



Greenness, air pollution, and temperature exposure effects in predicting premature mortality and morbidity: A small-area study using spatial random forest model

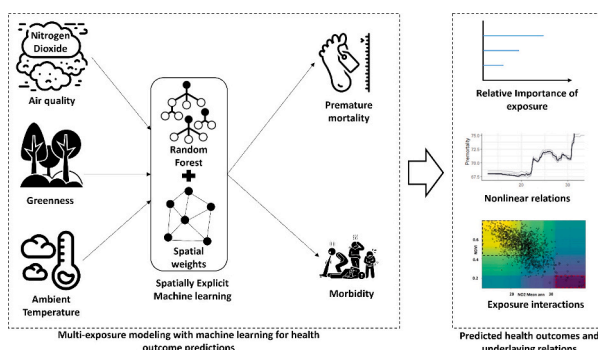
S.M. Labib^{*}

Department of Human Geography and Spatial Planning, Faculty of Geosciences, Utrecht University, the Netherlands

HIGHLIGHTS

- Effects of multiple environmental exposures on health modeled using machine learning
- Spatial dependency among exposure and health variables accounted for in the process
- Relative importance of air pollution, greenness & temperature exposure identified.
- Nonlinear exposure-response relations & interactions investigated for health outcomes.
- Air pollution indicated a greater influence on health than greenness & temperature exposure.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Lidia Mínguez Alarcon

Keywords:

Air pollution
Temperature
Exposure Assessment
Machine Learning
Greenness exposure
Public Health

ABSTRACT

Background: Although studies have provided negative impacts of air pollution, heat or cold exposure on mortality and morbidity, and positive effects of increased greenness on reducing them, a few studies have focused on exploring combined and synergetic effects of these exposures in predicting these health outcomes, and most had ignored the spatial autocorrelation in analyzing their health effects. This study aims to investigate the health effects of air pollution, greenness, and temperature exposure on premature mortality and morbidity within a spatial machine-learning modeling framework.

Methods: Years of potential life lost reflecting premature mortality and comparative illness and disability ratio reflecting chronic morbidity from 1673 small areas covering Greater Manchester for the year 2008–2013 obtained. Average annual levels of NO₂ concentration, normalized difference vegetation index (NDVI) representing greenness, and annual average air temperature were utilized to assess exposure in each area. These exposures were linked to health outcomes using non-spatial and spatial random forest (RF) models while accounting for spatial autocorrelation.

Results: Spatial-RF models provided the best predictive accuracy when accounted for spatial autocorrelation. Among the exposures considered, air pollution emerged as the most influential in predicting mortality and morbidity, followed by NDVI and temperature exposure. Nonlinear exposure-response relations were observed,

^{*} Vening Meineszgebouw A, Princetonlaan 8a, 3584 CB Utrecht, the Netherlands.

E-mail addresses: s.m.labib@uu.nl, labib.l.m@gmail.com.

and interactions between exposures illustrated specific ranges or sweet and sour spots of exposure thresholds where combined effects either exacerbate or moderate health conditions.

Conclusion: Air pollution exposure had a greater negative impact on health compared to greenness and temperature exposure. Combined exposure effects may indicate the highest influence of premature mortality and morbidity burden.

1. Introduction

Urban areas are increasingly facing the adverse health effects of increased air pollution, scarce greenness, and non-optimal temperature (Forouzanfar et al., 2016; Burkart et al., 2021; Barboza et al., 2021). In recent times, 55 % of the world's population has lived in urban areas; with a growing urbanization rate, it is expected to rise to 68 % by 2050 (United Nations, 2019). As a result, more people will live in areas that generally cause adverse health effects (Dadvand et al., 2012; Kasdagli et al., 2021). While several urban environmental health stressors have been identified and studied (Ohanyan et al., 2022a, 2022b; Nieuwenhuijsen et al., 2019), the health impact of air pollution, temperature, and greenness exposure has been widely researched in various domains of studies. Existing evidence indicates the negative health impact of high air pollution exposure and non-optimal temperature (Beelen et al., 2014; Gasparrini et al., 2022, 2015). In contrast, many studies noted the mostly positive health effects of increased exposure to urban greenery (Rojas-Rueda et al., 2019; Twohig-Bennett and Jones, 2018; Hunter et al., 2023). The synergies between these negative and positive environmental exposures are critical to understanding and (re)designing cities to ensure healthy places for urban citizens by reducing the disease burden in cities (Nieuwenhuijsen, 2020; Nieuwenhuijsen et al., 2022).

The 2015 Global Burden of Disease Study reported that air pollution was responsible for approximately 6.4 million deaths worldwide. It identified ambient air pollution as a significant environmental factor affecting both morbidity and mortality (Forouzanfar et al., 2016; Khomenko et al., 2021). Higher exposure to particulate matter (PM), nitrogen dioxide (NO₂), and ground-level ozone (O₃) are attributed to increased deaths and overall poor health conditions in the European region (Beelen et al., 2014; Dominski et al., 2021). Similarly, non-optimal temperatures were attributed to 1.69 million deaths globally in 2019 (Burkart et al., 2021). For England and Wales, Gasparrini et al. (2022) reported that each year, on average 791 excess deaths were attributable to heat, and 60,573 were attributable to cold. On the other hand, exposure to greenspace and urban greenery has been reported to lower mortality and morbidity and improve both general physical and mental health (Rojas-Rueda et al., 2019; James et al., 2016; Gascon et al., 2016a, 2016b). Barboza et al. (2021) reported meeting the WHO recommendation of access to greenspace (i.e., at least 0.5 ha of greenspace within 300 m of residences) could prevent 42,968 deaths per year in 31 European countries.

These findings on the adverse health effects of air pollution and non-optimal temperatures and the positive effects of greenspace have opened new avenues for studying the combined and interactive impacts of multiple environmental exposures. Consequently, several studies have investigated the health effects of multiple exposures on diverse populations and reported potential nonlinear relations between exposure variable and health outcomes (Kasdagli et al., 2021; Klompaker et al., 2021; Ji et al., 2020; Crouse et al., 2019; de Keijzer et al., 2017; Yitshak-Sade et al., 2017). Most of these studies investigating the health impact of one of these exposures considered other exposures as moderating, mediating, or confounding variables (Bloemsma et al., 2022; Jarvis et al., 2021; Crouse et al., 2019; de Keijzer et al., 2017). They frequently report that the observed relationships between air pollution, temperature, greenspace, and health may be mediated and moderated by the interactions among these exposure variables (Crouse et al., 2019; Ji et al., 2020; Dzhambov et al., 2018; Zhang et al., 2021; Denpetkul and Phosri, 2021). However, existing research that studies multi-exposure

interactions seldom considers spatial autocorrelation among exposures or health outcomes. In urban areas, there is often a co-location and clustering of multiple environmental exposures; for example, areas with low vegetation coverage may exhibit high concentrations of air pollution and high temperatures (Doiron et al., 2020; Yang et al., 2020; Dadvand et al., 2012). Such co-located, interactive exposures might be spatially correlated (de Keijzer et al., 2017; Browning and Rigolon, 2018; Verbeek, 2019; Elliott and Wartenberg, 2004). Failing to account for spatial autocorrelation when modeling the relationships between these exposures and health outcomes can lead to over- or underestimation of effect estimates, and exposure-response relation.

Furthermore, there is no universally accepted approach for incorporating spatial autocorrelation and nonlinearity when developing multi-exposure models. Previous studies have used several methods, such as spatial econometric models, spatial clustering, and geographically weighted regressions, in developing multi-exposure models while accounting for spatial effects (Iungman et al., 2021; Wang et al., 2023; Shuvo et al., 2021; Labib et al., 2021; de Keijzer et al., 2017). Additionally, studies have applied several statistical techniques to incorporate nonlinearity in the modeling process, such as applying splines, categorization of continuous variables and using generalized additive models. Recently, machine learning algorithms (e.g., Random Forest-RF, Artificial Neural Networks-ANN) have also been applied to study multi-exposure models due to their higher flexibility with highly correlated variables and no assumptions about the nature of the variables, while also accounting nonlinearity (Ohanyan et al., 2022a, b). Traditional spatial and statistical models, although capable of addressing some aspects of spatial autocorrelation, rely on strict assumptions about variable characteristics and data distribution (e.g., linearity of relationships), making them less suitable when these assumptions are not met (Wiemken and Kelley, 2020; Leist et al., 2022).

In contrast, machine learning models such as the random forest can capture nonlinear and non-additive associations without imposing strict assumptions about data type and distribution (Wiemken and Kelley, 2020; Seligman et al., 2018). Additionally, random forest models can account for complex interactions among variables simultaneously incorporating different types of variables (e.g., continuous, categorical). Considering such capabilities, it can be argued that such models can be an efficient modeling technique to investigate the influence of multiple exposures on health outcomes, where different exposures might have varying data distribution, complex interactions, and nonlinear relations. It should be noted that typical random forest models are not typically designed to incorporate spatial effects such as accounting for spatial autocorrelations. Hence previous studies overlooked such spatial aspects in the modeling process (Labib et al., 2023; Ohanyan et al., 2022a, 2022b; Liu et al., 2022). However, the recent development of a hybrid spatial random forest model has introduced a new approach to applying spatial regression with a random forest algorithm (Benito, 2021). Despite this advancement in spatially explicit random forest modeling, using such an approach to develop multi-exposure models remains unexplored.

Therefore, the aim of this present study was to investigate the associations of air pollution, greenspace, and temperature with mortality and morbidity by applying a spatial RF approach with a small-area ecological study to account for spatial effects in studying multi-exposure models also compare the results with non-spatial RF models. The specific objectives:

- To evaluate the relative importance of greenness, air pollution, and temperature exposures in predicting premature mortality and morbidity
- To model the exposure-response relations and interactions between greenness, air pollution, and temperature when predicting health outcomes

2. Materials and methods

2.1. Study area

This population-based small-area ecological study was based on data on mortality, morbidity, greenness, air pollution, and air temperature for the small areas (i.e., lower super output areas-LSOA) for the Greater Manchester City region (including ten metropolitan boroughs) in the United Kingdom. The city region has an area of 1277 km² consisting of 1673 LSOAs with a total population of 2,682,528 and a mean population of 1603 (standard deviation 394) (Labib et al., 2021). As the second largest urban conurbation in the UK, following London, the Greater Manchester City region is important for studying the health impacts of various environmental exposures, considering its growing population, gradual increase in air pollution, and temperature over the last few decades (Lindley and Walsh, 2005; Smith et al., 2011; Hyman et al., 2023); spatial variability of such exposures, and drives landscape patterns with varying greenness level across rural and urban gradient (Labib et al., 2021; Dennis et al., 2020). Additionally, the region is marked by sustained socio-spatial polarization, indicating higher levels of socioeconomic inequality and deprivation among neighborhoods compared to other provincial city regions in England (Hincks, 2015). These factors present a crucial opportunity to investigate the influence of multiple environmental exposures on population health in diverse urban neighborhoods.

