An explainable AI framework for spatiotemporal risk factor analysis in public health: a case study of cardiovascular mortality in South Korea

Eunjin Kang, Dongjin Cho, Siwoo Lee, Jungho Im, Dongwook Lee & Cheolhee Yoo

Taylor & Francis
Taylor & Francis Group

An explainable AI framework for spatiotemporal risk factor analysis in public health: a case study of cardiovascular mortality in South Korea

Eunjin Kang [a], Dongjin Cho [a,b], Siwoo Lee[a], Jungho Im [a,c,d], Dongwook Lee[e] and Cheolhee Yoo [f]

aDepartment of Civil, Urban, Earth, and Environmental Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea; bEnvironmental Planning Institute, Seoul National University, Seoul, South Korea; cGraduate School of Artificial Intelligence, UNIST, Ulsan, South Korea; dGraduate School of Carbon Neutrality, UNIST, Ulsan, South Korea; eDepartment of Occupational and Environmental Medicine, Inha University Hospital, Inha University, Incheon, South Korea; fDepartment of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

ABSTRACT

Understanding environmental disease risk factor analysis at the district level is essential for gaining valuable insights into regional disease variations, offering a broader perspective compared to individual-level studies. Recently, explainable artificial intelligence (XAI) has received increasing attention in the analysis of factors affecting public health. However, previous purely data-driven XAI-based risk factor analyses faced challenges in capturing regional effect of environmental variables, leading to confusion regarding key spatiotemporal risk factors. Therefore, this study proposes a framework that includes two complementary XAI-based risk factor analyses following two assumptions. Regionally rescaled environmental variables must account for the unequal effects on environmental factors, which are likely influenced by variations in adaptation capacity to weather conditions and differences in exposure-response relationships to air pollutants. District-level disease distribution highlights geographic disparity in sociodemographic vulnerability, whereas temporal variation in diseases by district underscores temporal environmental impacts. Based on these two hypotheses, we rescaled environmental variables using two complementary schemes: one that employs the district-level disease distribution as the target variable, and another that utilizes the temporal residual of the disease within each district. We evaluated this framework by analyzing the association between cardiovascular age-standardized mortality rate (CVD-ASMR) and various risk factors in South Korea from 2010 to 2019, using high-performing random forest and light gradient boosting models with additive Shapley explanation. Compared to previous purely data-driven XAI-based analyses, the proposed schemes achieved significantly better results in capturing regional exposure-response relationships. In two complementary schemes, the most explainable factor to districts with high CVD-ASMR was low education level related to sociodemographic vulnerability, whereas the most explainable factors to high temporal CVD-ASMR patterns were low greenness and high air pollution levels. In addition, the two complementary schemes enabled us to reasonably analyze the interaction effect of the two risk factors, i.e. temperature and air pollutants. Furthermore, high CVD-ASMR and its high temporal variation were observed in situations of high sociodemographic vulnerability with poor air quality. These findings provide insightful public health planning for sustainable cities and society by pinpointing high-risk areas and tailoring strategies to address regional environmental challenges.

1. Introduction

The acceleration of urbanization has led to uneven population growth and environmental degradation (Han et al. 2014; Sun et al. 2022). Heterogeneous levels of urban development led to local disparities, including environmental factors such as deteriorating air quality and increased susceptibility to heat, as well as sociodemographic factors such as education levels, access to health care, and income rates, thus contributing to regional health disparities (Kamal-Chaoui and Sanchez-Reaza 2012; Rosenberg et al. 2016; Seto, Parnell, and Elmqvist 2013). Numerous studies indicate that district-level disease studies can provide invaluable insights into the occurrence of mortality/morbidity and spatial variations, and capture the differential impact of urbanization, as individual-level studies are not always feasible (Conrad and Capewell 2012; Hoffmann et al. 2014;

Morgenstern 1995). Therefore, conducting risk factor analysis at the district level is crucial for a comprehensive understanding of sustainable public health.

The ever-growing threat of chronic and infectious diseases requires continuous advancements of public health strategies. By surpassing conventional statistical models, artificial intelligence (AI) has emerged as a powerful tool because it can learn complex associations between various factors and diseases from huge and complex datasets (Cho et al. 2021; Hossain, Uddin, and Khan 2021; Ross et al. 2016). AI techniques, such as random forest (RF) and light gradient boosting model (LGBM), aid in managing public health policies and fostering sustainable societies by generating geographical disease maps (Alvarez-Mendoza et al. 2020; Razavi-Termeh, Sadeghi-Niaraki, and Choi 2022), detecting disease signs (Manne 2021; Tayal et al. 2021), and predicting mortality and morbidity (Boudreault et al. 2024; Parikh et al. 2019). Despite the high performance of AI algorithms, their "black-box" nature poses difficulties (i.e. interpretability) in healthcare applications. To extend the application of AI to public health, explainable AI (XAI) algorithms have been used: Shapley additive explanations (SHAP), local interpretable model-agnostic explanations (LIME), anchors, and heatmaps (Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2016, 2018; Selvaraju et al. 2017). Unlike other algorithms, SHAP provides both global and local interpretability, enabling the analysis of positive or negative contributions of risk factors to individual predictions. SHAP has been successfully applied in various epidemiological settings to facilitate informed decision-making regarding treatment options, and improve risk detection (Alabi et al. 2022; Peng et al. 2021; Wang, Feng, and Qi 2021).

Despite the multispectral interpretability of SHAP, its purely data-driven application in risk factor studies has overlooked differences in regional exposure-response effects (Ke et al. 2023; Kim and Kim 2022; Ogata et al. 2021; Shimada-Sammori et al. 2023; Wen, Wang, and Shao 2024; Wertis et al. 2023; Yao et al. 2022). This oversight may lead to confusion regarding key spatiotemporal risk factors. For example, studies in a developed country (e.g. Japan) identified temperature as a key predictor of disease (Ke et al. 2023; Ogata et al. 2021; Shimada-Sammori et al. 2023), while studies in less urbanized cities placed emphasis on sociodemographic factors (Y. Kim and Kim 2022; Wen,

Wang, and Shao 2024). Notably, Wertis et al. (2023) highlighted that the scale of the analysis (multiple vs. single districts) impacts key risk factors. Because SHAP is data-driven and cannot automatically adjust for unequal regional effects on environmental factors, interpreting influential factors can be challenging. To better understand regional disease patterns and establish effective policies, it is crucial to comprehend these inconsistencies and develop solutions in SHAP-based analyses. To the best of our knowledge, there is no SHAP-based research that has proposed guidelines for understanding the influencing factors in different local conditions.

Cardiovascular diseases (CVD) are a leading cause of mortality and morbidity globally, attributable to 18.6 million deaths and 523 million cases in 2019 (Roth et al. 2020). In particular, South Korea has a high CVD mortality rate despite active management strategies (Arafa et al. 2021; Kim et al. 2015; Lee et al. 2021). Understanding the spatiotemporal association between risk factors and CVD could facilitate the prevention of CVD-related diseases and provide remedies for the increasing incidence of CVD. While several CVD-related risk factor analyses are based on model-specific variable importance, these metrics can make it difficult to understand positive/negative trait contributions to the target disease and individual sample contributions of input traits to the target disease (Dong et al. 2023; Hu, Liu, and Li 2020; Ji et al. 2020). Few studies identified risk factors for spatiotemporal CVD patterns using the XAI, but none addressed important categories of factors such as sociodemographic factors, air pollutants, and meteorological variables (Harford et al. 2022; Nakashima et al. 2021; Shimada-Sammori et al. 2023). Therefore, we applied the proposed XAI-based risk factor framework to understand the spatiotemporal association between CVD and various risk factors in South Korea from 2010 to 2019.

Our aim was to propose an XAI-based framework for analyzing various risk factors (i.e. environmental and sociodemographic factors) associated with the disease at the district level. This framework includes the concept of the regional effects of environmental factors. The key objectives of this study were to (1) design two XAI-based complementary schemes that reflect different local exposure-response relationships and adaptation effects to environmental factors; (2) systematically evaluate these two schemes in

a casestudy focusing on spatiotemporal CVD mortality patterns in South Korea; and (3) understand the contribution and interaction of these two complementary schemes based on spatiotemporal properties of the target variables.

2. Datasets

This framework was used to conduct a case study for risk factor analysis targeting age-standardized CVD mortality rate (CVD-ASMR) during 2010 2019 across 250 district-level divisions (si/gun/gu) in South Korea. South Korea is located between 124 and 131 °E longitude and 33 and 39 °N latitude (Figure 1), with approximately 51 million people according to the 2020 census. South Korea comprises 17 provincial-level divisions subdivided into multiple smaller entities, such as 250 si/gun/gu (as of 2019). We aligned the disease data and environmental and sociodemographic variables to 250 district-level divisions to conduct risk factor analysis. All features used in the case study and their brief descriptions are given in Table 1.

2.1. Standardized mortality rate

We used the 2010–2019 annual CVD-ASMR calculated using data from the National Health Insurance Service

(NHIS) in South Korea. The NHIS manages all residents health insurance medical information, including hospital visits (e.g. date of admission, date of discharge, and diagnosis codes) and personal information (e.g. gender, birth year, and residential address) (Song 2009). The vital status of participants was determined by Statistics Korea through NHIS, and death due to CVD was reported as I00–I99 of the International Classification of Diseases as the cause of death on the death certificate. To calculate CVD-ASMR by district and standardize it by gender (male and female) and age (0–15 years, 15–65 years, and >65 years), number of deaths due to CVD and population structure data for 250 districts were utilized. A population structure in South Korea from mid-2005 was used as the standard population structure for age-standardization.

