Two-Stage Bayesian Factor Analysis for Air Pollution Source Apportionment and Health Risk Assessment

Georges Bucyibaruta[1,*], Monica Pirani[2], Christina Mitsakou[3], David Green[4], Gary W. Fuller[4], Anja Tremper[4], Marta Blangiardo[2]

[1]Center for Equity in Global Surgery, University of Global Health Equity, Kigali, Rwanda
[2]MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK
[3]UK Health Security Agency, London, UK
[4] Environmental Research Group, MRC Centre for Environment and Health, School of Public Health, Imperial College London, London, UK
[*]Corresponding author gbucyibaruta@gmail.com

Abstract

Air pollution, especially particulate matter (PM), presents significant public health challenges and is associated with several Sustainable Development Goals (SDGs), notably SDG 3 (Good Health and Well-being) and SDG 11 (Sustainable Cities and Communities). Effective policy development requires robust statistical approaches to identify pollution sources and quantify their health impacts with appropriate uncertainty.

This study develops a novel two-stage Bayesian framework for air pollution source apportionment and health risk assessment, with the aim of quantifying the contribution of distinct particle sources to respiratory health outcomes in children, using particle number size distribution (PNSD) data from London.

In the first stage, we construct a Bayesian dynamic factor model with autoregressive components to infer latent pollution sources, incorporating non-negativity constraints and accounting for temporal dependence. In the second stage, we assess the relationship between source-specific exposures and respiratory hospital admissions in children via a Poisson regression model, explicitly propagating uncertainty from the source apportionment stage to the health model.

The model identifies four main sources: nucleation, traffic, urban activities, and secondary aerosols. Among these, traffic and secondary sources

exhibit the strongest and most consistent associations with increased respiratory hospital admissions. Importantly, models that do not account for uncertainty propagation tend to overestimate health risk associations, underscoring the value of the proposed Bayesian framework.

This work illustrates the advantages of integrating Bayesian methods for source apportionment and health effect estimation, with formal uncertainty propagation across model stages. The proposed framework enhances interpretability and supports evidence-based public health and environmental policy. It is readily extensible to other pollutants and settings, contributing to improved air quality management and progress toward global sustainability goals.
Keywords: Air pollution, source apportionment, Bayesian methods, two-stage model, uncertainty propagation

1 Introduction

Air pollution is a major public health concern, and there is extensive evidence of its role in a range of non-communicable diseases, especially respiratory and cardiovascular diseases (Evangelopoulos et al., 2019; Fuller, 2019; Fuller et al., 2022). Air pollution is addressed in several Sustainable Development Goals (SDGs), including SDG 3 (Good Health and Well-being), SDG 11 (Sustainable Cities and Communities), and SDG 13 (Climate Action), highlighting its impact on health, urban environments, and climate systems (WHO, 2025). In the UK, the Committee on the Medical Effects of Air Pollutants (COMEAP)[1] has been reviewing the toxicological and epidemiological evidence, advising the government and informing the research needs on the health effects associated with exposure to air pollution.

Particulate matter (PM), defined as any particle smaller than a certain size (e.g. PM_{10}, $PM_{2.5}$) has been recognised as one of the most dangerous pollutants (WHO, 2021). PM contains a mixture of contaminants from different sources, nevertheless the majority of epidemiological studies are still considering the totality of PM when assessing a link with health outcomes. Several recent studies have questioned whether a more meaningful approach is to studying the source composition of PM concentration, to account for the fact that different sources might have different harmful effects (see for instance Park et al. 2014; Hackstadt and Peng 2014; Pirani et al. 2015; Samoli et al. 2016; Boogaard et al. 2022; COMEAP 2022). This perspective also allows policy-makers to more directly connect research outcomes to interventions.

In this perspective, source apportionment (SA) refers to a class of methods aimed at partitioning pollution to the sources from which it was emitted. SA has been an active area of research in the last thirty years (Hopke, 2016; Yu

[1] https://www.gov.uk/government/groups/committee-on-the-medical-effects-of-air-pollutants-comeap

et al., 2022). Two main approaches can be considered: i) source-oriented deterministic models; and ii) receptor models (Viana et al., 2008). The former relies on knowledge of emissions and transport processes to predict the concentrations of pollutants at a receptor site; the latter is based on statistical procedures to identify and quantify the sources of pollutants based on the mixture of chemicals measured at the receptor sites. Receptor models have become popular as they do not require prior knowledge on the emission sources; additionally, software for this type of analysis is widely available.

To perform SA, compositional data with information on the different chemical components within PM concentration are needed to disentangle the total PM into sources. However, this type of data is expensive to collect and not widely available at monitoring site networks. As an alternative, measurement of particle number concentration (PNC) and related particle number size distribution (PNSD) have recently received much attention as a way to investigate PM sources (Hopke, 2016; Garcia-Marlès et al., 2024). This consists of measuring a wide range of particle sizes, ranging from 10nm to 2500nm diameter, spanning both ultrafine (< 100nm diameter) and fine (100 - 2500nm diameter) particle ranges. In particular, PNC represents the total number of particles per unit volume of air, often dominated by ultrafine particles (< 100 nm), while PNSD describes how particle numbers are distributed across particle size ranges, revealing distinct modes, nucleation, Aitken, and accumulation, that are associated with different sources and atmospheric processes (Garcia-Marlès et al., 2024).

Similar results have been reported when comparing sources from compositional data and PNSD, particularly for long-range transport sources (Gu et al., 2011). In the UK, Martínez-Hernández et al. (2025) proposed a functional factor model for hourly PNSD collected at a background station in Greater London, while Baerenbold et al. (2023) proposed a dependent Dirichlet process mixture model to apportion PNSD collected at a monitoring station near London Gatwick airport. Other studies relied on the popular Positive Matrix Factorization (PMF) to separate airborne PNSD into sources (Beddows et al., 2015; Beddows and Harrison, 2019; Tremper et al., 2022). Furthermore, a recent study applied PMF to analyze the air pollution mixture in four European cities, including London, identifying common sources of PNSD in cities while also distinguishing location-specific ones (Rivas et al., 2020).

A recent review on receptor modelling methods for air pollution SA, underlined that deterministic models do not provide accurate representations of the variability of species and concentrations observed in the atmosphere (Hopke, 2016), while other authors have recommended the use of a Bayesian approach to overcome this issue (Park et al., 2015; Molitor et al., 2016; Pirani et al., 2015). For example Hackstadt and Peng (2014) was one of the first studies to propose a Bayesian factor model for source attribution, using national databases to provide prior information on source emissions.

The Bayesian paradigm has also been advocated as a natural framework for

ensemble-based SA methods, which try to overcome limitations and uncertainty associated with individual standard SA models (Park and Oh, 2015; Tang et al., 2020), along with specific computational packages (Park et al., 2021). Bayesian approaches provide a principled way to incorporate prior knowledge, handle temporal and spatial correlations, enforce constraints such as nonnegativity, and propagate uncertainty in both model parameters and source contributions (Park et al., 2001; Rusanen et al., 2023; Dai et al., 2024; Frigeri et al., 2025). These contributions highlighted the need for additional work to better understand the potential of Bayesian approaches in receptor modelling. For instance, Park et al. (2001) included in the model an autoregressive component to account for temporal correlation in the data, while Park and Oh (2015) explicitly accounted for nonnegativity constraints on the source contributions and source compositions, in both parameter and model uncertainty estimation. Similarly, recent studies have demonstrated the flexibility of Bayesian frameworks in hierarchical models to integrate multiple sources of information and better quantify uncertainty in urban and regional air pollution source apportionment (Tang et al., 2020; Rusanen et al., 2023; Dai et al., 2024).

An additional and related challenge arises when evaluating the health effects of pollutant mixtures. The typical epidemiological approach consists of two-stages: the source profiles are estimated, and then they are plugged into a health regression model (see for instance Samoli et al. 2016; Bell et al. 2014; Rivas et al. 2021). Recently, statisticians have been advocating that uncertainty in the exposure estimates needs to be taken into account in the health-effect model. While several contributions have considered a joint approach, where the source profiles are estimated at the same time as their impact on health outcomes (Park et al., 2014; Molitor et al., 2016), this approach has been criticised as it potentially introduces feedback from the health endpoint into the source estimation (Szpiro and Paciorek, 2013).

Framed in this perspective, here we propose a two-stage model to apportion PNSD; the first stage allows for a data driven selection of the number of sources based on prior knowledge of sources characterization, and explicitly accounts for nonnegativity constraints on the source contributions and source compositions. Additionally, it includes an autoregressive component to account for temporal correlation in the data. The second stage links the apportioned sources to health outcomes; in particular, we evaluate the effect of the estimated sources on respiratory hospital admissions in children. To propagate the uncertainty from the first stage (SA) to the second stage (health model), we select a number of samples from the first stage, to represent the sources posterior multivariate distribution; we then run the second stage and combine the effect estimates using weights based on the likelihood of each of the selected first stage samples. We finally compare the results with that of a classic two-stage model which plugs-in point estimates only from the first stage.

The remainder of the paper is structured as follows: in section 2 we present the data and the proposed statistical approach. Then, in section 3, we describe

the results of the two-stage model, including the choice of the number of sources and their interpretation, as well as comparing the epidemiological results when we include or do not include the uncertainty propagation. Finally, in section 4 we provide some discussion points on the model specification, on the results and on the policy implications.

