Chemosphere 352 (2024) 141438

Contents lists available at ScienceDirect

Chemosphere

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

Implementation of a GEOAI model to assess the impact of agricultural land on the spatial distribution of PM2.5 concentration

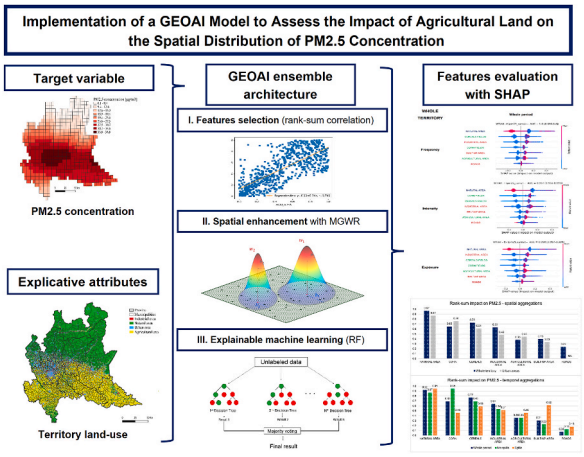
Lorenzo Gianquintieri [a,*], Daniele Oxoli [b], Enrico Gianluca Caiani [a,c], Maria Antonia Brovelli [b]

[a] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy
[b] Department of Civil and Environmental Engineering, Politecnico di Milano, Milan, Italy
[c] IRCCS Istituto Auxologico Italiano, Milan, Italy

HIGHLIGHTS

- GEOAI applied for the analysis of Atmospheric Composition global model data.
- First framework successfully assessing the impact of land-use on PM2.5 concentration.
- Agriculture causes PM2.5 exposition comparably to more studied urban sources.
- Corn and cereals cultures are correlated with PM2.5 concentration, rice don't.
- Effect of agriculture on PM2.5 concentration also recorded in urbanized areas.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling editor: R Ebinghaus

Keywords:
Pollution
PM
Agriculture
GEOAI
Environmental impact assessment

ABSTRACT

Air pollution is considered one of the major environmental risks to health worldwide. Researchers are making significant efforts to study it, thanks to state-of-art technologies in data collection and processing, and to mitigate its effect. In this context, while a lot is known about the role of urbanization, industries, and transport, the impact of agricultural activities on the spatial distribution of pollution is less studied, despite knowledge about emissions suggest it is not a secondary factor. Therefore, the aim of this study was to assess this impact, and to compare it with that of traditional polluting sources, harvesting the capabilities of GEOAI (Geomatics and Earth Observation Artificial Intelligence). The analysis targeted the highly polluted territory of Lombardy, Italy, considering fine particulate matter (PM2.5). PM2.5 data were obtained from the Copernicus-Atmosphere-Monitoring-Service and processed to infer time-invariant spatial parameters (frequency, intensity and exposure) of concentration across the whole period. An ensemble architecture was implemented, with three blocks: correlation-based features selection, Multiscale-Geographically-Weighted-Regression for spatial enhancement, and a final random forest classifier. Finally, the SHapley Additive exPlanation algorithm was applied to compute the relevance of the

* Corresponding author. Via Ponzio 34, Milano, MI, 20133, Italy.
E-mail address: lorenzo.gianquintieri@polimi.it (L. Gianquintieri).

https://doi.org/10.1016/j.chemosphere.2024.141438
Received 4 October 2023; Received in revised form 8 February 2024; Accepted 9 February 2024
Available online 15 February 2024

different land-use classes on the model. The impact of land-use classes was found significantly higher compared to other published models, showing that the insignificant correlations found in the literature are probably due to an unfit experimental setup. The impact of agricultural activities on the spatial distribution of PM2.5 concentration was comparable to the other considered sources, even when focusing only on the most densely inhabited urban areas. In particular, the agriculture's contribution resulted in pollution spikes rather than in a baseline increase. These results allow to state that public policymakers should consider also agricultural activities for evidence-based decision-making about pollution mitigation.

1. Introduction

Air pollution is considered, by the United Nations, one of the major environmental risks to health worldwide. Accordingly, it is addressed for mitigation in multiple United Nations Sustainable Development Goals, such as the 3.9 and 11.6 (Rafaj et al., 2018). As a matter of fact, the concentration in the troposphere of air pollutants and greenhouse gases increased exponentially in the last century, mainly due to anthropic activities including energy production, domestic heating, industrial processes, transport, and intensive farming.

Among all pollutants, one of the major concerns is Particulate Matter (PM) with a diameter smaller than 2.5 μm, PM2.5, also due to its longer residence time in the atmosphere, resulting in an increased risk of exposure for the population (Awe et al., 2022). Its correlation with severe health issues for humans is well assessed, and is in particular associated with increased morbidity and mortality of cardiopulmonary diseases (Xing et al., 2016).

The scientific community has been increasing its efforts to study and mitigate the phenomenon, with a huge production of literature related to different aspects, including emissions, chemical generation processes, diffusion and concentration levels. One of the main issues when dealing with PM2.5 is the need for a dense network of ground stations or sensors recording the concentration levels. To address the technical and economic challenges posed by this requirement, models for PM concentration were developed (Gardner-Frolick et al., 2022; Zaini et al., 2022; Mehmood et al., 2022, Gugnani and Singh, 2022; Gianquintieri et al., 2023b).

These models have gained momentum thanks to the diffusion of Geomatics and Earth Observation Artificial Intelligence (GEOAI) techniques, applying machine learning algorithms. Modelling PM concentration has several purposes, mainly for the interpolation of data in space or time and concentrations forecasting, but it is also widely adopted to study the generating sources through correlation analyses between environmental variables and the levels of PM concentration. The most commonly considered environmental factors (Gianquintieri et al., 2023b) are the meteorological variables (such as temperature, humidity, rainfalls, wind and more), and land-use classes, including built-up, roads, industrial areas, natural and agricultural lands.

While there is a wealth of knowledge about the emissions generated by urban areas, industries, and transport, less attention has been focused on agricultural activities, not usually considered among the traditional air pollution sources, despite being a significant contributor (McDuffie et al., 2020). In fact, agriculture is a well-known source of Ammonia (NH3) (Sapek, 2013) a pollutant associated to a low level of toxicity, also due to its short residence time in the atmosphere (Zhu et al., 2015); accordingly, in the European Union there is no concentration threshold for NH3 set by law. Still, its negative impact on human health has been addressed in the scientific literature (Higashiyama et al., 2007, Neghab et al. 2018), also in specific relation to agricultural activity (Loftus et al., 2015; Wyer et al., 2022), and it was recently associated with the increased velocity of diffusion of COVID-19 in Lombardy region, Italy, during the first pandemic peak (Gianquintieri et al., 2021). Moreover, NH3 is a precursor of secondary PM2.5: scientific literature suggests that the reduction of NH3 emissions represents a necessary action to reduce PM pollution (Erisman and Schaap, 2004; Wu et al., 2016). In addition to this indirect link, in which the contribution of agricultural activities to PM2.5 concentration passes through NH3 generation, a more direct association is due to residual burning activities, a factor that is receiving increasing attention from researchers (Jethva et al., 2019; Bray et al., 2019; Liu et al., 2020, Tongprasert and Ongsomwang, 2022), describing how this phenomenon could also affect urban areas (Liu et al., 2020).

As a result, little is known about the influence of agricultural land use on the spatial distribution of PM concentration levels. In 2022, according to a previous literature review (Gianquintieri et al., 2023b), 72 studies were published on modelling PM2.5 with advanced data analytics techniques, applying different correlation analyses between environmental variables and the target: among those, 35 (48.6%) included land-use in the explicative attributes, but only in 8 (11.1%) agricultural land-use class was explicitly referred to and analyzed. A stronger focus on this analysis object is therefore recommendable for research. Moreover, in some cases (Liu et al., 2020, Pu and Yoo, 2022), the focus on agricultural activity is on a time perspective rather than on a spatial one, limiting the analysis to specific time periods. An example of spatial correlation between agricultural land and air quality can be found in (Lambert et al., 2020), in which the analysis focused on the Great Plains of central USA, assessing the correspondence between the expansion of agricultural areas and the level of visible dust in the air, finding evidence of such phenomenon; however, to the best of our knowledge, no such kind of analysis has been performed in relation to PM.

Accordingly, based on literature analysis, a novel research question has emerged: what is the impact of agricultural land use on the spatial distribution of PM2.5 concentration over extended time periods, compared to other well-studied sources of pollution (such as urbanization, industry and transportation)? This is a very relevant topic for policy makers willing to mitigate the citizens' pollution exposure, by implementing evidence-based restriction policies that optimize the ratio between harm and benefit. Furthermore, in addition to a general assessment of concentration levels, a more focused analysis on urban areas is needed to target population exposure, and was therefore introduced in the framework.