2.2. Outcome definition

This study used two population-based health indicators indicating premature mortality and chronic morbidity levels at each LSOAs. The premature mortality data in this study refers to the measurement of years of potential life lost (YPLL) due to premature death before the age of 75, encompassing deaths from various causes, including diseases and external factors. The estimates were derived from mortality data obtained from the Office for National Statistics for the period between 2008 and 2012 (DCLG, 2015), and they represented all-cause mortality in each small area. It should be noted that YPLL accounted for the age of individuals who died unexpectedly at a younger age (before 75). YPLL values indicate the potential loss of life, productivity, and related socioeconomic loss owing to premature death, considering that individuals who faced premature death could have lived longer and contributed more to society (Gardner and Sanborn, 1990; Caraballo et al., 2023). Hence, it is a critical public health indicator to reflect the number and severity of health and socioeconomic burden due to premature deaths.

Furthermore, chronic morbidity data was incorporated, specifically measured through the Comparative Illness and Disability Ratio (CIDR), which was based on the English Indices of Multiple Deprivation (IMD) 2015 (DCLG, 2015). The CIDR indicator captures the prevalence of work-limiting morbidity and disability among individuals receiving benefits due to their inability to work caused by ill health. This data was obtained from the Department for Work and Pensions for the year 2013, and it provided a generalized measure of morbidity conditions. It is important to note that both health indicators have been standardized for age and sex, ensuring they are not susceptible to biases such as over-representing specific demographic groups or double counting. Further details of calculating these indicators can be found at DCLG (2015). Both outcome variables represented the overall state of mortality and morbidity in each small area, providing a generalized estimation in the

studied years. Previous studies also used such indicators to model the influence of environmental exposures (Dennis et al., 2020; Labib et al., 2021). It should be noted that health outcomes of 2008 to 2013 were best available data that ensured the consistency with all the exposure layers used in this study.

2.3. Exposure assessment

The main exposures were average greenness, air pollution concentration for nitrogen dioxide, and ambient air temperature in each LSOA. The availability of greenness at the small area scale was assessed using the Normalized Differential Vegetation Index (NDVI), a widely utilized measure derived from satellite imagery to indicate the relative abundance and spatial distribution of photosynthetically active vegetation. The NDVI has been extensively employed in research on greenness and its association with health outcomes (Martinez and Labib, 2023; Labib et al., 2020; Helbich, 2019). NDVI was estimated based on the reflectance measurements in the red (R) and the near-infrared (NIR) bands of the satellite sensor, and the values range from -1 to $+1$; in this case, higher values indicate a greater presence of greenness and a higher abundance of healthy vegetation (Rouse et al., 1974; Gascon et al., 2016a, 2016b). This study used Landsat 8 OLI (Operational Land Imager) images to estimate NDVI at 30 m spatial resolution for the year 2013. Google Earth Engine (GEE) was employed to identify images within the period spanning the first and last day of 2013, from which a composite image was generated. The composite image consisted of pixels with the median value across all identified pixels during this period, ensuring minimal cloud cover interference. A cloud removal mask was also applied to ensure the greenness data reflected vegetation presence on the ground. It should be noted that NDVI values for summer periods are generally considered in previous studies (Helbich, 2019; Gascon et al., 2016a, 2016b). However, due to considerable cloud coverage in the study area, NDVI values for a small period (e.g., summer months only) showed many missing pixels. Hence, a whole year was considered to ensure the least cloud cover pixels were used to extract NDVI values from the best available pixels. Additionally, such a yearly average is consistent with the air pollution and temperature exposures used in this study. In further processing, the negative values of NDVI were recoded as zero to exclude cells containing waterbodies. The average NDVI for each LSOA was calculated using the zonal statistics tool of QGIS (v23).

In this study, nitrogen dioxide (NO₂) has been considered as air pollutant, often commonly used in many health studies (Hyman et al., 2023; Bloemsma et al., 2022; de Keijzer et al., 2017). Annual concentration of nitrogen dioxide (NO₂) obtained from the Department for Environment, Food and Rural Affairs (DEFRA; <https://uk-air.defra.gov.uk/data/pcm-data>). The modeled NO₂ concentration was estimated at a 1×1 km resolution for the year 2013, represented in $\mu\text{g m}^{-3}$. The technical details of DEFRA model calibration and validation can be found at Brookes et al. (2014). The raw DEFRA data (a set of point locations) were converted to raster grid cells with 1×1 km using the vector-to-raster conversion tool (i.e., Rasterize) of QGIS. The annual NO₂ concentration grids were upscaled to LSOA by overlaying the grid to the LSOA polygon using the zonal statistics tool of QGIS to estimate the weighted mean of all the grid cells that have some parts within the given boundary of the LSOA (See Fig. 1a). The resulting area average NO₂ concentration was used in the analyses.

For each small area, the average air temperature has been estimated based on Land surface temperature (LST) for the year 2013. Landsat 8 OLI TIRS data were utilized within the period spanning the first and last day of 2013 to extract pixels with the median value of the thermal band. This approach considered the full year to account for both heat and cold-related mortality and morbidity (Gasparri et al., 2022). Additionally, it ensured that the exposure variable is a yearly estimate to match with the yearly measurement of health outcomes. The LST values were estimated at a spatial resolution of 30 m using the median pixel value of the

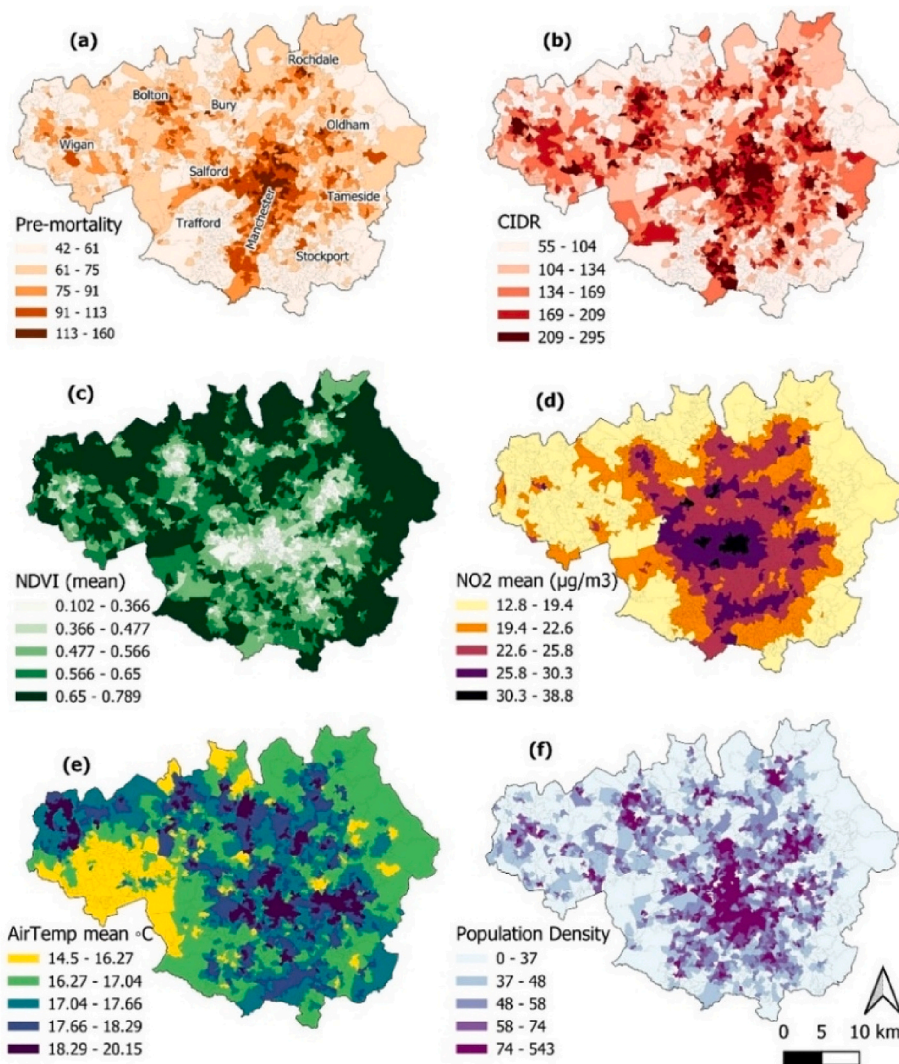


Fig. 1. Spatial pattern of health outcomes, exposures, and population density for the Greater Manchester region.

thermal band and a vegetation fraction-based emissivity algorithm outlined in Avdan and Jovanovska (2016). The 30×30 m grid LST values are upscaled to LSOA using zonal statistics to estimate the mean LST for the cells that has some parts within the given boundary of the LSOA. While LST exhibits a strong correlation with air temperature, it does not directly represent ambient air temperature (Jungman et al., 2023; Mutibwa et al., 2015). Therefore, the mean LST values were converted to mean air temperature using the methodology outlined by Marando et al. (2022). Marando and colleagues employed a linear regression model to estimate and validate air temperature based on LST and latitude for 601 Functional Urban Areas in Europe. Utilizing Eq. (1), the average air temperature for each Lower Super Output Area (LSOA) was estimated by considering the mean LST and the latitude of the LSOA centroid.

$$T_{air} = \beta_0 + \beta_1 * LST + \beta_2 * Latitude \quad (1)$$

Here, T_{air} is the average air temperature in °C. β_0 , β_1 , and β_2 are coefficients obtained from Marando et al. (2022).

2.4. Covariates

To account for socioeconomic conditions and other physical environmental contexts, various neighborhood-level indicators of deprivation and spatial indicators of the physical environment were considered

based on previous studies (Labib et al., 2021; Dennis et al., 2020; James et al., 2016). These indicators were collected from the Indices of Multiple Deprivation (IMD) scores (DCLG, 2015) and the Access to Healthy Assets and Hazards indicators (Daras et al., 2019). For each neighborhood, indicators related to socioeconomic status were collected from IMD sub-domain indicators. These indicators, including income deprivation, barriers to housing, and crime scores, served as proxies for income levels, the accessibility of housing and services, and crime rates. In addition, the physical environmental characteristics of each LSOA were obtained from Daras et al. (2019) and Dennis et al. (2020); these spatial indicators include the density of people, distance to nearest general medical practice (GPs), average distance to fast food outlets, and land use diversity (e.g., Shannon entropy). Several other covariates, such as education and health deprivation scores, were considered initially. However, they indicated high multicollinearity with income variables ($VIF > 5$). Therefore, these were excluded from the final modeling.