2 Material and methods

2.1 Data

2.1.1 Sampling site and PNSD data

The air quality data analysed in this study consist of PNSD that have been measured during the years 2014-2016. These PNSD data were collected from an urban background monitoring site located in a residential area of London (North Kensington, 51° 31′ 16″ North, 0° 12′ 48″ West). This station is part of both the London Air Quality Network and the national Automatic Urban and Rural Network, and is located on the grounds of Sion Manning School. The closest road, St Charles Square, is a quiet residential street located approximately 5 m from the monitoring site, within a predominantly residential area. The nearest major paved roads are the B450, 100 m to the east, and the heavily trafficked A40, 400 m to the south. Data were acquired using a Scanning Mobility Particle Sizer (SMPS; TSI 3080) coupled with a Condensation Particle Counter (CPC; TSI 3775) equipped with a long Differential Mobility Analyzer (DMA) and a diffusion dryer. These measurements adhered to the EUSAAR/ACTRIS protocol recommendations (Wiedensohler et al., 2012) and were subsequently corrected for diffusional losses. PNSD data were measured over 51 size bins, ranging between 17 and 604 nm. For this study, we aggregated the hourly data to obtain mean daily time series.
The North Kensington site was selected because it is a well-established urban background location widely used in previous aerosol studies to represent population exposure under typical London conditions (Beddows et al., 2015; Crilley et al., 2017; Harrison et al., 2019). Unlike roadside stations, it is minimally affected by local traffic turbulence and hence suitable for assessing temporal variations and source contributions at the city scale.

2.1.2 Gaseous pollutants and meteorological data

We collected daily time series concentration data for various gaseous pollutants, including carbon monoxide (CO; ppb), nitrogen oxides (NO_x; ppb), nitrogen dioxide (NO_2; ppb), nitric oxide (NO; ppb), ozone (O_3; ppb), and sulfur dioxide (SO_2; ppb), using the R package `openair` (Carslaw and Ropkins, 2012). Meteorological data, such as daily mean temperature (°C), atmospheric pressure (hPa), visibility (m), and relative humidity (%), were also obtained via the same package. Black carbon (BC_{IR}; $\mu g\,m^{-3}$ and BC_{UV}; $\mu g\,m^{-3}$) are measured through the magee aethalometer two wavelength at the North Kensington

monitoring station, while BC_{wood} ($\mu g\, m^{-3}$) is computed from the previous two measurements following Font et al. (2022). In addition, wind speed (m/s) and direction (angle) data were retrieved from the NOAA Integrated Surface Database for the London Heathrow site using the R package `worldmet` (Carslaw, 2023), as commonly done in previous studies (Harrison et al., 2019). These data were chosen because the Heathrow station provides high-quality, continuous meteorological records with standardized quality control, and its location offers reliable representation of regional air mass movement across Greater London, being minimally influenced by local urban structures or microclimatic effects.

2.1.3 Hospitalisation data

The data related to hospital admission for respiratory diseases (ICD 10 codes: J00–J99) among children (<16 years of age) were available from the registry of the Hospital Episode Statistics within the UK Small Area Health Statistics Unit (SAHSU) and cover all children resident within the Greater London boundaries. For the purpose of this analysis, we assumed that children residing across Greater London were uniformly exposed to the urban background air pollution conditions represented by the North Kensington monitoring site. This assumption is consistent with previous time-series epidemiological studies, where temporal variations in regional background concentrations are considered to dominate over intra-urban spatial differences (Crilley et al., 2017; Harrison et al., 2019).

2.2 Two-stages modelling approach

Our modelling approach consists of two stages. In the first stage, we specify a dynamic factor model to estimate the latent air particles sources and their associated uncertainty. In the second stage we link the identified pollution sources to the number of hospital admissions for respiratory causes in children in Greater London.

The two-stages model is specified in details in the following sections. Both stages are implemented using an MCMC algorithm within the `rjags` package (Finley, 2013) in R version 4.2.1. For the first stage we considered 2 chains with 50000 iterations as burnin and 50000 as posterior sample thinned by 10, and 10000 as adaptation sample which required approximately 2.5 hours per chain on a standard desktop computer. Convergence of the Bayesian models was assessed by inspecting trace plots and calculating the Gelman-Rubin R-hat statistic for key parameters. For the second stage we considered 2 chains with 40000 iterations as burnin and 10000 as posterior distribution, thinned by 20, so that the final posterior sample was 1000 across the two chains, which required approximately 1.5 hours per chain on a standard desktop computer. The convergence to the stationary distribution was evaluated using traceplots and the Rhat analytical tool (Gelman et al., 2013). The R scripts used for sources

apportionment are available on GitHub at `https://github.com/Georges3/Source-apportionment`.

2.2.1 Bayesian dynamic factor model for source apportionment

We propose a *Bayesian dynamic multivariate receptor model* to estimate PNSD source profiles and their temporal contributions. Let

$$X_t = (X_{t1}, X_{t2}, \ldots, X_{tD})$$

be a vector of particle size bins of dimension D measured at time t, for $t = 1, \ldots, T$. The model is defined as:

$$X_t = F_t \boldsymbol{\Delta} + \epsilon_t, \tag{1}$$

$$F_t \sim N(\rho F_{t-1}, \sigma_{\text{mass}}) I(F_t > 0), \tag{2}$$

$$\epsilon_t \sim N(0, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_D^2), \tag{3}$$

with the initial condition $F_1 \sim N(0, \sigma_f) I(F_1 > 0)$. While, $\epsilon_t = (\epsilon_{t1}, \ldots, \epsilon_{tD})$ represents the measurement error. In this formulation, $F_t = (F_{t1}, \ldots, F_{tK})$ represents the *source contribution vector* at time t ($t = 1, \ldots, T$) for each of the K unobserved sources, while $\boldsymbol{\Delta}$ denotes the *source composition matrix* of dimension $K \times D$. Each row of $\boldsymbol{\Delta}$, denoted $\Delta_k = (\Delta_{k1}, \ldots, \Delta_{kD})$, corresponds to the relative contributions of each size bin to source k. The term $\epsilon_t = (\epsilon_{t1}, \ldots, \epsilon_{tD})$ accounts for measurement errors.

For model identifiability, we followed the conditions described by Park et al. (2021): (i) each row of $\boldsymbol{\Delta}$ contains at least $K - 1$ zero elements; (ii) for each $k = 1, \ldots, K$, the rank of $\Delta(k)$ is $K - 1$; and (iii) $\sum_{d=1}^{D} \Delta_{kd} = 1$ for each k. Unlike traditional receptor models that assume static source contributions, our model incorporates a *first-order autoregressive process* (AR(1)) for the latent factors F_t, governed by the parameter $|\rho| < 1$. This extension explicitly captures the temporal autocorrelation inherent in time-resolved PNSD data, providing a more realistic representation of dynamic source behavior.

Following Park et al. (2021), for $\boldsymbol{\Delta}$, a point mass at zero is assigned to $K(K-1)$ preselected elements to satisfy the identifiability constraint (i). For the (nonzero) free elements of Δ, we assign a truncated multivariate normal distribution. The truncation restricts values to be non-negative, in accordance with the physical reality that particle source contributions cannot take negative values:

$$\text{vec}(\Delta^+) \sim N_{DK-K(K-1)}(\mu_0, \Sigma_0) I(\text{vec}(\Delta^+) \geq 0),$$

where $\text{vec}(\Delta^+)$ denotes the vector of the free elements of Δ (for which zeros are not preasssigned) stacked columnwise; $p = DK - K(K-1)$ is the number of free elements in Δ; μ_0 is an p-dimensional vector and Σ_0 is an $p \times p$-dimensional diagonal matrix, which are prespecified mean vector and diagonal covariance matrix, respectively. Incompatible *bins* were selected based on prior knowledge

from previous studies (Rivas et al., 2020) and the physical-chemical characteristics of the sources: for example, particles in the nucleation mode are unlikely to contribute to larger particle size bins typical of traffic or secondary sources. This ensures that each source profile is distinguishable and interpretable. Hence, bins incompatible with a given source k are fixed at zero, while the remaining elements are initialized with nonnegative values (typically between 0 and 1); we used 0.5 as a neutral starting point.

To complete the Bayesian specification, prior distributions are assigned to all unknown parameters. The variance components, from equation (3), follow inverse-Gamma priors:

$$\begin{aligned}\sigma_d^{-2} &\sim \text{Gamma}^{-1}(1, 0.01), d = 1, \ldots, D,\\ \sigma_f^{-2} &\sim \text{Gamma}^{-1}(1, 0.01),\\ \sigma_{\text{mass}}^{-2} &\sim \text{Gamma}^{-1}(1, 0.01),\end{aligned}$$

which are weakly informative and allow the data to dominate the posterior inference while maintaining inferential stability and the autoregressive parameter follows a uniform prior,

$$\rho \sim \text{Uniform}(0, 1),$$

reflecting minimal prior knowledge about temporal autocorrelation (Mezzetti et al., 2023). Overall, this specification balances prior information and identifiability constraints, extending traditional PMF by embedding temporal dependence in the latent source factors within a fully Bayesian framework.