2.2. Environmental factors

Air pollutants, meteorological conditions, and the normalized difference vegetation index (NDVI) were used to determine environmental factors. The annual average concentrations of air pollutants (PM_{10}, SO_2, NO_2, CO, and O_3) were obtained from the AIRKOREA website (https://www.airkorea.or.kr/web) under the control of the National Institute of Environmental
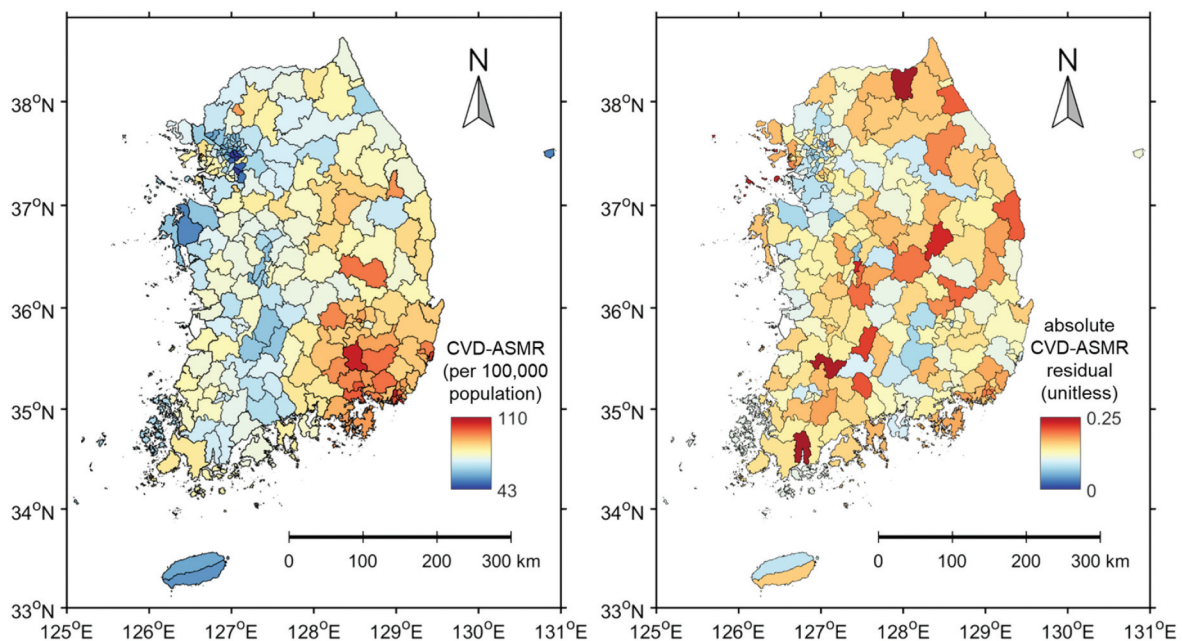
Figure 1. Study area maps depicting CVD-ASMR and the temporal CVD-ASMR residual. These values were calculated at the district level and averaged over the period 2010−2019; specifically, the district map of temporal CVD-ASMR residuals displays the average absolute CVD-ASMR residuals across 250 districts (si/gun/gu).

Table 1. Summary of target and input variables, used in a case study of spatiotemporal CVD-ASMR risk factor analysis from 2010–2019 in South Korea.

Name	Abbreviations	Units	Data sources
A target variable			
Cardiovascular disease Standardized mortality rate	CVD-ASMR	%	National Health Insurance Service
Input variables			
Annual mean PM_{10}	PM_{10}	$\mu g/m^3$	Ministry of Environment
Annual mean O_3	O_3	ppm	
Annual mean NO_2	NO_2	ppm	
Annual mean CO	CO	ppm	
Annual mean SO_2	SO_2	ppm	
Annual mean temperature	Temp	°C	Korea
Annual amount of precipitation	precipitation	mm	Meteorological Administration
Number of heatwave days	heatwave	day	
Number of cold wave days	coldwave	day	
Annual mean NDVI	NDVI	-	NASA, MODIS
Distance from urban areas	distance_urban	Decimal degree	
Sick bed	sickbed	Beds per 1,000 people	Statistics Korea
Doctor	doctor	Doctors per 1,000 people	
Smoking rate	smoke	%	National health insurance service
Percentage of population aged 65 and above	age65	%	Ministry of the Interior and Safety
Rate of elderly living alone	solitary	%	Statistics Korea
High school graduate or less rate	under_hs	%	
Number of disabled persons	disable	Number of persons with disabilities per 100,000 people	

Research. In South Korea, ground-based air quality stations are classified into urban, roadside, rural, and background sites to effectively monitor air quality nationwide. We used 343 urban air quality stations to account for local air quality and avoid unexpected emissions. Daily meteorological data, including average temperature (°C) and precipitation (mm), were derived from 516 automatic weather stations and 93 automated synoptic observation systems of the Korea Meteorological Administration (KMA).

As air pollution and meteorological data are primarily derived from ground stations, we employed spatial interpolation methods, such as ordinary kriging and inverse distance weighting (IDW), to estimate environmental hazards at unsampled locations (J. Li and Heap 2014). Ordinary kriging was chosen for air pollutants and temperature due to their spatial autocorrelation, which is influenced by their distribution patterns (S. Y. Kim et al. 2014; Rivera-González et al. 2015; Shtiliyanova et al. 2017; Yang, Wang, and August 2004). For precipitation, IDW was applied to capture the characteristics of heavy rainfall in specific regions (F. W. Chen and Liu 2012; Keblouti, Ouerdachi, and Boutaghane 2012). Notably, the accuracy of spatial interpolation is sensitive to the density of ground stations within the study area. To ensure reliability, we validated these methods for representing district-

level environmental factors using five-fold cross-validation (Table S1).

To account for extreme temperature events, we included annual heat wave and cold wave days, calculated using spatially interpolated daily average maximum and minimum temperatures (J. Chen et al. 2020; Son, Bell, and Lee 2014). According to the KMA, annual heatwave days were calculated when the daily maximum temperature was above 33°C. The annual cold wave day was calculated when the daily minimum temperature was less than − 12°C. To consider the influence of greenery on disease, we used NDVI, obtained from the Terra Satellite-Moderate Resolution Imaging Spectroradiometer (MODIS) MOD13A1 product at a spatial resolution of 500 m (https://search.earthdata.nasa.gov/). The 16-day composite MODIS NDVI was averaged annually. Finally, we obtained the district-level values for the average value of environmental variables in 250 administrative units using the "Zonal Statistics" function in ArcGIS.

2.3. Sociodemographic factors

Sociodemographic factors were used to consider regional vulnerability: percentage with a high school diploma or less, number of disabilities per 100,000 people, percentage of elderly living alone, number

of hospital beds and doctors per 1,000 population, smoking rate, distance from the nearest urban area, and percentage of people aged 65 or older. The percentage with a high school diploma or less, the number of disabilities per 100,000 population, the percentage of elderly living alone, and the number of hospital beds and doctors per 1,000 population were obtained from the open-source Korean Statistical Information System (https://kosis.kr.) of the National Statistical Office in the period 2010–2019. Since the education level data is generated every five years, the following years' data was replaced: Data from 2010 was used for 2010–2014, and data from 2015 was used for 2015–2019. The smoking rate was obtained from the open-source NHIS (http://nhiss.nhis.or.kr:8087/intro/index.do) to consider the regional medical characteristics. The distance from the nearest urban area was calculated using the "Euclidean distance" function in ArcGIS with urban and built-up information from the MODIS Land Cover Type (MCD12Q1) (https://search.earthdata.nasa.gov/) on board Terra and Aqua satellites. When each sociodemographic factor was missing, this was imputed by applying population proportions to the data from the year closest to the missing value.

3. Methodology

The proposed framework aims to conduct XAI-based risk factor analysis to analyze spatiotemporal disease patterns at the district level based on different variables (e.g. sociodemographic and environmental). We assumed that regional risk levels from environmental factors vary, likely due to differences in adaptation capacity to weather conditions (Kim et al. 2019; Wu et al. 2024) and local exposure-response relationships to air pollutants (Li, Jin, and Kan 2019; Papadogeorgou and Dominici 2020). To explore the different spatiotemporal relationship between various factors and diseases, we proposed two complementary schemes targeting spatial and temporal disease variations, respectively.

  (a) Proposed scheme (P1): Examine the association between district-level disease distribution and diverse risk factors to highlight those related to spatial disease variation.
  (b) Proposed scheme (P2): Investigate the association between temporal disease variations by district and various factors to highlight the impact of risk factors on temporal disease variations.

Based on two complementary concepts, we created a district-level spatiotemporal dataset that includes various input variables and a target disease. To assess the effectiveness of our proposed complementary schemes, we compared their model performance and SHAP analysis results with a previously established purely data-driven XAI strategy (comparison scheme, C). This comparison scheme used raw input data (risk factors and disease targets) without any rescaling or preprocessing. Subsequently, we selected the optimal AI models for application in the XAI analysis. Finally, we performed a SHAP-based risk factor analysis to evaluate the contribution of different traits to the predicted CVD-ASMR, with a focus on the spatial and temporal variability of the disease. Figure 2 illustrates the proposed framework for identifying associations between various risk factors and diseases at the district level.

3.1. Assumption of the proposed framework

The proposed framework considers the regional effect of environmental factors on health outcomes for districts consisting of neighborhood factors (e.g. age, gender, parental education, emissions, and unhealthy behaviors). Previous meta-analyses and time-analyses found that the short- and long-term effects of environmental factors (e.g. air pollutants and temperature) on hospitalizations and/or mortality vary depending on regional characteristics within the same country and between different countries worldwide (Bell and Dominici 2008; Chen et al. 2017; Lee et al. 2022; Lee et al. 2016; Tan et al. 2021). For example, residents of areas exposed to a one-unit increase in particulate matter (PM) concentrations (e.g. Europe and North America) would be at higher risk of heart disease and acute respiratory illnesses than residents of areas exposed to constant higher PM concentrations (e.g. China) (Li, Jin, and Kan 2019; Tan et al. 2021). Residents of northern cities in South Korea are more susceptible to a similar rise in temperature increase than residents of southern cities, where the average temperature is relatively higher (Kim et al., 2019).
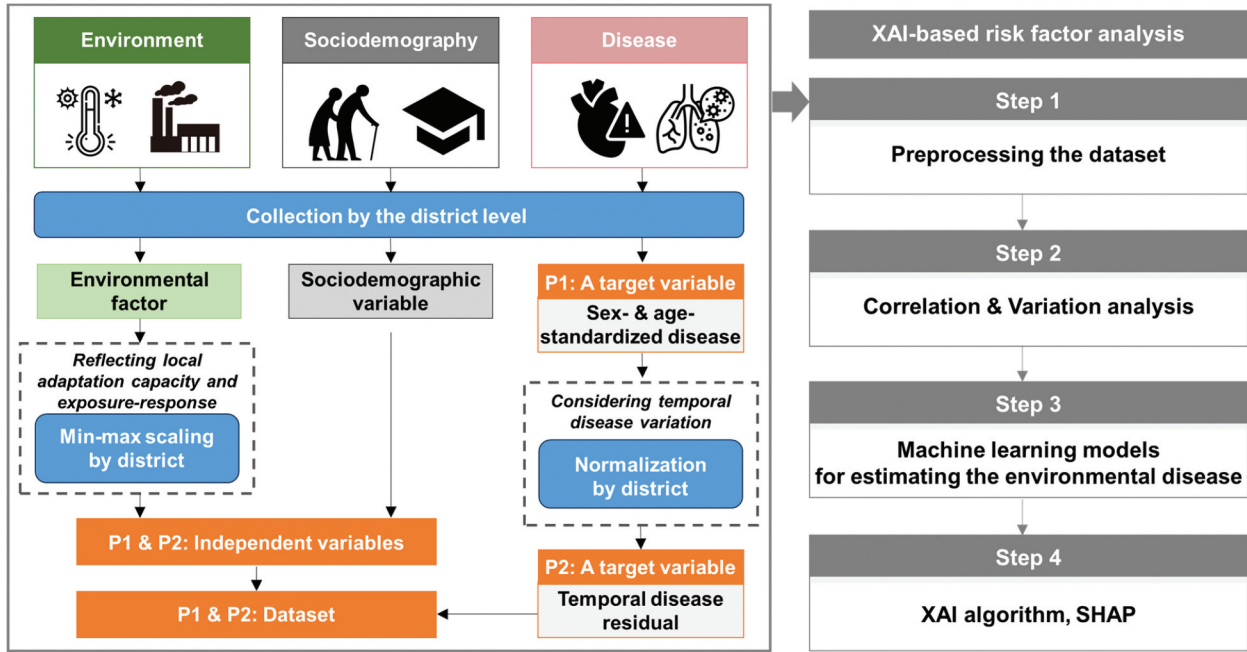
Figure 2. Workflow for the proposed xai-based spatiotemporal risk factor analysis in public health.

