POLITECNICO MILANO 1863

Exposure Dynamics in Lombardy: Geospatial Analysis of Presence Data & Pollution Fields

Authors: Ilenia Di Battista, Leonardo Marchesin, Elettra Zuliani
Github repository: https://github.com/LeoMarche24/GDEH_Project

Doctoral course: Geospatial Data Science for Environmental Health
Professor: Prof. Lorenzo Gianquintieri
Academic year: 2025

---

1. Introduction

Air pollution is one of the most pressing global challenges, with consequences that demand urgent attention. Understanding the spatio-temporal distribution and implications of atmospheric pollutants is essential for public health impacts. For instance, particulate matter with a diameter smaller than 10 micrometres (PM_{10}), is a dangerous component of the atmosphere, and its effects shall be investigated, particularly in highly populated and industrialised areas. As the studies on air quality increase in precision and effectiveness, environmental datasets are becoming increasingly large and complex, calling for advanced mathematical modelling techniques to extract meaningful insights.
Therefore, the aim of this work is to adapt an advanced mathematical framework to the analysis of population exposure to PM_{10}, with the goal of providing valuable evidence to support environmental monitoring strategies and policies designed to mitigate the adverse effects of air pollution.
The Lombardy region, situated in the Po Valley, is one of the most critical areas for air quality in Europe, due to both geographical and anthropogenic factors. The region's morphology is particularly unfavourable: the Po Valley is landlocked, since enclosed by the Alps to the North and West, and the Apennines to the South, with a unique opening to the Adriatic Sea on the East. This creates limited atmospheric circulation, which worsens air quality. During winter, this issue is even intensified by thermal inversion, a phenomenon that occurs in areas with low air circulation, when the ground is cold. From the anthropogenic view point Lombardy is characterised by large urban centres, such as the metropolitan city of Milan, where high levels of heating and traffic are significant contributors to atmospheric pollution. Moreover, the region serves as Italy's primary industrial and production hub, hosting numerous factories that contribute significantly to emissions, thereby increasing pollutant concentrations in the air. As the most populated region in Italy, Lombardy is particularly fragile to the air quality effects on the population. Furthermore, it is one of the most densely populated regions in both Italy and Europe. Indeed, the municipality of *Bresso*, which lies in the neighborhood of Milan, is considered the most densely populated municipality in Italy. All these characteristics make Lombardy particularly vulnerable to air pollution, as also highlighted by the European Environmental Agency (EEA) [5], which reports that concentrations exceeding the European daily limit value for PM_{10}, set to 50 $\mu g/m^3$, occur predominantly in Italy, and specifically within the Po Valley.
In this work, we analyse population exposure to PM_{10} [9] by combining pollutant concentrations, estimated with a spatio-temporal smoothing model, with presence data describing how many people are in a given place at a given time. In this framework, exposure reflects the pollutant levels actually experienced by individuals according to their spatial and temporal distribution. This concept is particularly relevant in a region such as Lombardy, where high pollution

levels intersect with intense population density and mobility. Rather than considering air quality as a purely environmental variable, exposure provides a direct link to potential health risks, as it quantifies how pollution interacts with the presence and movement of people across the territory. By integrating both the environmental and demographic dimensions, the analysis aims to capture not only areas where pollution is high, but also where and when it most directly affects the population.

We focus our analysis on the period going from 13^{th} December 2021 to 13^{th} January 2022 in order to analyse possible patterns during the Christmas holidays. The PM_{10} data are collected daily from the 66 monitoring stations, managed by Agenzia Regionale per la Protezione dell'Ambiente (ARPA) [2], and located on the territory on the basis of population density. Figure 1 illustrates PM_{10} concentration data, measured in $[\mu g/m^3]$, for four selected days within the study period. The highest levels are consistently found in the metropolitan area of Milan, followed by the cities of Brescia and Bergamo and the industrial zones of the Po Valley, whereas the mountain regions show significant lower values, highlighting a complex spatial pattern. Substantial temporal variations are also evident: for example, concentrations are very high on 1^{st} January, in contrast with the uniformly low levels observed across the region on 6^{th} January, likely due to intense precipitation events.