This Bayesian dynamic factor model therefore extends traditional PMF and receptor modelling frameworks by embedding temporal dependence in the latent source factors, enabling both source identification and time-evolving source contribution estimation in a unified probabilistic structure.
Although only $\mathbf{X}_t$ is directly observed, estimation of the latent source contributions $\mathbf{F}_t$ and source profiles $\boldsymbol{\Delta}$ is feasible within the Bayesian hierarchical framework through data augmentation and iterative posterior sampling. Specifically, the model treats $\mathbf{F}_t$ and $\boldsymbol{\Delta}$ as unobserved random variables linked to the observations through the likelihood $\mathbf{X}_t = \mathbf{F}_t\boldsymbol{\Delta} + \boldsymbol{\epsilon}_t$. Inference proceeds by iteratively updating the conditional posterior distributions of all unknown quantities—$\mathbf{F}_t$, $\boldsymbol{\Delta}$, ρ, σ_{mass}, and $\boldsymbol{\Sigma}$—using a Markov Chain Monte Carlo (MCMC) algorithm. This allows simultaneous estimation of the source profiles and their temporal contributions from the measured particle size distributions. Conceptually, this approach follows the receptor modelling framework, in which the observed concentrations are expressed as mixtures of unobserved sources, while the Bayesian formulation enables probabilistic quantification of uncertainty and incorporation of prior knowledge and identifiability constraints. The joint estimation of $\mathbf{F}_t$ and $\boldsymbol{\Delta}$ thus proceeds by alternately sampling from their respective full conditional distributions until convergence, as implemented in the `rjags` package.

2.2.2 Selection of the number of sources

The number of sources K needs to be defined *a priori*. A common approach consists of running multiple models with different values of K and inspecting the resulting sources in terms of temporal profiles, correlation with other gaseous pollutants (e.g., O_3, NO_2), and meteorological conditions (e.g., temperature, wind speed, and direction) to choose the most 'reasonable' configuration. In our study, we considered a set of models M_g, with $g = 3, 4, 5, 6$, corresponding to different choices of K—specifically $K = 3, 4, 5, 6$—which were informed by previous studies on particle number size distribution source apportionment (Beddows and Harrison, 2019; Rivas et al., 2020). This allowed us to systematically compare model performance across different plausible source configurations and select the most appropriate number of latent sources based on model fit criteria.
It is worth noting that Bayesian nonparametric approaches can relax this assumption by inferring K directly from the data. Examples include Dirichlet Process mixture models and infinite factor models (Grazian, 2023; Wade, 2023; Baerenbold et al., 2023; Aiello et al., 2025), which have recently been explored in the context of environmental and air pollution applications. These approaches offer a flexible framework for automatically determining the effective number of latent sources, while requiring more complex modelling and higher computational resources. Here, we employ an alternative principled approach, by means of two criteria: the posterior Bayesian R squared and the Bayesian p-value.

Following Gelman et al. (2019), Bayesian R^2 generalizes the classical R^2 to the Bayesian setting by partitioning total variance into explained and residual components. The explained variance (V) reflects systematic patterns captured by the model, while the residual variance (var_{res}) captures uncertainty due to measurement error and parameter estimation. The ratio $R^2 = V/(V + \text{var}_{\text{res}})$ thus represents the proportion of variability explained by the model, fully accounting for uncertainty.

We denote the observed data by X, the model under consideration by M_g, and the vector of unknown model parameters by $\boldsymbol{\theta}$. After fitting the model using MCMC, we obtain m posterior draws of the parameters: $\boldsymbol{\theta}_i$ for $i = 1, \ldots, m$. For each draw $\boldsymbol{\theta}_i$, we generate a vector of predicted values $\mathbf{X}^{\text{pred}_i}$ of length $L = T \times D$ from the posterior predictive distribution, representing one possible realization of the model given the observed data. Here, $L = T \times D$ denotes the total number of observed data points, i.e., all time points across all particle size bins.

The **explained variance** V quantifies the variation in the posterior mean predictions across observations:

$$V = \frac{1}{L-1} \sum_{l=1}^{L} \left(\bar{X}_l^{\text{pred}} - \bar{\bar{X}}^{\text{pred}} \right)^2,$$

where $\bar{X}_l^{\text{pred}} = \frac{1}{m}\sum_{i=1}^{m} X_l^{\text{pred}_i}$ is the posterior mean prediction for observation l, and $\bar{\bar{X}}^{\text{pred}} = \frac{1}{L}\sum_{l=1}^{L} \bar{X}_l^{\text{pred}}$ is the overall mean of the posterior predictive means.

The residual variance var_{res} captures the variability of the predictions around the posterior mean for each observation:

$$\text{var}_{\text{res}} = \frac{1}{L-1}\frac{1}{m-1}\sum_{l=1}^{L}\sum_{i=1}^{m}\left(X_l^{\text{pred}_i} - \bar{X}_l^{\text{pred}}\right)^2.$$

The **Bayesian** R^2 is then calculated as:

$$\text{Bayesian } R^2 = \frac{V}{V + \text{var}_{\text{res}}},$$

which represents the fraction of total variability in the observed data explained by the model while accounting for uncertainty in parameter estimates. By evaluating this quantity on all posterior draws, we obtain a distribution of Bayesian R^2 values, providing a comprehensive measure of the fit of the model.

We also calculate posterior predictive p-values (Gelman et al., 2013), which quantify the probability, given the model, that a predicted observation exceeds the observed value. Specifically, for each element $l = 1, \ldots, L$ and for each posterior draw $i = 1, \ldots, m$, we compute

$$p_l = P(X_l \leq X_l^{\text{pred}_i} | M_g) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(X_l^{\text{pred}_i} \geq X_l),$$

where $\mathbf{1}(\cdot)$ is the indicator function, equal to 1 if the condition is true and 0 otherwise. The quantity p_l represents the proportion of predictive posterior draws that are greater than or equal to the observed data point X_l.

The set of values p_l, $l = 1, \ldots, L$, was visualized in a histogram to assess the fit of the model. If the model adequately represents the data, the histogram should be approximately uniform; deviations from uniformity indicate potential model mis-specification.

2.2.3 Health model

In the second stage, we link the estimated sources to health data through a Poisson regression model. We considered a lag window of 0–5 days for exposure to capture both immediate and delayed effects of particle exposure on respiratory hospitalisations. This choice is consistent with previous time-series studies (Rivas et al., 2021; Atkinson et al., 2010) and reflects the biological plausibility that air pollution can trigger both acute responses and delayed exacerbations of respiratory conditions. This stage assesses the association between sources and hospital admissions for respiratory causes in children through the following

specification (which from now on we refer to as the **health model**):

$$Y_t \sim \text{Pois}(\lambda_t E_t) \tag{4}$$
$$\log(\lambda_t) = \beta_0 + F_t\boldsymbol{\beta} + \sum_c s(Z_{tc}, \psi_c) + \delta_{I_t}$$

where Y_t is the health outcome at time t; β_0 is the general baseline risk of health outcome; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ is the vector of parameters describing the effect of each source-specific exposure on the risk of hospitalisation for respiratory conditions;

The nonlinear functions $s(Z_{tc}, \psi_c)$ in the health model represent smooth effects of daily average temperature and time, included to adjust for seasonality and long-term trends. These functions were modelled using *low-rank thin plate regression splines*, which provide a flexible and computationally efficient approximation to full thin plate splines. Thin plate splines define smooth functions by minimizing a bending-energy penalty, and their low-rank formulation avoids the need to choose knot locations explicitly while ensuring numerical stability and reducing the risk of overfitting (Wood, 2017). Following standard practice in environmental time-series modelling, we selected modest basis dimensions, i.e. four basis functions for temperature and nine for time, guided by preliminary exploratory analysis and consistent with methodological recommendations for avoiding excessive flexibility in temporal smooths (Crainiceanu et al., 2005; Perperoglou et al., 2019; Pirani et al., 2016; Blangiardo et al., 2019). These smooth terms act as confounder adjustments and influence the source–health effect estimates only indirectly through improved model fit. Finally, the term δ_{I_t} is an indicator variable that classifies the days of the week according to workday or holiday.

The priors for the health model parameters were specified as follows: $\beta_0 \sim N(0, 1)$,and $\boldsymbol{\beta} \sim N_K(0, 1)$. On the log-relative risk scale, these priors are weakly informative, roughly constraining most prior mass to relative risks between $\exp(-2) \approx 0.14$ and $\exp(2) \approx 7.4$, which is a reasonable range for epidemiological studies (Gelman et al., 2008) especially in the non communicable disease context. At the same time, this choice prevents extreme values in the posterior, while allowing the estimates to be data driven. In the context of our Poisson regression linking source-specific particulate exposures to paediatric hospital admissions, these priors provide sufficient flexibility to capture plausible health effects while ensuring numerical stability and avoiding unrealistic associations. The same $N(0, 1)$ prior was assigned to the weekday/holiday indicator term δ, reflecting a weakly informative assumption that the day-of-week effect is unlikely to induce very large changes in relative risk, but still allowing the data to inform its estimation.

2.3 Uncertainty propagation

An important aspect of the model consists of the propagation of the uncertainty on the apportioned sources (first stage) into the health model (second stage).

In this context, we explored two approaches; in the first we estimated source-specific health effects by plugging in the posterior mean of the F_t in the health model (called **plug-in** model). This is the approach typically used in environmental epidemiology, consisting in synthesizing the posterior distribution of the exposure through the posterior mean and then plugging this in the health model (see for instance Rivas et al. (2021)). The second approach takes advantage of the whole posterior distribution of F_t generated in stage one and employs the following steps:

1. We selected a sample of dimension $I = 1000$ from the posterior distribution for the matrix F, chosen after having explored the density plots to ensure no bimodality as well as fixing the Monte Carlo error to be less than 5% of the posterior standard deviation for all the parameters. This leads to 1000 vectors with length $T \times K$ and we calculate the joint posterior density using a Normal approximation $\mathbf{F}_i \mid \mathbf{X} \sim MVN(\hat{\mu_F}, \hat{\Sigma}_F)$ with $\hat{\mu_F}$ and $\hat{\Sigma}_F$ the posterior mean vector and variance-covariance matrix.