Coherently, the aim of this study was to implement a GEOAI model capable of assessing the impact of agricultural lands on the spatial distribution of PM2.5 concentration levels, both in general and specifically in urban areas, and to compare it with that of traditional sources (i.e., urbanization, industry and transportation) on the basis of available land-use information. The analysis was conducted on a specific, highly polluted target territory (Lombardy region, in northern Italy), applying state-of-the-art technologies in data sources (such as the Copernicus project, allowing for a continuous mapping, thus partly solving the issue of limited recording stations), data processing (with machine learning algorithms building an ensemble GEOAI architecture) and data analysis (i.e. SHapley Additive exPlanation, SHAP).

2. Methods

2.1. Study setting

The geographic area included in the study was the Lombardy region, in northern Italy, accounting for roughly 10 million residents over a surface of 23′863 km^2. The environmental characteristics vary consistently across its territory, which includes strongly urbanized areas,

industries, large agricultural land, and natural landscapes (woods, mountains, and frequent water basins). A continuous mapping of the land use characteristics was obtained from the Lombardy region land use map, openly released as a vector layer (scale 1:5000, production year 2018) from the regional geoportal (https://tinyurl.com/cuwj7auh). Moreover, agricultural areas were further reclassified into crop types by intersection with the Lombardy region agricultural land use vector map (https://tinyurl.com/rafdxkkp). Road infrastructure areas were extracted from the Lombardy region topographic database (https://tinyurl.com/musuh2yj). The classes of land use were aggregated into three main categories: I) built-up area (as a total, and separately for buildings, industries, and streets), II) agricultural area (as a total, and separately for rice, corn, and cereals cultures) and III) natural area. The further subdivision into subclasses, in particular for different cultures, constitutes a step forward compared to previous analyses. A complete description of the attributes included in the analysis, with their corresponding data sources, can be found in (Gianquintieri et al., 2023a).

The considered pollutant was the particulate matter <2.5 μm (PM2.5), known to be a major hazard to human health. Its concentration values expressed as [μg/m3] for the Lombardy region were retrieved from the Copernicus Atmosphere Monitoring Service (CAMS): specifically, the ensemble median from CAMS European air quality forecasts (analysis dataset at surface level) was used (https://ads.atmosphere.copernicus.eu). Data are openly distributed as multi-temporal grids (NetCDF format), with a spatial resolution of 0.1° and a time resolution of 1 h.

The temporal interval taken into consideration spanned from May 2020 to December 2021, with weekly average values for PM2.5 concentration, thus resulting in a time-series of 85 total records. To enhance the insight into the impact of agricultural activity, three different temporal aggregation strategies were performed: I) the whole period, II) SPILL: months in which agricultural fertilizer spills are performed (March, April, October and November), III) NO SPILL: months in which no spill of agricultural fertilizer is performed (January, February, May, June, July, August, September and December).

From the spatial point of view, data were computed on a regular grid, obtained by downscaling the CAMS grid cells overlapping the Lombardy region, for a total of 748 squared cells of approximately 5.5 × 5.5 km, as reported in Fig. 1. Variables representing the different land-use classes and crop types were computed as % fraction of the total area in each grid cell, performing a geographical intersection between vector polygons (data sources overlaying the grid), and computing the ratio between the obtained surfaces. The values of explicative attributes were computed for each cell, using three different approaches of spatial windowing: considering each cell alone, considering the average of the surrounding 8 cells (thus focusing only on the edge conditions regardless of the point of measurement), and considering the average among both the central and the surrounding 8 cells. Moreover, two different aggregation strategies were implemented for the selection of the geographical area of interest: first, the analysis was performed on the whole study region; to increase, compared to previous works, the capability to focus on citizens' PM2.5 exposure, a second analysis was performed including only the urban areas, defined as cells with more than 25% of their surface covered by built-up land use classes (including buildings, streets, industries, and other infrastructures), thus resulting in 94 cells (12.6% of the total).

2.2. Target definition

In the analyzed time interval, the environmental characteristics of the territory, and its land-use in particular, varied minimally so that they were considered constant. However, a simple average for the entire period would result in a strong loss of informative content. Therefore, to avoid this problem, three different indicators were computed:

- Pollution frequency (*f*): ratio (in the range 0–1) of records in which the concentration of PM2.5 overcomes the European legal threshold
- Pollution intensity (*I*): magnitude of the pollution concentration excess above the European legal threshold, computed as the third quartile of the distribution of values $(I_{wj} - th) / th$, with *th* being the legal threshold, and I_{wj} the concentration of PM2.5 for the j-th week in the analysis period
- Pollution cumulative exposure (*Ex*): product between the frequency and the intensity

This target transformation is therefore a simple but innovative approach aimed at solving a well-assessed issue in this research topic.

The European legal threshold is set to 25 μm/m^3 as daily average. Unfortunately, as no weekly threshold is given, a weekly reference limit of 14.29 μm/m^3 was defined based on exceeding the daily threshold in 4 days out of seven.

The three indicators were computed for each cell, separately considering the three different temporal aggregation strategies (whole period, SPILL, NO SPILL), and assumed as targets to be correlated with the land use characteristics of the territory.

2.3. Model implementation

A data processing model was implemented to study the correlation between land-use characteristics and the concentration of PM2.5 pollution. Among the possible choices (addressed in Gianquintieri et al., 2023b), a random forest classifier analyzed with SHapley Additive exPlanation (SHAP) (Lundberg and Lee, 2017) was selected, and embedded in an ensemble architecture with multiple functional blocks. In particular, two additional functional blocks were included in series, before the random forest: the first for attributes selection, and the following one for spatial enhancement.

The proposed algorithm for attributes selection consisted in a threshold equal to 0.5 on the uni-variate Spearman's correlation coefficient between each attribute and the PM2.5 concentration (similarly to

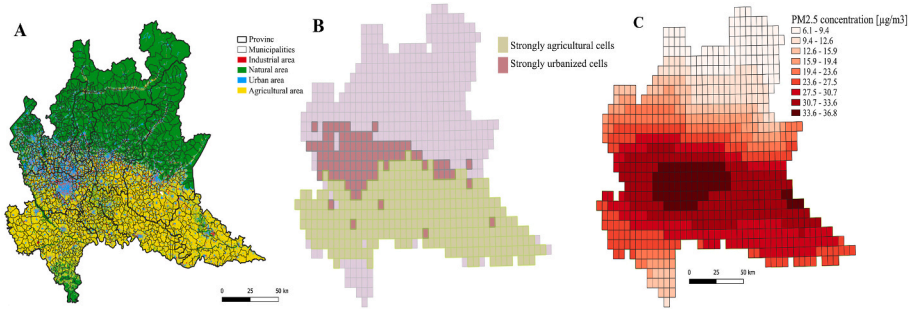


Fig. 1. Mapping of the target territory (Lombardy region, northern Italy), with its land-use classification (panel A); its division into 748 squared cells of approximatively 5.5 × 5.5 km, highlighting the most urbanized and agricultural areas (panel B); example of color-coded concentration of PM2.5 attributed to each cell (panel C).

what was proposed in (Gianquintieri et al., 2023a)); the target variable is represented by the daily concentration for each cell, also considering the three different approaches of spatial windowing (single cell, surrounding cells, whole block) previously described in 2.1, Separately for each attribute, the approach resulting in the highest correlation value was selected and maintained also for the following analyses.

For spatial enhancement, the selected method was the Multiscale Geographically Weighted Regression (MGWR) [Fotheringham et al., 2016, Oshan et al., 2019], a technique that models the spatially varying relationship between the target and the explicative attributes on a local level, considering multiple spatial scales and optimizing the final choice. This method can be applied for multiple purposes, including attributes selection and modelling itself, but in this case it was applied as a spatial filter to highlight local relationships. To prevent including attributes with spatial collinearity, a threshold of 0.8 on the collinearity matrix was set, to progressively remove collinear attributes once detected. The considered targets for the MGWR algorithm (to be correlated with maintained attributes) are the transformed time-invariant targets (f, I and Ex, as described in 2.2). In particular, the selected outputs from MGWR were:

- Bandwidth: the number of cells around each record in the dataset to be considered to maximize the correlation
- Weights: a numerical weight assigned for each attribute on each cell included in the bandwidth, separately for each record in the dataset, to maximize the local correlation

These two parameters were used to re-map the whole spatially transformed dataset, thus enhancing the local spatial relationship.

In summary, the final model was constituted by an ensemble GEO-AI architecture, composed of three blocks in series (Fig. 2):

I) Spearman's rank-based correlation for attributes selection
II) MGWR for spatial enhancement through local filtering and re-mapping
III) Random Forest model with SHapley Additive exPlanation (SHAP), for the analysis of the impact of land use characteristics on PM2.5 concentration.