The presence data employed in this study were provided by PoliS-Lombardia in collaboration with Vodafone Business Italy, and are not publicly available. For each day and for each of the 198 areas in Lombardy considered in their analysis, the dataset records the number of people present, inferred from aggregated mobile phone traffic. These data are obtained through Vodafone Analytics, which processes billions of daily geolocation signals from the mobile network and applies statistical procedures to generate anonymised and aggregated estimates of population distribution. In compliance with the General Data Protection Regulation (GDPR), no individual-level information is accessible, and all outputs are strictly aggregated before being released. The measurements refer exclusively to Vodafone users. However, data are projected on

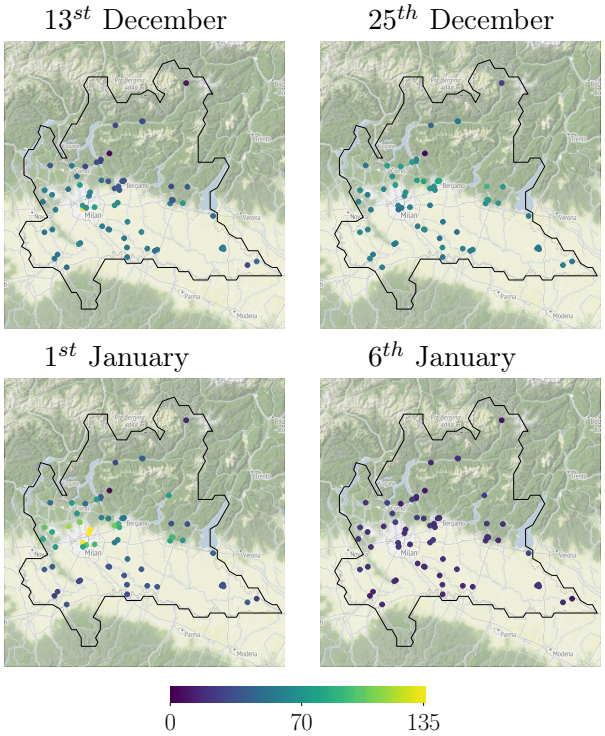

Figure 1: PM_{10} data collected on 13^{th}, 25^{th} December 2021, and 1^{st}, 6^{th} January 2022, measured in $[\mu g/m^3]$.

the whole population, as an estimate of the total presences in the recorded area and time.

Although approximate, these presence data offer a robust proxy for the spatial and temporal distribution of the population. By capturing residential and mobility patterns, they provide insights into how population density varies across time and space. This is particularly relevant in a region such as Lombardy, where commuting, urban-rural exchanges, and seasonal variations strongly influence the actual number of people exposed in a given area.

Ultimately, this study aims to provide a comprehensive framework that integrates environmental and demographic data, offering new insights into the dynamics of air pollution exposure in Lombardy and creating the conditions for more informed environmental health policies.

2. Spatio-Temporal Smoothing Model

In order to analyse the spatio-temporal distribution of PM_{10} concentration we employ a non-parametric spatio-temporal regression model. The model belongs to the family of Spatial Re-

gression models with Partial Differential Equation regularisation, reviewed by Sangalli in [13]. This family of methods enables the analysis of complex spatial and spatio-temporal phenomena by incorporating information about the spatial domain of interest, which may be irregularly shaped, non-convex, or curved; and information about the data representing the phenomenon under study. The model does not make any parametric assumption on the covariance structure, as the spatial and spatio-temporal dependence is modelled through the regularising terms, nor does it assume any separability in space-time allowing for great flexibility. Importantly, the proposed model can deal with irregularly sampled data and can accurately handle problems affected by severe missing-data patterns. This is a crucial modelling feature when dealing with space-time data, especially in environmental sciences. In fact, data recorded by sensors, like meteorological and climate control units and environmental measuring stations, often exhibit missing entries, due to malfunctioning of the device or other specific conditions. Specifically, in our study on PM_{10} distribution, 2.7% of the spatio-temporal data are missing. Although this is not a large proportion, it still requires appropriate handling.