2. We synthesize the I values into a computationally more manageable number G; we considered the deciles of the densities $\mathbf{w} = (w_1, \ldots, w_i, \ldots, w_I)$, with $w_i = f(\mathbf{F} = \mathbf{F}_i \mid \mathbf{X})$ which guarantee a limited number of re-running of the second stage, while allowing us to explore the whole distribution.

3. We run the health model $G = 9$ times, each one selecting F as the matrix of posterior values where the joint density corresponds to $w_g, g = 1, \ldots, G$.

While step 1-3 are in line with previous work on uncertainty propagation in two-stage models (see for example Blangiardo et al. 2016; Lee et al. 2017; Villejo et al. 2023), here we combine each $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_K\}$ using a density-weighted mixture of the resulting stage 2 posteriors of the G runs, each weighted by the w_g value (**weighted approach**). Note that each weight w_g is $w_g \propto p(F_g \mid \mathbf{X})$ and $\mathbf{w}$ are normalised, so that pooling the stage 2 betas results in a model averaging strategy.

Additionally, we considered a **unweighted** approach, which consists of combining the posterior distribution for the vector of $\boldsymbol{\beta}$ from each of the the G samples all with equal weights. This is the approach used in Blangiardo et al. (2016); Lee et al. (2017); Villejo et al. (2023). Accounting for the uncertainty in the estimated source contributions F_t is essential, as neglecting it may lead to overly narrow credible intervals and biased estimates in the health model. By propagating the full posterior distribution of F_t, our approach ensures that the uncertainty in the source apportionment stage is appropriately reflected in the health effect estimates, leading to more robust and interpretable inference.

Mathematically, the weighted approach can be summarized as an approxi-

mation to the marginal posterior of the health effects,

$$p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}) = \int p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{F})\, p(\mathbf{F} \mid \mathbf{X})\, d\mathbf{F} \approx \sum_{g=1}^{G} w_g\, p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{F}_g),$$

$$w_g \propto p(\mathbf{F}_g \mid \mathbf{X}), \quad \sum_{g=1}^{G} w_g = 1.$$

where $\{\mathbf{F}_g\}_{g=1}^{G}$ are G representative matrices selected from the posterior of F, and $p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{F}_g)$ denotes the posterior distribution of the health model conditional on F_g.

3 Results

Figure 1 plots the correlation matrix across bin sizes, showing a strong dependence from bins close in size, with values even larger than 0.97. This stresses the importance of considering bin size dependence in the model formulation.

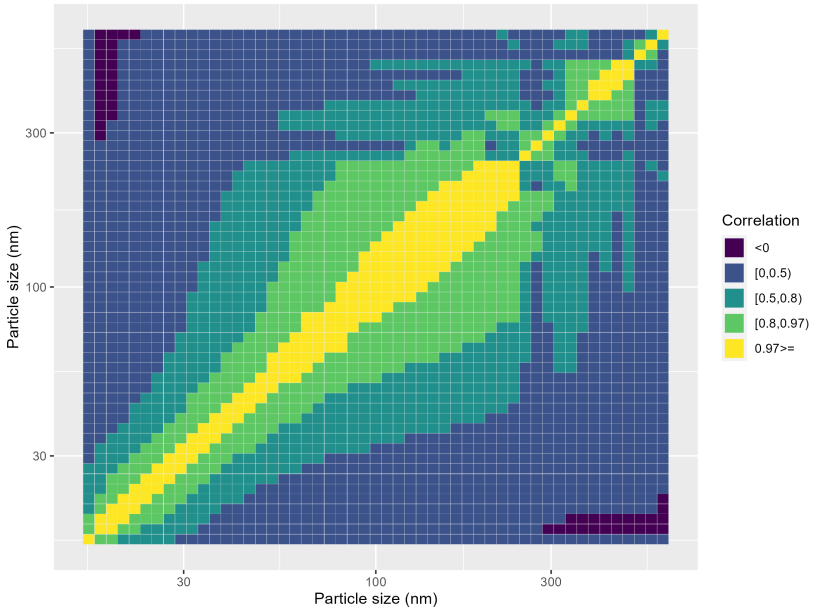


Figure 1: Correlation matrix across particle size bins. It refers to the inter-bin correlations across particle number size distribution (PNSD) measurements. Strong correlations (often exceeding 0.97) are observed between adjacent size bins, highlighting the need to account for bin-size dependence in the model formulation.

The daily paediatric hospitalisations for respiratory causes range from 20 to 221 in the study period, with an average of 92.6 which we used as E_t in the



Figure 2: Exploratory plots of mean daily temperature and paediatric hospital admissions (left); time series of paediatric hospital admissions (right). The red line shows a fitted spline highlighting non-linearity, used to guide model specification.

health model. As illustrated in Figure 2, hospital admissions exhibit a clear seasonal pattern and vary nonlinearly with daily temperature.

3.1 Model selection

We ran the multivariate factor model multiple times, specifying a different number of factors (sources), ranging from 3 to 6, with the aim of finding the model most supported by the data. Figure 3 presents the plot of the Bayesian p-values and the R^2 values for each model. The ones with 4 and 5 sources have the same value for R^2 but the p-values are more in line with a uniform distribution for the model with 4 sources. To better distinguish between these two solutions, we also considered the plausibility of the profiles, i.e. their ability to cover the entire range of particle sizes, as well as the correlation with other pollutants and meteorological variables. We did not find substantial differences in terms of correlation (Figure 5 here and Figure 1 in Supplementary Material); however, considering 5 factors seems to lead to a split of one source found in the 4 factor solution into two, resulting in a large overlap with another one. Due to this and the slightly better behaviour from a statistical perspective, we chose the solution with 4 sources as our final model.

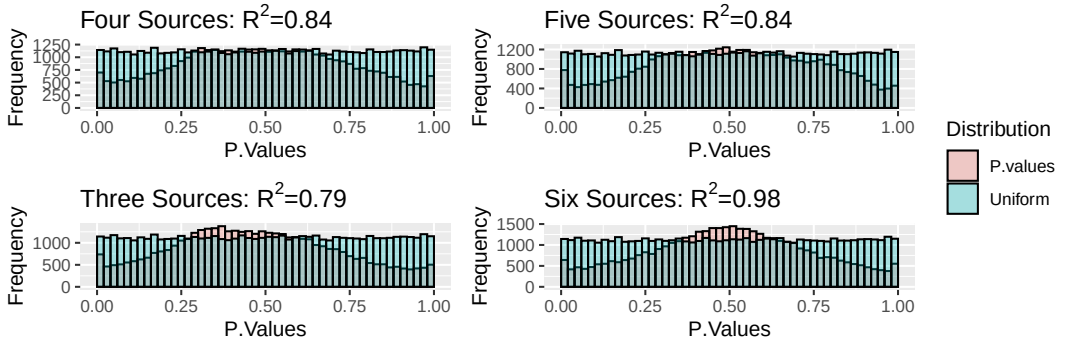


Figure 3: Model selection metrics: Model selection diagnostics for the Bayesian dynamic factor model. The figure shows Bayesian p-values and R^2 values for models with 3 to 6 sources. The 4-source model provides the best balance between model fit and posterior predictive performance, with Bayesian p-values closest to a uniform distribution and plausible source profiles consistent with particle size coverage and correlations with other pollutants and meteorological variables.

3.2 Bayesian dynamic factor model results

The next step consists in interpreting the sources, giving them a label based on their profiles, their temporal patterns and their correlation with other gasses and meteorological variables. We present source profiles in Figure 4, consisting of the proportion of particle concentration by bin sizes for each source.

Source 1 is characterized by a high concentration of particles in the smaller bins, ranging from 19.5 to 42.5 nm (mode equals to 19.5nm). We label this source as **nucleation**, referring to the process by which tiny particles form from gaseous precursors in the atmosphere. It could represent emissions from Heathrow Airport, particularly due to the prevalence of strong south-westerly winds (Figure 7). This source is more prevalent during summer and winter (Figure 6), but it does not show a strong correlation with any of the gases or meteorological variables (Figure 5). The estimated autoregressive parameter for Source 1, ρ_1, was 0.957 (95% CI: 0.933 – 0.983).

Source 2 has the highest particle concentration of sized around 48-563nm (mode equals to 273.90 nm). This source shows strong seasonality, with peaks in March-April and then in September-October (Figure 6) and has the strongest correlation with NO_x, NO_2 and CO, but also a negative correlation with visibility and O_3. It is characterised by a weak wind coming from East (Figure 7). We label this **secondary**, as it could represent particles that are not directly emitted but are formed in the atmosphere through chemical reactions between primary pollutants and other atmospheric compounds, for instance from reac-

tions between sulfur dioxide, nitrogen oxides, and ammonia. The estimated autoregressive parameter for source 2, ρ_2, was of 0.979 (95%CI: 0.962 - 0.996). Additionally, the observed correlations, seasonality, and wind direction suggest that secondary may include contributions from long-range transport, such as agricultural emissions or continental Europe.

Source 3 has the highest particle concentration of bins sized around 20.55-93.05nm (mode equals to 45.30nm), with values close to 0 from 100nm and above. This source is slightly lower during winter and shows a moderate correlation with CO, NO_x, SO_2 and atmospheric pressure. Looking at the wind, we can see that it is characterised by low speed and a North-East direction (Figure 7). We label this **traffic**, as is likely to represent the emissions coming from vehicles. The estimated autoregressive parameter for source 3, ρ_3, was of 0.965 (95%CI: 0.942 - 0.988).