Previous XAI-based modeling studies did not consider the regional effect of district-level environmental variables (Ke et al. 2023; Y. Kim and Kim 2022; Ogata et al. 2021; Shimada-Sammori et al. 2023; Wen, Wang, and Shao 2024; Wertis et al. 2023; Yao et al. 2022). Moreover, these studies focused primarily on overall mortality/morbidity rates at the district level, making it difficult to capture temporal variation in the disease amidst significant spatial variation across multiple districts, especially on a nationwide scale. Therefore, this study hypothesized that considering regional exposure-response relationships to environmental factors helps to find influencing factors in different contextual backgrounds in XAI-based risk factor analysis. In addition, regional disease residuals can help identify influencing factors in time series within XAI-based spatiotemporal modeling.

3.2. Preprocessing the dataset

Environmental and sociodemographic factors and disease data were collected in each administrative region. When accounting for local exposure-response relationships to environmental factors, environmental variables that arise at the district level (e.g. meteorological variables and air pollutants) were rescaled to remove their spatial characteristics and highlight their temporal correlation with a target disease. Finally, min-max scaling was applied to obtain regional environmental values within the common range (0–1) for all administrative units (Equation (1)).

$$z_{i,j} = \frac{x_{i,j} - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{1}$$

where, $z_{i,j}$ is the rescaled value of the environmental variable at time j in district i, $x_{i,j}$ is the environmental value at time j in district i, and $\min(x_i)$ and $\max(x_i)$ correspond to the minimum and maximum values, respectively, of the environmental variables in district i within the study period.

Age- and gender-standardized disease data are recommended as target variables for understanding overall disease patterns. To understand temporal disease fluctuations in public health, we calculated the temporal disease residuals per district, i.e. the extent of the difference between the temporal disease residual and the average disease progression in the district's time series. Residual disease is standardized by district because it can be relative across regions (Eq. (2)).

$$s_{i,j} = \frac{y_{i,j} - \text{mean}(y_i)}{\text{mean}(y_i)} \tag{2}$$

where, $s_{i,j}$ is the temporal residual disease by district at time j in district i, $y_{i,j}$ is the disease value at time j in district i, and $\text{mean}(y_i)$ corresponds to the average disease value in district i during the study period.

Finally, we constructed a dataset comprising 2,500 samples from 250 districts over 10 years. Specifically, P1 included rescaled environmental variables and raw sociodemographic variables as independent variables, with CVD-ASMR as the target. P2 had the same independent variables as P1, but used the temporal residual of CVD-ASMR by district as the target. In contrast, C utilized raw environmental and sociodemographic variables, with CVD-ASMR as the target, without any rescaling. Table S2 summarizes the rescaling status of input and target variables across the proposed and comparison schemes.

3.3. Machine learning models for XAI-based spatiotemporal risk factor analysis

This framework recommends applying various machine learning models to determine the optimal model for XAI-based spatiotemporal risk factor analysis. We include three different algorithms: ElasticNet, RF, and LGBM, which are widely used in various fields of data mining and statistics. ElasticNet is a linear regression method that combines the advantages of both Lasso and Ridge techniques to improve the generalization of traditional linear regression models (Zou and Hastie 2005). Lasso penalizes the sum of the absolute values of the coefficients (L1 norm), while Ridge penalizes the sum of the squared coefficients (L2 norm). These penalties help to mitigate overfitting by addressing feature multicollinearity and refining the selection of regression coefficients. ElasticNet incorporates both L1 and L2 penalties, and its two key parameters – the penalty amount and the ratio of L1 to L2 penalties – allow for flexible adjustment. In this study, ElasticNet was implemented using the "ElasticNet" function from Python's scikit-learn module, with parameters optimized for best performance.

RF is an ensemble model widely used to address the limitations of conventional decision trees. RF builds multiple independent trees using randomly sampled subsets of training data and input features at each tree's node (Breiman 2001). This approach reduces output variance and bias by averaging predictions across all trees. RF offers robust performance and is less sensitive to parameter changes compared to other methods (Lee, Kim, and Park 2023; Odebiri et al. 2020). The model was implemented using the "RandomForestRegressor" function from Python's

scikit-learn module, with optimization focusing on two key hyperparameters: the number of trees (n_estimators) and the number of features considered at each split (max_features).

The LGBM is an advanced version of the Gradient Boosted Decision Tree (GBDT), a sequential ensemble technique introduced by Ke et al. (2017). GBDT iteratively prioritizes data samples with the highest loss when building subsequent trees. LGBM improves upon this by using a leaf-wise growth strategy, which can significantly reduce memory usage and computation time compared to the level-wise approach of traditional GBDT models. This makes LGBM particularly effective for handling large datasets (Cho et al. 2022; Liu et al. 2023). In this study, LGBM was implemented using the "lgb" function from Python's lightgbm module. Given the model's sensitivity to parameter tuning, multiple parameters were optimized, including the number of trees (n_estimators), the proportion of features used in each tree (colsample_bytree), maximum tree depth (max_depth), maximum number of leaves (num_leaves), subsampling rate (subsample), and the learning rate (learning_rate) is increased.

Bayesian optimization was applied to determine the optimal hyperparameters for ElasticNet, RF, and LGBM. This method is widely used to optimize data-driven models, as it can identify optimal parameters more efficiently than existing optimization tools, such as random and grid searches (César de Sá et al. 2022; Kang et al. 2023; Ma et al. 2022). The technique employs a Gaussian process to update its posterior distribution when selecting the next set of hyperparameters, thereby expediting the model optimization process. The hyperparameter settings for the Bayesian-optimized ElasticNet, RF, and LGBM are summarized in Table S3. To determine the optimal hyperparameters for XAI-based analysis, five-fold cross-validation was performed on the training dataset.

3.4. XAI algorithm, SHAP

The contribution of the input variables was estimated using SHAP based on game theory (Lundberg and Lee 2017). The SHAP method provides the extent to which each variable contributes positively and negatively to individual predictions compared to tree-based feature importance (Jang et al. 2021). The SHAP value was calculated as the average response of the

remaining variables when a particular factor was excluded (Equation. (3)).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right] \quad (3)$$

where, ϕ_i represents the SHAP value for the input factor i, f denotes the prediction model, F indicates the set of all input variables, S represents one of all possible subsets except i, x_S denotes the value of the variable within the subset S, and $\frac{|S|!(|F|-|S|-1)!}{|F|!}$ is the normalized weight sum that takes into account the number of subsets without i.

Compared to other XAI techniques (e.g. LIME), the SHAP method provides a global view by analyzing all predictions to understand the overall feature contribution and the interaction between two features. This comprehensive understanding is achieved through diverse SHAP plots in various fields including epidemiology (Gao et al. 2022; Kang et al. 2023; Y. Kim and Kim 2022; Zhu, Geiß, and So 2024). A SHAP summary plot provides the overall significance of the feature importance and highlights how each feature negatively/positively contributes to the prediction. A SHAP dependency plot shows how interactions between two factors increase/decrease the predicted result. We performed SHAP analysis using the Python implementation of SHAP called "TreeexExplainer."

3.5. Model evaluation

To evaluate the optimal models for XAI-based analysis, five-fold cross-validation was conducted. This method randomly divided the entire dataset into five subsets, using four subsets for training and one for validation in each iteration. The model's accuracy was assessed using the coefficient of determination (R^2; Eq. (4)), root mean square error (RMSE; Eq. (5)), mean absolute error (MAE; Eq. (6)), and index of agreement (IOA; Eq. (7)). The IOA serves as a standardized indicator of prediction errors, where higher values denote greater model accuracy (Willmott 1981).

$$R^2 = \left(\frac{\sum_{i=1}^{n} (O-\bar{O})(P-\bar{P})}{\sqrt{\sum_{i=1}^{n} (O-\bar{O})^2 \sum_{i=1}^{n} (P-\bar{P})^2}} \right)^2 \quad (4)$$

$$RMSE = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (P-O)^2} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |P-O| \quad (6)$$

$$IOA = 1 - \frac{\sum_{i=1}^{n} (P-O)^2}{\sum_{i=1}^{n} \left(|O-\bar{O}| + |P-\bar{O}| \right)^2} \quad (7)$$

where, n is the total sample size; O and P are the referred and predicted disease mortality/morbidity, respectively; and $\bar{O}$ and $\bar{P}$ are the average values of referred and predicted disease mortality/morbidity, respectively.

4. Results

4.1. Variation and correlation analysis

Before building the framework, we examined simple statistical analyses regarding CVD-ASMR and influencing factors. The CVD-ASMR showed a higher spatial variance, whereas the CVD-ASMR residual by district showed a greater temporal variance (Figure 3). This finding indicates that, in the spatiotemporal models, P1 and C were designed to address risk factors related to the spatial variance of the target, whereas P2 was tailored to capture the risk factors related to the temporal variance of the target. In the correlation analysis, sociodemographic variables were highly correlated with CVD-ASMR, whereas environmental variables were strongly correlated with the CVD-ASMR residuals by district (Figure 4). These results indicate that district-level CVD-ASMR distribution was more closely linked to sociodemographic variables with geographic disparity, although CVD-ASMR residuals by district were associated with environmental variables with higher temporal fluctuations.