2.5. Machine learning models

Greenness, NO_2 concentration, and air temperature were linked to premature mortality and morbidity rates using two types of Random Forest (RF) regression models. RF is an ensemble machine learning technique based on bootstrap aggregation/bagging, as outlined in Breiman (2001). Many decision trees are generated by randomly

selecting subsets of predictors and observations from bootstrapped and original sample data. The predictions generated by these individual trees are then averaged to produce the final predicted value (Cutler et al., 2012; Breiman, 2001). For instance, in the context of this research, exposure, covariates, and health data were bootstrapped to create a replica of the original sample data, and then from this expanded sample (including both original and bootstrapped data), random rows were drawn to create “bags” of data. These bags of data were randomly placed into different decision trees to predict the health outcome variable. By taking the average predicted value of the health outcome from each decision tree, the final predicted value for the health outcome was estimated. During the training process of individual decision trees, a random subset of data was selected, and some observations from the bagged data were left out or considered “out-of-bag” (OOB). These OOB observations served as validation data to assess the prediction accuracy of the trained model (Cutler et al., 2012; Breiman, 2001). The predicted values from all OOB observations were compared to the actual outcome values (e.g., health outcomes) to evaluate prediction errors (e.g., Root Mean Square Error-RMSE), R-square, and the importance of predictor variables in predicting health outcomes. Permutation importance was utilized to determine the importance of predictor variables. Each predictor variable was individually subjected to random permutation (holding other variables constant), and the model's accuracy was re-computed using the OOB data. Variables that exhibited the greatest average decrease in accuracy after permutation was considered the most significant in terms of importance (Boehmke and Greenwell, 2019).

This study applied the aforementioned standard RF modeling approach to model the relations between multiple exposures and health outcomes. While this standard RF modeling approach is one of the most widely used machine learning algorithms in many fields, in essence, such RF models are non-spatial (Non-spatial RF) and are not adopted to model spatially explicit patterns (Hengl et al., 2018; Liu et al., 2022; Benito, 2021). Environmental exposures and public health outcome variables usually have spatial patterns due to the co-location and spatial clustering of multiple exposures and health outcomes (Labib et al., 2020; de Keijzer et al., 2017; Elliott and Wartenberg, 2004). Thus, not accounting for spatial patterns during the calibration of RF model parameters can potentially lead to sub-optimal predictions and systemic over or under-prediction of the effect of exposures on health outcomes due to the presence of spatial autocorrelation in residuals. Spatial autocorrelation of Non-spatial RF model residuals was tested using Moran's I statistics at multiple distance thresholds (i.e., 0, 100, 300, 500) to represent neighbors at multiple scales. If present, a spatially explicit RF (Spatial-RF) modeling approach was fitted to account for spatial autocorrelation.

Spatial-RF is an extension of standard non-spatial RF modeling to account for spatial autocorrelation. These models incorporate synthetic “Spatial predictors” into the standard non-spatial RF to help the model understand spatial dependence in training data (Benito, 2021). Spatial predictors were created based on a distance matrix representing neighbors among small area records (Hengl et al., 2018; Benito, 2021). This study applied Moran's Eigenvectors Maps (MEMs) approach to create spatial predictors. MEMs are orthogonal vectors of the double-centered distance matrix representing spatial weights between neighbors and reflecting the effect of spatial proximity on each other (Dray et al., 2006, 2012; Griffith, 1996). The MEMs operation created a few hundred synthetic spatial predictors (as no maximum was selected). However, a sequential optimization was applied to reduce dimensionality by selecting first “n” spatial predictors that minimize the spatial correlation of the residuals and maximize R-squared (Benito, 2021). The spatial predictors that do not affect model performance and reduce spatial correlation (e.g., no reduction of Moran's-I or Moran's-I of the spatial predictor equal or lower than 0) were removed.

For non-spatial and spatial RF comparisons, this study developed and tested four random forest models, as outlined in Table 1. All RF models were tuned for hyperparameters, such as numbers of decision trees,

Table 1

Outline of the RF models calibrated in this study.

SL	Model structure	Model type
Model-1	Premature mortality = Exposures + Covariates	Non-spatial RF
Model-2	Premature mortality = Exposures + Covariates + Synthetic spatial predictors	Spatial RF
Model-3	Morbidity Ratio = Exposures + Covariates	Non-spatial RF
Model-4	Morbidity Ratio = Exposures + Covariates + Synthetic spatial predictors	Spatial RF

minimum node size of the terminal nodes, and the number of variables considered for each split in the decision nodes (mtry) were optimized using five-fold spatial cross-validation (SCV) to ensure the models were not overfitted during the training process. During the training process using SCV, the dataset was divided into training and test sets (75 % and 25 % respectively).

To interpret the outputs of the machine learning models, notably the importance of independent variables and exposure-response relations, several explainable artificial intelligences or XAI techniques such as permutation importance (Altmann et al., 2010), partial dependence, and two-way interaction plots (Greenwell, 2017) have been utilized in this study. These techniques allowed opening the black box nature of machine learning algorithms and provided increased interpretability of the models by visualizing the relative importance of exposure variables and relationships among health outcomes and exposure variables (Leist et al., 2022; Petch et al., 2022; Zhao et al., 2021; Feng et al., 2022). All the model training, tests, and visualization were conducted using the spatialRF package (version 1.1.3; Benito, 2021) in R version (4.3.0).

3. Results

3.1. Descriptive analyses

This study includes 1673 small areas representing approximately 2.7 million residents. The geographical distributions of exposure and health outcome variables analyzed in this study showed spatial patterns as demonstrated in Fig. 1.

Out of the ten districts with the city region (Fig. 1a), central Manchester showed a higher level of premature mortality and morbidity (i.e., CIDR) ratio, along with high population density, low greenness, high air pollution, and temperature exposure compared to other districts. Each district central (e.g., Bolton center) area also exhibited similar trends (Fig. 1). The peripheries indicated contrasting patterns. Overall, the city region illustrated clear spatial clustering patterns in health outcomes and exposure variables, with central areas having higher health burdens, high air pollution, temperature, and low greenness exposure (Fig. 1 a-f). Such patterns indicate these variables potentially have high spatial autocorrelation, which might be crucial in modeling the relations.

For the studied small areas, the average premature mortality (years

Table 2

Health outcomes and exposure characteristics.

Variable	Minimum	Maximum	Mean	Std. deviation
Pre-mortality rate (Years of potential life lost)	42.369	159.592	74.473	18.325
Morbidity (Comparative Illness and Disability Ratio-CIDR)	54.566	295.388	144.299	45.415
Concentration of NO ₂ in µg/m ³	12.825	38.773	22.656	3.736
NDVI (unit less)	0.102	0.789	0.538	0.115
Air temperature in °C	14.501	20.146	17.373	0.842

of potential life lost) was 74.47 years, and the average CIDR ratio was 144.29 for the study period (Table 2). The CIDR ratio is relatively higher relative to other areas in England. A rate of 100 is the England average; mean values above 100 indicated a higher level of acute morbidity in the majority of the small areas within the studied region. The annual average NO₂ concentration for 2013 was 22.656 µg/m³, with a maximal concentration of 38.773 µg/m³. The average NO₂ concentration during the study period was higher than the current WHO guideline for NO₂ (i. e., 10 µg/m³). The mean NDVI values for the small areas ranged between 0.12 and 0.789, with an average of 0.538 for all small areas (Table 2). The high values of NDVI are primarily observed in the outskirts of the city region (Fig. 1c). The average annual temperature was 17.37 °C, with a yearly maximal temperature of 20.14 °C. The temperature values mostly showed fewer variations among studied small areas (Fig. 1e). Additionally, mean NDVI showed a strong negative correlation with NO₂ concentration ($r = -0.624$, $p < 0.01$) and a weak negative correlation with temperature ($r = -0.369$, $p < 0.01$) (details in Table S1, Supplementary material). NO₂ concentration showed a weak positive correlation with temperature ($r = 0.39$, $p < 0.01$). NO₂ concentration and temperature indicated moderate ($r = 0.49$, $p < 0.01$) and weak ($r = 0.239$, $p < 0.01$) positive correlations with premature mortality correspondingly and weak ($r = 0.355$, $p < 0.01$) and very weak ($r = 0.154$, $p < 0.01$) positive correlations with morbidity ratio. In contrast, NDVI exposure showed moderate ($r = -0.429$, $p < 0.01$) and weak ($r = -0.335$, $p < 0.01$) negative correlations with premature mortality and morbidity ratio. Lastly, premature mortality demonstrated a very strong positive correlation with the morbidity ratio. All correlations were statistically significant at $p < 0.01$ (Table S1). Descriptive statistics of covariates can be found in Table S2, Supplementary material.

3.2. Importance of exposures on predicting health outcomes

The model evaluation results of multi-exposure non-spatial and spatial random forest models are presented in Table 3. Table 3 shows for the premature mortality models, compared to non-spatial RF, the spatial RF model has higher R² and lower RMSE and no significant spatial autocorrelation. In contrast, for morbidity models, the non-spatial RF indicated higher R² and lower RMSE; however, the residual of the non-spatial RF showed significant spatial autocorrelation (Table 3), violating the independence assumption. The details of all model's autocorrelation results and final hyperparameters can be found in Table S3, supplementary material. Based on these evaluations, it can be argued that including spatial predictors in the RF models eliminated residual autocorrelation, thus improving the model's predictive performance and reliability. Therefore, Spatial RF models are considered the primary models to explain the exposure-response relationships in this study. Nonetheless, a comparison with the non-spatial RF models is also discussed.

The variable importance results for the multi-exposure final models are illustrated in Fig. 2. Across all models, income deprivation, barriers to housing, and crime scores showed greater relative importance in predicting premature mortality and morbidity ratio than any exposure variables. Regarding the exposure variables for premature mortality prediction (Fig. 2a, b), NO₂ concentration exhibited the highest relative

importance, followed by NDVI, and temperature showed the least importance in both non-spatial and spatial models. However, compared to the non-spatial RF (model-1), the spatial RF (model-2) illustrated a slight decrease in the relative importance of some exposure variables. Specifically, the relative importance of NO₂ concentration decreased from 6.073 in model-1 (Fig. 2a) to 5.574 in model-2 (Fig. 2b), indicating that accounting for spatial autocorrelation led to a slight reduction in the predictive relevance of NO₂ concentration. A similar trend was observed for NDVI. In contrast, the relative importance of temperature increased from 1.319 in model-1 to 1.647 in model-2. Additionally, in the spatial RF model, the relative importance of NDVI and temperature exposure increased compared to other controlling variables, such as GP distance and land use diversity, compared to what was observed in the non-spatial RF model (Fig. 2a, b). These results suggest that controlling for spatial autocorrelation might influence the relative importance of different exposure variables in predicting premature mortality.