2.4. Data and software availability

Data processing was performed with Python (v3.7) programming language, while graphical representations were implemented in QGIS. All the code used for the analysis is publicly available on GitHub (https://github.com/gisgeolab/D-DUST/tree/WP4). Input data for the analysis, including CAMS datasets and Lombardy region land use maps, are open and accessible at the links and references provided in Section 2.1. Sample analysis-ready data, together with documentation on data pre-processing, were also published on Zenodo (https://doi.org/10.5281/zenodo.6906903).

3. Results

3.1. Attributes selection

As previously described, relevant attributes were selected using as parameter the Spearman's uni-variate correlation coefficient, and subsequently removing collinear attributes. The correlation was computed between each attribute (separately) and the daily time-series of PM2.5 concentration level for each grid cell included in the two spatial aggregation protocols. Detailed results are provided in Table 1. When analyzing the whole territory, the best approach resulted in considering the whole block (central cell + surroundings), while for urban areas there was an almost equal number of attributes for which the best strategy was either to consider the entire block or the surroundings alone. Overall, Spearman's correlation coefficients (R) resulted higher when analyzing the whole territory compared to urban areas only (with the exception of the natural area), coherently with the reduced sample size in the latter. Few attributes were discarded as resulting in R values below the 0.5 threshold: the share of rice fields for both the whole territory and urban areas, and the urbanization level for the urban areas. The highest correlation values were found for natural areas (negative), followed by built-up areas (with industry as the main component), and slightly lower values for agricultural areas (with corn as the main component in the whole territory, and cereals for urban areas only).

3.2. Spatial enhancement

As previously described (2.4), the following collinearities among attributes were identified and removed from MGWR analysis: urbanization level when considering the whole territory, and road density when focusing on urban areas only. The resulting optimal bandwidths for the remaining attributes by the MGWR algorithm are reported in the following Table 2. Values ranged between 7 and 19 cells, with the majority of bandwidths limited to 11 adjacent cells (likely due, at least in part, to the fact that attributes are already computed with a spatial windowing on the surrounding cells, as described in 2.1).

3.3. GEOAI modelling

The final computational block was represented by a random forest classifier, trained with land-use characteristics to identify areas (regular grid cells) with an increased risk of having high concentrations of PM2.5 pollution. It is important to highlight that this model was not aimed at prediction, but it was applied in order to infer knowledge about the correlation between the attributes and the target variable. Therefore, the final assessment was made by the SHAP algorithm, which evaluates the magnitude of the contribution of each attribute (both locally and

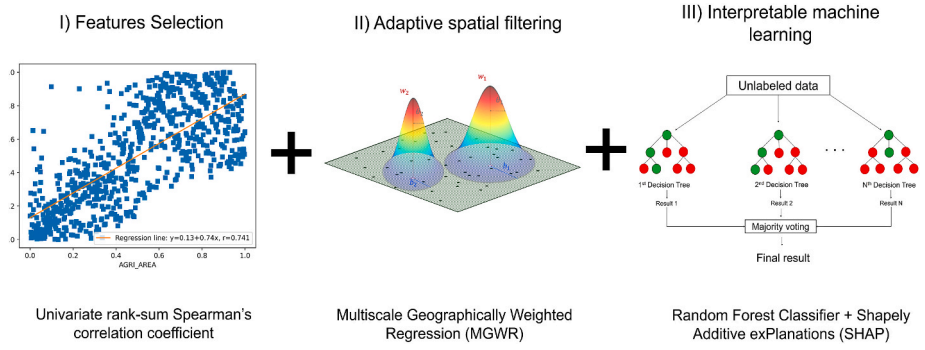


Fig. 2. Graphical representation of the three functional blocks put in series in the ensemble GeoAI architecture implemented to study the impact of different land-use classes on the spatial distribution of the risk of PM2.5 pollution concentration.

Table 1
Results of the Spearman's rank-based correlation analysis between land-use characteristics and PM2.5 concentration values, separately for the different protocols of spatial and temporal data aggregation (see text for details).

Spearman's R (spatial windowing)*	Whole territory			Urban areas		
	All times	No-spills	Spills	All times	No-spills	Spills
Agricultural area	.767 (B)	.768 (B)	.760 (B)	0.543 (S)	0.552 (S)	0.517 (S)
Cereals fields	.731 (S)	.730 (S)	.729 (S)	0.610 (B)	0.604 (B)	0.623 (B)
Corn fields	.899 (B)	.900 (B)	.892 (B)	0.589 (S)	0.597 (S)	0.566 (S)
Rice fields	.228 (B)	.223 (B)	.236 (B)	0.344 (B)	0.358 (B)	0.315 (B)
Built-up area	.800 (B)	.793 (B)	.817 (B)	0.552 (S)	0.538 (S)	0.579 (S)
Urbanized area	.717 (B)	.709 (B)	.736 (B)	0.244 (S)	0.222 (S)	0.286 (S)
Industrial area	.862 (B)	.857 (B)	.874 (B)	0.725 (B)	0.719 (B)	0.734 (B)
Roads density	.775 (B)	.767 (B)	.791 (B)	0.508 (S)	0.493 (S)	0.530 (S)
Natural area	-.861 (B)	-.862 (B)	-.857 (B)	-0.909 (B)	-0.918 (B)	-0.882 (B)

*(C) = highest value obtained considering a single cell.
(S) = highest value obtained considering the 8 surrounding cells around the target one.
(B) = highest value obtained considering both the central and the surrounding cells.

globally) in the training of the algorithm.

This allows the ranking of the attributes according to their contribution, identifying the most relevant ones, by a graphical representation of the results (SHAP plot [Lundberg et al., 2020]), separately for each experimental set-up: 3 different targets (frequency, intensity, and exposure to pollution), in 3 different time-frames (whole analysis period, SPILL, NO SPILL), in 2 different areas (whole territory, or urban areas only), for a total of 18 different combinations. Relevant results are reported in the following Fig. 3(a–d), where each row in the chart corresponds to an attribute whose relevance decreases from top to bottom. The graphical representation is composed of point clouds, where the thickness of the cloud represents the frequency of points in that position, and each point position on the X axis represents the magnitude of the contribution (either positive or negative with respect to the origin of the axis) for that specific attribute for each record in the training set; points are also gradually colored from red to blue, with red corresponding to the highest values of the attribute and blue to the lowest values (normalized value in the 0–1 range for computational purposes). Additionally, the AUC (with 95% C.I.) of the ROC curve resulting from a ten-fold cross-validation protocol is reported.

To provide a quantitative comparison of the previous results, the rank of attributes in the SHAP evaluation across the different protocols was compared by dividing the sum of ranks by the maximal possible value, thus obtaining a scaled representation in the 0–1 range (with 1 being the maximal possible impact, and 0 no impact at all). This analysis was run separately for the two spatial aggregation strategies and for the three temporal strategies. Results are reported in Fig. 4: for both spatial and temporal perspectives, it was possible to identify the share of the natural area as the most impactful attribute in the model, followed by the amount of cereals and corn fields (almost comparable, with different results depending on the protocols), and by the share of industrial area. Lower impact was related to overall agricultural area, built-up area (with the exception of no-spills periods) and roads length.

4. Discussion

4.1. Methodological issues

From the methodological point of view, the most important issue was represented by the need to compute a correlation between constant explicative attributes and a target time-series. If a direct approach is implemented (thus trying to model the impact of different variables directly with the time-series of PM2.5 concentration levels), the attributes that variate with similar frequency (such as meteorological variables) could result much more relevant, thus 'hiding' the impact of land-use and strongly reducing its contribution to the model (Araki et al., 2022). This issue is one of the main reasons why the models developed in literature struggle to assess the actual impact of land-use (Araki et al., 2022; Wu and Song, 2022; Su et al., 2022; Cheewinsiriwat et al., 2022). To overcome this limit, it was necessary to define a single fixed value representing the pollution concentration for each cell across the whole period of analysis (according to the different aggregation strategies), inferring it from the time-series. This decision led to the computation of the three target indicators (intensity, frequency and exposure), a simple but innovative approach to tackle this issue.

Regarding the choice of the model architecture, the latest research has shown the potential of GEO-AI, considering the spatially explicit task, through, in particular, machine learning algorithms. Considering the aim of our analysis (i.e., the correlation between land-use characteristics and pollution concentration), the most widely adopted approach was to implement a random forest classifier (Gianquintieri et al., 2023b). Such an approach, while guaranteeing effectiveness, is not only simple and agile from the computational point of view, but it represents also a good choice in terms of interpretability, a crucial characteristic for a machine learning model that is intended to be used for correlation analysis rather than application on an external dataset (e.g., prediction). In particular, a widely adopted method to inspect the impact of each variable on the model is the SHapley Additive exPlanation (SHAP), an algorithm based on cooperative games theory that

Table 2
Land-use characteristics included in the analysis of correlation with PM2.5 in the different experimental set-ups (see text for details), with their optimal bandwidth (number of adjacent cells to be taken in consideration) for spatial enhancement as identified by Multiscale Geographically Weighted Regression (MGWR).

