datasets. Considering hourly data for 7 years and 51 particle size bins (dimensions), after fitting the basis functions, our method takes around 5 s to run on a standard laptop.

We have assumed that the PNSD data are a sequence of continuous functions expressed in terms of L basis functions. However, in practice, the observed data is a discrete version. Thus, the user needs to select the number of basis functions L. One approach is to set L 'large enough' and tune the smoothing parameter, i.e. instead of selecting L, one selects the smoothing parameter. This is the approach we have followed here. We used a common λ parameter that might compromise the fit for extreme particle size concentrations; however, this did not affect the interpretation of the identified sources. This might not apply to a different application context and different criteria might be used, such as using a specific λ for each of the curves.

One limitation of our proposal is the absence of uncertainty quantification in the number of sources. Although we have described criteria for the selection of sources, this is limited. Another limitation is related to the covariates incorporation. One could incorporate relevant information to disentangle the PNSD data, such as wind speed, time series of different pollutants, and meteorology data. These ideas will be investigated in future projects. Our future project also includes a more 'realistic' model that incorporates a dynamic number of sources. It is natural to think that the number and type of sources can change over time. The latter can be attributed to many reasons: (1) changes in human behaviour, (2) technological advances (such as electric vehicles), and (3) external factors (e.g. specific meteorological events such as Sahara dust, changes in urban environments due to new constructions, control strategies of air pollution, etc.). So, apart from having a method that can handle large and high-dimensional datasets, it would be desirable to have a method with a dynamic number of sources, i.e. K depending on time. This dynamical approach could also incorporate the information from the norms of the original PNSD data, which currently are not considered in the model.

We conclude that our method provides a new approach to apportion PNSD data. Compared to existing methods, our method takes into account the time dependence and can handle large datasets. Although we can directly apply our methodology to long-period hourly data, we recommend performing a sensitivity analysis to decide whether or not the number of sources remains the same during the whole period.

To increase the impact of this research, we developed an R package `SAfda` that can be downloaded from Github https://github.com/I-MH/SAfda. This requires the R package `robCompositions` (Templ et al., 2011).

Acknowledgments

The authors thank Anja Tremper of the Environmental Research Group of Imperial College London for sharing the data and for the useful feedback. Also, the authors thank Renata Talska, Jitka Machalova and Karel Hron of Palacky University Olomouc, and Alessandra Menafoglio of Politecnico di Milano for sharing R codes for the continuous representation of y_t on the Bayes space. Finally, the authors are grateful to the two referees for their valuable comments that helped improve the article.

Conflicts of interest: None declared.

Funding

Open access to this publication was funded through Lancaster University. M.B. and M.P. acknowledge support from the Medical Research Council Grant MR/T044713/1. Their work is also partly supported by the MRC Centre for Environment and Health, which is funded by the Medical Research Council (MR/S019669/1, 2019–2024).

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

Data availability

The data used in this paper can be downloaded here https://uk-air.defra.gov.uk/networks/site-info?site_id=KC1. Alternatively, the data can be obtained from the authors upon request.

Appendix A. Smoothing parameter and number of sources

The value λ was selected using generalized cross-validation (Ramsay & Silverman, 2005). Then, we select the value of λ that minimizes the average GCV among all curves. We perform this analysis in a subset of 5,000 particle number size distribution curves only to reduce computational cost (no significant difference was found with more curves). Figure A1 shows the results. The optimal value is $\lambda = 0.997$.

The value of $\widehat{K}$ was selected according to the ratio criteria described in Section 6.1. Figure A2 shows the corresponding plots for each data period.

Appendix B. Estimated sources in simulation study

To illustrate the estimation accuracy of the true sources in the simulation study (Section 5), we select the noise level medium ($M = 5{,}000$) and plot the functional boxplots (Sun & Genton, 2011) of the estimated sources over the 100 replicates (Figures B1 and B2). The estimation for Source 2 (with lower dispersion) was able to recover the location of the mode but over estimates the dispersion of the density.