For morbidity ratio, health outcome, income deprivation, barriers to housing, and crime scores were also the most important predictors for predicting morbidity in each small area. Among the exposure variables, in the non-spatial RF model, NDVI exposure exhibited the highest relative importance, followed by NO₂ concentration exposure, and temperature showed the least relative importance for morbidity (Fig. 2c). However, when accounting for spatial autocorrelation, the results changed. In the spatial RF model (Fig. 2d), NO₂ concentration indicated the highest predictive importance, followed by NDVI, while temperature consistently showed the least predictive importance for morbidity. Interestingly, the shift from the non-spatial RF (model-3) to the spatial RF (model-4) revealed an increase in the relative importance of all exposure variables in predicting morbidity ratio. Additionally, the relative importance of the exposure variable also changed in the spatial RF model; specifically, NO₂ exposure became a stronger predictor of morbidity compared to NDVI exposure. Moreover, controlling for spatial autocorrelation might demonstrate an increase or decrease in the relative importance of the exposure variable because each exposure might have distinct spatial patterns and how they relate to outcome variables.

3.3. Exposure-response relations

Fig. 3 presents the predicted exposure-response relationships for spatial and non-spatial RF models. Evidently, both model types depict nonlinear associations between exposures and health outcomes. In particular, the spatial RF models indicate smoother nonlinear relations and slightly attenuated exposure influence on health outcomes compared to the non-spatial RF models. Notably, in the non-spatial RF model, the major effects of NDVI exposure on premature mortality were observed between average NDVI range of 0.2–0.4, while in the spatial RF model, such effects were observed between average NDVI value of 0.2–0.5 (Fig. 3 a, b). Conversely, for morbidity outcomes, the primary impact of NDVI ranges between 0.3 and 0.5 (non-spatial RF) and 0.3–0.6 (spatial RF) (Fig. 3g, h). Similarly, for NO₂ concentration, it is clear that the main effects of NO₂ exposure on premature mortality were observed between the range of 20–30 µg/m³ (Fig. 3 c, d). However, the spatial RF curve showed a gradual increase in premature mortality as increasing NO₂ concentration, and above 30, it indicated the largest increase, compared to the non-spatial RF model, which revealed more inconsistent effects of NO₂ concentration on premature deaths. Similar relations for morbidity health outcomes were also observed for NO₂ exposure and morbidity ratio. Finally, for temperature exposure, the exposure-response relation indicated a “U” shape relation (Fig. 3 e, f, k, l). Increased mortality and morbidity were observed at lower temperature exposures below 16 degrees. As temperature increased up to 17–18 °C, mortality and morbidity were reduced. However, as the temperature exposure increased beyond 18 °C, there was a gradual increase in premature deaths and morbidity ratio, and a further rise in temperature exposure indicated higher negative health consequences. Moreover, accounting for spatial autocorrelation in a machine-learning model can

Table 3
Exposure and health outcome model evaluations.

Model	R-square (OOB)	RMSE (OOB)	Moran's-I (p-value)
Model-1 (Non-spatial RF, pre-mortality)	0.671	10.51	0.03 (0.0001)
Model-2 (Spatial RF, pre-mortality)	0.725	9.61	0.001(0.116)
Model-3 (Non-spatial RF, morbidity)	0.889	15.098	0.033 (0.0001)
Model-4 (Spatial RF, morbidity)	0.834	18.531	0.001 (0.093)

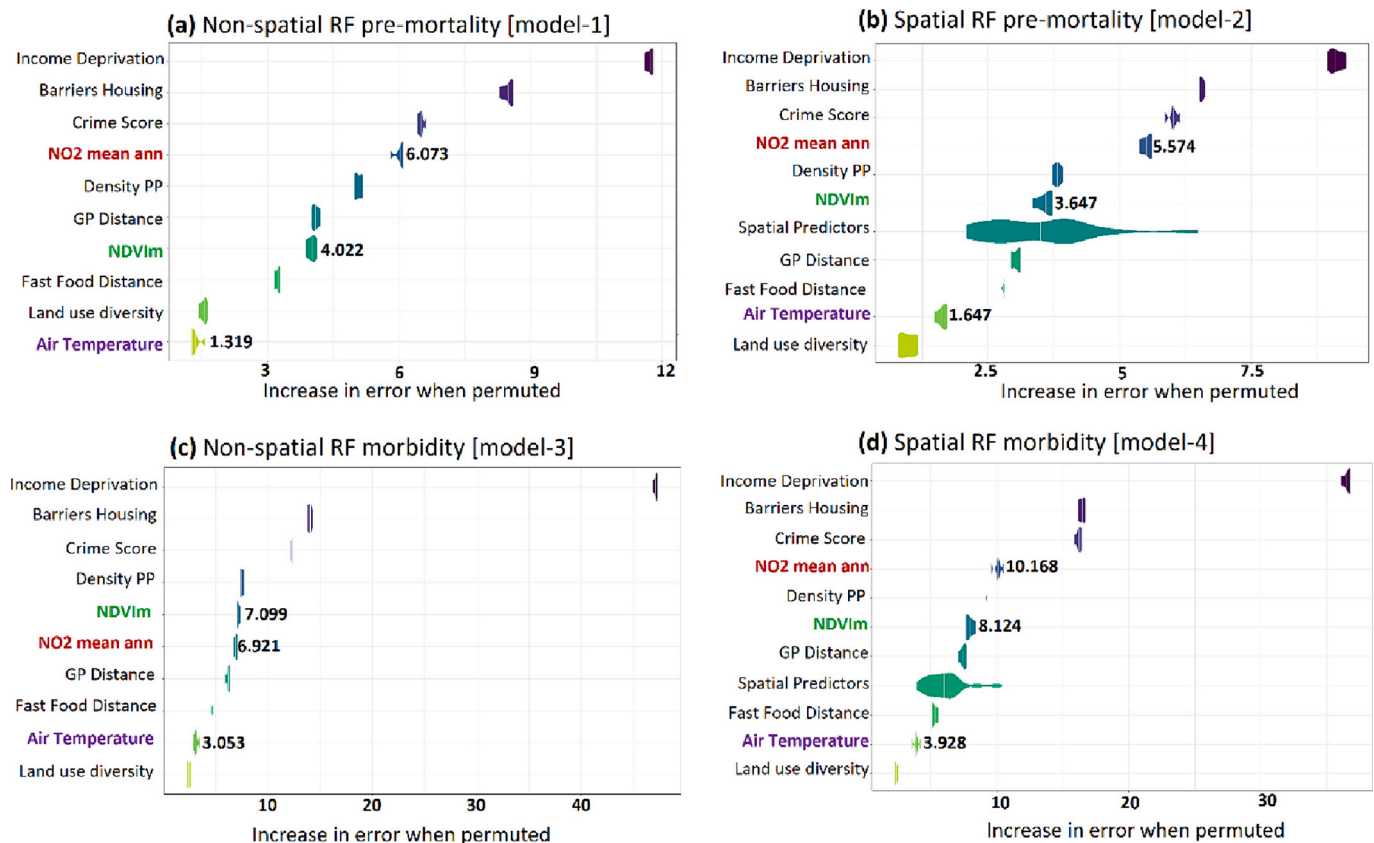


Fig. 2. Permutation-based variable importance for predictors over fivefold cross-validation; higher values indicate prediction error increases if the variable drops from the model when permeated with all other variables. Spatial predictors are the *Synthetic spatial variables* included in the model to account for spatial autocorrelation in non-spatial model residual.

lead to varying exposure relations as nonlinear trends might differ based on the model's adjustment for spatial autocorrelations.

3.4. Exposure interaction effects

Two-way interactions between exposures based on the spatial RF-modeled prediction surface have been presented in Fig. 4. The prediction surface indicates interactions between two exposures in the prediction process and different zones of health risk or benefits based on these exposures. For instance, Fig. 4a illustrates the interactions between NDVI and NO₂ concentration exposure for premature mortality, and it indicates very high risks (red-dotted box) of premature mortality for NDVI values <0.2 and NO₂ concentration > 30 µg/m³. By contrast, NDVI>0.5 and NO₂ concentration < 20 µg/m³ indicate the lowest risk exposure zone (black-dotted box). Other predicted values of premature mortality indicate the intermediate zones of interactions between NDVI and NO₂ concentration, reflecting the relative influence of these exposures on predicting premature mortality. Areas with high NO₂ concentrations between 25 and 30 µg/m³ indicated lower premature mortality predictions if the average NDVI exposures were >0.4. Such interactions suggest that greening might moderate the health impact of higher air pollution exposure in certain areas. For morbidity, similar patterns existed for NDVI and NO₂ concentration (Fig. 4b), although the high and low-risk and intermediate exposure zones are slightly different than those of premature mortality outcome.

The interactions between NDVI and mean air temperature also indicate several high-risk and low-risk exposure thresholds for both premature mortality and morbidity ratios, as indicated in Fig. 4c, d. It is clear that temperatures between 17 and 18 °C and average NDVI >0.5 are associated with lower predicted mortality and morbidity. Notably, certain predicted values on the prediction surface indicate in areas with

high-temperature exposure (e.g., 19 °C), NDVI >0.6 might lower the mortality and morbidity burden, thus illustrating the potential health benefits of cooling associated with higher greenness. Furthermore, interaction plots between NO₂ concentration and mean air temperature illustrate similar high, low risk, and intermediate exposure thresholds related to health outcome predictions (Fig. 4 e, f). Interestingly, for the intermediate interaction zones between these exposures, areas with very high NO₂ concentration (>30 µg/m³) and average air temperature exposure between 17 and 18 °C predicted a slightly lower health impact for both health outcomes. As a whole, the high and low-risk exposure zones pinpoint the sweet and sour spots of exposure interactions, thresholds, and intermediate zones of interaction effects of exposures on health outcomes are critical to better understand the potential moderating effects of one exposure to another. The interaction results of the non-spatial RF models also indicated similar patterns and can be found in the supplementary document Fig. S1.