Let $\{\mathbf{p}_i\}_{i=1,\dots,n}$ be a set of n spatial locations over a bounded spatial domain $\mathcal{D} \subset \mathbb{R}^2$, and $\{t_j\}_{j=1,\dots,m}$ be a set of m time points in a time interval $[0,T] \subset \mathbb{R}$. Let $\{y_{ij}\}_{i=1,\dots,n,\, j=1,\dots,m}$ be the realisations of a real random variable Y_{ij}, measured at the space-time location $(\mathbf{p}_i, t_j)$. For convenience, we define the set of indices corresponding to observed data, as

$$O = \{(i,j) : y_{ij} \text{ is observed}, \\ i = 1,...,n,\ j = 1,...,m\},$$

denoting by $|O|$ its cardinality. Thus, we model $\{y_{ij} : (i,j) \in O\}$ as noisy observations of an underlying spatio-temporal smooth function $f(\mathbf{p},t)$ on $\mathcal{D} \times [0,T]$:

$$y_{ij} = f(\mathbf{p}_i, t_j) + \epsilon_{ij} \quad (i,j) \in O, \tag{1}$$

where $\{\epsilon_{ij} : (i,j) \in O\}$ are independently distributed residuals with zero mean and constant variance σ^2.

We estimate the unknown field f by minimising the following penalised functional

$$J(f) = \sum_{(i,j)\in O} (y_{ij} - f(\mathbf{p}_i, t_j))^2 + \\ + \lambda_D \int_0^T \int_{\mathcal{D}} (\Delta f)^2 d\mathbf{p}\, dt + \\ + \lambda_T \int_0^T \int_{\mathcal{D}} \left(\frac{\partial^2 f}{\partial t^2}\right)^2 d\mathbf{p}\, dt, \tag{2}$$

where $\lambda_D > 0$ and $\lambda_T > 0$ are two positive smoothing parameters weighting the two penalty terms in space and time, thereby ensuring smoothness of the solution along both dimensions. These parameters are selected using the Generalised Cross-Validation (GCV) criterion. In particular, we choose (λ_D, λ_T) by minimising the GCV score, defined as

$$GCV(\lambda_D, \lambda_T) = \sum_{(i,j)\in O} \frac{(y_{ij} - \hat{f}(\mathbf{p}_i, t_j))^2}{|O| - df}, \tag{3}$$

where $\hat{f}$ is the estimate obtained by the model and *df* are the corresponding effective degrees of freedom of the model.

It is worth noting that the minimisation of functional (2) constitutes an infinite-dimensional problem, which does not admit a closed-form solution. Therefore, we adopt a convenient finite-dimensional discretisation of the estimation problem. Specifically, as detailed in Arnone et al. [1], we employ the Finite Element Method in space and cubic B-spline interpolation in time to obtain a finite-dimensional estimate. The choice of discretisation fineness represents a trade-off between accuracy and computational cost: the spatial and temporal bases should be sufficiently rich to capture the localised features of the signal, whereas overly fine bases are unnecessary and increase computational burden, as discussed in [3]. The proposed method is implemented in the `R/C++` library `fdaPDE`, available on GitHub [10].

Figure 2 displays the estimated spatio-temporal PM_{10} concentration field obtained from the proposed regression model, corresponding to the observed data in Figure 1. The smoothed fields confirm that Milan and its surrounding area exhibit the highest pollutant concentrations, followed by Brescia and parts of the Po Valley. In contrast, the northern mountain areas consistently experience cleaner air, although an in-