Appendix C. Sources with CI

Using the bootstrap procedure described in Section 4, we build the confidence bootstrap regions (Figure C1). We use $\ell = 100$ bootstrap replicates and bootstrap series of $T_{\text{boot}} = 10{,}000$ length. Overall, the first two Sources show less uncertainty, while sources 3 and 4 tend to increase uncertainty when selecting only four sources.

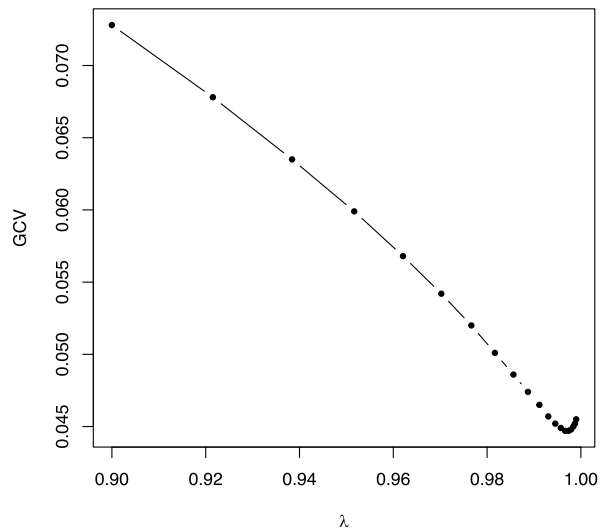

Figure A1. Profile cost function for λ selection using a subsample of 5,000 PNSD curves.

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

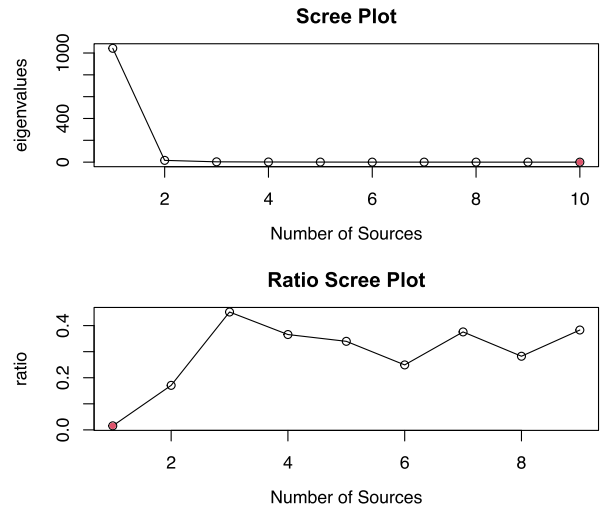

Figure A2. Selecting the number of sources based on eigenvalue ratio criteria.