In the proposed models, most input variables showed higher spatiotemporal correlations than those in C (Figure 4). While temporal correlations were consistent for all schemes, the spatial correlations of environmental variables were lower in the proposed schemes due to rescaling of environmental variables by each administrative district (Fig. S1). Furthermore, the spatial association between CVD-ASMR residuals by district and sociodemographic variables showed lower correlations as CVD-ASMR was rescaled to emphasize regional temporal changes. Nevertheless, the highest
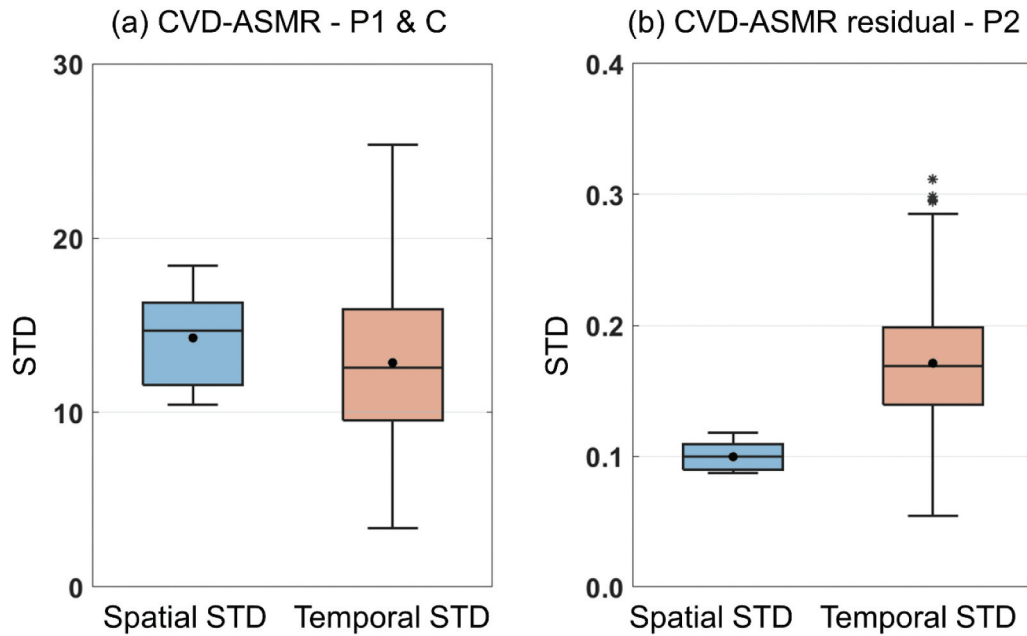
Figure 3. Boxplot of spatial and temporal standard deviations (STD) for CVD-ASMR and CVD-ASMR residual, by district, during the study period.
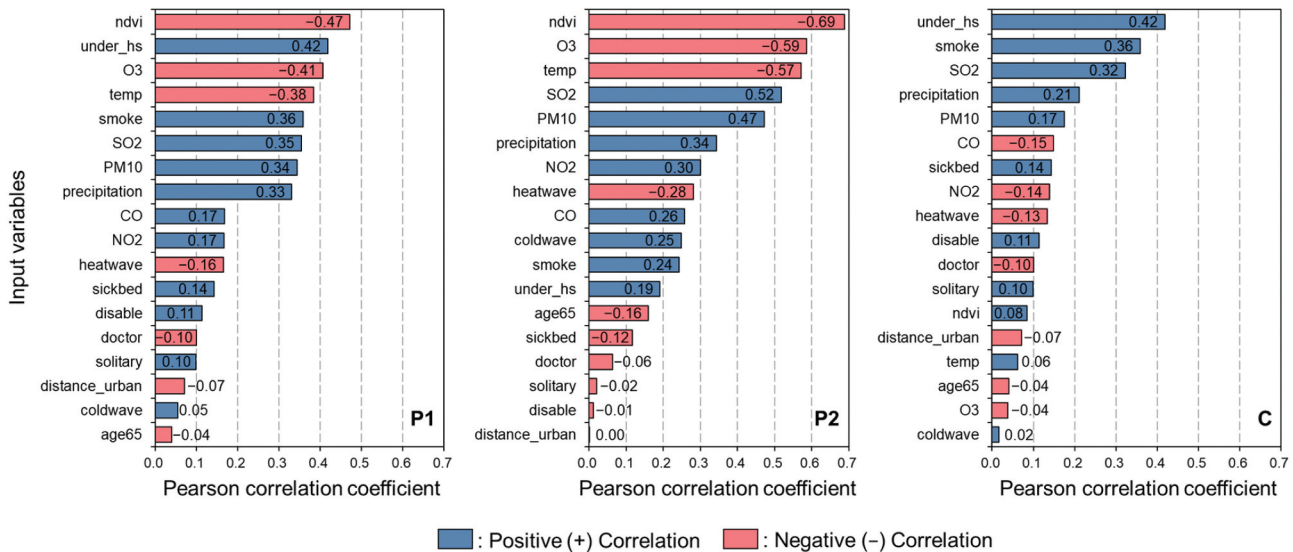
Figure 4. Bar graph illustrating the spatiotemporal correlation coefficients between input and target variables during the study period, as computed using the Pearson's test. Blue indicates positive correlation coefficients, while red indicates negative correlation coefficients.

spatiotemporal *R* values, particularly at P2, indicated a stronger association between environmental variables and CVD-ASMR temporal variations than those with sociodemographic variables.

4.2. Model performance evaluation

The non-linear RF and LGBM models exhibited significantly superior performance compared to ElasticNet,

a linear model (Table 2 and Figure 5). ElasticNet exhibited similar training and validation performances for all schemes, suggesting that it prevents overfitting compared to RF and LGBM. However, it showed the lowest R^2 (0.49–0.78) and IOA (0.80–0.93) and the highest MAE (6.5–9.8) and RMSE (8.4–12.7). In contrast, RF and LGBM had significantly higher accuracies than ElasticNet, with R^2 and IOA of approximately 1 in training and R^2 of 0.66–0.80 and IOA of 0.87–0.94 in

Table 2. Mean (standard deviation) of the four evaluation metrics (i.e. R^2, MAE, RMSE, and IOA) derived from five times training performance in random five-fold cross-validation, based on ElasticNet, RF, and LGBM in three proposed schemes.

	ElasticNet			RF			LGBM		
Metric	P1	P2	C	P1	P2	C	P1	P2	C
R^2	0.56	0.78	0.50	0.97	0.98	0.97	1.00	0.99	0.99
	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.02)
MAE	9.1	6.5	9.8	3.0	2.2	2.8	0.2	0.6	1.0
	(0.09)	(0.06)	(0.14)	(0.04)	(0.01)	(0.01)	(0.25)	(0.94)	(1.23)
RMSE	11.7	8.4	12.6	3.9	2.9	3.7	0.4	0.9	1.3
	(0.10)	(0.11)	(0.14)	(0.04)	(0.03)	(0.02)	(0.39)	(1.27)	(1.59)
IOA	0.84	0.93	0.80	0.99	0.99	0.99	1.00	1.00	1.00
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)

Figure 5. Scatterplots of random five-fold cross-validation results of ElasticNet, RF, and LGBM in three schemes. Color gradient from blue to red indicates increasing sample density. Red solid and gray dashed lines represent the regression and identity lines (y=x), respectively.

validation. When analyzing the spatial and temporal variability estimation performance of CVD-ASMR (Fig. S2), RF and LGBM outperformed ElasticNet, and better captured both spatial and temporal variability with higher IOA and lower RMSE.

Among the three schemes, P1 and C exhibited similar performances; however, P2 outperformed the other two (Table 2 and Figure 5). P2 was superior in estimating the CVD-ASMR time series, showing superiority in capturing the regional maximum and minimum values (Table 3). ElasticNet exhibited improved accuracy in P2 because of the stronger linear relationships between the input variables and the target compared to those in the other schemes. Nevertheless, RF and LGBM demonstrated higher accuracies in the proposed schemes with R^2 of 0.66–0.80 and IOA of 0.87–0.94. Therefore, this study conducted a risk factor analysis using SHAP values of RF and LGBM with reliable performance for all three schemes.

4.3. SHAP application for risk factor analysis

The main input variables for each scheme were similar in terms of spatiotemporal correlation coefficients (Figures 4 and 6). For example, sociodemographic factors (e.g. under_hs and smoking) were the most important factors in estimating the CVD-ASMR. Environmental factors (e.g. NDVI) were the most important factors in P2 (Figure 6). Additionally, their contribution signs were comparable to the spatiotemporal signs of the R values. For example, under_hs and smoking positively contributed to CVD-ASMR, whereas NDVI contributed negatively to residual CVD-ASMR by district. These results indicated that higher CVD-ASMR was related to higher sociodemographic vulnerability, and a higher CVD-ASMR residual score by district was linked to greater environmental

hazards. The highest-ranked input variables were comparable across schemes in RF and LGBM. The RF and LGBM showed similar positive and negative contribution signs for each scheme.

Several variables show opposite contribution signs between C and the proposed scheme (Figure 6). The NDVI and temperature positively contributed to the prediction in C, whereas they negatively contributed to the prediction in the proposed schemes. CO, NO_2, and extreme meteorological conditions negatively contributed to the prediction of C, although the proposed schemes mitigated their negative contribution patterns and the distinct positive contribution of CO. These variables were rescaled by administrative districts to reflect temporal correlations more closely than spatial ones (i.e. local exposure-response relationships to environmental factors).

NDVI and temperature showed distinct opposite interactions with air pollutants between the proposed schemes and C (Figures 7, 8, S3, and S4). The proposed schemes had comparable interactions with air pollutants, with P2 having a more pronounced interaction than P1. In the proposed schemes, higher NDVI correlated with lower predictions and interacted with lower air pollutant concentrations. Conversely, higher NDVI contributed to higher predictions for C, interacting with fewer air pollutants. In scheme C, the regions with high NDVI were rural, whereas the regions with low NDVI were urban, suggesting that the NDVI serves as a distinguishing factor between rural and urban areas. When the temperature was lower and the air pollutant concentration was higher, their interactions contributed to better predictions in the proposed schemes (Figure 8). In contrast, temperature showed a U-shaped relationship with CVD-ASMR, with no clear interaction with air pollutants in scheme C.