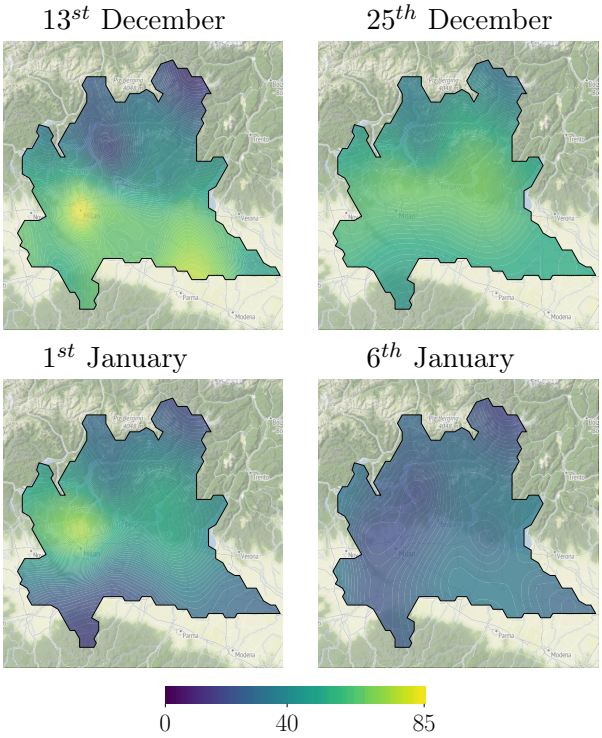

Figure 2: PM_{10} smooth field estimated by the proposed Spatio-Temporal Regression Model on 13^{th}, 25^{th} December 2021, and 1^{st}, 6^{th} January 2022.

crease in pollution is noticeable on 25^{th} December, likely related to the increased tourism activity in the mountains region. On 6^{th} January, by comparison, the entire region benefited from lower pollution levels, in agreement with the data in Figure 2.

3. Exposure Analysis

3.1. Objectives

In this section we present the analysis of the population exposure to PM_{10} in Lombardy, combining the pollutant concentration estimates displayed in Figure 2 with the population presence data [7, 15]. The spatial smoothing procedure described in Section 2 provides a reliable basis for estimating exposure, capturing pollution patterns even in areas where ground monitoring stations are sparse or absent. The analysis thus shifts from purely environmental modelling to an assessment of how estimated pollution levels translate into actual population-level exposure. The smoothing estimation is integrated with the monitoring system of mobile phone data, as described in Section 1, to provide an estimate of the exposure within the Lombardy region.

3.2. Geolocation and Data Management

To link presence data with the pollution field, we apply a dedicated geolocation procedure. Presence data, provided by PoliS-Lombardia, consists of unique aggregated records associated with predefined areas, e.g. *Alta Bergamasca*, *Brianza* (198 in total), which do not always correspond to official administrative boundaries. We use the `tidygeocoder` [4] package in `R` [11] to perform geocoding via OpenStreetMap [8], assigning each record an estimate of its latitude and longitude. Each geocoded point is then assigned to the municipality polygon in which it lies, yielding a total of 177 municipalities for which presence data are available. Note that whenever a municipality was referred to multiple times by different areas (e.g., area of *Milano* and *Nord Milano*), the values of the presences have been summed.

Presence data were coded with a specified coordinate reference system (CRS), specifically $CRS = 4326$. On the other hand, the shape files of the Lombardy provinces and municipalities were in the standard $CRS = 32632$. Therefore, presence data have been transformed via QGIS [14], to monitor the transformation. After this initial step, the whole procedure has been developed in `R` [11] via the geographic information system incorporated in the `tidygeocoder` package [4].

To match the spatial granularity of the presence data, we aggregate the PM_{10} concentration field at the municipality level computing its area-average concentration. We perform this procedure only for the municipalities with available presence data. Let the days be indexed by $j = 1, \dots, I$, and municipalities by $c \in \{Abbiategrasso, \dots, Volta\ Mantovana\}$. For each municipality c, the average smoothed field is computed as

$$S_j^c := \frac{1}{|c|} \int_c f_j(p)\, dp,$$

where $|c|$ denotes the area of the municipality and the integral is performed over its spatial domain. Algorithmically, since the smoothed field is evaluated over a grid of points, the integral is approximated as the average of the pollution