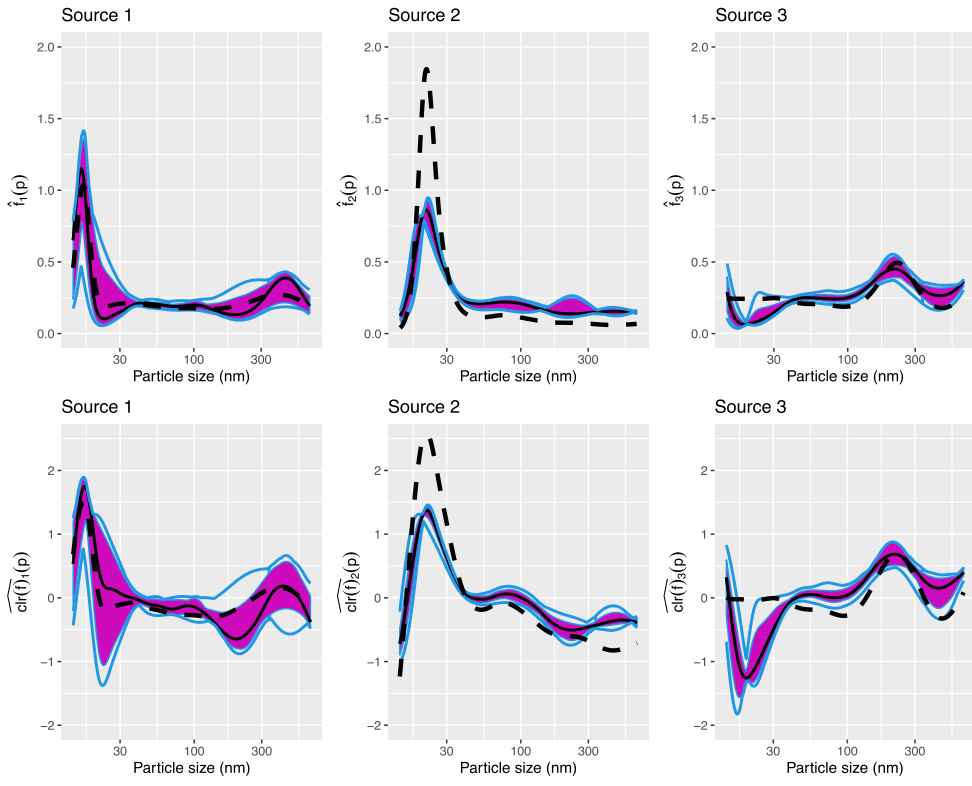

Figure B1. Functional boxplots for estimated sources in Bayes space (Top) and clr densities (Bottom) from Section 5 with $M = 5{,}000$ and $T = 240$. The continuous black line represents the median curve, and the dashed black line is the true source.

Downloaded from https://academic.oup.com/jrssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

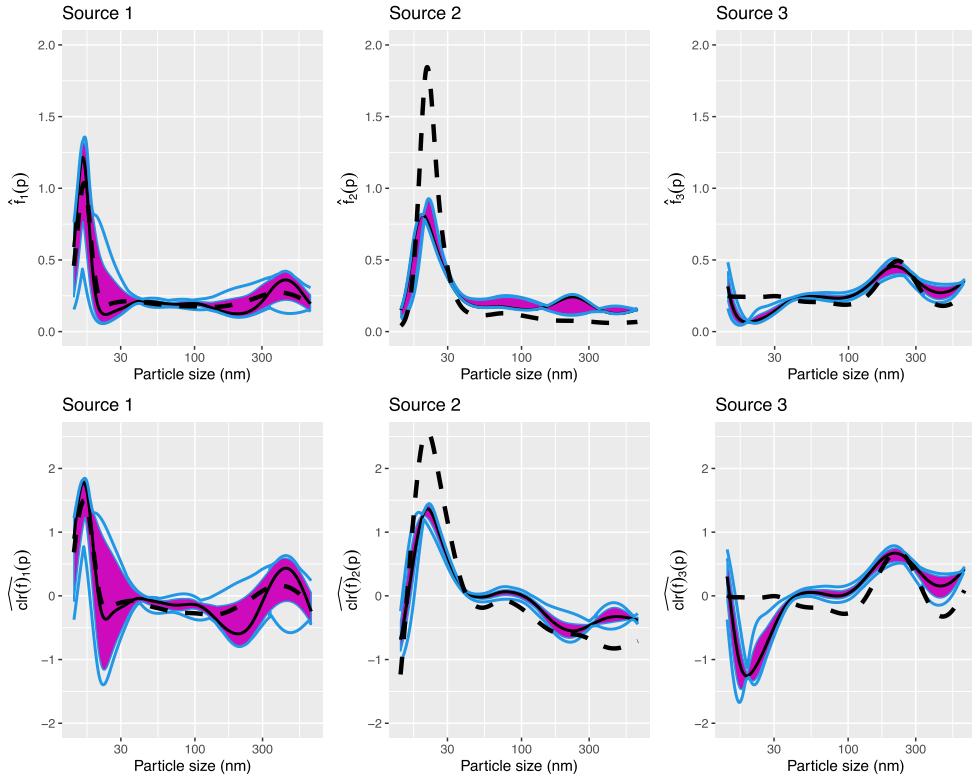

Figure B2. Functional boxplots for estimated sources in Bayes space (Top) and clr densities (Bottom) from Section with $M = 5{,}000$ and $T = 240$. The continuous black line represents the median curve, and the dashed black line is the true source.