4. Discussion

4.1. Main findings

This study presented the application of random forest models in predicting premature mortality and morbidity conditions based on multiple environmental exposures (i.e., NO₂, NDVI, and Air temperature) while accounting for spatial autocorrelation for an ecological study design over a large city region (i.e., Greater Manchester). To the best of the author's knowledge, no prior study has employed spatial RF models in similar contexts and study designs. This study indicates that while the conventional non-spatial RF models adequately predict health outcomes for small areas, such models might provide suboptimal results due to spatial autocorrelation in model residuals, which might influence

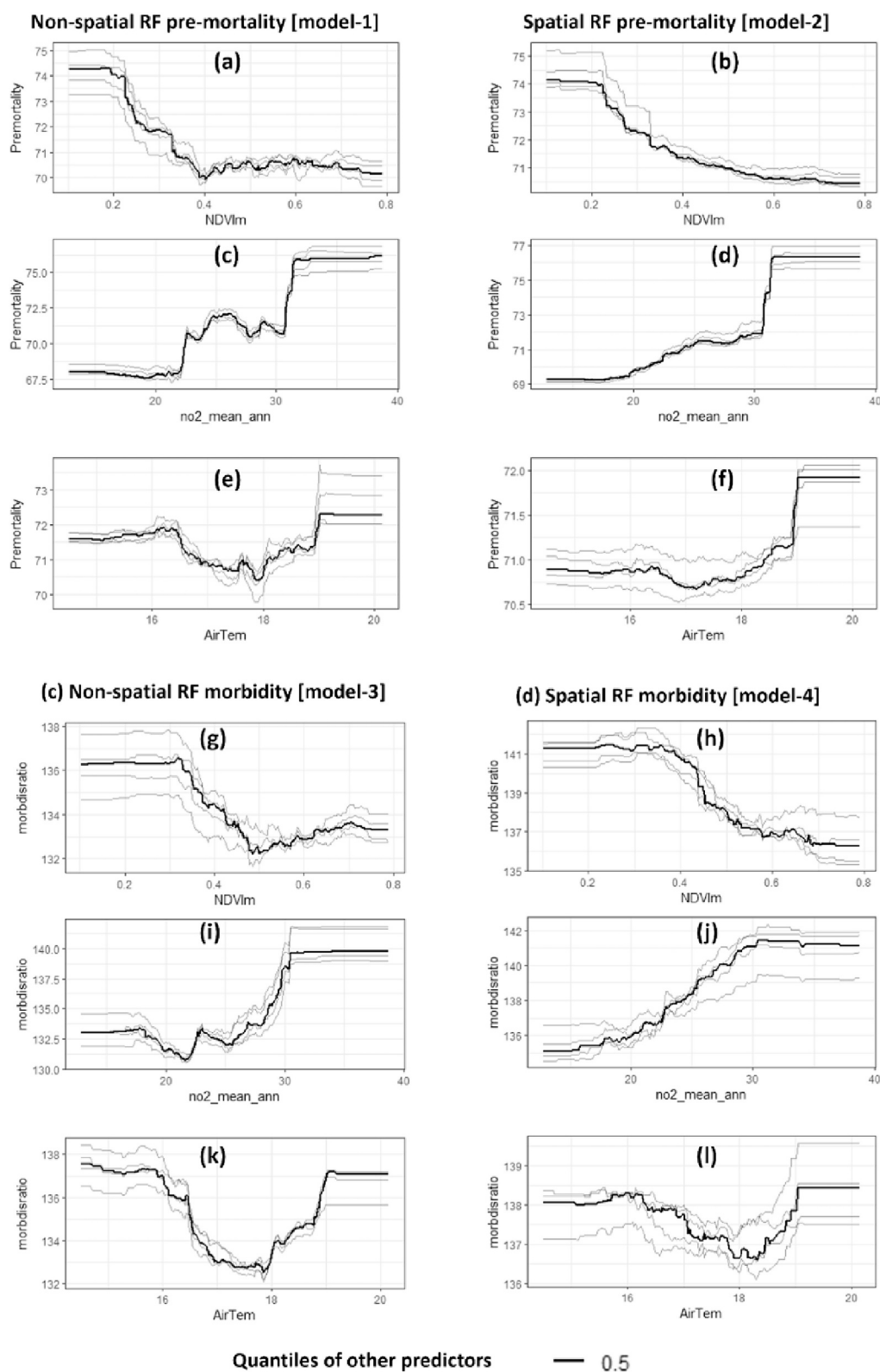


Fig. 3. Predicted exposure-response relations for non-spatial and spatial RF models based on partial dependence plots. The relations have been evaluated on a five-fold spatial cross-validation approach, illustrated as grey lines. All other variables in the model were set to quantiles when predicting the exposure-response relationships.

exposure-response relations, induce greater prediction errors, and lower R^2 (see Table 3). This study demonstrates when environmental exposure variables that are spatially explicit and exhibit spatial patterns are used in predicting or analyzing health outcomes, ignoring the spatial

autocorrelation in the modeling process (e.g., non-spatial RF model) results in suboptimal predictions.

Additionally, using explanatory approaches to infer the feature importance or exposure-response relations based on such a suboptimal

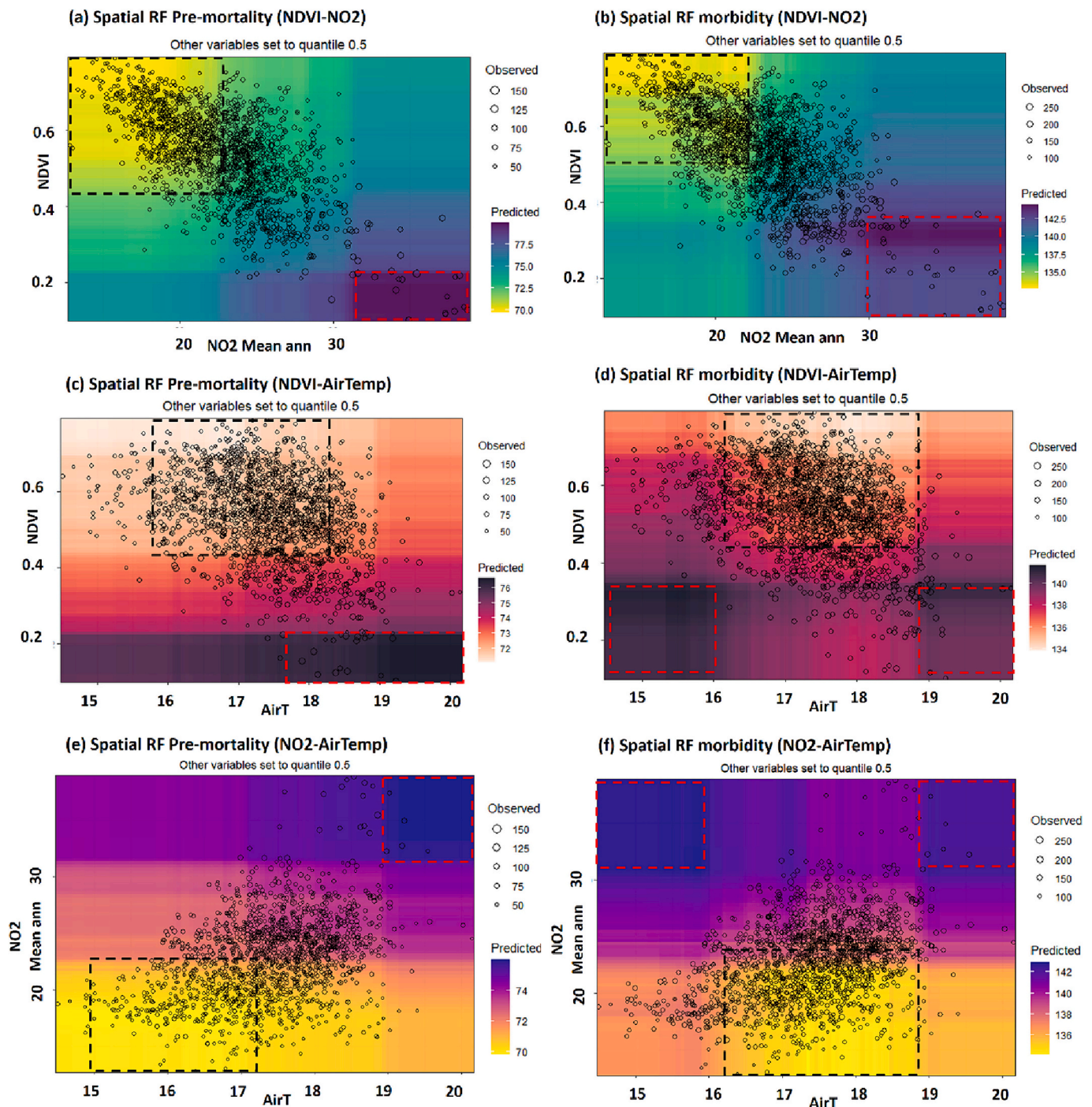


Fig. 4. Two-way interaction effects of varying exposure metrics in predicting health outcomes. The black boxes indicate low and red boxes indicate high health risk zones. In-between areas of interactions are noted as intermediate effect moderation zones.

model may lead to under or over-estimation of the relative importance of exposure on health outcome and might illustrate inconsistent exposure-response relations, as highlighted in [Sections 3.2 and 3.3](#), respectively. Therefore, studies using machine learning models using spatially explicit exposure or health data should consider testing spatial autocorrelation and potentially use spatially explicit machine learning models to obtain greater reliability and accuracy in predicting as well as explaining health outcomes. Several recent studies have applied machine learning algorithms in predicting health outcomes ([Choi et al., 2023](#); [Ohanyan et al., 2022a, 2022b](#); [Wei et al., 2022](#); [Gatti et al., 2020](#)); unfortunately, these studies did not investigate the potential issues

related to spatial dependency in their model specification and did not adjust their models accordingly. Such omission of spatial autocorrelation might have affected the observed associations and predictions they obtained. The effects of omission of spatial dependency in the model specification have already been identified in previous studies where authors discussed that ignoring spatial autocorrelations would affect the predictions and, in some cases, could nullify the observed associations in traditional statistical models ([de Keijzer et al., 2017](#); [Hodges and Reich, 2010](#)).

The models of this study illustrated that the relative importance of yearly NO₂ concentration, NDVI, and air temperature exposure in

predicting premature mortality and morbidity can slightly differ between non-spatial and spatial RF models. Among the exposures considered, NO₂ concentration indicated the higher feature importance in explaining health outcomes, usually followed by NDVI exposure in the spatial RF models. In the non-spatial RF model, the relative importance of NDVI was slightly above the NO₂ exposure; however, when spatial autocorrelation is accounted for in the spatial RF model, the relative importance of NO₂ exposure becomes stronger. Air temperature indicated the least impact on health outcomes for this study area during the period of the study in all the models. These relative importance of exposure variables are somewhat consistent with previous multi-exposure studies, where NO₂ exposure also indicated greater influence on health outcomes compared to greenness exposures (de Keijzer et al., 2017; Kasdagli et al., 2021; Avellaneda-Gómez et al., 2022). However, these studies applied traditional statistical models and often ignored the presence of potential interaction effects of multiple exposures. The low relative importance of air temperature exposure in explaining health outcomes may reflect the lower variability of temperature distribution observed in the study area during the study period. Hence, more longitudinal studies covering large spatial areas with multiple exposures simultaneously included in the models are required to better understand the relative impact of varying exposures and their relations with health outcomes.