MGWR optimal bandwidth ('NA' if removed for collinearity, '/' if excluded before)		Whole territory			Urban areas		
		All times	No-spills	Spills	All times	No-spills	Spills
Agricultural area	Frequency	9	7	7	11	11	9
	Intensity	7	7	9	7	11	11
	Exposure	7	11	7	9	11	11
Cereals fields	Frequency	9	7	7	7	7	7
	Intensity	7	7	9	7	7	7
	Exposure	7	11	7	9	7	7
Corn fields	Frequency	7	7	7	7	11	7
	Intensity	7	7	7	7	11	11
	Exposure	7	13	7	7	11	7
Rice fields	Frequency	/	/	/	/	/	/
	Intensity	/	/	/	/	/	/
	Exposure	/	/	/	/	/	/
Built-up area	Frequency	9	9	11	13	11	11
	Intensity	7	7	9	7	11	13
	Exposure	9	11	9	11	11	11
Urbanized area	Frequency	NA	NA	NA	/	/	/
	Intensity	NA	NA	NA	/	/	/
	Exposure	NA	NA	NA	/	/	/
Industrial area	Frequency	9	7	9	11	11	7
	Intensity	7	7	9	7	7	7
	Exposure	11	11	11	9	11	11
Roads density	Frequency	9	9	9	NA	NA	NA
	Intensity	7	7	9	NA	NA	NA
	Exposure	11	11	7	NA	NA	NA
Natural area	Frequency	7	11	9	9	11	13
	Intensity	7	7	7	7	11	19
	Exposure	7	11	7	11	11	11

computes the contribution of each variable (both locally, on single records, and globally) on the model predictions. Furthermore, according to literature suggestions, when dealing with pollution modelling, the performance of a machine learning algorithm can be enhanced by including it in an ensemble architecture (Gianquintieri et al., 2023b), with multiple functional blocks.

It is worth noting that some details of the model's implementation required specific attention. First of all, since Spearman's coefficient is rank-based, some correlations other than linear would also result in high values. In order to include more complex correlations (i.e., bimodal), it was decided to adopt a significantly low threshold of 0.5 for Spearman's coefficient. Concerning the MGWR application, it must be pointed out that the algorithm suffers from collinearity between the explicative attributes: therefore, it was necessary to assess eventual collinearities, and remove some of the attributes. This was done with a threshold of 0.8 on the collinearity matrix, progressively removing attributes that had the strongest collinearities, one by one, until none over the threshold was detected. For the implementation of the Random Forest model, given the time-invariance of the attributes, and the consequent choice of remapping the target to set-up a time-invariant spatial analysis, the available dataset was composed of only 748 records, one for each cell, which is very limited for training a machine learning model. To overcome this limitation, a two-step solution was adopted:

- Task simplification: the regression task having as target the PM2.5 concentration was transformed into a binary classification task, labelling as either 1 or 0 cells with increased (or not) risk of having pollution concentrations over threshold. Boundary values were set as 0.1 for f, 0.25 for I, and 0.025 for Ex.
- Data augmentation: each record was repeated in the dataset for a number of times proportional to the distance of the target from the boundary value, adding a random noise (with a max amplitude of 10% of the original signal) on all the attributes. As a result, the dataset dimension was increased to a range of 10′000–25′000 records (depending on the target and on temporal and spatial aggregation

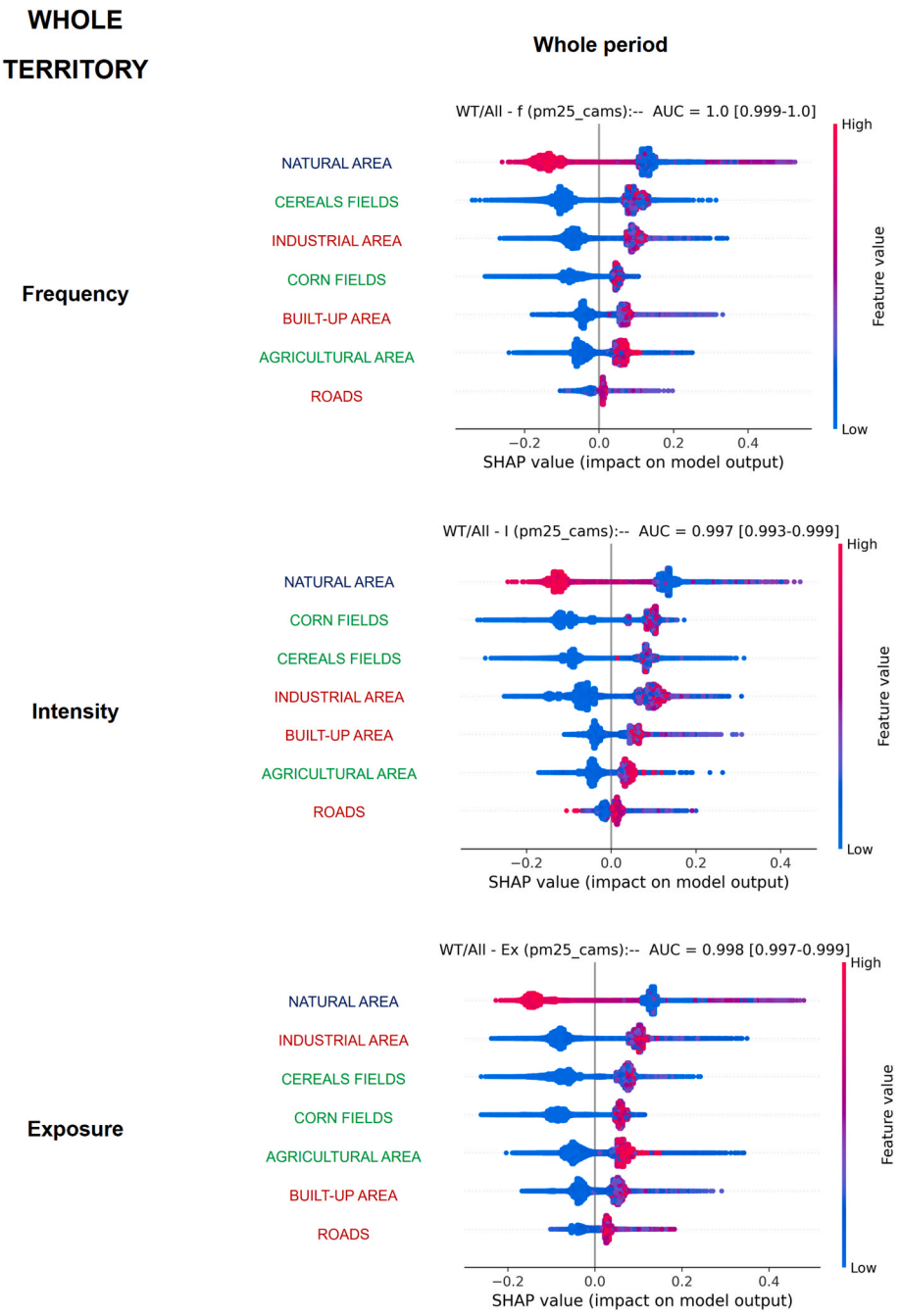


Fig. 3a. Results of the SHAP algorithm evaluating the impact of different land-use classes on a random forest classifier evaluating the risk of PM2.5 pollution concentration (either in frequency, intensity or exposure, see text for details) in the whole territory of Lombardy region, Italy, from May 2020 to December 2021.

strategies), thus providing a stronger impact on higher values of PM2.5 concentration despite the binarization of the target.

Finally, to increase robustness, all Random Forest results were averaged across a ten-fold cross-validation splitting.

4.2. Main results

Considering the whole territory, the only non-relevant attribute resulted to be the share of rice fields, confirming preliminary findings by uni-variate analysis on the same dataset (Gianquintieri et al., 2023a), while the amount of the urbanized area had to be discarded for collinearity. Concerning bandwidths, values were in general small, with a maximal value of 13 cells and an average of 8.3 (first and third quartile 7–9), suggesting a local relationship prevailing on global dynamics, as confirmed by other references in literature (Wu and Song, 2022; Pu and Yoo, 2022). In the RF model, the natural area resulted to be the most impactful variable in almost all set-ups, except for frequency during spills periods and intensity during no-spills periods (Fig. 3b), in which the highest relevance is assigned to built-up area and corn fields, respectively. Considering rank-sum analysis (Fig. 4), the agricultural area resulted slightly less impactful, yet comparable, with respect to the built-up environment; however, this was reverted if focusing on the single components, as both cereals and corn fields resulted more relevant than industrialized areas. However, it is worth noticing that this relation changes depending on the set-up. Finally, the least relevant attribute was the density of roads.