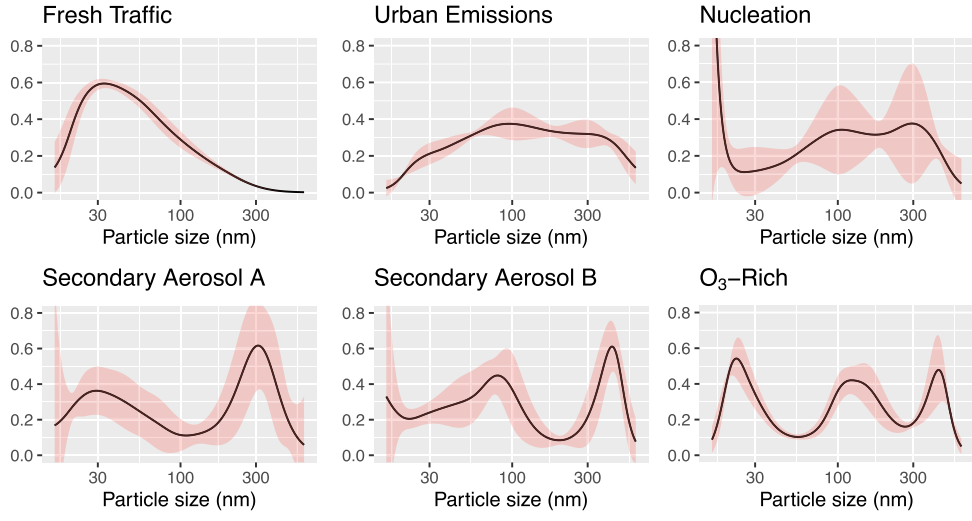

Figure C1. Sources with Bootstrap bands for the data, period 2011–2018.

Downloaded from https://academic.oup.com/jrssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

References

Al-Kindi S. G., Brook R. D., Biswal S., & Rajagopalan S. (2020). Environmental determinants of cardiovascular disease: Lessons learned from air pollution. *Nature Reviews Cardiology*, *17*(10), 656–672. https://doi.org/10.1038/s41569-020-0371-2

Baerenbold O., Meis M., Martínez-Hernández I., Euán C., Burr W. S., Tremper A., Fuller G., Pirani M., & Blangiardo M. (2023). A dependent Bayesian Dirichlet process model for source apportionment of particle number size distribution. *Environmetrics*, *34*(1), 1–18. https://doi.org/10.1002/env.v34.1

Beddows D. C. S., Harrison R. M., Green D. C., & Fuller G. W. (2015). Receptor modelling of both particle composition and size distribution from a background site in London, UK. *Atmospheric Chemistry and Physics*, *15*(17), 10107–10125. https://doi.org/10.5194/acp-15-10107-2015

Bishop H. F. J., & Anna Bornioli M. A. (2022). Effectiveness of London's ultra low emission zone in reducing air pollution: A pre-and post-comparison of NO_2 and PM_{10} levels. *Journal of Environmental Health*, *85*(1), 16–23. https://www.neha.org/july-aug-2022

Carslaw D. (2023). *Worldmet: Import surface meteorological data from NOAA integrated surface database (ISD)*. R Package Version 0.9.8.