Source 4 has a peak in the particle concentration of sized between 31.65 and 254.90(mode equals to 75.05nm), but is characterised by a large variability. This source is lower in Spring and correlated moderately with NO_2 (positively) and with O_3 (negatively). It is characterised by Easterly wind (Figure 7) and we label this **urban**, as this behavior is consistent with emissions arising from urban activities besides transportation, for instance including heating and cooking, waste burning, energy production, constructions and demolitions and chemical uses in households and businesses. The estimated autoregressive parameter for source 4, ρ_4, was of 0.974 (95%CI: 0.955 - 0.994).

3.3 Health model: phase 2

Figure 8 presents the point estimates and 95% credible intervals (95%CI) for the log relative risk of the four sources and lag 0 to 5. The green lines represent the results from the model without uncertainty (plugged-in), where we included the posterior mean of the daily sources contributions as exposure, while the orange lines visualise the results for the model with uncertainty propagated, obtained repeating the analysis 9 times, each time drawing and weighting the values of the four sources from their distributions obtained from the phase 1 model (weighted).

Overall, there is also a shift towards 0 for most of the effects when uncertainty propagation is considered. Secondary sources appear to have more consistent and positive effects, with higher coefficients across lags, particularly for lag 4 when the model estimates an increase of 0.7% (95%CI 0.2%-1%) in the risk of hospitalisation, for one standard deviation increase in the source concentration. The remaining sources show the largest difference between the model without and with uncertainty. For instance, in the former, nucleation has a positive association with risk of hospitalisations across all lags, while when uncertainty is considered evidence of an association remains only for lag 1 and lag 4, with 0.4% (95%CI 0.1%-0.6%) and 0.39% (95%CI 0.1%-0.7%) increases in the risk of hospital admissions for one sd increase in the source concentration respectively. Higher concentration of the urban source is associated with lower risk of hospitalisation in the model without uncertainty across all the lags, while

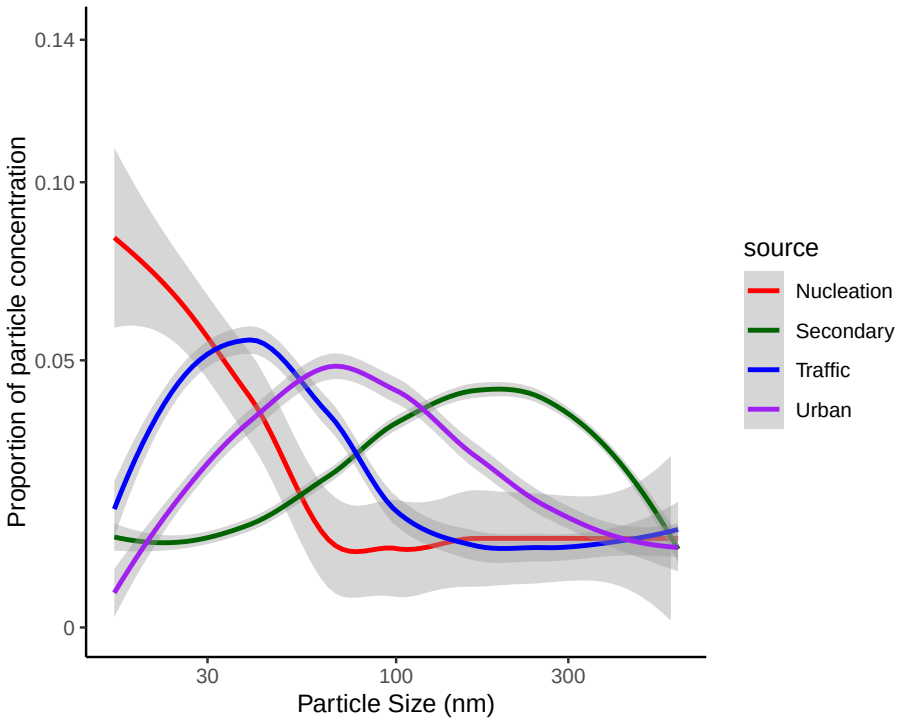


Figure 4: Results of Particle size distribution for the 4 sources identified by the model.

when uncertainty is propagated from the stage 1 model there is no evidence of an association anymore, except for lag 2 where the effect switch from negative to positive, equal to 0.38% (95%CI 0.07%-0.57%) increase in the risk of hospital admission for one standard deviation increase in the concentration.
These results highlight the temporal dynamics of exposure effects, explicitly evaluating lagged associations. By considering lags 0–5, the model captures both immediate (lag 0) and short-delayed effects (e.g., secondary sources at lag 4), directly addressing the timing of health impacts.
The comparison of the weighted model with the unweighted one, obtained combining the full posterior distributions of each element of the $\boldsymbol{\beta}$ using the G samples from the first stage model but without weighting them, is presented in Figure 9. Most of the point estimates are in agreement with the weighted model ones, except for Nucleation, where across all lags the unweighted model show negative estimates, while the weighted ones are mostly on the positive side.Nevertheless, across all the sources and lags, the unweighted approach is characterised by a large uncertainty, with likely values of $\boldsymbol{\beta}$ estimates ranging from -0.05 and 0.05. Incorporation of stage 1 uncertainty is crucial to avoid overestimation or misleading effects, particularly for sources with weaker sig-

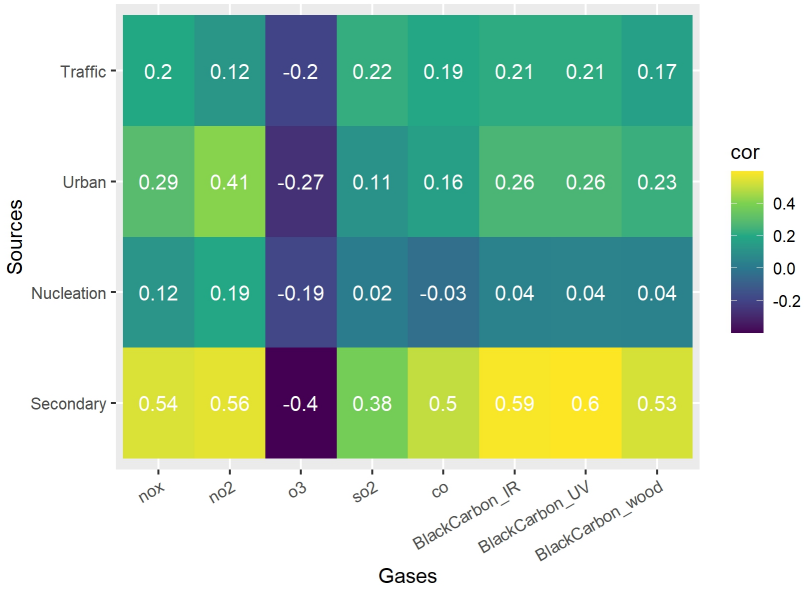


Figure 5: Correlation between the estimated sources and other air pollutants.

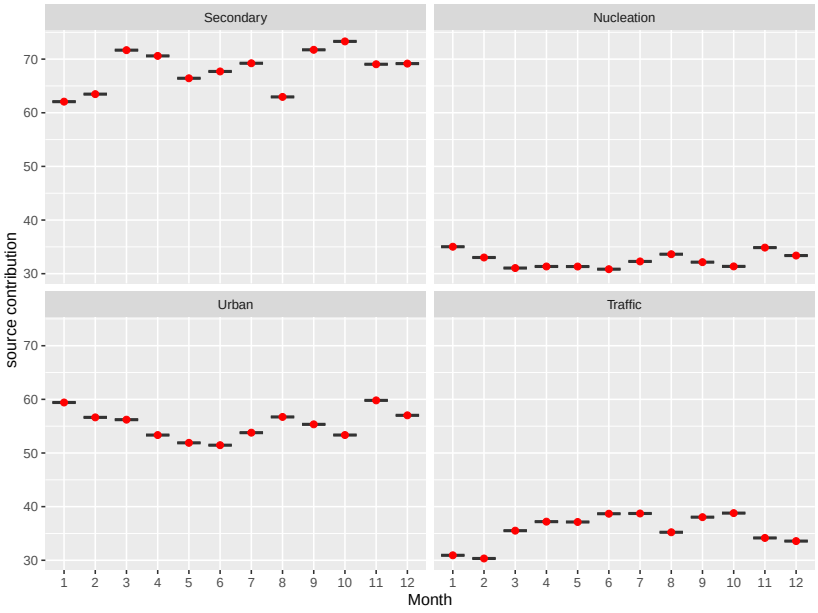


Figure 6: Estimated sources and average contributions per month.