Table 3. Four evaluation metrics for the maximum/minimum values of the target factor in each district, derived using random five-fold cross-validation, based on ElasticNet, RF, and LGBM in three proposed schemes.

| Maximum value	ElasticNet			RF			LGBM		
	P1	P2	C	P1	P2	C	P1	P2	C
R^2	0.44	0.81	0.43	0.64	0.79	0.74	0.68	0.77	0.73
MAE	12.6	11.2	15.0	12.0	10.0	13.6	11.0	10.0	12.0
RMSE	16.3	13.3	18.7	15.1	12.4	16.3	13.8	12.5	14.5
IOA	0.69	0.84	0.66	0.74	0.86	0.75	0.81	0.86	0.81
Minimum value	ElasticNet			RF			LGBM		
	P1	P2	C	P1	P2	C	P1	P2	C
R^2	0.33	0.62	0.43	0.36	0.60	0.53	0.48	0.57	0.49
MAE	9.1	8.4	12.3	8.6	7.2	10.5	7.7	7.0	8.8
RMSE	11.4	10.5	14.4	10.8	9.3	12.4	9.8	9.3	11.0
IOA	0.66	0.78	0.62	0.67	0.80	0.68	0.75	0.80	0.72

RF
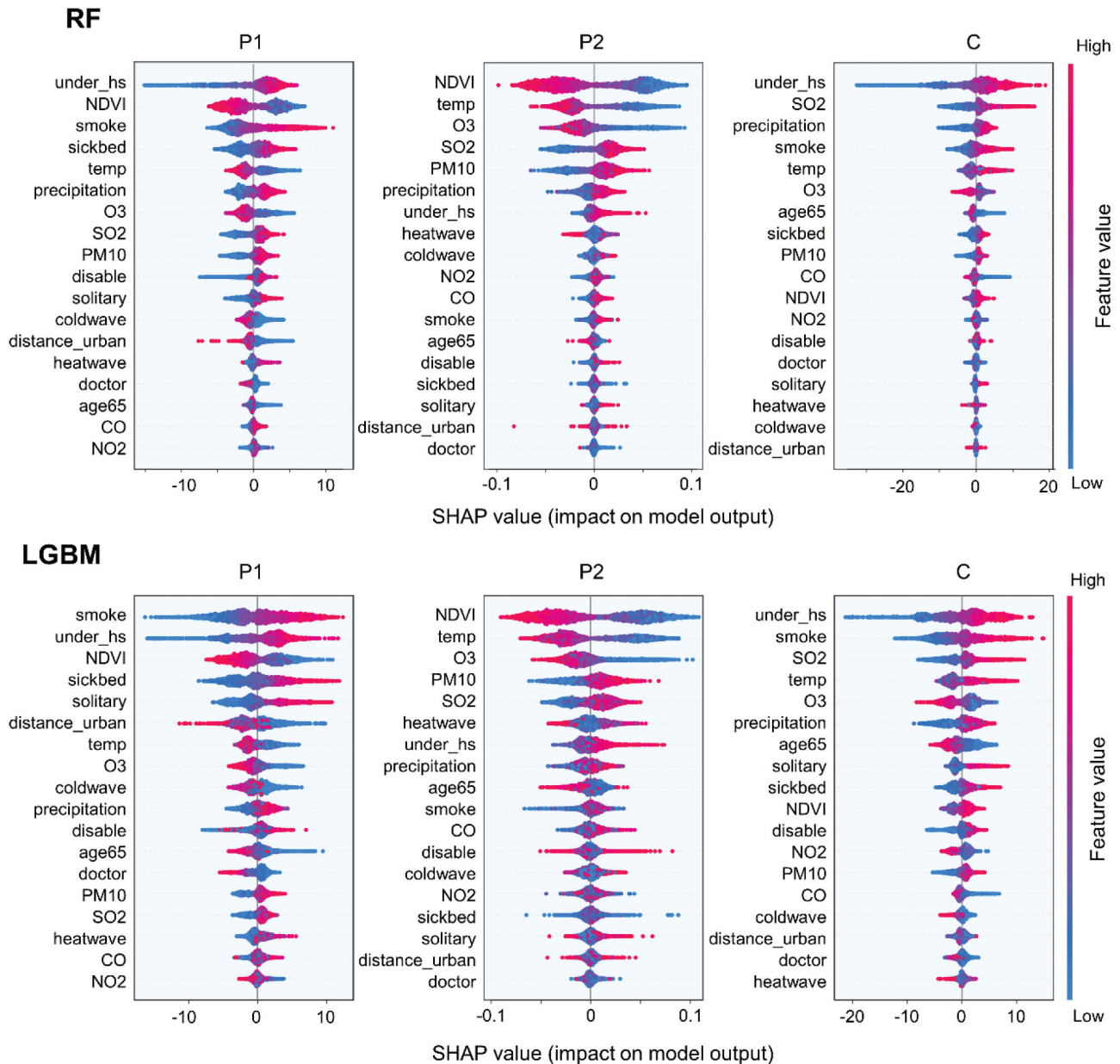
LGBM

Figure 6. SHAP summary plots were created for all input variables, ranked by global feature importance in RF and LGBM for all schemes (P1, P2, and C). Each point on the plot represents a sample, and these points are colored according to the corresponding feature values, ranging from low (blue) to high (pink). A positive SHAP value indicates an increase in the target variable above the expected value (i.e. the mean target variable), whereas a negative SHAP value indicates a decrease in the target variable below the expected value.

The interaction between under_hs and PM_{10} primarily contributed to overall CVD-ASMR and its temporal residual (Figure 9). Although environmental variables were regionally rescaled by removing its spatial characteristics, the spatial distribution of SHAP values for PM_{10} was similar to the CVD-ASMR distribution (Figures 2 and 9). Their strong positive interaction led to high CVD-ASMR.

In P2, dynamic temporal variation of CVD-ASMR was shown along with time series of PM_{10}. The spatial pattern of the SHAP value for under_hs was not as similar to that of the temporal residual of CVD-ASMR as the SHAP value for PM_{10} in P1. This was due to the high dispersion of yearly SHAP values for under_hs, which resulted in a smooth spatial map. However, the SHAP value
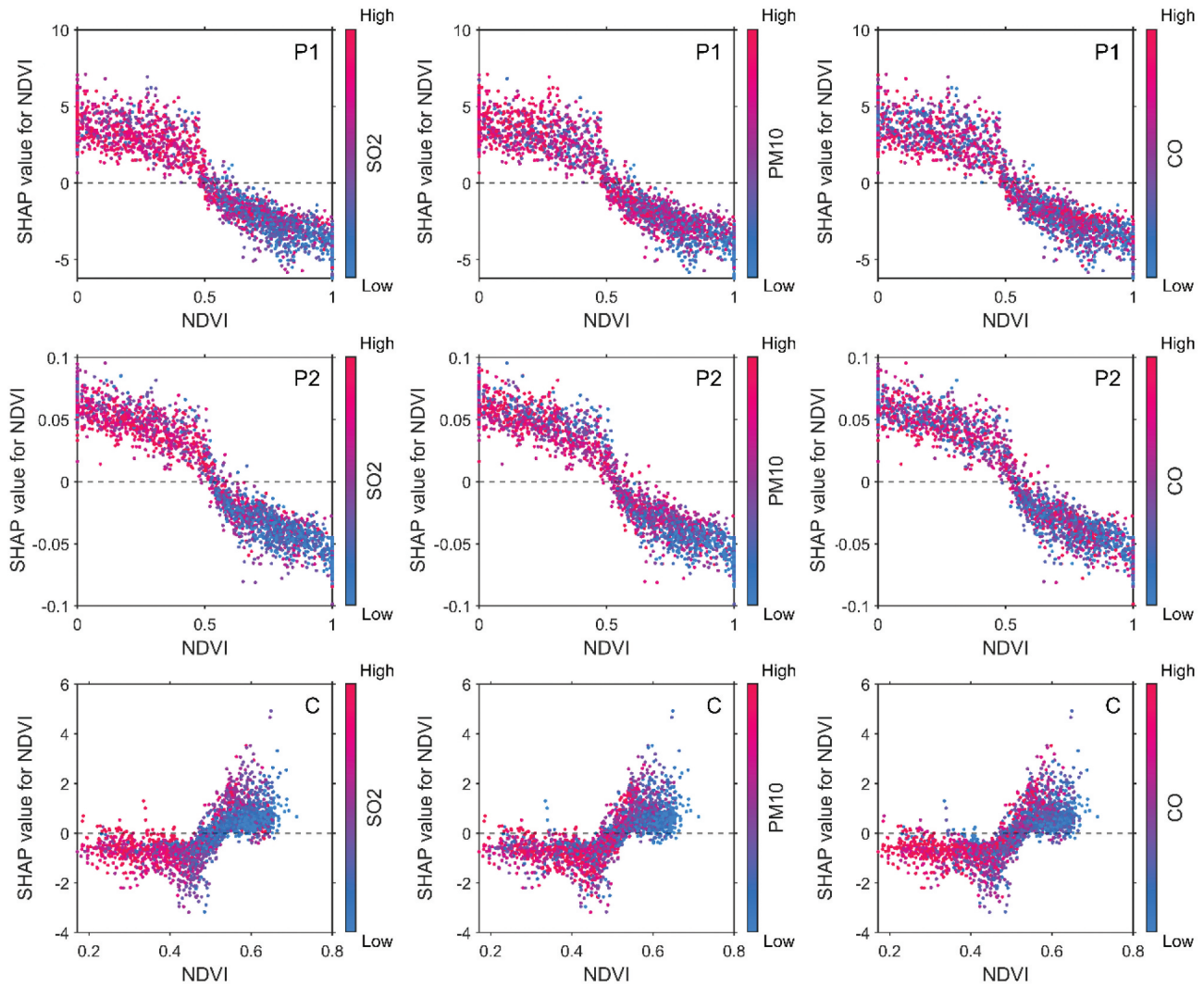
Figure 7. Rf-based SHAP values indicate the contribution of NDVI to the predicted target variables, while interacting with SO_2, PM_{10}, and CO in all schemes.

of under_hs affected the temporal variation of CVD-ASMR to the time series of PM_{10}. A high temporal variation was observed in regions with high sociodemographically vulnerable areas