Focusing on urban areas alone, the urbanization level resulted in

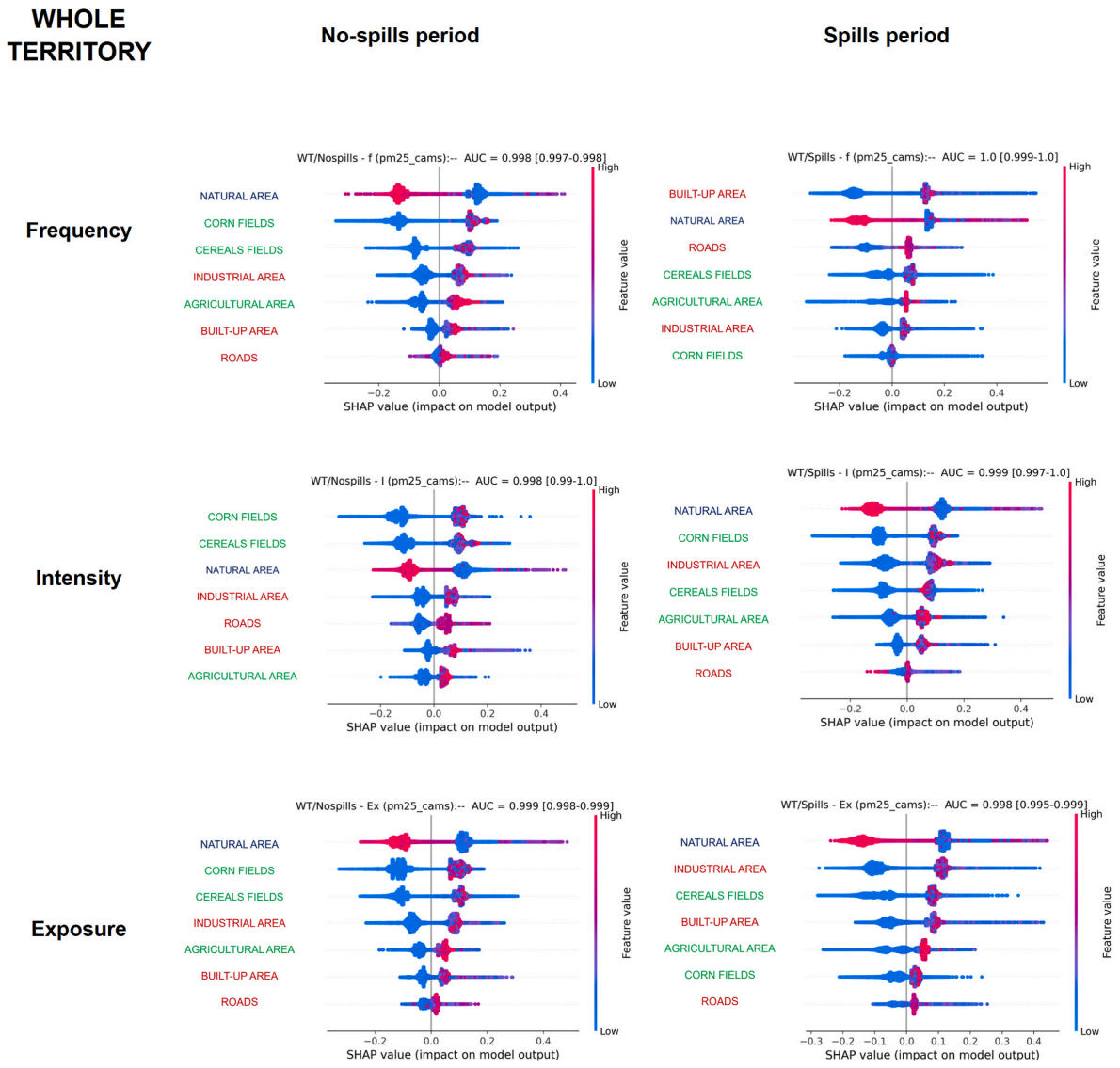


Fig. 3b. Results of the SHAP algorithm evaluating the impact of different land-use classes on a random forest classifier evaluating the risk of PM2.5 pollution concentration (either in frequency, intensity or exposure, see text for details) in the whole territory of Lombardy region, Italy, from May 2020 to December 2021, either in periods when no manure is spilled on crops (left) or when this operation is performed (right).

being non-significant, which can be expected considering that all the target territory was above a certain threshold in terms of urbanization, strongly reducing the variability in this parameter. Optimal bandwidths were a little wider, with a max value of 19 cells and an average of 9.6 (quartiles 7–11), still confirming the prevalence of local relationships over global ones but showing that the impact of the surroundings was more relevant when focusing on densely populated areas. As a result, the primary role of natural area in the model was confirmed, while the relation between agricultural and built-up environment showed in this case a prevalence of the first over the latter, as confirmed by the single components, with cereals and corn fields having larger impact compared to the industrialized area. Again, it is important to point out that this depends on the set-up. Noticeably, considering the whole period, industries seemed to be more relevant in terms of the frequency of pollution events, while agricultural activities were more likely to generate the most intense peaks.

An additional consideration must be pointed out with regard to time aggregation protocols. The chosen subdivision distinguishes periods in which manure spills were performed or not: contrary to possible assumptions, the impact of agricultural activity resulted more relevant in no-spills periods. A possible explanation is that the contribution of agricultural activities to PM pollution is not due to manure spills, but to other operations such as residual burning, as suggested in the literature; however, additional investigations of these phenomena will be needed, eventually with the possible availability of precise recorded timing for the different agricultural operations.

4.3. Comparison with the literature

Our results seem to be, at first sight, in contrast with previous findings in the literature. However, when considering more in detail the suitability of the applied methodologies for the specific purpose of comparing the impact of different land-uses, previous findings appear less supported by a robust approach.

For example, in (Cheewinsiriwat et al., 2022), agricultural land did not significantly correlate with PM2.5 concentration. However, it is possible to hypothesize that the implemented framework, based on a simple linear correlation with an optimized circular buffer around ground stations, was not robust enough to generate significant results. This hypothesis is corroborated by the non-significance also recorded for the urban area, and by the overall very low correlation coefficients found associated with each class of land-use. As a matter of fact, this