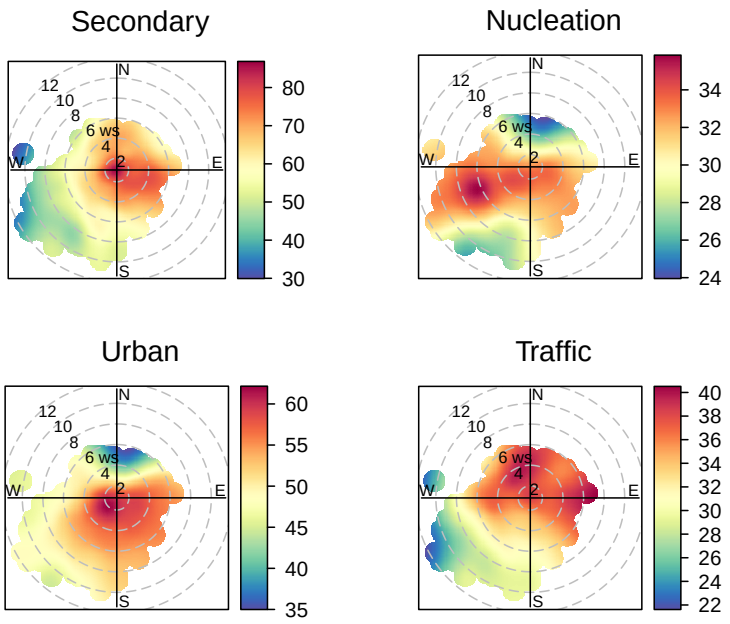


Figure 7: Polar plots showing how the estimated sources are affected by daily wind direction and wind velocity. The concentration scale of each source is reported in the legend.

nals. Posterior probability visualizations provide an intuitive interpretation of the likelihood of positive associations across lags.

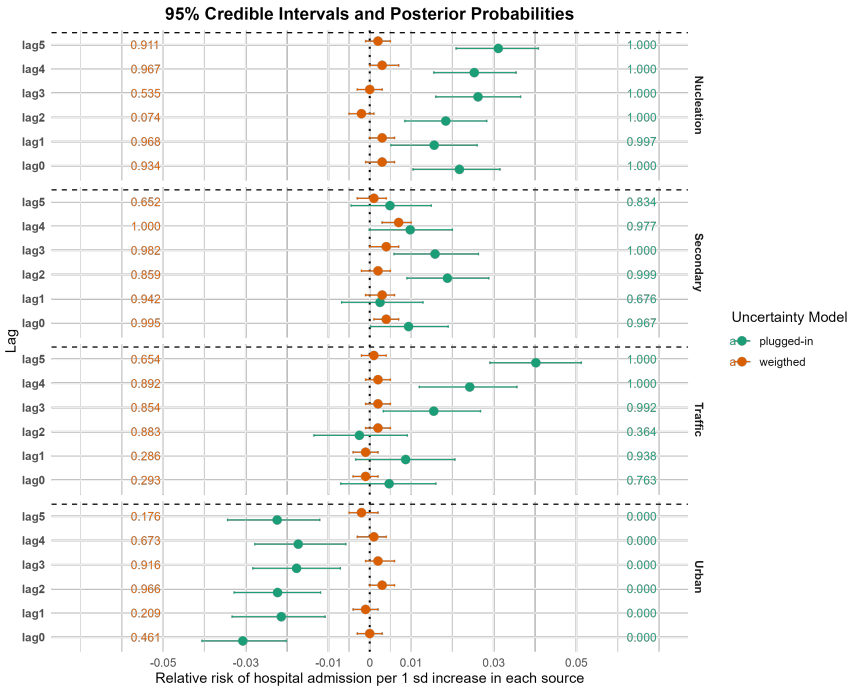


Figure 8: Results of the stage 2 model (plugged-in and weighted). The figure shows the log transformed relative risks and 95%CI for the lags 0 to 5 across the four air pollution sources for the model without uncertainty from stage 1 (plugged-in) (green lines) and with uncertainty propagated through a weighted combination of the posterior distributions (orange lines). Additionally, on the left and of the right end sides we present the posterior probability that the corresponding effect is above 0, indicating a positive association between the source and the risk of paediatric hospital admission in London.

4 Discussion

This paper proposes a novel approach to SA and health risk assessment of air pollution using a two-stage Bayesian factor analysis. The methodology is applied to apportion PNSD and assess their health impacts. The first stage employs a Bayesian dynamic factor model with autoregressive components to estimate latent air pollution sources while enforcing nonnegativity constraints on source contributions and compositions. The second stage links these sources to respiratory hospital admissions in children through a Poisson regression model, with uncertainty from the first stage fully propagated into the health model.

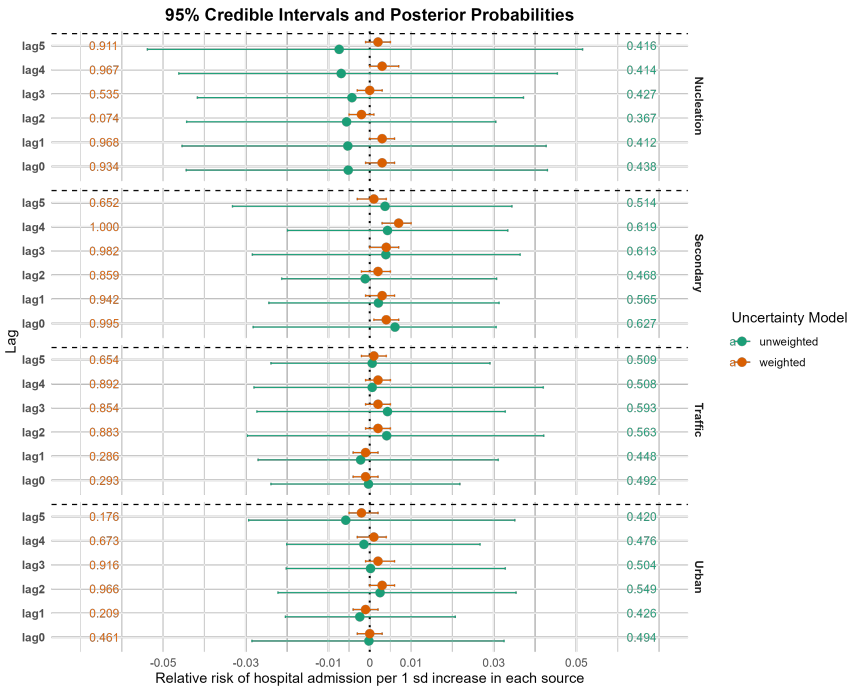


Figure 9: Results of the stage 2 model (unweighted and weighted). The figure shows the log transformed relative risks and 95%CI for the lags 0 to 5 across the four air pollution sources for the model where the uncertainty from stage 1 is incorporated through a simple pooling of the posterior distributions for each element in the $\boldsymbol{\beta}$ vector (unweighted - green lines) and through a weighted combination of the posterior distributions (weighted - orange lines). Additionally, on the left and of the right end sides we present the posterior probability that the corresponding effect is above 0, indicating a positive association between the source and the risk of paediatric hospital admission in London.

Our two-stage Bayesian framework integrates uncertainty propagation from source apportionment into health effect estimation, providing a coherent structure for quantifying the contribution of specific particle sources to respiratory hospital admissions. The model also accounts for short-term exposure lags (0–5 days) to capture both immediate and delayed effects.
COMEAP in its statement published in 2022 (COMEAP, 2022) concluded that although the toxicological and epidemiological evidence shows that different constituents are likely to have different toxicological actions, there was no consistent indication of specific particle components that are more toxic than others; therefore they recommended research to address some outstanding issues, such as conducting epidemiological studies incorporating source apportionment data and designs of studies that can distinguish the effects of correlated exposures.

This study aimed to address some of these by examining the intersection of air quality, health, and pollution sources, providing a science-driven pathway for achieving multiple SDGs while maintaining a strong focus on urban public health.

In support of SDG 3, which focuses on good health and well-being, it highlights the impact of air pollution on child health. It also contributes to SDG 11, aimed at promoting sustainable cities and communities, by identifying key urban pollution sources and informing targeted mitigation strategies. The implementation of policies aiming to improve air quality (e.g., electric cars) has been affecting the particle composition in the atmosphere, and it would be important to understand the different health impacts of the various particle components. Lastly, the research emphasizes the disproportionate impact of air pollution on vulnerable populations, particularly children in urban areas, thereby contributing to SDG 10, which seeks to reduce inequalities.

Air pollution, especially PM, poses significant health risks, yet traditional epidemiological studies often overlook the heterogeneity of PM sources. By using PNSD data, which captures both ultrafine and fine particles, this study addresses a critical gap in source identification. A key advantage of the Bayesian approach is that it integrates uncertainty directly into the estimation of source profiles and contributions. In contrast, methods like PMF require the analyst to provide a priori estimates of measurement uncertainty to run the model. Also, while post hoc techniques such as bootstrapping can assess the stability of source profiles from PMF, in the Bayesian approach, we rely on an integrated treatment of uncertainty taking advantage of the full posterior distribution, which enhances the reliability of health impact estimates.

Through this analysis, four main pollution sources were identified: nucleation, traffic, urban activities, and secondary sources, each distinguished by unique particle size profiles and temporal patterns. These are in line with previous literature (Rivas et al., 2020; Pandolfi et al., 2020; Garcia-Marlès et al., 2024). Among the apportioned sources, traffic and secondary factors exhibit the strongest and most consistent associations with respiratory health risks. Notably, the inclusion of uncertainty in the model highlights potential overestimates in traditional approaches that fail to incorporate uncertainty propagation.

These findings align with previous research on PNSD and health effects. Rivas et al. (2021) assessed the impact of ultrafine sources on daily mortality in Barcelona, Helsinki, London, and Zurich. In London, they found associations between nucleation, secondary sources, and traffic emissions with respiratory outcomes. However, in our study, when models were run without propagation of uncertainty, an unexpected inverse relationship (negative estimates) was found for urban sources, which disappeared when uncertainty was considered, emphasizing the importance of accounting for uncertainty in epidemiological models. Braniš et al. (2010) analyzed the association of particle number and $PM_{2.5}$ concentrations with mortality and cardiorespiratory hospital admissions in Prague while Halonen et al. (2009) examined associations between different particle size modes and mortality/hospital admissions among 65+ in Helsinki, Finland. While in these studies no association was found between air pollution

and daily mortality, Braniš et al. (2010) identified a link between concentration of submicron particles 15–487nm and respiratory admissions.