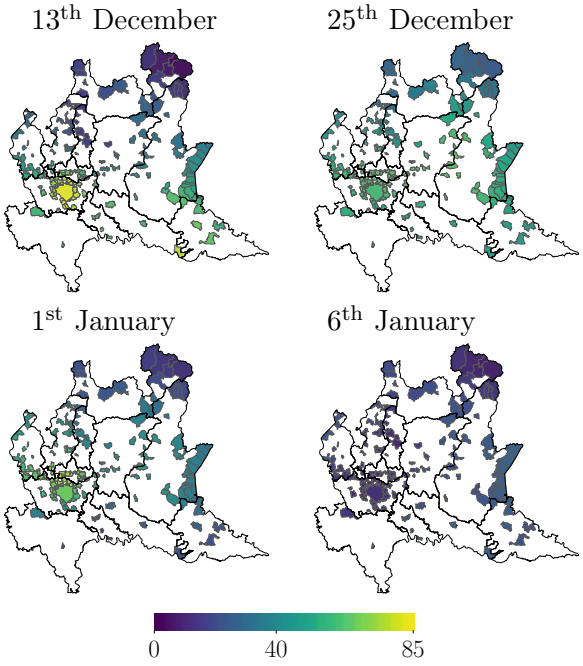

Figure 3: Aggregation of the PM_{10} smoothed field across the municipalities with available presence data.

field on all the grid points within the municipality of interest. The selection of the points inside the municipality has been performed with a clipping procedure, combining the coordinate of the smoothed field (already given in $CRS = 32632$), and the polygon of the municipality. The same geographic information system [4] has been adopted for this task.

Figure 3 shows the mean PM_{10} concentration field aggregated at the municipality level, for some days within the study period. As we already discussed, the aggregation is performed only for municipalities with available presence data, leaving substantial areas of the region without coverage, as illustrated in the figure.

3.3. Exposure Metric

We now turn to the exposure analysis, where we combine the daily aggregated mean values of PM_{10} concentration with the corresponding daily presence measurements. We conduct the analysis at the provincial level, as many municipalities lack available presence data, as we see from Figure 3.

For each province, we compute a weighted average exposure, where weights reflects the relative presence distribution across municipalities [6, 16]. This procedure is repeated for each day within the study period, thus capturing both spatial and temporal variations in exposure across the Lombardy region.

For days indexed by $j = 1, \dots, I$ and provinces $k \in \{Bergamo, \dots, Varese\}$, we evaluate the exposure E_j^k. Let S_j^c denote the average smoothed PM_{10} concentration in municipality c on day j, and let U_j^c denote the corresponding presence counts. We then define the exposure in province k on day j as

$$E_j^k = \frac{\sum_c S_j^c \times U_j^c}{\sum_c U_j^c}. \tag{4}$$

Figure 4 illustrates the exposure levels across Lombardy provinces for four representative days within the study period. The patterns are consistent with the smoothed concentration field reported in Figure 2. In particular, a clear spatial gradient emerges: the provinces of *Milano* and *Monza* register the most critical exposure levels, with *Monza* even surpassing *Milano* on 1st January. On the contrary, both provinces show markedly lower exposure levels on 25th December, plausibly reflecting reduced emissions and a smaller resident population due to holiday travel. Coherently, on this same day, the exposure in mountain provinces increases relative to the rest of the period, thereby highlighting a pronounced temporal variation as well.

To further investigate this temporal variation, Figure 5 displays the temporal trends of PM_{10} exposure, capturing both day-to-day fluctuations in exposure dynamics and differences across provinces. This representation confirms the spatial heterogeneity and temporal evolution of PM_{10} exposure already highlighted in Figure 4. In addition, we observe comparable dynamics among certain provinces. For example, *Milano* and *Monza*, shown in violet and dark red, display closely aligned temporal trajectories, while provinces in the Po Valley, represented in green, shows similar patterns as well. In the next section, we further investigate these groups' dynamic.