Carslaw D. C., & Ropkins K. (2012). Openair—An R package for air quality data analysis. *Environmental Modelling & Software*, *27–28*, 52–61. https://doi.org/10.1016/j.envsoft.2011.09.008

Castel S., & Burr W. S. (2021). Assessing statistical performance of time series interpolators. *Engineering Proceedings*, *5*(1), 57. https://doi.org/10.3390/engproc2021005057

Chan T. W., & Mozurkewich M. (2007). Application of absolute principal component analysis to size distribution data: Identification of particle origins. *Atmospheric Chemistry and Physics*, *7*(3), 887–897. https://doi.org/10.5194/acp-7-887-2007

Dutton S. J., Vedal S., Piedrahita R., Milford J. B., Miller S. L., & Hannigan M. P. (2010). Source apportionment using positive matrix factorization on daily measurements of inorganic and organic speciated $PM_{2.5}$. *Atmospheric Environment*, *44*(23), 2731–2741. https://doi.org/10.1016/j.atmosenv.2010.04.038

Euán C., Ombao H., & Ortega J. (2018). The hierarchical spectral merger algorithm: A new time series clustering procedure. *Journal of Classification*, *35*(1), 71–99. https://doi.org/10.1007/s00357-018-9250-5

Fuller R., Landrigan P. J., Balakrishnan K., Bathan G., Bose-O'Reilly S., Brauer M., Caravanos J., Chiles T., Cohen A., Corra L., Cropper M., Ferraro G., Hanna J., Hanrahan D., Hu H., Hunter D., Janata G., Kupka R., Lanphear B., & Yan C. (2022). Pollution and health: A progress update. *The Lancet Planetary Health*, *6*(6), e535–e547. https://doi.org/10.1016/S2542-5196(22)00090-0

Gao J., Yuan Z., Liu X., Xia X., Huang X., & Dong Z. (2016). Improving air pollution control policy in China—A perspective based on cost–benefit analysis. *Science of The Total Environment*, *543*, 307–314. https://doi.org/10.1016/j.scitotenv.2015.11.037

Harrison R. M., Beddows D. C., Alam M. S., Singh A., Brean J., Xu R., Kotthaus S., & Grimmond S. (2019). Interpretation of particle number size distributions measured across an urban area during the FASTER campaign. *Atmospheric Chemistry and Physics*, *19*(1), 39–55. https://doi.org/10.5194/acp-19-39-2019

Hime N. J., Marks G. B., & Cowie C. T. (2018). A comparison of the health effects of ambient particulate matter air pollution from five emission sources. *International Journal of Environmental Research and Public Health*, *15*(6), 1206. https://doi.org/10.3390/ijerph15061206

Holman C. (1999). 8—Sources of air pollution. In S. T. Holgate, J. M. Samet, H. S. Koren, & R. L. Maynard (Eds.), *Air pollution and health* (pp. 115–148). Academic Press.

Hopke P. K., Chen Y., Rich D. Q., Mooibroek D., & Sofowote U. M. (2023). The application of positive matrix factorization with diagnostics to BIG DATA. *Chemometrics and Intelligent Laboratory Systems*, *240*, Article 104885. https://doi.org/10.1016/j.chemolab.2023.104885

Hopke P. K., Feng Y., & Dai Q. (2022). Source apportionment of particle number concentrations: A global review. *Science of The Total Environment*, *819*, Article 153104. https://doi.org/10.1016/j.scitotenv.2022.153104

Hron K., Menafoglio A., Templ M., Hrůzová K., & Filzmoser P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*, *94*, 330–350. https://doi.org/10.1016/j.csda.2015.07.007

Loomis D., Grosse Y., Lauby-Secretan B., Ghissassi F. E., Bouvard V., Benbrahim-Tallaa L., Guha N., Baan R., Mattock H., Straif K., & On behalf of the International Agency for Research on Cancer Monograph Working Group IARC (2013). The carcinogenicity of outdoor air pollution. *The Lancet Oncology*, *14*(13), 1262–1263. https://doi.org/10.1016/S1470-2045(13)70487-X

Machalová J., Hron K., & Monti G. (2016). Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics*, *43*(8), 1419–1435. https://doi.org/10.1080/02664763.2015.1103706