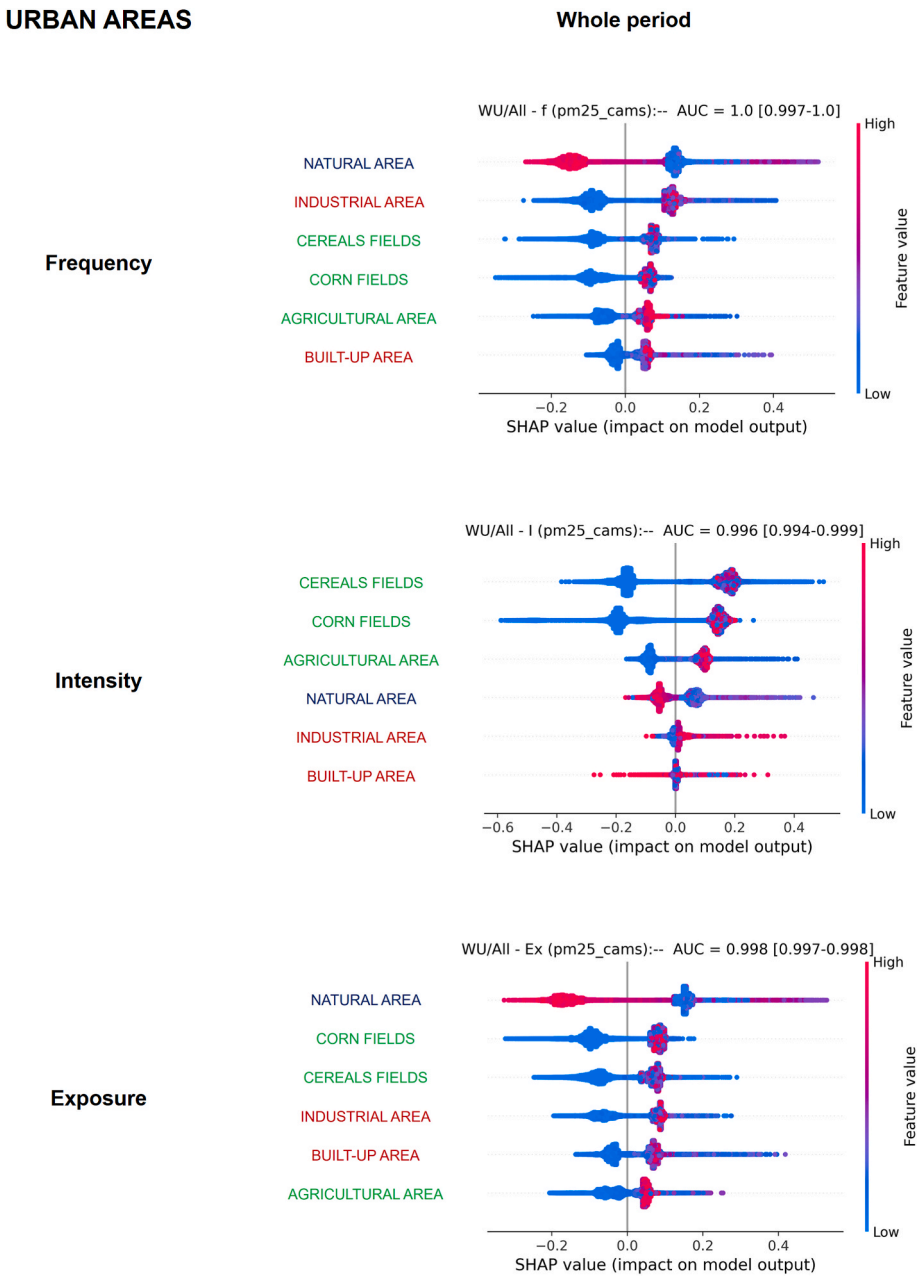


Fig. 3c. Results of the SHAP algorithm evaluating the impact of different land-use classes on a random forest classifier evaluating the risk of PM2.5 pollution concentration (either in frequency, intensity or exposure, see text for details) in the most densely inhabited urban areas of Lombardy region, Italy, from May 2020 to December 2021.

correlation analysis also included meteorological factors, and the selected methodology seems to be oriented to the comparison of time-series measurements, as suggested by the much higher correlations found for this latter category.

As previously highlighted in section 2.2, a model having a time-series value as the target is not a robust choice to assess the impact of time-invariant attributes, especially if other time-variant attributes are simultaneously included. This issue was pointed out very clearly in (Araki et al., 2022), aiming at the implementation of a predictive model, with an intermediate step of feature selection that allows the comparison of the impact of different land-use classes, along with meteorological variables. Despite the order of the impact of land-use classes being similar to the one obtained in our study (close prevalence of built-up over agricultural land, and smaller impact of roads), all these time-invariant attributes have very poor impact on the overall modelling, with a clear prevalence of time-variant attributes.

Another relevant example of how the experimental set-up has a strong impact on the results is reported in (Wu and Song, 2022). Here, a land-use regression analysis (targeting PM2.5) resulted in a very different order compared to our results, with the built-up area as the most relevant factor, followed by natural area, then agricultural lands, and finally industrial areas. However, all reported correlation coefficients were very low, with the highest value (built-up) being 0.234, down to 0.033 for the industrial area. Also, considering that their aim was to implement a predictive model, such a low impact of land-use on PM2.5 concentration appears unlikely realistic. These results were probably due to a limited number of recording stations available, moreover all located in urban areas.

Similarly, in (Su et al., 2022) the correlation analysis mixed land-use classes with time-variant environmental factors (meteorological

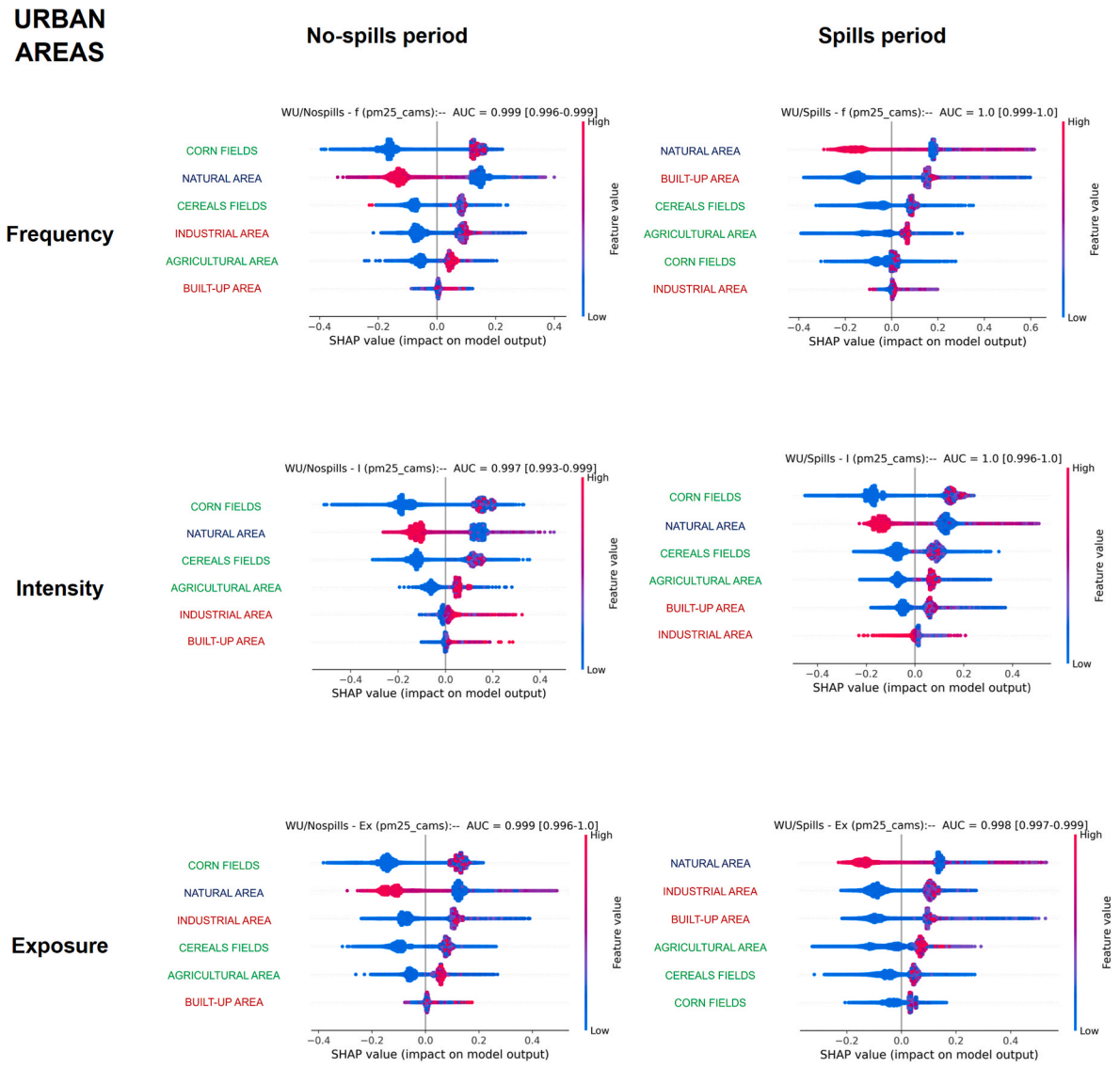


Fig. 3d. Results of the SHAP algorithm evaluating the impact of different land-use classes on a random forest classifier evaluating the risk of PM2.5 pollution concentration (either in frequency, intensity or exposure, see text for details) in the most densely inhabited urban areas of Lombardy region, Italy, from May 2020 to December 2021, either in periods when no manure is spilled on crops (left) or when this operation is performed (right).

variables). Noticeably, in this case one of the land-use classes resulted significant, with a comparable impact to other time-variant factors; even more noticeably, such land-use class represented roads density, which in contrast resulted the least impactful attribute in our model. This may be due to the different adopted set-up, a hypothesis corroborated by the fact that in that study also the impact of crop lands resulted higher compared to that of natural areas (another opposite result compared to our model), despite both having very limited impact. Yet, it is also possible that these discrepancies were due to actual differences in the territory under observation.

4.4. *Research advancements*

To the best of our knowledge, this is the first study successfully implementing a GEOAI model capable of effectively comparing the impact of different land-use classes on the spatial distribution of concentration of PM2.5, solving the main issues that hindered previous research.