5. Discussion

This study provides the first evidence from the XAI technique and detailed spatiotemporal district-level nationwide data: the unequal effects of environmental factors on district-level health. Earlier studies rarely examined sociodemographic and environmental factors with varying spatiotemporal patterns simultaneously and often overlooked the importance of its local exposure-response relationships when analyzing environmental variables. Additionally, these studies

only addressed spatial patterns of disease, disregarding temporal disease variations by district (Ke et al. 2023; Kim and Kim 2022; Ogata et al. 2021; Shimada-Sammori et al. 2023; Wen, Wang, and Shao 2024; Wertis et al. 2023; Yao et al. 2022). To comprehensively understand influencing factors from spatiotemporal perspectives, we propose two complementary schemes for two objectives: the overall spatiotemporal disease pattern and the residual disease by district. We recommend 1) disease distribution at the district level itself, characterized by dynamic spatial variations, to highlight the geographical disparities in sociodemographic vulnerability, and 2) disease residuals per district, characterized by fluctuating temporal variations, to highlight environmental factors related to disease and different neighborhood characteristics.
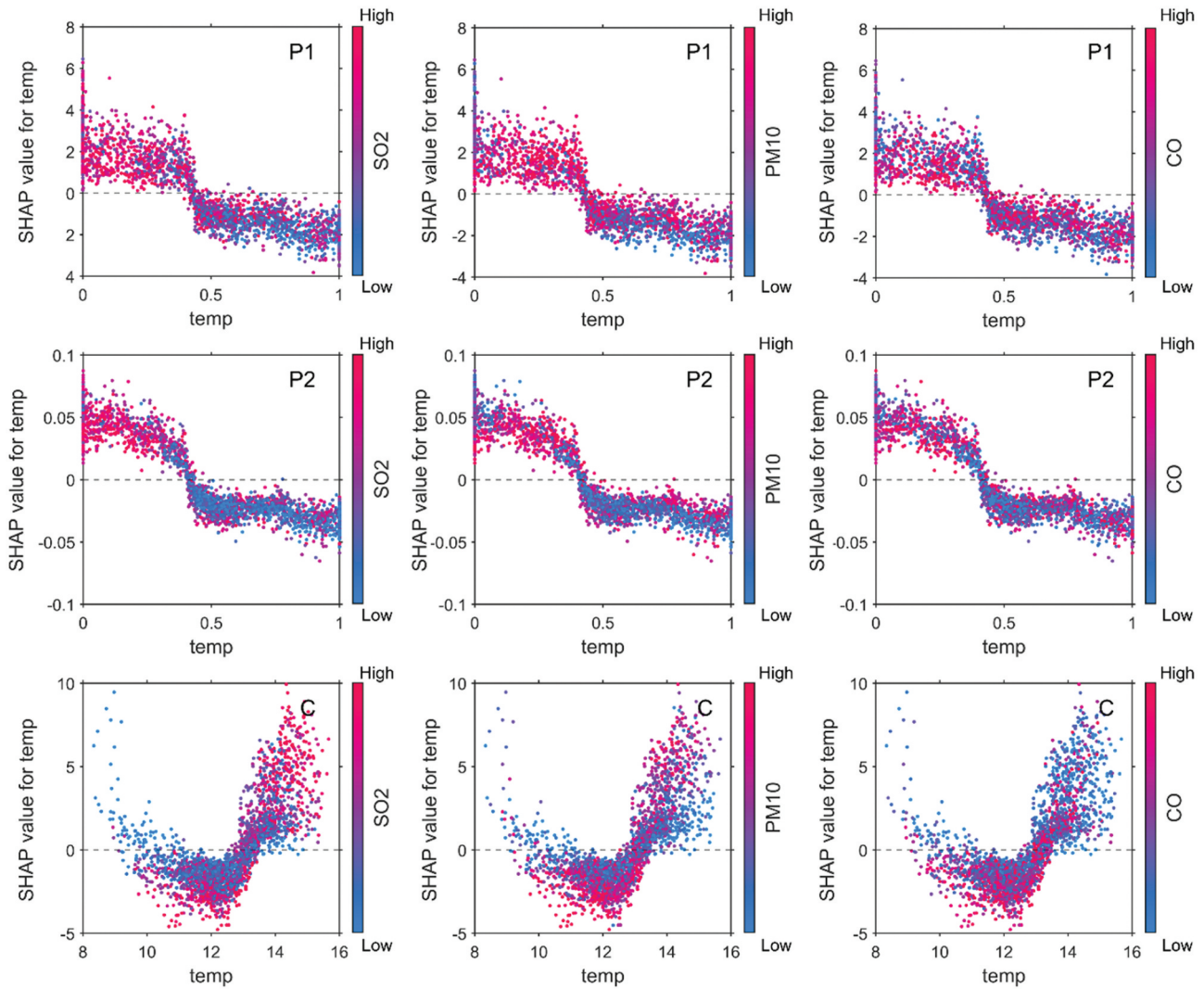
Figure 8. Rf-based SHAP values indicate the contribution of temperature to the predicted target variables, while interacting with SO_2, PM_{10}, and CO in all schemes.

To indirectly validate the proposed complementary schemes, simple statistical analyses provided valuable insights into the overall relationship between the risk factors and targets in the AI dataset (Chelgani 2021; Fatahi et al. 2022; Zhang et al. 2024). Targets with greater spatial variance showed stronger R values with sociodemographic vulnerability variables, whereas targets with higher temporal variance had larger correlation coefficients with environmental factors. Multidimensional statistical analyses revealed that spatiotemporal correlations were based on spatial rather than temporal patterns. These results contradicted established medical knowledge and, in time series analysis, found primarily positive associations between air pollutants and diseases, including CVD mortality (Kim

et al. 2021; Park et al. 2021; Pranata et al. 2020). The rescaled environmental factors can be identified based on their temporal association rather than their spatial association with district-level disease patterns based on different regional environments.

In the XAI-based analysis, sociodemographic variables positively contributed to the objectives of all schemes, implying significant impacts on the geographical disparities in sociodemographic vulnerability (e.g. lower education and unhealthy behavior) (Bevan et al. 2023; Khan et al. 2021). Environmental variables showed a significant contribution to temporal disease variation by district, suggesting that temporal environmental stressors are related to temporal disease variation despite community resilience
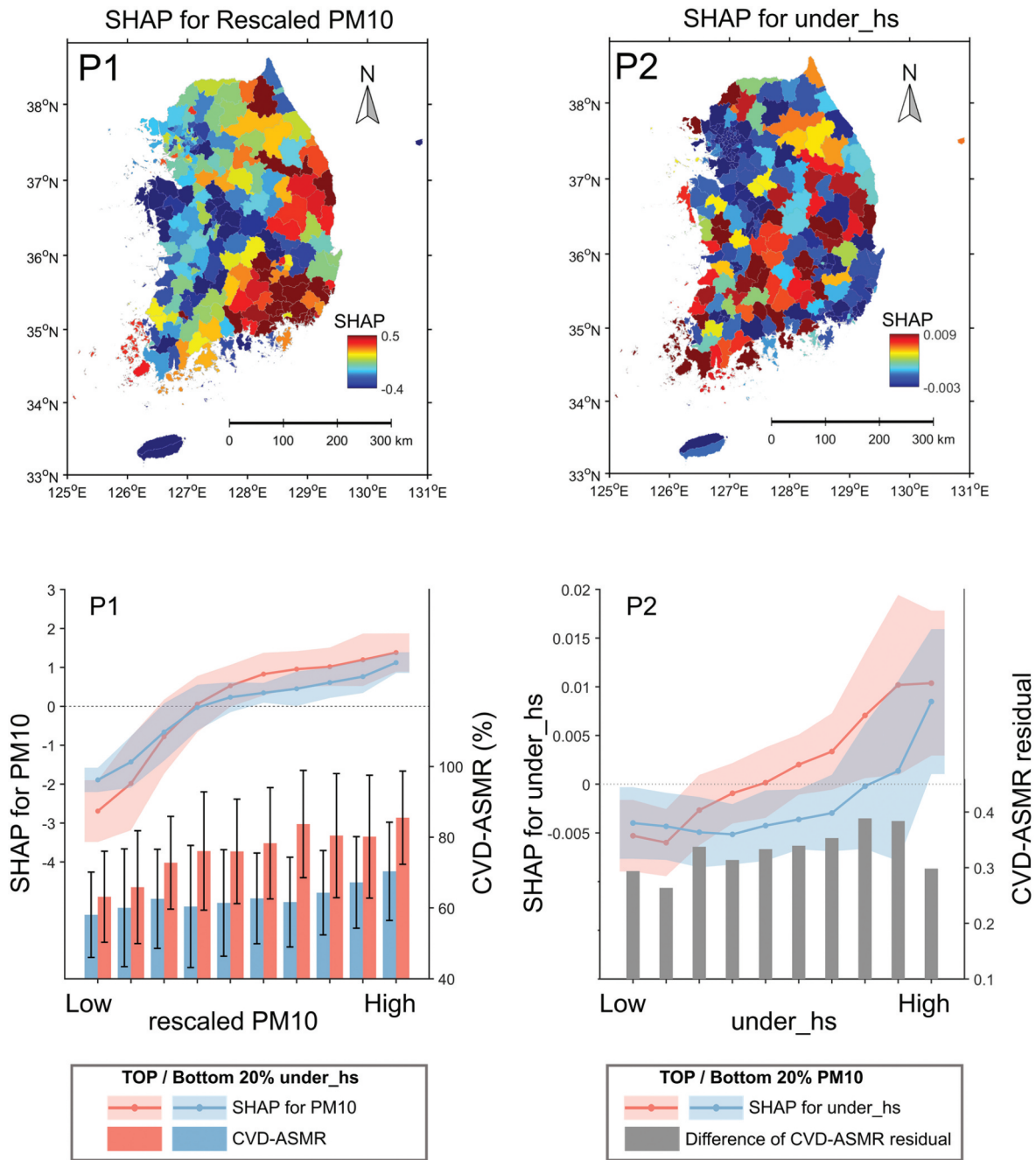
Figure 9. Spatial distribution of SHAP values for PM_{10} in P1 and under_hs in P2 (up). An area graph showing the mean (line graph) and standard deviation (shaded area) of SHAP values for rescaled PM_{10} (left) and under_hs (right) values for the top (pink) and bottom (blue) 20% of under_hs (left) and PM_{10} (right), alongside a bar graph for the mean and standard deviation of CVD-ASMR (left) and a bar graph for the difference between magnitudes of CVD-ASMR residual in situations of high (top 20%) and low (bottom 20%) PM_{10} (right).