3.4. Further interpretation

While the daily exposure time series in Figure 5 provided a first overview of temporal dynamics across provinces, the superposition of multiple functional lines limited the interpretability of spatio-temporal patterns. To further investi-

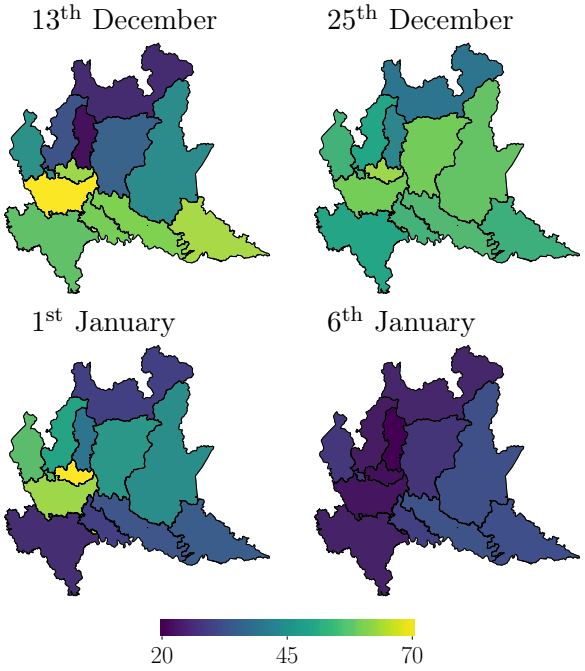

Figure 4: Population exposure to PM_{10} evaluated for the Lombardy provinces in four days during the period.

gate this aspect, we apply a Functional Principal Component Analysis (fPCA), treating each provincial exposure curve as a function and applying the well-established PCA procedure in the functional domain [12]. The decomposition proved to be highly efficient, with the first two components alone explaining about 92% of the total variability, thus offering a reliable and concise summary of the main exposure dynamics.
Figure 6 displays the first two components of the fPCA. Firstly, the mean of the exposure functions (the black line in both pictures) reflects the dynamic pointed out in Figure 4. In fact, the general magnitude of the functions decreases in the holiday period, as indicated by the mean. The principal components in the functional domain can be seen as a variation on the mean. The blue line stands for the behaviour of a functional datum which has a positive score in that component, while the red one stands for a negative value of the score.
The first component can be interpreted as a measure of the general magnitude of exposure, where we notice a particular difference between positive and negative effects especially during the working days, namely the starting and ending point of our time domain. Provinces with a high value of the first component are linked to

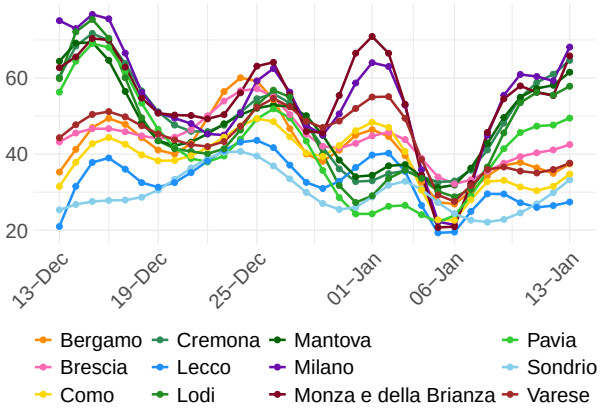

Figure 5: Temporal trends of the population exposure to PM_{10} for each province in Lombardy.

a general level of exposure which is above the mean, especially in the working days.
In contrast, the second component highlights what we refer to as a *holiday effect*. A positive score on this axis is associated with provinces that experience higher exposure levels during the Christmas and New Year period, with respect to the mean. On the contrary, negative or low scores characterise provinces where exposure is not so significant during the same interval. In this sense, the fPCA offers a way to separate baseline exposure levels from temporal fluctuations, such as those associated with festive periods.
Figures 7 and 8 display the values of the two principal components associated to each province. This analysis illustrates a clear spatial structure. For instance, the four provinces located along the Po Valley, bordering the river in the southern part of Lombardy, exhibit high values on the first component but low on the second, reflecting their consistently high background exposure while experiencing limited exposure during the holiday period compared to the other provinces. In another case, *Milano* and *Monza* score high on both components, the first highlighting their generally elevated exposure, and the second indicating that their exposure remains above the regional average even during the holiday period, although the temporal trends in Figure 5 reveal a slight reduction in *Milano* and *Monza* compared to working days.
Overall, the fPCA approach proves valuable in providing a compact and interpretable representation of exposure dynamics, allowing us to separate the contribution of general pollution levels