Our findings underscore the need for targeted mitigation strategies, particularly for vehicular emissions and secondary pollutants. The proposed methodological framework is adaptable to other regions and pollutants, providing a versatile tool for air quality management and health risk assessment.

Like most studies, this research has several limitations. One key consideration is the reliance on specific modelling assumptions. Although uncertainty has been accounted for, some significant associations may still be due to chance. The exposure model has been run on daily time series, while the typical approach in the source apportionment context would be to use hourly data, which are able to fully exploit the temporal variability in the air pollution concentration. However, our choice was dictated by the health data, which are only available at a daily resolution, making it not feasible to model hourly data. This could have an impact on the correlation between the sources and the pollutants or meteorological variables, resulting in values closer to zero. Another limitation pertains to exposure misclassification, as the data are derived from a single monitoring station, which may not fully capture the spatial heterogeneity of ultrafine particles. However, this should be interpreted within the context of the time-series design, which focuses on estimating associations driven by temporal rather than spatial variability in exposure and health outcomes. Nevertheless, this limitation could still affect the accuracy of the estimated health risks. In addition, the temporal and spatial variability of air pollution further complicates the interpretation of health risk estimates, particularly in large urban environments where pollution levels fluctuate considerably.

To address these limitations, future studies should (i) develop an approach which allows to estimate sources using hourly data which can then be upscaled at the daily level in the epidemiological model; (ii) incorporate data from multiple monitoring stations to better account for spatial variability. Additionally, expanding the analysis to include additional health outcomes, such as asthma and cardiovascular diseases, could provide a more comprehensive understanding of the health effects of air pollution. Furthermore, investigating long-term exposure effects and interactions between different particle sources may yield deeper insights into their cumulative impacts on public health. The proposed modelling framework could also be extended to incorporate more advanced temporal structures, such as distributed lag models (DLMs) or distributed lag non-linear models (DLNMs), to better capture cumulative and correlated exposure–response relationships (Gasparrini et al., 2010). Finally, simulation studies should be designed to evaluate the performance of the proposed two-stage Bayesian approach under controlled, realistic data-generating scenarios. Such analyses would help assess the model's ability to recover known source contributions and health effect estimates, and to explore different strategies for choosing the latent factors, such as weighted sampling directly from the posterior distribution.

This paper advances air pollution research by integrating cutting-edge Bayesian techniques with real-world health data. The findings provide valuable insights

for policymakers and establish a strong foundation for future interdisciplinary studies in environmental health.

Acknowledgments

The work of the UK Small Area Health Statistics Unit is overseen by UK Health Security Agency (UK HSA). SAHSU is part of the MRC Centre for Environment and Health supported by the UK Medical Research Council, Grant number: MR/L01341X/1), and the National Institute for Health Research (NIHR) through its Health Protection Units (HPRUs) at Imperial College London in Environmental Exposures and Health and in Chemical and Radiation Threats and Hazards, and through Health Data Research UK (HDR UK) in the Social and Environmental Determinants of Health Driver Programme. This paper does not necessarily reflect the views of UKHSA, NIHR or the Department of Health and Social Care.

This work uses data provided by patients and collected by the NHS as part of their care and support. Hospital Episode Statistics data are copyright © 2025, re-used with the permission of NHS England. All rights reserved.

All authors acknowledge infrastructure support for the Department of Epidemiology and Biostatistics provided by the NIHR Imperial Biomedical Research Centre (BRC).

Conflicts of Interest

The authors declare no conflict of interest.

Contributions

GB: methodology; analysis; writing (draft and review) MP: methodology; supervision; writing (draft and review); funding acquisition CM: writing (review); funding acquisition DG: data; writing (review); funding acquisition GF: data; writing (review); funding acquisition AT: data; writing (review); funding acquisition MB: methodology; supervision; writing (draft and review); funding acquisition

Ethics

All data used in this study were taken from existing datasets and no original data from individuals or human samples were collected. Hospital admission data are already held by SAHSU, supplied by NHS digital; data use is covered by approval from national Research Ethics Services(reference 17/LO/0846) and by the Health Research Authority's Confidentiality and Advisory Group (CAG) approval for section 251 support(reference 14/CAG/1039).

SAHSU holds approvals both from the London - South East Research Ethics Committee (22/LO/0256) and from the Health Research Authority - Confidentiality Advisory Group (20/CAG/0028).

Patient consent for publication

Not required.

Data availability

Due to confidentiality of the patient and ethical restrictions, the health data used in this study may be available only under specific conditions. However, the air pollution-related data sets analyzed during the current study are available from the corresponding author upon reasonable request.

References

Aiello, L., R. Argiento, S. Legramanti, and L. Paci (2025). Bayesian nonparametric clustering for spatio-temporal data, with an application to air pollution. *arXiv preprint arXiv:2505.24694*.

Atkinson, R. W., G. W. Fuller, H. R. Anderson, R. M. Harrison, and B. Armstrong (2010). Urban ambient particle metrics and health: a time-series analysis. *Epidemiology 21*(4), 501–511.

Baerenbold, O., M. Meis, I. Martínez-Hernández, C. Euán, W. S. Burr, A. Tremper, G. Fuller, M. Pirani, and M. Blangiardo (2023). A dependent Bayesian Dirichlet process model for source apportionment of particle number size distribution. *Environmetrics 34*(1), e2763.

Beddows, D. and R. M. Harrison (2019). Receptor modelling of both particle composition and size distribution from a background site in London, UK–a two-step approach. *Atmospheric Chemistry and Physics 19*(7), 4863–4876.

Beddows, D., R. M. Harrison, D. Green, and G. Fuller (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.

Bell, M. L., K. Ebisu, B. P. Leaderer, J. F. Gent, H. J. Lee, P. Koutrakis, Y. Wang, F. Dominici, and R. D. Peng (2014). Associations of PM2.5 constituents and sources with hospital admissions: analysis of four counties in Connecticut and Massachusetts (USA) for persons $\geq$ 65 years of age. *Environmental health perspectives 122*(2), 138–144.

Blangiardo, M., F. Finazzi, and M. Cameletti (2016). Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. *Spatial and spatio-temporal epidemiology 18*, 1–12.

Blangiardo, M., M. Pirani, L. Kanapka, A. Hansell, and G. Fuller (2019). A hierarchical modelling approach to assess multi pollutant effects in time-series studies. *PLoS One 14*(3), e0212565.

Boogaard, H., A. Patton, R. Atkinson, J. Brook, H. Chang, D. Crouse, J. Fussell, G. Hoek, B. Hoffmann, R. Kappeler, M. Kutlar Joss, M. Ondras, S. Sagiv, E. Samoli, R. Shaikh, A. Smargiassi, A. Szpiro, E. Van Vliet, D. Vienneau, J. Weuve, F. Lurmann, and F. Forastiere (2022). Long-term exposure to traffic-related air pollution and selected health outcomes: A systematic review and meta-analysis. *Environment International 164*, 107262.

Braniš, M., J. Vyškovská, M. Malỳ, and J. Hovorka (2010). Association of size-resolved number concentrations of particulate matter with cardiovascular and respiratory hospital admissions and mortality in Prague, Czech Republic. *Inhalation toxicology 22*(sup2), 21–28.

Carslaw, D. (2023). *worldmet: Import surface meteorological data from NOAA integrated surface database (ISD)*. R package version 0.9.8.

Carslaw, D. C. and K. Ropkins (2012). openair — an R package for air quality data analysis. *Environmental Modelling & Software 27–28*(0), 52–61.

COMEAP (2022). Particulate air pollution: health effects of exposure.

Crainiceanu, C. M., D. Ruppert, and M. P. Wand (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of statistical software 14*, 1–24.

Crilley, L. R., F. Lucarelli, W. J. Bloss, R. M. Harrison, D. C. Beddows, G. Calzolai, S. Nava, G. Valli, V. Bernardoni, and R. Vecchi (2017). Source apportionment of fine and coarse particles at a roadside and urban background site in London during the 2012 summer clearflo campaign. *Environmental Pollution 220*, 766–778.

Dai, T., Q. Dai, J. Yin, J. Chen, B. Liu, X. Bi, J. Wu, Y. Zhang, and Y. Feng (2024). Spatial source apportionment of airborne coarse particulate matter using PMF-Bayesian receptor model. *Science of The Total Environment 917*, 170235.

Evangelopoulos, D., K. Katsouyanni, H. Walton, and M. Williams (2019). Personalising the health impacts of air pollution: Interim statistics summary for a selection of statements. *King's College London*.

Finley, A. O. (2013). Using jags in r with the rjags package.

Font, A., K. Ciupek, D. Butterfield, and G. Fuller (2022). Long-term trends in particulate matter from wood burning in the United Kingdom: Dependence on weather and social factors. *Environmental Pollution 314*, 120105.

Frigeri, M., V. Berrocal, and A. Guglielmi (2025). Bayesian source apportionment of spatio-temporal air pollution data. *arXiv preprint arXiv:2510.27551*.

Fuller, G. (2019). *The invisible killer: The rising global threat of air pollution-and how we can fight back*. Melville House.