In particular:

- the limited availability of recording ground-stations was tackled by considering new state-of-art concentration models, in particular the Copernicus Atmosphere Monitoring Service, which enabled a continuous mapping of the target variable over the territory; moreover, a data enhancement strategy (see section 4.1) was successfully implemented to furtherly increase the dataset dimension, making it suitable for machine learning application.
- land-use can be approximated, in the considered temporal frame, as a time-invariant characteristic of the territory: coherently, it was necessary to infer a target variable that represented the spatial concentration of PM2.5, regardless of the time-series point values. In order to account for different definitions of concentration, three measurements, based on the comparison of recorded values with the legal threshold, were computed: frequency, intensity, and exposure. In this way, it was possible to effectively study the spatial distribution of PM2.5 concentration in relation to the land-use, regardless of its temporal fluctuations.
- In previous studies, the identified correlations were usually very low, which is incompatible with the assessed knowledge about PM2.5 emissions. To obtain significant information, a state-of-the-art data processing architecture was implemented, including features

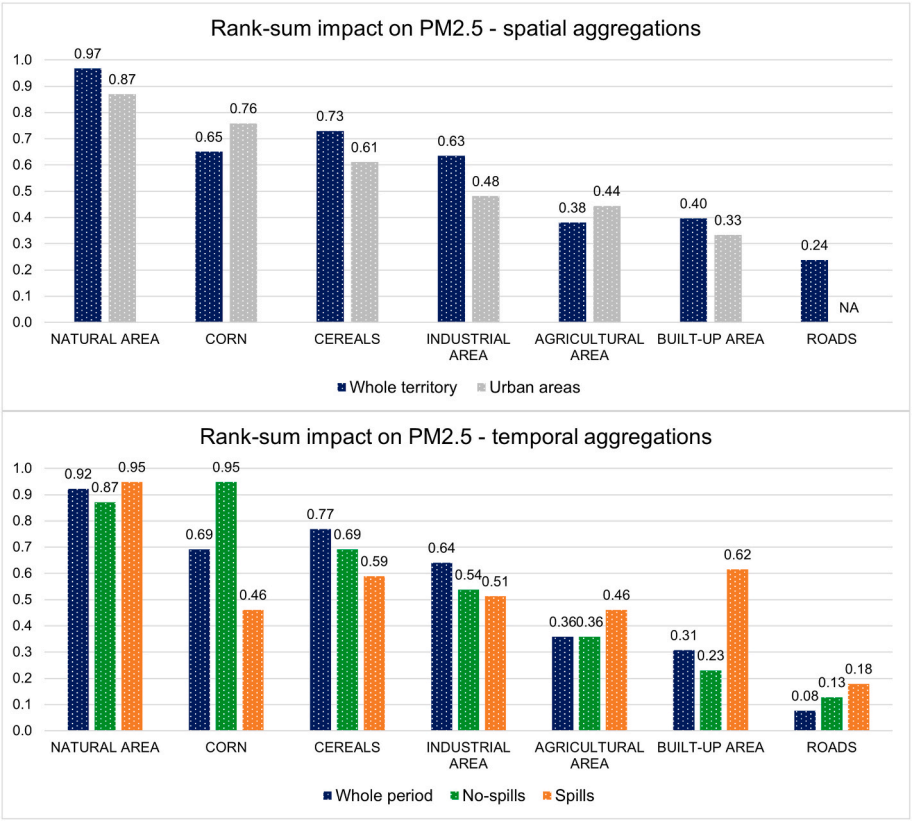


Fig. 4. Cumulated results (with rank-sum approach) of the SHAP algorithm evaluating the impact of different land-use classes on a random forest classifier evaluating the risk of PM2.5 pollution concentration in the territory of Lombardy region, Italy, with different approaches of spatial or temporal aggregation (see text for details).

selection, a GEOAI spatial enhancement method (capable of highlighting local relations), and a final step of interpretable machine learning. This approach is different from the traditional correlation analysis frameworks, as it exploits the capabilities of a selected machine learning algorithm, and it evaluates how it uses the different attributes in the 'training', i.e., how the algorithm learns to predict an output (spatial distribution of PM2.5 concentration) on the basis of a given input (land-use characteristics).

It is worth noticing that the random forest classifiers (one for each set-up) resulted in very high values of AUC (Area Under Curve of the ROC, Receiver Operating Characteristic), as reported in Fig. 3, all very close to 1 (representing the perfect classification): this clearly indicates a strong overfitting on the training data, mainly due to the intermediate step of spatial enhancement through MGWR. While this condition could represent a problem for the application of the trained random forest classifiers, it is perfectly suitable for the purpose of this study, in which the classifiers are not trained to be applied on additional data but are only used to inspect how the different explicative attributes contribute to and impact on the training of the model. Therefore, while poor performance could be expected on a different data set, the significance of the attributes on the analyzed data set was maximized.

With this methodology, it was possible to effectively assess that the impact of agricultural activity on the spatial distribution of PM2.5 concentration, both in general and in urban areas, is comparable (if not superior, depending on the experimental set-up) to that of more traditional sources such as urbanization, industry and transports. This finding is in line with well-assessed knowledge regarding emissions, but, to the best of our knowledge, it was never confirmed in terms of spatial distribution of pollution concentration.

4.5. *Limits and future developments*

The main limitation of this work is relevant to data availability. Despite the adopted mitigation strategies, the dimension of the dataset still represents a critical issue. While a sufficient dimension for an effective implementation was reached, its enlargement would be beneficial for future research to increase the statistical robustness of the results.

On top of data quantity, a matter of data quality must be considered: despite a preliminary analysis showed that CAMS modelling currently represents the best option as a proxy measurement, ground-stations remain the gold standard when it comes to measuring ground-truth about pollution concentration.

An additional issue is specifically relevant to land-use analysis, as this is one of the multiple features included in the development of CAMS: the risk is that the findings are somehow reverse-engineering the original measurement. However, this was considered acceptable for two reasons: first, the land-use evaluation adopted for CAMS is generated from a different source than that applied in our model and, secondarily, CAMS measurements are not used directly, but are consistently processed in order to infer the actual model target.

Finally, specifically concerning agricultural areas, better results could be obtained by knowing in detail the timeline of the different activities across the territory, in particular about manure spills and residual burnings; additionally, a precise distinction between actual agriculture and farming facilities (possibly distinguishing according to the farmed species), currently not included in the model, could provide additional useful insights.

5. Conclusions

As air pollution, and PM2.5 in particular, constitutes a major risk for human health, data-driven evidence is critical in the identification of factors that contribute to its concentration, in order to guide preventive measures. In particular, agricultural activity is an understudied element, especially in comparison with more traditional sources such as urbanization, industrialization and transports. This field of research was recently enhanced by new technologies for data collection (satellite imagery) and data processing (AI), but the assessment of the impact of land-use on the concentration of PM2.5 was inconclusive, due to the experimental set up of previous works, resulting in non-significant correlations, which were interpreted by researchers as a lack of cause-effect relationship.

In this context, we hypothesized that a different experimental set up, addressing and solving the issues emerged from previous works, could instead verify what is the impact of agricultural activity on PM2.5 spatial concentration, comparing it with the traditional sources. In fact, the proposed GEOAI architecture, composed by three blocks (correlation-based features selection, MGWR spatial enhancement, and random forest classifier) successfully allowed the comparison of the impact of agricultural land with other land-use classes (built-up and natural environment) and sub-classes on the spatial distribution of PM2.5 concentration, considering the territory of Lombardy region (northern Italy) as a study case. Satellite monitoring (CAMS data) was used as the origin of pollution measurement, thus overcoming the limited availability of ground-stations. The novel data processing pipeline, specifically implemented for this task, was able to overcome the main issues identified in previous studies, such as the stationarity of land-use compared to fluctuations of PM concentration (originally recorded as a time-series), hence resulting in a purely spatial analysis.

Our model demonstrates a significantly higher sensitivity to land-use classes compared to other published models, which were originally developed for different purposes. This allowed us to provide compelling evidence, in opposition to previous knowledge, that agricultural activities have a comparable, if not superior, impact on the spatial distribution of PM2.5 concentration with respect to other frequently studied sources, such as urbanization, industry, and transportation, particularly when analyzing densely populated urban areas. Therefore, this study is a step forward from the current state-of-art, and its higher robustness compared to previous works puts the presented results ahead of established knowledge.

In particular, the contribution of agricultural activities appears more related to pollution spikes rather than to an increase in the baseline. Further research is needed to specifically identify the activities that generate these concentration peaks (probably residual burnings, according to literature knowledge about emissions). In line with previous research (Wyer et al., 2022, Lambert et al., 2020), these results show that public policymakers should also consider agricultural activities for evidence-based decision-making about pollution mitigation, to optimize the ratio between the harms and benefits of restriction policies.

Funding

This study was performed within the framework of the D-DUST project (Data-driven moDelling of particUlate with Satellite Technology aid), which received support from Fondazione Cariplo (project ID: 2020-4022).

CRediT authorship contribution statement

Lorenzo Gianquintieri: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Daniele Oxoli:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Supervision, Writing – review & editing. **Enrico Gianluca Caiani:** Supervision, Writing – review & editing. **Maria Antonia Brovelli:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

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

Data availability

Link to public GitHub repository with code and data included in the text.

References

Araki, S., Shimadera, H., Hasunuma, H., Yoda, Y., Shima, M., 2022. Predicting daily PM2.5 exposure with spatially invariant accuracy using Co-existing pollutant concentrations as predictors. Atmosphere 13 (5), 782. https://doi.org/10.3390/atmos13050782.