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

Martínez-Hernández I., Gonzalo J., & González-Farías G. (2022). Nonparametric estimation of functional dynamic factor model. *Journal of Nonparametric Statistics*, *34*(4), 895–916. https://doi.org/10.1080/10485252.2022.2080825

Marval J., & Tronville P. (2022). Ultrafine particles: A review about their health effects, presence, generation, and measurement in indoor environments. *Building and Environment*, *216*, Article 108992. https://doi.org/10.1016/j.buildenv.2022.108992

Moreno-Ríos A. L., Tejeda-Benítez L. P., & Bustillo-Lecompte C. F. (2022). Sources, characteristics, toxicity, and control of ultrafine particles: An overview. *Geoscience Frontiers*, *13*(1), Article 101147. https://doi.org/10.1016/j.gsf.2021.101147

Mukherjee A., & Agrawal M. (2017). World air particulate matter: Sources, distribution and health effects. *Environmental Chemistry Letters*, *15*(2), 283–309. https://doi.org/10.1007/s10311-017-0611-9

Murray C. J. L., Aravkin A. Y., Zheng P., Abbafati C., Abbas K. M., Abbasi-Kangevari M., Abd-Allah F., Abdelalim A., Abdollahi M., Abdollahpour I., Abegaz K. H., Abolhassani H., Aboyans V., Guimarães Abreu L., Abrigo M. R. M., Abualhasan A., Abu-Raddad L. J., Abushouk A. I., Adabi M., & Lim S. S. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. *The Lancet*, *396*(10258), 1223–1249. https://doi.org/10.1016/S0140-6736(20)30752-2

Norris G., Duvall R., Brown S., & Bai S. (2014). *EPA positive matrix factorization (PMF) 5.0 fundamentals and user guide*. US Environmental Protection Agency EPA/600/R-14/108.

Ohlwein S., Kappeler R., Kutlar Joss M., Künzli N., & Hoffmann B. (2019). Health effects of ultrafine particles: A systematic literature review update of epidemiological evidence. *International Journal of Public Health*, *64*(4), 547–559. https://doi.org/10.1007/s00038-019-01202-7

Paatero P., & Tapper U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*(2), 111–126. https://doi.org/10.1002/env.v5:2

Park E. S., Hopke P. K., Oh M.-S., Symanski E., Han D., & Spiegelman C. H. (2014). Assessment of source-specific health effects associated with an unknown number of major sources of multiple air pollutants: A unified Bayesian approach. *Biostatistics*, *15*(3), 484–497. https://doi.org/10.1093/biostatistics/kxu004

Pirani M., Best N., Blangiardo M., Liverani S., Atkinson R. W., & Fuller G. W. (2015). Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. *Environment International*, *79*, 56–64. https://doi.org/10.1016/j.envint.2015.02.010

Ramsay J. O., & Silverman B. W. (2005). *Functional data analysis* (2nd ed.). Springer series in statistics. Springer.

Rice G., & Shang H. L. (2017). A plug-in bandwidth selection procedure for long-run covariance estimation with stationary functional time series. *Journal of Time Series Analysis*, *38*(4), 591–609. https://doi.org/10.1111/jtsa.v38.4

Rivas I., Beddows D. C., Amato F., Green D. C., Järvi L., Hueglin C., Reche C., Timonen H., Fuller G. W., Niemi J. V., Pérez N., Aurela M., Hopke P. K., Alastuey A., Kulmala M., Harrison R. M., Querol X., & Kelly F. J. (2020). Source apportionment of particle number size distribution in urban background and traffic stations in four European cities. *Environment International*, *135*, Article 105345. https://doi.org/10.1016/j.envint.2019.105345

Sun X., Wang H., Guo Z., Lu P., Song F., Liu L., Liu J., Rose N. L., & Wang F. (2020). Positive matrix factorization on source apportionment for typical pollutants in different environmental media: A review. *Environmental Science: Processes & Impacts*, *22*(2), 239–255. https://doi.org/10.1039/C9EM00529C