This study used partial dependence plots to identify nonlinear exposure-response relations in a multi-exposure model (Fig. 3). The results illustrated the nonlinear exposure-response for all the exposure variables considered to explain mortality and morbidity outcomes, although the curves showed slightly different patterns between non-spatial and spatial RF models. The nonlinear exposure patterns for greenness, NO₂ concentration, and ambient temperature metrics align with prior research (Ohanyan et al., 2022a; Labib et al., 2023; Ji et al., 2020; Zhu et al., 2017). Notably, for temperature exposure, the “U” shaped exposure-response curve observed in this study was also found in Gasparrini et al. (2015) and Curriero et al. (2002). While the previous studies used traditional statistical models applying various splines, this study underscores the ability of machine learning algorithms to discern nonlinear trends within exposure-response functions, even within cross-sectional and ecological study designs. It should be noted that compared to the traditional statistical models, for the machine learning algorithm, such as random forest, the partial dependence plots (PDP) construct the relations exposure-response using the tuned model to predict health outcomes based on a single exposure variable showing the marginal effect while holding other variables constant (Greenwell, 2017; Petch et al., 2022). In this study, the curves are created based on PDP over fivefold cross-valuation to ensure the identified nonlinear relations are not overfitted to the observed data only. Hence, the curves illustrate more localized variations, which are often averaged out when splines are used to explain such relations.

Nonlinearity in exposure-response is pivotal for comprehending critical exposure thresholds, as illustrated in this study. The traditional statistical approaches, such as the Cox proportional hazard and logistic models, often use exposure intervals (e.g., 0.1 increment in NDVI, interquartile range of NO₂ concentration) to identify the nonlinear effect between the lowest and highest exposure ranges (Zhang et al., 2021; Klompaker et al., 2021; Ji et al., 2020; Yitshak-Sade et al., 2017); however, such categorization of continuous exposure values could lead to loss of information concerning the intermediate exposure-response relations. By contrast, the exposure-response curves generated by the explainable machine learning approaches (e.g., variable importance, partial dependence plot, Shapley Additive exPlanations-SHAP) do not need categorization of exposure variables and can provide a nuanced representation of sharp changes in nonlinear exposure-response relations (Ren et al., 2023; Nohara et al., 2022; Elshawi et al., 2019; Wiemken and Kelley, 2020).

One of the major contributions of the current study is to investigate the interaction effects of multiple exposure variables in predicting

premature mortality and morbidity levels. The two-way interaction plots identified the sweet and sour zones of exposures, illustrating the prediction of low and high levels of premature mortality and morbidity underpinning the effects of multiple exposures (details in Fig. 4). The intermediate zones of exposure interactions are crucial to explore the moderating effects of one exposure to another. For instance, in areas with high NO₂ concentration (>30 µg/m³), higher or moderate NDVI values (> 0.5) or presence of optimal air temperature (around 17–18 °C) indicated lower predicted mortality and morbidity. Such interaction effects are critical to identifying how much increase or decrease of a specific exposure may result in effect modification in other exposure. The results of effect modifications of one exposure to another found in this study are mostly consistent with several previous studies using conventional statistical models. For instance, Klompaker et al. (2021) noted that the strength of the association of NO₂ exposure on health outcomes across greenness tertiles might be modified, as NO₂ concentrations were lower with increasing greenness. Ji et al. (2022) indicated due to moderation in the co-exposure model, NO₂ exposure appeared to be harmful in places of colder climates. Zhang et al. (2021) showed higher greenness might protect against high-temperature exposure. The result of this study is not in line with Xu et al. (2023), where the authors did not observe any significant interaction between NO₂ and greenness exposure. However, it should be noted that some of these studies often indicated modification effects based on statistical tests. By contrast, the prediction surface used in this study does not provide an exact percentage for modification; instead, it focuses on prediction while accounting for multiple exposures together. Nonetheless, the interaction-based predictions identified in this study are reasonable considering the current understanding of pathways of how greenness can moderate the adverse impact of higher air pollution by reducing air pollutants or by providing cooling benefits, as discussed in several studies (Markelych et al., 2017; Nieuwenhuijsen et al., 2017; lungman et al., 2023).

4.2. Strengths and implications of this study

In addition to the empirical evidence, this study has several crucial methodological strengths and implications for future research in applying machine learning approaches in studying environment health relations. First, it is one of the foremost studies that considered applying an ensemble machine learning algorithm incorporating spatial dependency among multiple exposures and health outcomes to model nonlinear exposure-response relations and identify interactions between exposures. Several traditional statistical approaches are available to account for nonlinearity or spatial autocorrelation. However, most of these traditional statistical models are limited due to various assumptions (e.g., linearity, multicollinearity) about the data they use within the modeling process; these assumptions often become more difficult to meet when high dimensional data are used to model nonlinear relations, and when the data has correlated/co-located observations (Wiemken and Kelley, 2020; Leist et al., 2022). This study illustrated how a hybrid ensemble machine learning algorithm can account for nonlinearity in relations, spatial dependency, and interaction effect identification. Such an approach provides an innovative modeling structure for future studies dealing with several of these issues in their exposure and health modeling process.

Second, this study applied several visualization techniques (e.g., partial dependence plots) for the interpretability of the modeled relations. Such interpretability is critical to understanding the underlying relations observed in the machine learning algorithm. This can increase the current understanding of complex nonlinear and interactive relations among diverse exposures and health outcomes. Using explainable artificial intelligence or XAI tools to open the black boxes of machine learning can become critical outputs for epidemiological studies to decode and predict complex relations between multiple exposures and health outcomes, which traditional statistical models might be unable to investigate with higher accuracy than machine learning

models (Leist et al., 2022; Zhao et al., 2021).

Finally, the method of this study can be applied to individual cohort data to study health outcomes and multiple exposures while accounting for spatial autocorrelation. Such an application would provide more reliable model development by incorporating more individual confounder variables to ensure the observed relations are more reliable. Once such models are developed on large cohort data, the tuned models can be used for health impact assessment for counterfactual exposure scenarios, such as reduced air pollution, increased greenness, and optimum temperature exposures. As machine learning models often outperform traditional statistical models in terms of predictive accuracy (Wiemken and Kelley, 2020; Leist et al., 2022; Feng and Jiao, 2021), models trained on extensive cohort data would become critical resources to study future exposure conditions under varying policy scenarios with higher confidence.

4.3. Limitations and future developments

The results of this study are limited by its ecological design, as such design is always affected by ecological fallacy. Additionally, ecological design might have resulted in exposure misclassification and higher aggregation of exposure at a large spatial scale, which might have affected the exposure variations observed in small area geography (Blakely and Woodward, 2000; Wang et al., 2017). However, the study area had a large coverage of different small areas (e.g., core urban areas with high density and peripheral regions with low density), thus providing opportunities to investigate spatially varying exposure and health outcomes. Although several commonly used confounders were included in this study based on previous research, these covariates were only area-level measures, which are unable to capture individual variations of exposure and health behaviors (Feng and Jiao, 2021; Labib et al., 2021). In particular, some of these covariates worked as surrogate measures of multiple risk factors, such as barriers to housing, including housing multiple conditions that might influence individual health conditions. However, the models were not adjusted for individual housing conditions, smoking, food consumption, or physical activity behaviors. Another crucial limitation of this study is the use of a cross-section dataset; hence no causality could be inferred from the modeled relations. Furthermore, the RF models used in this study are unable to provide confidence interval and significance values of the variable importance and predictions due to their deterministic approach to calibrating the model. Therefore, the results cannot be directly compared to previous studies applying traditional statistical models to determine effect size and significance. However, it should be noted that the RF models were trained and tested over a fivefold spatial cross-validation and hyperparameter tuning to ensure confidence in the prediction. Finally, the spatial RF models are computationally intensive as they add many synthetic spatial predictions and often take longer to train and test the models over multiple cross-validations.

5. Conclusion

In conclusion, this study presents a new way of predicting and explaining premature mortality and morbidity using machine learning models, which can account for spatial dependency among the variables. Ignoring spatial dependency in machine learning models can provide suboptimal predictions and under or overestimation of the relative importance of exposure effects on health outcomes. Using robust spatial random forest models, this study showed yearly average air pollution exposure is relatively more important in predicting premature mortality and morbidity than greenness and temperature exposure. The nonlinear exposure-response relations showed usually increasing NO₂ concentration, lower greenness, and too-high or too-low temperature exposure associated with higher mortality and morbidity conditions. Additionally, examining interactions between multiple exposures reveals specific ranges where combined effects either exacerbate or moderate health

outcomes, highlighting critical thresholds for exposure impacts.

CRedit authorship contribution statement

S.M. Labib: Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

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

Data will be made available on request.

Acknowledgment

The author would like to thank the reviewers and editor of this paper for their constructive comments and suggestions. Thanks to Prof. Mark Nieuwenhuijsen (ISGlobal) and Prof James Woodcock (University of Cambridge) for their initial comments on the work presented in this manuscript in the workshop hosted at University of Cambridge.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.scitotenv.2024.172387>.