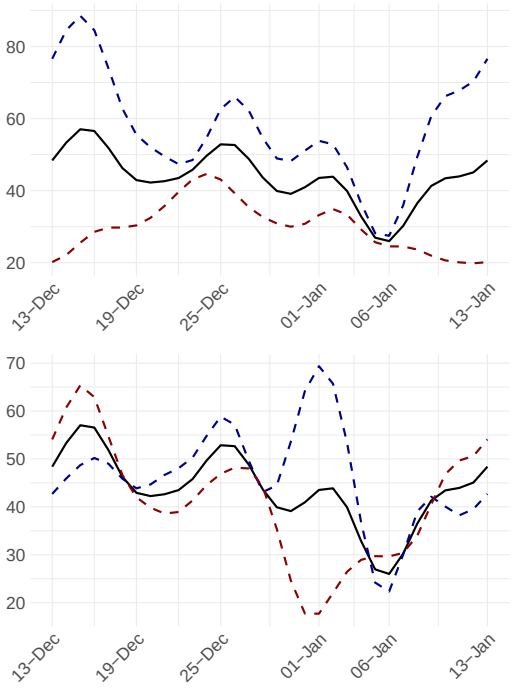

Figure 6: First (top panel) and second (bottom panel) principal component of the fPCA. The solid black line is the mean of the functions, the blue dashed line is the effect of a positive score and the red dashed line is the effect of a negative score.

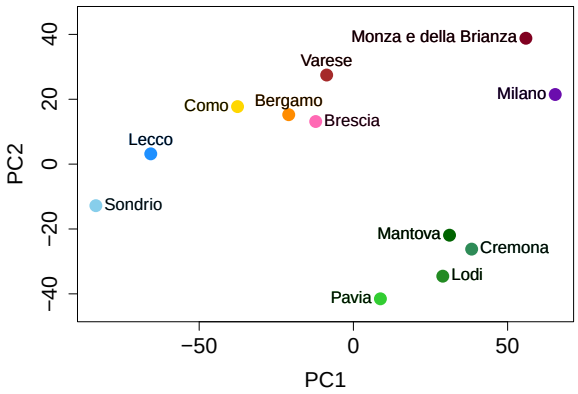

Figure 7: Provincial scores of the first two principal components of the fPCA represented in the plane PC1-PC2.

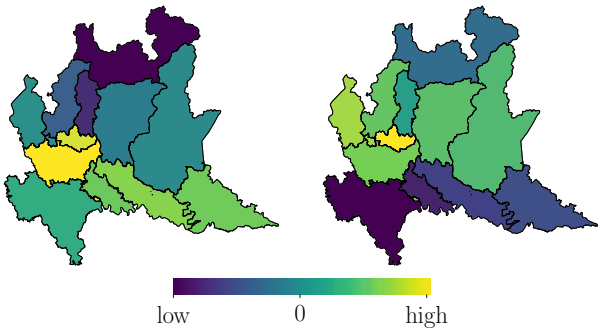

Figure 8: Spatial distribution of the first two principal components scores of the fPCA.

from temporal anomalies linked to specific periods, thus offering insights that are not immediately apparent from a direct inspection of the time series alone.

4. Conclusions

In this work, we presented a spatio-temporal exposure analysis combining smoothed pollution fields with presence data. The proposed modelling approach allowed us to account for both spatial and temporal variability, offering a perspective that may be useful for environmental health applications and for supporting policy development. By integrating presence data derived from mobility assessments, we aimed to capture daily fluctuations and provide a more dynamic picture of population exposure.