Awe, Y., Larsen, B., Sanchez-Triana, E., 2022. The Global Health Cost of PM2.5 Air Pollution: A Case for Action beyond 2021. World Bank Group, United States of America. Retrieved from. https://policycommons.net/artifacts/2232589/the-global-health-cost-of-pm25-air-pollution/2990492.

Bray, C.D., Battye, W.H., Aneja, V.P., 2019. The role of biomass burning agricultural emissions in the indo-gangetic plains on the air quality in New Delhi, India. Atmospher. Environ. 218 https://doi.org/10.1016/j.atmosenv.2019.116983.

Cheewinsiriwat, P., Duangyiwa, C., Sukitpaneenit, M., Stettler, M.E.J., 2022. Influence of land use and meteorological factors on PM2.5 and PM10 concentrations in Bangkok, Thailand. Sustainability 14 (9), 5367. https://doi.org/10.3390/su1409536.

Erisman, J.W., Schaap, M., 2004. The need for ammonia abatement with respect to secondary PM reductions in Europe. Environ. Pollut. 129, 159–163. https://doi.org/10.1016/j.envpol.2003.08.042.

Fotheringham, A.S., Yang, W., Kang, W., 2016. Multi-scale geographically weighted regression. Ann. Assoc. Am. Geogr. 107, 1247–1265. https://doi.org/10.1080/24694452.2017.1352480.

Gardner-Frolick, R., Boyd, D., Giang, A., 2022. Selecting data analytic and modeling methods to support air pollution and environmental justice investigations: a critical review and guidance framework. Environ. Sci. Technol. 56 (5), 2843–2860. https://doi.org/10.1021/acs.est.1c01739.

Gianquintieri, L., Brovelli, M.A., Pagliosa, A., Bonora, R., Sechi, G.M., Caiani, E.G., 2021. Geospatial correlation analysis between air pollution indicators and estimated speed of COVID-19 diffusion in the Lombardy region (Italy). Int. J. Environ. Res. Publ. Health 18, 12154. https://doi.org/10.3390/ijerph182212154.

Gianquintieri, L., Oxoli, D., Caiani, E.G., Brovelli, M.A., 2023a. Land use influence on ambient PM2.5 and ammonia concentrations: correlation analyses in the Lombardy region, Italy. AGILE GIScience Ser. 4, 26. https://doi.org/10.5194/agile-giss-4-26-2023.

Gianquintieri, L., Oxoli, D., Caiani, E.G., Brovelli, M.A., 2023b. State-of-art in modelling particulate matter (PM) concentration: a scoping review of aims and methods. Environ. Devel. Sustain. under review (last update Feb 08 2024).

Gugnani, V., Singh, R.K., 2022. Analysis of deep learning approaches for air pollution prediction. Multimed. Tool. Appl. 81 (4), 6031–6049. https://doi.org/10.1007/s11042-021-11734-x.

Higashiyama, H., Yoshimoto, D., Okamoto, Y., Kikkawa, H., Asano, S., Kinoshita, M., 2007. Receptor-activated Smad localisation in Bleomycin-induced pulmonary fibrosis. J. Clin. Pathol. 60, 283–289. https://doi.org/10.1136/jcp.2006.037606.

Jethva, H., Torres, O., Field, R., Lyapustin, A., Gautam, R., Kayetha, V., 2019. Connecting crop productivity, residue fires, and air quality over northern India. Nat. Scientif. Rep. 9 (1) https://doi.org/10.1038/s41598-019-52799-x.

Lambert, A., Hallar, A.G., Garcia, M., Strong, C., Andrews, E., Hand, J.L., 2020. Dust impacts of rapid agricultural expansion on the Great Plains. Geophys. Res. Lett. 47 (20) https://doi.org/10.1029/2020GL090347.

Liu, T., He, G., Lau, A.K.H., 2020. Statistical evidence on the impact of agricultural straw burning on urban air quality in China. Sci. Total Environ. 711 https://doi.org/10.1016/j.scitotenv.2019.134633.

Loftus, C., Yost, M., Sampson, P., Torres, E., Arias, G., Breckwich Vasquez, V., Hartin, K., Armstrong, J., Tchong-French, M., Vedal, S., Bhatti, P., 2015. Ambient ammonia exposures in an agricultural community and pediatric asthma morbidity. Epidemiology 26 (6), 794–801. https://doi.org/10.1097/EDE.0000000000000368.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, p. 30. In: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 2522–5839. https://doi.org/10.1038/s42256-019-0138-9.

McDuffie, E.E., Smith, S.J., O'Rourke, P., Tibrewal, K., Venkataraman, C., Marais, E.A., Zheng, B., Crippa, M., Brauer, M., Martin, R.V., 2020. A global anthropogenic

emission inventory of atmospheric pollutants from sector-and fuel-specific sources (1970–2017): an application of the Community Emissions Data System (CEDS). Earth Syst. Sci. Data 12 (4), 3413–3442. https://doi.org/10.5194/essd-12-3413-2020.

Mehmood, K., Bao, Y., Saifullah, Cheng, W., Khan, M.A., Siddique, N., Abrar, M.M., Soban, A., Fahad, S., Naidu, R., 2022. Predicting the quality of air with machine learning approaches: current research priorities and future perspectives. J. Clean. Prod. 379 https://doi.org/10.1016/j.jclepro.2022.134656.

Neghab, M., Mirzaei, A., Kargar Shouroki, F., Jahangiri, M., Zare, M., Yousefinejad, S., 2018. Ventilatory disorders associated with occupational inhalation exposure to nitrogen trihydride (ammonia). Ind. Health 56 (5), 427–435. https://doi.org/10.2486/indhealth.2018-0014.

Oshan, T.M., Li, Z., Kang, W., Wolf, L.J., Fotheringham, A.S., 2019. mgwr: a Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. ISPRS Int. J. Geo-Inf. 8 (6), 269. https://doi.org/10.3390/ijgi8060269.

Pu, Q., Yoo, E.H., 2022. A gap-filling hybrid approach for hourly PM2.5 prediction at high spatial resolution from multi-sourced AOD data. Environ. Pollut. 315 https://doi.org/10.1016/j.envpol.2022.120419.

Rafaj, P., Kiesewetter, G., Gül, T., Schöpp, W., Cofala, J., Klimont, Z., Purohit, P., Heyes, C., Amann, M., Borken-Kleefeld, J., Cozzi, L., 2018. Outlook for clean air in the context of sustainable development goals. Global Environ. Change 53, 1–11. https://doi.org/10.1016/j.gloenvcha.2018.08.008.

Sapek, A., 2013. Ammonia emissions from non-agricultural sources. Pol. J. Environ. Stud. 22 (1), 63–70. ISSN: 1230-1485.

Su, Z., Lin, L., Chen, Y., Hu, H., 2022. Understanding the distribution and drivers of PM2.5 concentrations in the yangtze river delta from 2015 to 2020 using random forest regression. Environ. Monit. Assess. 194 (4) https://doi.org/10.1007/s10661-022-09934-5.

Tongprasert, P., Ongsomwang, S., 2022. A suitable model for spatiotemporal particulate matter concentration prediction in rural and urban landscapes, Thailand. Atmosphere 13 (6). https://doi.org/10.3390/atmos13060904.

Wu, P., Song, Y., 2022. Land use quantile regression modeling of fine particulate matter in Australia. Rem. Sens. 14 (6), 1370. https://doi.org/10.3390/rs14061370.

Wu, Y., Gu, B., Erisman, J.W., Reis, S., Fang, Y., Lu, X., Zhang, X., 2016. PM2.5 pollution is substantially affected by ammonia emissions in China. Environ. Pollut. 218, 86–94. https://doi.org/10.1016/j.envpol.2016.08.027.

Wyer, K., Kelleghan, D., Blanes-Vidal, V., Schauberger, G., Curran, T., 2022. Ammonia emissions from agriculture and their contribution to fine particulate matter: a review of implications for human health. J. Environ. Manag. 323 https://doi.org/10.1016/j.jenvman.2022.116285.

Xing, Y.F., Xu, Y.H., Shi, M.H., Lian, Y.X., 2016. The impact of PM2.5 on the human respiratory system. J. Thorac. Dis. 8 (1), 69–74. https://doi.org/10.3978/j.issn.2072-1439.2016.01.19.

Zaini, N., Ean, L.W., Ahmed, A.N., Malek, M.A., 2022. A systematic literature review of deep learning neural network for time series air quality forecasting. Environ. Sci. Pollut. Control Ser. 29 (4), 4958–4990. https://doi.org/10.1007/s11356-021-17442-1.

Zhu, L., Henze, D.K., Bash, J.O., Cady-Pereira, K.E., Shephard, M.W., Luo, M., Capps, S. L., 2015. Sources and impacts of atmospheric NH3: current understanding and frontiers for modeling, measurements, and remote sensing in North America. Curr. Pollut. Rep. 1, 95–116.