References

- Altmann, A., Tološi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics* 26 (10), 1340–1347.
- Avdan, U., Jovanovska, G., 2016. Algorithm for automated mapping of land surface temperature using LANDSAT 8 satellite data. *J. Sens.* 2016, 1–8.
- Avellaneda-Gómez, C., Vivanco-Hidalgo, R.M., Olmos, S., Lazzano, U., Valentín, A., Milà, C., Ambrós, A., Roquer, J., Tonne, C., 2022. Air pollution and surrounding greenness in relation to ischemic stroke: a population-based cohort study. *Environ. Int.* 161, 107147.
- Barboza, E.P., Cirach, M., Khomenko, S., Iungman, T., Mueller, N., Barrera-Gómez, J., Rojas-Rueda, D., Kondo, M., Nieuwenhuijsen, M., 2021. Green space and mortality in European cities: a health impact assessment study. *Lancet Planet. Health* 5 (10), e718–e730.
- Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z.J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., Vineis, P., 2014. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. *Lancet* 383 (9919), 785–795.
- Benito, B.M., 2021. spatialRF: Easy Spatial Regression With Random Forest. R Package Version 1.1.0. <https://doi.org/10.5281/zenodo.4745208>.
- Blakely, T.A., Woodward, A.J., 2000. Ecological effects in multi-level studies. *J. Epidemiol. Community Health* 54 (5), 367–374.
- Bloemsma, L.D., Wijga, A.H., Klompaker, J.O., Hoek, G., Janssen, N.A., Lebret, E., Brunekreef, B., Gehring, U., 2022. Green space, air pollution, traffic noise and mental wellbeing throughout adolescence: findings from the PIAMA study. *Environ. Int.* 163, 107197.
- Boehmke, B., Greenwell, B.M., 2019. *Hands-on Machine Learning with R*. CRC press.
- Breiman, L., 2001. Random Forests. *Machine Learning*, 45, pp. 5–32.
- Brookes, D.M., et al., 2014. Technical Report on UK Supplementary Assessment Under The Air Quality Directive (2008/50/EC), The Air Quality Framework Directive (96/62/EC) and Fourth Daughter Directive (2004/107/EC) for 2014. 210. Available at: https://uk-air.defra.gov.uk/library/reports?report_id=993.
- Browning, M.H., Rigolon, A., 2018. Do income, race and ethnicity, and sprawl influence the greenspace-human health link in city-level analyses? Findings from 496 cities in the United States. *Int. J. Environ. Res. Public Health* 15 (7), 1541.
- Burkart, K.G., Brauer, M., Aravkin, A.Y., Godwin, W.W., Hay, S.I., He, J., Iannucci, V.C., Larson, S.L., Lim, S.S., Liu, J., Murray, C.J., 2021. Estimating the cause-specific relative risks of non-optimal temperature on daily mortality: a two-part modelling approach applied to the Global Burden of Disease Study. *Lancet* 398 (10301), 685–697.
- Caraballo, C., Massey, D.S., Ndumele, C.D., Haywood, T., Kaleem, S., King, T., Liu, Y., Lu, Y., Nunez-Smith, M., Taylor, H.A., Watson, K.E., 2023. Excess mortality and years of potential life lost among the black population in the US, 1999–2020. *JAMA* 329 (19), 1662–1670.

- Choi, E.S., Lee, J.S., Hwang, Y., Lee, K.S., Ahn, K.H., 2023. Association between early preterm birth and maternal exposure to fine particulate matter (PM₁₀): a nation-wide population-based cohort study using machine learning. *PLoS One* 18 (8), e0289486.
- Crouse, D.L., Pinaut, L., Balram, A., Brauer, M., Burnett, R.T., Martin, R.V., Van Donkelaar, A., Villeneuve, P.J., Weichenthal, S., 2019. Complex relationships between greenness, air pollution, and mortality in a population-based Canadian cohort. *Environ. Int.* 128, 292–300.
- Curriero, F.C., Heiner, K.S., Samet, J.M., Zeger, S.L., Strug, L., Patz, J.A., 2002. Temperature and mortality in 11 cities of the eastern United States. *Am. J. Epidemiol.* 155 (1), 80–87.
- Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random Forests. *Methods and applications, Ensemble machine learning*, pp. 157–175.
- Dadvand, P., de Nazelle, A., Triguero-Mas, M., Schembari, A., Cirach, M., Amoly, E., Figueras, F., Basagaña, X., Ostro, B., Nieuwenhuijsen, M., 2012. Surrounding greenness and exposure to air pollution during pregnancy: an analysis of personal monitoring data. *Environ. Health Perspect.* 120 (9), 1286–1290.
- Daras, K., Green, M.A., Davies, A., Barr, B., Singleton, A., 2019. Open data on health-related neighbourhood features in Great Britain. *Sci. Data* 6 (1), 107.
- DCLG (Department for Communities and Local Government), 2015. English indices of deprivation [computer file]. Available online: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015> Or https://assets.publishing.service.gov.uk/media/5a7f24b240f0b62305b85578/English_Indices_of_Deprivation_2015_-_Technical-Report.pdf. Licensed Under: <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3>.
- Dennis, M., Cook, P.A., James, P., Wheat, C.P., Lindley, S.J., 2020. Relationships between health outcomes in older populations and urban green infrastructure size, quality and proximity. *BMC Public Health* 20 (1), 1–15.
- Denpetkul, T., Phosri, A., 2021. Daily ambient temperature and mortality in Thailand: estimated effects, attributable risks, and effect modifications by greenness. *Sci. Total Environ.* 791, 148373.
- Doiron, D., Setton, E.M., Shairsingh, K., Brauer, M., Hystad, P., Ross, N.A., Brook, J.R., 2020. Healthy built environment: spatial patterns and relationships of multiple exposures and deprivation in Toronto, Montreal and Vancouver. *Environ. Int.* 143, 106003.
- Dominski, F.H., Branco, J.H.L., Buonoanno, G., Stabile, L., da Silva, M.G., Andrade, A., 2021. Effects of air pollution on health: a mapping review of systematic reviews and meta-analyses. *Environ. Res.* 201, 111487.
- Dray, S., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol. Model.* 196 (3–4), 483–493.
- Dray, S., Péllissier, R., Couteron, P., Fortin, M.J., Legendre, P., Peres-Neto, P.R., Bellier, E., Bivand, R., Blanchet, F.G., de Cáceres, M., Dufour, A.B., 2012. Community ecology in the age of multivariate multiscale spatial analysis. *Ecol. Monogr.* 82 (3), 257–275.
- Dzhambov, A., Hartig, T., Markevych, I., Tilov, B., Dimitrova, D., 2018. Urban residential greenspace and mental health in youth: different approaches to testing multiple pathways yield different conclusions. *Environ. Res.* 160, 47–59.
- Elliott, P., Wartenberg, D., 2004. Spatial epidemiology: current approaches and future challenges. *Environ. Health Perspect.* 112 (9), 998–1006.
- Elshawi, R., Al-Mallah, M.H., Sakr, S., 2019. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med. Inform. Decis. Mak.* 19 (1), 1–32.
- Feng, C., Jiao, J., 2021. Predicting and mapping neighborhood-scale health outcomes: a machine learning approach. *Comput. Environ. Urban. Syst.* 85, 101562.
- Feng, S., Meng, Q., Guo, B., Guo, Y., Chen, G., Pan, Y., Zhou, J., Xu, J., Zeng, Q., Wei, J., Xu, H., 2022. Joint exposure to air pollution, ambient temperature and residential greenness and their association with metabolic syndrome (MetS): a large population-based study among Chinese adults. *Environ. Res.* 214, 113699.
- Forouzanfar, M.H., Afshin, A., Alexander, L.T., Anderson, H.R., Bhutta, Z.A., Biryukov, S., Brauer, M., Burnett, R., Cercy, K., Charlson, F.J., Cohen, A.J., 2016. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* 388 (10053), 1659–1724.
- Gardner, J.W., Sanborn, J.S., 1990. Years of potential life lost (YPLL)—what does it measure? *Epidemiology* 1 (4), 322–329.
- Gascon, M., Cirach, M., Martínez, D., Dadvand, P., Valentín, A., Plasència, A., Nieuwenhuijsen, M.J., 2016a. Normalized difference vegetation index (NDVI) as a marker of surrounding greenness in epidemiological studies: the case of Barcelona city. *Urban For. Urban Green.* 19, 88–94.
- Gascon, M., Triguero-Mas, M., Martínez, D., Dadvand, P., Rojas-Rueda, D., Plasència, A., Nieuwenhuijsen, M.J., 2016b. Residential green spaces and mortality: a systematic review. *Environ. Int.* 86, 60–67.
- Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., Tobias, A., Tong, S., Rocklöv, J., Forsberg, B., Leone, M., 2015. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *Lancet* 386 (9991), 369–375.
- Gasparrini, A., Masselot, P., Scottichini, M., Schneider, R., Mistry, M.N., Sera, F., Macintyre, H.L., Phalkey, R., Vicedo-Cabrera, A.M., 2022. Small-area assessment of temperature-related mortality risks in England and Wales: a case time series analysis. *Lancet Planet. Health* 6 (7), e557–e564.
- Gatti, R.C., Velichevskaya, A., Tateo, A., Amoroso, N., Monaco, A., 2020. Machine learning reveals that prolonged exposure to air pollution is associated with SARS-CoV-2 mortality and infectivity in Italy. *Environ. Pollut.* 267, 115471.
- Greenwell, B.M., 2017. pdp: an R package for constructing partial dependence plots. *R J.* 9 (1), 421.
- Griffith, D.A., 1996. Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Can. Geogr.* 40 (4), 351–367.
- Helbig, M., 2019. Spatiotemporal contextual uncertainties in green space exposure measures: exploring a time series of the normalized difference vegetation indices. *Int. J. Environ. Res. Public Health* 16 (5), 852.
- Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* 6, e5518.
- Hincks, S., 2015. Neighbourhood change and deprivation in the Greater Manchester city-region. *Environ. Plan. A* 47 (2), 430–449.
- Hodges, J.S., Reich, B.J., 2010. Adding spatially-correlated errors can mess up the fixed effect you love. *Am. Stat.* 64 (4), 325–334.
- Hunter, R.F., Nieuwenhuijsen, M., Fabian, C., Murphy, N., O'Hara, K., Rappe, E., Sallis, J. F., Lambert, E.V., Duenas, O.L.S., Sugiyama, T., Kahlmeier, S., 2023. Advancing urban green and blue space contributions to public health. *Lancet Public Health* 8 (9), e735–e742.
- Hymán, S., Zhang, J., Andersen, Z.J., Cruickshank, S., Møller, P., Daras, K., Williams, R., Topping, D., Lim, Y.H., 2023. Long-term exposure to air pollution and COVID-19 severity: a cohort study in Greater Manchester, United Kingdom. *Environ. Pollut.* 327, 121594.
- Iungman, T., Khomenko, S., Nieuwenhuijsen, M., Barboza, E.P., Ambrós, A., Padilla, C., Mueller, N., 2021. The impact of urban and transport planning on health: assessment of the attributable mortality burden in Madrid and Barcelona and its distribution by socioeconomic status. *Environ. Res.* 196, 110988.
- Iungman, T., Cirach, M., Marando, F., Barboza, E.P., Khomenko, S., Masselot, P., Quijal-Zamorano, M., Mueller, N., Gasparrini, A., Urquiza, J., Heris, M., 2023. Cooling cities through urban green infrastructure: a health impact assessment of European cities. *Lancet* 401 (10376), 577–589.
- James, P., Hart, J.E., Banay, R.F., Laden, F., 2016. Exposure to greenness and mortality in a nationwide prospective cohort study of women. *Environ. Health Perspect.* 124 (9), 1344–1352.
- Jarvis, I., Davis, Z., Sbihi, H., Brauer, M., Czekajlo, A., Davies, H.W., Gergel, S.E., Guhn, M., Jerrett, M., Koehoorn, M., Oberlander, T.F., 2021. Assessing the association between lifetime exposure to greenspace and early childhood development and the mediation effects of air pollution and noise in Canada: a population-based birth cohort study. *Lancet Planet. Health* 5 (10), e709–e717.
- Ji, J.S., Zhu, A., Lv, Y., Shi, X., 2020. Interaction between residential greenness and air pollution mortality: analysis of the Chinese Longitudinal Healthy Longevity Survey. *Lancet Planet. Health* 4 (3), e107–e115.
- Ji, J.S., Liu, L., Zhang, J., Kan, H., Zhao, B., Burkart, K.G., Zeng, Y., 2022. NO₂ and PM_{2.5} air pollution co-exposure and temperature effect modification on premature mortality in advanced age: a longitudinal cohort study in China. *Environ. Health* 21 (1), 97.
- Kasdagli, M.I., Katsouyanni, K., de Hoogh, K., Lagiou, P., Samoli, E., 2021. Associations of air pollution and greenness with mortality in Greece: an ecological study. *Environ. Res.* 196, 110348.
- de Keijzer, C., Agis, D., Ambrós, A., Arévalo, G., Baldasano, J.M., Bande, S., Barrera-Gómez, J., Benach, J., Cirach, M., Dadvand, P., Ghigo, S., 2017. The association of air pollution and greenness with mortality and life expectancy in Spain: a small-area study. *Environ. Int.* 99, 170–176.
- Khomenko, S., Cirach, M., Pereira-Barboza, E., Mueller, N., Barrera-Gómez, J., Rojas-Rueda, D., de Hoogh, K., Hoek, G., Nieuwenhuijsen, M., 2021. Premature mortality due to air pollution in European cities: a health impact assessment. *Lancet Planet. Health* 5 (3), e121–e134.
- Klompaker, J.O., Hart, J.E., James, P., Sabath, M.B., Wu, X., Zanobetti, A., Dominici, F., Laden, F., 2021. Air pollution and cardiovascular disease hospitalization—are associations modified by greenness, temperature and humidity? *Environ. Int.* 156, 106715.
- Labib, S.M., Lindley, S., Huck, J.J., 2020. Spatial dimensions of the influence of urban green-blue spaces on human health: a systematic review. *Environ. Res.* 180, 108869.
- Labib, S.M., Lindley, S., Huck, J.J., 2021. Estimating multiple greenspace exposure types and their associations with neighbourhood premature mortality: a socioecological study. *Sci. Total Environ.* 789, 147919.
- Labib, S.M., Lindley, S., Huck, J.J., 2023. Nonlinear associations between urban greenness exposures and neighborhood level years of potential life lost: a study in Greater Manchester. *Sci. Talks* 6, 100218.
- Leist, A.K., Klee, M., Kim, J.H., Rehkopf, D.H., Bordas, S.P., Muniz-Terrera, G., Wade, S., 2022. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Sci. Adv.* 8 (42) (eabk1942).
- Lindley, S.J., Walsh, T., 2005. Inter-comparison of interpolated background nitrogen dioxide concentrations across Greater Manchester, UK. *Atmos. Environ.* 39 (15), 2709–2724.
- Liu, X., Kounadi, O., Zurita-Milla, R., 2022. Incorporating spatial autocorrelation in machine learning models using spatial lag and eigenvector spatial filtering features. *ISPRS Int. J. Geo Inf.* 11 (4), 242.
- Marando, F., Heris, M.P., Zulian, G., Udías, A., Mentaschi, L., Chrysoulakis, N., Parastatidis, D., Maes, J., 2022. Urban heat island mitigation by green infrastructure in European Functional Urban Areas. *Sustain. Cities Soc.* 77, 103564.
- Markevych, I., Schoierer, J., Hartig, T., Chudnovsky, A., Hystad, P., Dzhambov, A.M., De Vries, S., Triguero-Mas, M., Brauer, M., Nieuwenhuijsen, M.J., Lupp, G., 2017. Exploring pathways linking greenspace to health: theoretical and methodological guidance. *Environ. Res.* 158, 301–317.
- Martinez, A., Labib, S.M., 2023. Demystifying normalized difference vegetation index (NDVI) for greenness exposure assessments and policy interventions in urban greening. *Environ. Res.* 220, 115155.