(Xiong et al. 2019; Zhou et al. 2010). In our case study, NDVI and temperature contributed negatively to the objectives, whereas air pollutants and extreme meteorological conditions contributed positively to the objectives of the proposed scheme. The results align with findings from previous epidemiological studies (Alahmad et al. 2023; J. Chen et al. 2020; Kim

et al. 2021; Yeager, Smith, and Bhatnagar 2020). District-level environmental factors, including air quality, temperature, and green space, act as contextual determinants of local public health (Hankey and Marshall 2017; Mahakalkar et al. 2024). For instance, greenness mitigates air, light, and noise pollution through its barrier and absorption properties,

contributing to a lower CVD risk (Hedblom et al. 2017; Yeager, Smith, and Bhatnagar 2020). Cold exposure increases sympathetic nervous system activity, leading to vasoconstriction and elevated cardiac oxygen demand, while also triggering hypercoagulability due to increased blood viscosity and clotting factor abnormalities (Greaney, Kenney, and Alexander 2016; Stewart et al. 2017). Additionally, poor air quality induces oxidative stress that originates in the lungs and extends to systemic vascular oxidative stress reactions (Pranata et al. 2020).

The proposed schemes show a beneficial interaction between higher NDVI and lower air pollutants, indicating improved air quality and high greenness (Heo and Bell 2019; Yitshak-Sade, Kloog, and Novack 2017). Harmful interactions between lower temperatures and higher levels of air pollutants were observed, suggesting a compounding effect of detrimental factors on the disease (Hansel, McCormack, and Kim 2016; Son et al. 2011). High CVD-ASMR was observed with interactions between educated vulnerable areas during the high PM_{10} period (Deguen and Zmirou-Navier 2010). The dynamic temporal variation of CVD-ASMR was linked to interactions with sociodemographically disadvantaged areas and dynamic temporal variation of PM_{10}, reflecting regional health resilience. These findings implicated the important role of our proposed model in reducing regional disparities and improving temporal resilience within cities to reduce the impact of spatiotemporal health patterns at the district level.

Despite the valuable insights into spatiotemporal district-level risk factors, our study has limitations in establishing causal relationships between the identified factors and health outcomes due to the characteristics of an ecological design. Furthermore, due to challenges in data collection, this framework only conducted a risk factor analysis for CVD-ASMR in South Korea during 2010–2019. Further case studies are needed to explore the application of this framework to other chronic diseases (e.g. respiratory diseases and cancer) and infectious diseases (e.g. COVID-19), other geographical areas, and specific time periods (e.g. weekly or monthly). By integrating socioeconomic information and accounting for unmeasured confounding variables, researchers could provide comprehensive insights into regional vulnerability and sensitivity levels. In this case study, economic data (e.g. gross domestic product) could not be used due to their coarse spatial resolution.

Additionally, we recommend incorporating information on daytime and nighttime populations to achieve a more realistic spatial population distribution than relying solely on residence information (Rauch et al. 2021; Yoo et al. 2023).

6. Conclusion

This study proposed two XAI-based complementary schemes for analyzing influencing factors in spatiotemporal district-level datasets. Unlike previous studies that overlooked local exposure-response relationships to environmental factors, this study reflected regionally different impacts of environmental factors. Rather than focusing on spatiotemporal disease patterns like previous studies, this study used additional objectives to identify the risk factors associated with temporal disease variations. Two XAI-based complementary schemes provided an in-depth interpretation of the disease-related estimation model based on comprehensive and detailed feature contributions. According to the proposed framework for spatiotemporal modeling of CVD-ASMR patterns in South Korea, RF and LGBM outperformed ElasticNet in XAI-based risk factor analysis. Compared to previous risk factor analysis approaches, our SHAP analysis revealed that low greenness and temperature and high air pollutant levels were the most explainable factors for high CVD-ASMR, reflecting local exposure-response effects. The interaction between lower NDVI, temperature, and higher air pollutant levels contributed to an increase in CVD-ASMR. In addition, two complementary schemes showed different characteristics: P1, which focused on spatial disease variation showed a strong association with sociodemographic vulnerability variables, whereas P2, which focused on temporal disease variation showed a strong association with environmental variables. In addition, the two complementary schemes enabled us to reasonably analyze the interaction effect of the two different categorized risk factors (e.g. temperature and air pollutants). Furthermore, high CVD-ASMR and its temporal variation were observed in situations of high sociodemographic vulnerability with poor air quality. These results demonstrated the need for rescaling of environmental factors based on local exposure-response relationships, as well as two complementary schemes that provide multi-dimensional insights for XAI-based spatiotemporal risk factor analysis at the

district level. Public health surveillance systems have recently been introduced to monitor a wide range of health conditions, risk factors, and public health issues. The XAI-based risk factor analysis can identify data-driven potential risk factors, aiding in the screening and discovery of underlying causes behind suddenly occurring public health issues. Furthermore, this novel framework is expected to enable stakeholders and researchers to develop more sophisticated diagnostic tools for spatiotemporal district-level risk factor analysis. These insights provide a valuable foundation for policymakers and environmentalists to design targeted interventions that address issues such as degraded air quality and sudden temperature fluctuations, particularly in regions with high sociodemographic vulnerability.

Disclosure statement

Funding

ORCID

Eunjin Kang http://orcid.org/0000-0001-7738-4899
Dongjin Cho http://orcid.org/0000-0001-6795-6451
Jungho Im http://orcid.org/0000-0002-4506-6877
Cheolhee Yoo http://orcid.org/0000-0002-3922-2300

Data availability statement

The raw data used in this study, excluding CVD-ASMR and smoking rate data, are publicly available from the following sources: The processed data and code generated in this study are available upon request from the corresponding authors.

- Air pollutants: https://www.airkorea.or.kr/web/.
- Meteorological variables: https://apihub.kma.go.kr/.
- NDVI and land cover variables: https://search.earthdata.nasa.gov/search.
- Sociodemographic variables: https://kosis.kr/index/index.do."

References

Alabi, R. O., A. Almangush, M. Elmusrati, I. Leivo, and A. A. Mäkitie. 2022. "An Interpretable Machine Learning Prognostic System for Risk Stratification in Oropharyngeal Cancer." *International Journal of Medical Informatics* 168:104896. https://doi.org/10.1016/j.ijmedinf.2022.104896.

Alahmad, B., H. Khraishah, D. Royé, A. M. Vicedo-Cabrera, Y. Guo, S. I. Papatheodorou, S. Achilleos, et al. 2023. "Associations Between Extreme Temperatures and Cardiovascular Cause-Specific Mortality: Results from 27 Countries." *Circulation* 147 (1): 35–46. https://doi.org/10.1161/CIRCULATIONAHA.122.061832.

Alvarez-Mendoza, C. I., A. Teodoro, A. Freitas, and J. Fonseca. 2020. "Spatial Estimation of Chronic Respiratory Diseases Based on Machine Learning Procedures—An Approach Using Remote Sensing Data and Environmental Variables in Quito, Ecuador." *Ecuador Applied Geography* 123:102273. https://doi.org/10.1016/j.apgeog.2020.102273.

Arafa, A., H. H. Lee, E. S. Eshak, K. Shirai, K. Liu, J. Li, N. S. Anni, S. Y. Shim, H. C. Kim, and H. Iso. 2021. "Modifiable Risk Factors for Cardiovascular Disease in Korea and Japan." *Korean Circulation Journal* 51 (8): 643–655. https://doi.org/10.4070/kcj.2021.0121.

Bell, M. L., and F. Dominici. 2008. "Effect Modification by Community Characteristics on the Short-Term Effects of Ozone Exposure and Mortality in 98 US Communities." *American Journal of Epidemiology* 167 (8): 986–997. https://doi.org/10.1093/aje/kwm396.

Bevan, G., A. Pandey, S. Griggs, J. E. Dalton, D. Zidar, S. Patel, S. U. Khan, K. Nasir, S. Rajagopalan, and S. Al-Kindi. 2023. "Neighborhood-Level Social Vulnerability and Prevalence of Cardiovascular Risk Factors and Coronary Heart Disease." *Current Problems in Cardiology* 48 (8): 101182. https://doi.org/10.1016/j.cpcardiol.2022.101182.

Boudreault, J., A. Ruf, C. Campagna, and F. Chebana. 2024. "Multi-Region Models Built with Machine and Deep Learning for Predicting Several Heat-Related Health Outcomes." *Sustainable Cities and Society* 115:105785. https://doi.org/10.1016/j.scs.2024.105785.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

César de Sá, N., M. Baratchi, V. Buitenhuis, P. Cornelissen, and P. M. van Bodegom. 2022. "AutoML for Estimating Grass Height from ETM+/OLI Data from Field Measurements at a Nature Reserve." *GIScience & Remote Sensing* 59 (1): 2164–2183. https://doi.org/10.1080/15481603.2022.2152304.

Chelgani, S. C. 2021. "Estimation of Gross Calorific Value Based on Coal Analysis Using an Explainable Artificial Intelligence." *Machine Learning with Applications* 6:100116. https://doi.org/10.1016/j.mlwa.2021.100116.

Chen, F. W., and C. W. Liu. 2012. "Estimation of the Spatial Rainfall Distribution Using Inverse Distance Weighting (IDW) in the Middle of Taiwan." *Paddy and Water Environment* 10 (3): 209–222. https://doi.org/10.1007/s10333-012-0319-1.

Chen, H., Y. Lin, Q. Su, and L. Cheng. 2017. "Spatial Variation of Multiple Air Pollutants and Their Potential Contributions to All-Cause, Respiratory, and Cardiovascular Mortality Across China in 2015–2016." *Atmospheric Environment* 168:23–35. https://doi.org/10.1016/j.atmosenv.2017.09.006.

Chen, J., M. Zhou, J. Yang, P. Yin, B. Wang, C. Q. Ou, and Q. Liu. 2020. "The Modifying Effects of Heat and Cold Wave Characteristics on Cardiovascular Mortality in 31 Major Chinese Cities." *Environmental Research Letters* 15 (10): 105009. https://doi.org/10.1088/1748-9326/abaea0.

Cho, D., D. Bae, C. Yoo, J. Im, Y. Lee, and S. Lee. 2022. "All-Sky 1 Km MODIS Land Surface Temperature Reconstruction Considering Cloud Effects Based on Machine Learning." *Remote Sensing* 14 (8): 1815. https://doi.org/10.3390/rs14081815.

Cho, S. M., P. C. Austin, H. J. Ross, H. Abdel-Qadir, D. Chicco, G. Tomlinson, C. Taheri, et al. 2021. "Machine Learning Compared with Conventional Statistical Models for Predicting Myocardial Infarction Readmission and Mortality: A Systematic Review." *The Canadian Journal of Cardiology* 37 (8): 1207–1214. https://doi.org/10.1016/j.cjca.2021.02.020.

Conrad, D., and S. Capewell. 2012. "Associations Between Deprivation and Rates of Childhood Overweight and Obesity in England, 2007–2010: An Ecological Study." *BMJ Open* 2 (2): e000463. https://doi.org/10.1136/bmjopen-2011-000463.