Sun Y., & Genton M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, *20*(2), 316–334. https://doi.org/10.1198/jcgs.2011.09224

Talská R., Menafoglio A., Hron K., Egozcue J. J., & Palarea-Albaladejo J. (2020). Weighting the domain of probability densities in functional data analysis. *Stat*, *9*(1), e283. https://doi.org/10.1002/sta4.v9.1

Talská R., Menafoglio A., Machalová J., Hron K., & Fišerová E. (2018). Compositional regression with functional response. *Computational Statistics & Data Analysis*, *123*, 66–85. https://doi.org/10.1016/j.csda.2018.01.018

Templ M., Hron K., & Filzmoser P. (2011). robCompositions: An R-package for robust statistical analysis of compositional data, Chapter 25. In *Compositional Data Analysis* (pp. 341–355). John Wiley and Sons, Ltd. https://doi.org/10.1002/9781119976462.ch25

Thurston G. D., Ito K., & Lall R. (2011). A source apportionment of U.S. fine particulate matter air pollution. *Atmospheric Environment*, *45*(24), 3924–3936. https://doi.org/10.1016/j.atmosenv.2011.04.070

Tremper A. H., Jephcote C., Gulliver J., Hibbs L., Green D. C., Font A., Priestman M., Hansell A. L., & Fuller G. W. (2022). Sources of particle number concentration and noise near London Gatwick Airport. *Environment International*, *161*, Article 107092. https://doi.org/10.1016/j.envint.2022.107092

Uria-Tellaetxe I., & Carslaw D. C. (2014). Conditional bivariate probability function for source identification. *Environmental Modelling & Software*, *59*, 1–9. https://doi.org/10.1016/j.envsoft.2014.05.002

Van den Boogaart K. G., Egozcue J. J., & Pawlowsky-Glahn V. (2014). Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics*, *56*(2), 171–194. https://doi.org/10.1111/anzs.12074

Downloaded from https://academic.oup.com/jrsssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025

Wang X., Liu E., Yan M., Zheng S., Fan Y., Sun Y., Li Z., & Xu J. (2023). Contamination and source apportionment of metals in urban road dust (Jinan, China) integrating the enrichment factor, receptor models (FA-NNC and PMF), local Moran's index, Pb isotopes and source-oriented health risk. *Science of The Total Environment*, *878*, Article 163211. https://doi.org/10.1016/j.scitotenv.2023.163211

Watson J. G., Zhu T., Chow J. C., Engelbrecht J., Fujita E. M., & Wilson W. E. (2002). Receptor modeling application framework for particle source apportionment. *Chemosphere*, *49*(9), 1093–1136. https://doi.org/10.1016/S0045-6535(02)00243-6

Wiedensohler A., Birmili W., Nowak A., Sonntag A., Weinhold K., Merkel M., Wehner B., Tuch T., Pfeifer S., Fiebig M., Fjäraa A. M., Asmi E., Sellegri K., Depuy R., Venzac H., Villani P., Laj P., Aalto P., Ogren J. A., & Bastian S. (2012). Mobility particle size spectrometers: Harmonization of technical standards and data structure to facilitate high quality long-term observations of atmospheric particle number size distributions. *Atmospheric Measurement Techniques*, *5*(3), 657–685. https://doi.org/10.5194/amt-5-657-2012

Wu W., Jin Y., & Carlsten C. (2018). Inflammatory health effects of indoor and outdoor particulate matter. *Journal of Allergy and Clinical Immunology*, *141*(3), 833–844. https://doi.org/10.1016/j.jaci.2017.12.981

Zeng Y., Cao Y., Qiao X., Seyler B. C., & Tang Y. (2019). Air pollution reduction in China: Recent success but great challenge for the future. *Science of The Total Environment*, *663*, 329–337. https://doi.org/10.1016/j.scitotenv.2019.01.262

Downloaded from https://academic.oup.com/jrssc/article/74/1/229/7820614 by Imperial College London Library user on 06 March 2025