Nevertheless, some limitations should be noted. Presence data represent only an approximation of the actual population distribution, as they are based on Vodafone mobile phone users and exclude other groups. Moreover, the geolocation of presence records is based on approximations, since the reference areas do not always align with administrative boundaries. Concerning the smoothing of PM_{10} field, the analysis is affected by the relatively small number of monitoring stations, namely 66 across the whole region. Although the spatio-temporal smoothing improves reliability, uncertainties remain, especially in areas distant from monitoring sites.

Despite these limitations, the analysis highlighted meaningful patterns in exposure to PM_{10}. Figures 4 and 5 show that the metropolitan provinces of *Milano* and *Monza* experience the highest exposure levels, but also a notable reduction during Christmas holidays, reflecting the tendency of residents to leave the city and the associated decrease in emissions. Similar reductions related to vacation were observed in the Po Valley, possibly due to reduced industrial activity, while mountain provinces, though generally less exposed, showed a slight incasement during the same period, consistent with seasonal

activity induced by vacations.
The functional principal component analysis (fPCA) of provincial time series further clarified these dynamics. The first component distinguished provinces with consistently higher concentrations, such as *Milano* and *Monza*, and the ones in the Po Valley, from those with lower levels, such as *Sondrio* and *Lecco*. The second component captured a holiday-related effect, with Po Valley provinces stressing the pronounced reduction in exposure levels in the area during the holiday period.
Overall, the study demonstrates the value of robust mathematical modelling for environmental exposure analysis, contributing to a better understanding of air pollution dynamics and their potential health implications. Future research could benefit from testing alternative sources of presence data, extending the analysis to different spatial and temporal domains, and including multiple pollutants to develop a more comprehensive assessment of air quality.

5. Acknowledgements

This project was developed with the support of Artificial Intelligence (AI) tools. AI was used for: generating and refactoring R code, improving code readability and documentation, drafting and revising text. All results, analyses, and interpretations were reviewed and validated by the authors.

References

[1] E. Arnone, L.M. Sangalli, and A. Vicini. Smoothing spatio-temporal data with complex missing data patterns. *Statistical Modelling*, 23:327–356, 2023.

[2] ARPA Lombardia. Agenzia regionale per la protezione dell'ambiente della lombardia, 2025. Accessed: 2025-09-24.

[3] L. Azzimonti, F. Nobile, L. M. Sangalli, and P. Secchi. Mixed finite elements for spatial regression with pde penalization. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):305–335, 2014.

[4] Jesse Cambon. *tidygeocoder: Geocoding Made Easy*, 2024. R package version 1.0.5.

[5] European Environment Agency. Particulate matter - pm10, 2025.

[6] Ivan C. Hanigan, Geoffrey G. Morgan, and Fay H. Johnston. A comparison of methods for calculating population exposure estimates of daily weather for health research. *International Journal of Health Geographics*, 5(1):38, 2006.

[7] Y. Liu, X. Chen, and Y. Zhang. Population exposure to pm10 in china: A model-based estimation. *International Journal of Environmental Research and Public Health*, 10(1):210–222, 2013.

[8] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org. `https://www.openstreetmap.org`, 2017.

[9] World Health Organization. Exposure assessment. In *Principles for the Assessment of Risks to Human Health from Exposure to Chemicals*. WHO, 1999.

[10] A. Palummo, E. Arnone, A. Clemente, L. M. Sangalli, J. Ramsay, and L. Formaggia. fdapde: Physics-informed spatial and functional data analysis. GitHub, `https://github.com/fdaPDE/fdaPDE`, 2025.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[12] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2nd edition, 2005.

[13] L.M. Sangalli. Spatial regression with partial differential equation regularisation. *International Statistical Review*, 89:505–531, 2021.

[14] QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2025.

[15] James D. Wilson. *Exposure Assessment*. Wiley, 2021.

[16] L. Zhao, Y. Chen, X. Wu, and J. Li. Improving assessment of population exposure and health risks: An integrated framework combining environmental monitoring and demographic data. *Journal of Exposure Science & Environmental Epidemiology*, 2024.