Fuller, R., P. J. Landrigan, K. Balakrishnan, G. Bathan, S. Bose-O'Reilly, M. Brauer, J. Caravanos, T. Chiles, A. Cohen, L. Corra, et al. (2022). Pollution and health: a progress update. *The Lancet Planetary Health 6*(6), e535–e547.

Garcia-Marlès, M., R. Lara, C. Reche, N. Pérez, A. Tobías, M. Savadkoohi, D. Beddows, I. Salma, M. Vörösmarty, T. Weidinger, C. Hueglin, N. Mihalopoulos, G. Grivas, P. Kalkavouras, J. Ondracek, N. Zikova, J. V. Niemi, H. E. Manninen, D. C. Green, A. H. Tremper, M. Norman, S. Vratolis, E. Diapouli, K. Eleftheriadis, F. J. Gómez-Moreno, E. Alonso-Blanco, A. Wiedensohler, K. Weinhold, M. Merkel, S. Bastian, B. Hoffmann, H. Altug, J.-E. Petit, P. Acharja, O. Favez, S. M. D. Santos, J.-P. Putaud, A. Dinoi, D. Contini, A. Casans, J. A. Casquero-Vera, S. Crumeyrolle, E. Bourrianne, M. V. Poppel, F. E. Dreesen, S. Harni, H. Timonen, J. Lampilahti, T. Petäjä, M. Pandolfi, P. K. Hopke, R. M. Harrison, A. Alastuey, and X. Querol (2024). Source apportionment of ultrafine particles in urban Europe. *Environment International 194*, 109149.

Gasparrini, A., B. Armstrong, and M. G. Kenward (2010). Distributed lag non-linear models. *Statistics in medicine 29*(21), 2224–2234.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A., B. Goodrich, J. Gabry, and A. Vehtari (2019). R-squared for Bayesian regression models. *The American Statistician*.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su (2008). A weakly informative default prior distribution for logistic and other regression models.

Grazian, C. (2023). A review on bayesian model-based clustering. *arXiv preprint arXiv:2303.17182*.

Gu, J., M. Pitz, J. Schnelle-Kreis, J. Diemer, A. Reller, R. Zimmermann, J. Soentgen, M. Stoelzel, H.-E. Wichmann, A. Peters, et al. (2011). Source apportionment of ambient particles: comparison of positive matrix factorization analysis applied to particle size distribution and chemical composition data. *Atmospheric Environment 45*(10), 1849–1857.

Hackstadt, A. J. and R. D. Peng (2014). A Bayesian multivariate receptor model for estimating source contributions to particulate matter pollution using national databases. *Environmetrics 25*(7), 513–527.

Halonen, J. I., T. Lanki, T. Yli-Tuomi, P. Tiittanen, M. Kulmala, and J. Pekkanen (2009). Particulate air pollution and acute cardiorespiratory hospital admissions and mortality among the elderly. *Epidemiology 20*(1), 143–153.

Harrison, R. M., D. Beddows, M. S. Alam, A. Singh, J. Brean, R. Xu, S. Kotthaus, and S. Grimmond (2019). Interpretation of particle number size distributions measured across an urban area during the faster campaign. *Atmospheric Chemistry and Physics 19*(1), 39–55.

Hopke, P. K. (2016). Review of receptor modeling methods for source apportionment. *Journal of the Air & Waste Management Association 66*(3), 237–259.

Lee, D., S. Mukhopadhyay, A. Rushworth, and S. K. Sahu (2017). A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health. *Biostatistics 18*(2), 370–385.

Martínez-Hernández, I., C. Euán, W. S. Burr, M. Meis, M. Blangiardo, and M. Pirani (2025). Modelling particle number size distribution: a continuous approach. *Journal of the Royal Statistical Society Series C: Applied Statistics 74*(1), 229–248.

Mezzetti, M., C. P. Ryan, P. Balestrucci, F. Lacquaniti, and A. Moscatelli (2023). Bayesian hierarchical models and prior elicitation for fitting psychometric functions. *Frontiers in Computational Neuroscience 17*, 1108311.

Molitor, J., E. Coker, M. Jerrett, B. Ritz, A. Li, et al. (2016). Part 3. modeling of multipollutant profiles and spatially varying health effects with applications to indicators of adverse birth outcomes. *Research report (Health Effects Institute)* (183 Pt 3), 3–47.

Pandolfi, M., D. Mooibroek, P. Hopke, D. van Pinxteren, X. Querol, H. Herrmann, A. Alastuey, O. Favez, C. Hüglin, E. Perdrix, et al. (2020). Long-range and local air pollution: what can we learn from chemical speciation of particulate matter at paired sites? *Atmospheric Chemistry and Physics 20*(1), 409–429.

Park, E. S., P. Guttorp, and R. C. Henry (2001). Multivariate receptor modeling for temporally correlated data by using MCMC. *Journal of the American Statistical Association 96*(456), 1171–1183.

Park, E. S., P. K. Hopke, M.-S. Oh, E. Symanski, D. Han, and C. H. Spiegelman (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.

Park, E. S., E.-K. Lee, and M.-S. Oh (2021). Bayesian multivariate receptor modeling software: Bnfa and bayesmrm. *Chemometrics and Intelligent Laboratory Systems 211*, 104280.

Park, E. S. and M.-S. Oh (2015). Robust Bayesian multivariate receptor modeling. *Chemometrics and Intelligent Laboratory Systems 149*, 215–226.

Park, E. S., E. Symanski, D. Han, and C. Spiegelman (2015). Part 2. development of enhanced statistical methods for assessing health effects associated with an unknown number of major sources of multiple air pollutants. *Research report (Health Effects Institute)* (183 Pt 1-2), 51–113.

Perperoglou, A., W. Sauerbrei, M. Abrahamowicz, and M. Schmid (2019). A review of spline function procedures in R. *BMC Medical Research Methodology 19*(1), 46.

Pirani, M., N. Best, M. Blangiardo, S. Liverani, R. W. Atkinson, and G. W. Fuller (2015). Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. *Environment International 79*, 56–64.

Pirani, M., A. Panton, D. A. Purdie, and S. K. Sahu (2016). Modelling macronutrient dynamics in the Hampshire Avon river: A Bayesian approach to estimate seasonal variability and total flux. *Science of the Total Environment 572*, 1449–1460.

Rivas, I., D. C. Beddows, F. Amato, D. C. Green, L. Järvi, C. Hueglin, C. Reche, H. Timonen, G. W. Fuller, J. V. Niemi, et al. (2020). Source apportionment of particle number size distribution in urban background and traffic stations in four European cities. *Environment International 135*, 105345.

Rivas, I., L. Vicens, X. Basagaña, A. Tobías, K. Katsouyanni, H. Walton, C. Hüglin, A. Alastuey, M. Kulmala, R. M. Harrison, et al. (2021). Associations between sources of particle number and mortality in four European cities. *Environment International 155*, 106662.

Rusanen, A., A. Björklund, M. Manousakas, J. Jiang, M. T. Kulmala, K. Puolamäki, and K. R. Daellenbach (2023). A novel probabilistic source apportionment approach: Bayesian auto-correlated matrix factorization. *Atmospheric Measurement Techniques Discussions 2023*, 1–28.

Samoli, E., R. W. Atkinson, A. Analitis, G. W. Fuller, D. Beddows, D. C. Green, I. S. Mudway, R. M. Harrison, H. R. Anderson, and F. J. Kelly (2016). Differential health effects of short-term exposure to source-specific particles in london, uk. *Environment International 97*, 246–253.

Szpiro, A. A. and C. J. Paciorek (2013). Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics 24*(8), 501–517.

Tang, J.-H., S.-C. C. Lung, and J.-S. Hwang (2020). Source apportionment of pm2.5 concentrations with a Bayesian hierarchical model on latent source profiles. *Atmospheric Pollution Research 11*(10), 1715–1727.

Tremper, A. H., C. Jephcote, J. Gulliver, L. Hibbs, D. C. Green, A. Font, M. Priestman, A. L. Hansell, and G. W. Fuller (2022). Sources of particle number concentration and noise near London Gatwick Airport. *Environment International 161*, 107092.

Viana, M., T. Kuhlbusch, X. Querol, A. Alastuey, R. Harrison, P. Hopke, W. Winiwarter, M. Vallius, S. Szidat, A. Prevot, C. Hueglin, H. Bloemen, P. Wahlin, R. Vecchi, A. Miranda, A. Kasper-Giebl, W. Maenhaut, and R. Hitzenberger (2008). Source apportionment of particulate matter in Europe: a review of methods and results. *Journal of Aerosol Science 39*, 827–849.

Villejo, S. J., J. B. Illian, and B. Swallow (2023). Data fusion in a two-stage spatio-temporal model using the INLA-SPDE approach. *Spatial Statistics 54*, 100744.

Wade, S. (2023). Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A 381*(2247), 20220149.

WHO (2021). *WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide.* Geneva, Switzerland: World Health Organization.

WHO (2025). Air pollution and sustainable development goals (sdgs).

Wiedensohler, A., W. Birmili, A. Nowak, A. Sonntag, K. Weinhold, M. Merkel, B. Wehner, T. Tuch, S. Pfeifer, M. Fiebig, et al. (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.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Yu, L., W. Liu, X. Wang, Z. Ye, Q. Tan, W. Qiu, X. Nie, M. Li, B. Wang, and W. Chen (2022). A review of practical statistical methods used in epidemiological studies to estimate the health effects of multi-pollutant mixture. *Environmental Pollution 306*, 119356.