- Mutiibwa, D., Strachan, S., Albright, T., 2015. Land surface temperature and surface air temperature in complex terrain. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.* 8 (10), 4762–4774.
- Nieuwenhuijsen, M.J., 2020. Urban and transport planning pathways to carbon neutral, liveable and healthy cities; a review of the current evidence. *Environ. Int.* 140, 105661.
- Nieuwenhuijsen, M.J., Khreis, H., Triguero-Mas, M., Gascon, M., Dadvand, P., 2017. Fifty shades of green. *Epidemiology* 28 (1), 63–71.
- Nieuwenhuijsen, M.J., Agier, L., Basagaña, X., Urquiza, J., Tamayo-Uria, I., Giorgis-Allemand, L., Robinson, O., Siroux, V., Maitre, L., de Castro, M., Valentin, A., 2019. Influence of the urban exposome on birth weight. *Environ. Health Perspect.* 127 (4), 047007.
- Nieuwenhuijsen, M.J., Barrera-Gómez, J., Basagaña, X., Cirach, M., Daher, C., Pulido, M. F., Jungman, T., Gasparrini, A., Hoek, G., de Hoogh, K., Khomenko, S., 2022. Study protocol of the European Urban Burden of Disease Project: a health impact assessment study. *BMJ Open* 12 (1), e054270.
- Nohara, Y., Matsumoto, K., Soejima, H., Nakashima, N., 2022. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput. Methods Prog. Biomed.* 214, 106584.
- Ohanyan, H., Portengen, L., Kaplani, O., Huss, A., Hoek, G., Beulens, J.W., Lakerveld, J., Vermeulen, R., 2022a. Associations between the urban exposome and type 2 diabetes: results from penalized regression by least absolute shrinkage and selection operator and random forest models. *Environ. Int.* 170, 107592.
- Ohanyan, H., Portengen, L., Huss, A., Traini, E., Beulens, J.W., Hoek, G., Lakerveld, J., Vermeulen, R., 2022b. Machine learning approaches to characterize the obesogenic urban exposome. *Environ. Int.* 158, 107015.
- Petch, J., Di, S., Nelson, W., 2022. Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Can. J. Cardiol.* 38 (2), 204–213.
- Ren, X., Mi, Z., Georgopoulos, P.G., 2023. Socioexposomics of COVID-19 across New Jersey: a comparison of geostatistical and machine learning approaches. *J. Expo. Sci. Environ. Epidemiol.* 1–11.
- Rojas-Rueda, D., Nieuwenhuijsen, M.J., Gascon, M., Perez-Leon, D., Mudu, P., 2019. Green spaces and mortality: a systematic review and meta-analysis of cohort studies. *Lancet Planet. Health* 3 (11), e469–e477.
- Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring vegetation systems in the Great Plains with ERTS. *NASA Spec. Publ.* 351 (1), 309.
- Seligman, B., Tuljapourkar, S., Rehkopf, D., 2018. Machine learning approaches to the social determinants of health in the health and retirement study. *SSM-Pop. Health* 4, 95–99.
- Shuvo, F.K., Mazumdar, S., Labib, S.M., 2021. Walkability and greenness do not walk together: investigating associations between greenness and walkability in a large metropolitan city context. *Int. J. Environ. Res. Public Health* 18 (9), 4429.
- Smith, C.L., Webb, A., Levermore, G.J., Lindley, S.J., Beswick, Y.K., 2011. Fine-scale spatial temperature patterns across a UK conurbation. *Clim. Chang.* 109 (3–4), 269–286.
- Twohig-Bennett, C., Jones, A., 2018. The health benefits of the great outdoors: a systematic review and meta-analysis of greenspace exposure and health outcomes. *Environ. Res.* 166, 628–637.
- United Nations, 2019. Population Division. In: *World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420)*. United Nations, New York.
- Verbeek, T., 2019. Unequal residential exposure to air pollution and noise: a geospatial environmental justice analysis for Ghent, Belgium. *SSM-Pop. Health* 7, 100340.
- Wang, F., Wang, J., Gelfand, A., Li, F., 2017. Accommodating the ecological fallacy in disease mapping in the absence of individual exposures. *Stat. Med.* 36 (30), 4930–4942.
- Wang, R., Grekousis, G., Maguire, A., McKinley, J.M., Garcia, L., Rodgers, S.E., Hunter, R.F., 2023. Examining the spatially varying and interactive effects of green and blue space on health outcomes in Northern Ireland using multiscale geographically weighted regression modeling. *Environ. Res. Commun.* 5 (3), 035007.
- Wei, H., Sun, J., Shan, W., Xiao, W., Wang, B., Ma, X., Hu, W., Wang, X., Xia, Y., 2022. Environmental chemical exposure dynamics and machine learning-based prediction of diabetes mellitus. *Sci. Total Environ.* 806, 150674.
- Wiemken, T.L., Kelley, R.R., 2020. Machine learning in epidemiology and health outcomes research. *Annu. Rev. Public Health* 41 (1), 21–36.
- Xu, S., Marcon, A., Bertelsen, R.J., Benediktssdottir, B., Brandt, J., Engemann, K., Frohn, L.M., Geels, C., Gislason, T., Heinrich, J., Holm, M., 2023. Long-term exposure to low-level air pollution and greenness and mortality in Northern Europe. The Life-GAP project. *Environ. Int.* 181, 108257.
- Yang, C., Li, R., Sha, Z., 2020. Exploring the dynamics of urban greenness space and their driving factors using geographically weighted regression: a case study in Wuhan Metropolis, China. *Land* 9 (12), 500.
- Yitshak-Sade, M., Kloog, I., Novack, V., 2017. Do air pollution and neighborhood greenness exposures improve the predicted cardiovascular risk? *Environ. Int.* 107, 147–153.
- Zhang, H., Liu, L., Zeng, Y., Liu, M., Bi, J., Ji, J.S., 2021. Effect of heatwaves and greenness on mortality among Chinese older adults. *Environ. Pollut.* 290, 118009.
- Zhao, Y., Wood, E.P., Mirin, N., Cook, S.H., Chunara, R., 2021. Social determinants in machine learning cardiovascular disease prediction models: a systematic review. *Am. J. Prev. Med.* 61 (4), 596–605.
- Zhu, J., Zhang, X., Zhang, X., Dong, M., Wu, J., Dong, Y., Chen, R., Ding, X., Huang, C., Zhang, Q., Zhou, W., 2017. The burden of ambient air pollution on years of life lost in Wuxi, China, 2012–2015: a time-series study using a distributed lag nonlinear model. *Environ. Pollut.* 224, 689–697.