Deguen, S., and D. Zmirou-Navier. 2010. "Social Inequalities Resulting from Health Risks Related to Ambient Air Quality—A European Review." *European Journal of Public Health* 20 (1): 27–35. https://doi.org/10.1093/eurpub/ckp220.

Dong, W., I. Motairek, K. Nasir, Z. Chen, U. Kim, Y. Khalifa, D. Freedman, S. Griggs, S. Rajagopalan, and S. G. Al-Kindi. 2023. "Risk Factors and Geographic Disparities in Premature Cardiovascular Mortality in US Counties: A Machine Learning Approach." *Scientific Reports* 13 (1): 2978. https://doi.org/10.1038/s41598-023-30188-9.

Fatahi, R., H. Nasiri, E. Dadfar, and S. Chehreh Chelgani. 2022. "Modeling of Energy Consumption Factors for an Industrial Cement Vertical Roller Mill by SHAP-XGBoost: A"conscious Lab" Approach." *Scientific Reports* 12 (1): 7543. https://doi.org/10.1038/s41598-022-11429-9.

Gao, J., Q. Meng, L. Zhang, and D. Hu. 2022. "How Does the Ambient Environment Respond to the Industrial Heat Island Effects? An Innovative and Comprehensive Methodological Paradigm for Quantifying the Varied Cooling Effects of Different Landscapes." *GIScience & Remote Sensing* 59 (1): 1643–1659. https://doi.org/10.1080/15481603.2022.2127463.

Greaney, J. L., W. L. Kenney, and L. M. Alexander. 2016. "Sympathetic Regulation During Thermal Stress in Human Aging and Disease." *Autonomic Neuroscience* 196:81–90. https://doi.org/10.1016/j.autneu.2015.11.002.

Han, L., W. Zhou, W. Li, and L. Li. 2014. "Impact of Urbanization Level on Urban Air Quality: A Case of Fine Particles (PM$_{2.5}$) in Chinese Cities." *Environmental Pollution* 194:163–170. https://doi.org/10.1016/j.envpol.2014.07.022.

Hankey, S., and J. D. Marshall. 2017. "Urban Form, Air Pollution, and Health." *Current Environmental Health Reports* 4: 491–503. https://doi.org/10.1007/s40572-017-0167-7.

Hansel, N. N., M. C. McCormack, and V. Kim. 2016. "The Effects of Air Pollution and Temperature on COPD." *COPD: Journal of Chronic Obstructive Pulmonary Disease* 13 (3): 372–379. https://doi.org/10.3109/15412555.2015.1089846.

Harford, S., H. Darabi, S. Heinert, J. Weber, T. Campbell, P. Kotini-Shah, E. Markul, K. Tataris, T. V. Hoek, and M. Del Rios. 2022. "Utilizing Community Level Factors to Improve Prediction of Out of Hospital Cardiac Arrest Outcome Using Machine Learning." *Resuscitation* 178:78–84. https://doi.org/10.1016/j.resuscitation.2022.07.006.

Hedblom, M., I. Knez, Å. Ode Sang, and B. Gunnarsson. 2017. "Evaluation of Natural Sounds in Urban Greenery: Potential Impact for Urban Nature Preservation." *Royal Society Open Science* 4 (2): 170037. https://doi.org/10.1098/rsos.170037.

Heo, S., and M. L. Bell. 2019. "The Influence of Green Space on the Short-Term Effects of Particulate Matter on Hospitalization in the U.S. for 2000–2013." *Environmental Research* 174:61–68. https://doi.org/10.1016/j.envres.2019.04.019.

Hoffmann, R., G. Borsboom, M. Saez, M. Mari Dell'olmo, B. Burström, D. Corman, C. Costa, et al. 2014. "Social Differences in Avoidable Mortality Between Small Areas of 15 European Cities: An Ecological Study." *International Journal of Health Geographics* 13 (1): 8–11. https://doi.org/10.1186/1476-072X-13-8.

Hossain, M. E., S. Uddin, and A. Khan. 2021. "Network Analytics and Machine Learning for Predictive Risk Modelling of Cardiovascular Disease in Patients with Type 2 Diabetes." *Expert Systems with Applications* 164:113918. https://doi.org/10.1016/j.eswa.2020.113918.

Hu, L., B. Liu, and Y. Li. 2020. "Ranking Sociodemographic, Health Behavior, Prevention, and Environmental Factors in Predicting Neighborhood Cardiovascular Health: A Bayesian Machine Learning Approach." *Preventive Medicine* 141:106240. https://doi.org/10.1016/j.ypmed.2020.106240.

Jang, E., Y. J. Kim, J. Im, and Y. G. Park. 2021. "Improvement of SMAP Sea Surface Salinity in River-Dominated Oceans Using Machine Learning Approaches." *GIScience & Remote Sensing* 58 (1): 138–160. https://doi.org/10.1080/15481603.2021.1872228.

Ji, J., L. Hu, B. Liu, and Y. Li. 2020. "Identifying and Assessing the Impact of Key Neighborhood-Level Determinants on Geographic Variation in Stroke: A Machine Learning and Multilevel Modeling Approach." *BMC Public Health* 20 (1): 1–12. https://doi.org/10.1186/s12889-020-09766-3.

Kamal-Chaoui, L., and J. Sanchez-Reaza. 2012. "Urban Trends and Policies in OECD Countries." *In: Organisation for Economic Cooperation and Development (OECD)* Regional Development Working Papers, Paris, France.

Kang, E., S. Park, M. Kim, C. Yoo, J. Im, and C. K. Song. 2023. "Direct Aerosol Optical Depth Retrievals Using MODIS Reflectance Data and Machine Learning Over East Asia." *Atmospheric Environment* 309:119951. https://doi.org/10.1016/j.atmosenv.2023.119951.

Ke, D., K. Takahashi, J. Y. Takakura, K. Takara, and B. Kamranzad. 2023. "Effects of Heatwave Features on Machine-Learning-Based Heat-Related Ambulance Calls Prediction Models in Japan." *Science of the Total Environment* 873:162283. https://doi.org/10.1016/j.scitotenv.2023.162283.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu. 2017. "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 3146–3154. Long Beach, CA, USA. December 4–9, 2017.

Keblouti, M., L. Ouerdachi, and H. Boutaghane. 2012. "Spatial Interpolation of Annual Precipitation in Annaba-Algeria-Comparison and Evaluation of Methods." *Energy Procedia* 18:468–475. https://doi.org/10.1016/j.egypro.2012.05.058.

Khan, S. U., Z. Javed, A. N. Lone, S. S. Dani, Z. Amin, S. G. Al-Kindi, S. S. Virani, et al. 2021. "Social Vulnerability and Premature Cardiovascular Mortality Among US Counties, 2014 to 2018." *Circulation* 144 (16): 1272–1279. https://doi.org/10.1161/CIRCULATIONAHA.121.054516.

Kim, A., S. J. Yoon, Y. A. Kim, and E. J. Kim. 2015. "The Burden of Acute Myocardial Infarction After a Regional Cardiovascular Center Project in Korea." *International Journal for Quality in Health Care* 27 (5): 349–355. https://doi.org/10.1093/intqhc/mzv064.

Kim, H., G. Byun, Y. Choi, S. Kim, S. Y. Kim, and J. T. Lee. 2021. "Effects of Long-Term Exposure to Air Pollution on All-Cause Mortality and Cause-Specific Mortality in Seven Major Cities of South Korea: Korean National Health and Nutritional Examination Surveys with Mortality Follow-Up." *Environmental Research* 192:110290. https://doi.org/10.1016/j.envres.2020.110290.

Kim, H., H. Kim, G. Byun, Y. Choi, H. Song, and J. T. Lee. 2019. Difference in Temporal Variation of Temperature-Related Mortality Risk in Seven Major South Korean Cities Spanning 1998–2013." *Science of The Total Environment* 656:986–996.

Kim, S. Y., S. J. Yi, Y. S. Eum, H. J. Choi, H. Shin, H. G. Ryou, and H. Kim. 2014. "Ordinary Kriging Approach to Predicting Long-Term Particulate Matter Concentrations in Seven Major Korean Cities." *Environmental Health and Toxicology* 29:e2014012. https://doi.org/10.5620/eht.e2014012.

Kim, Y., and Y. Kim. 2022. "Explainable Heat-Related Mortality with Random Forest and SHapley Additive exPlanations (SHAP) Models." *Sustainable Cities and Society* 79:103677. https://doi.org/10.1016/j.scs.2022.103677.

Lee, H. H., S. M. J. Cho, H. Lee, J. Baek, J. H. Bae, W. J. Chung, and H. C. Kim. 2021. "Korea Heart Disease Fact Sheet 2020: Analysis of Nationwide Data." *Korean Circulation Journal* 51 (6): 495–503. https://doi.org/10.4070/kcj.2021.0097.

Lee, K., B. Kim, and S. Park. 2023. "Evaluating the Potential of Burn Severity Mapping and Transferability of Copernicus EMS Data Using Sentinel-2 Imagery and Machine Learning Approaches." *GIScience & Remote Sensing* 60 (1): 2192157. https://doi.org/10.1080/15481603.2023.2192157.

Lee, W., M. Choi, M. L. Bell, C. Kang, J. Jang, I. Song, Y. O. Kim, K. Ebi, and H. Kim. 2022. "Effects of Urbanization on Vulnerability to Heat-Related Mortality in Urban and Rural Areas in South Korea: A Nationwide District-Level Time-Series Study." *International Journal of Epidemiology* 51 (1): 111–121. https://doi.org/10.1093/ije/dyab148.

Lee, W. H., J. Y. Choo, J. Y. Son, and H. Kim. 2016. "Association Between Long-Term Exposure to Air Pollutants and Prevalence of Cardiovascular Disease in 108 South Korean Communities in 2008–2010: A Cross-Sectional Study." *Science of the Total Environment* 565:271–278. https://doi.org/10.1016/j.scitotenv.2016.03.163.

Li, J., and A. D. Heap. 2014. "Spatial Interpolation Methods Applied in the Environmental Sciences: A Review." *Environmental Modelling & Software* 53:173–189. https://doi.org/10.1016/j.envsoft.2013.12.008.

Li, X., L. Jin, and H. Kan. 2019. "Air Pollution: A Global Problem Needs Local Fixes." *Nature* 570 (7762): 437–439. https://doi.org/10.1038/d41586-019-01960-7.

Liu, H., T. Liao, Y. Wang, X. Qian, X. Liu, C. Li, S. Li, et al. 2023. "Fine-Grained Wetland Classification for National Wetland Reserves Using Multi-Source Remote Sensing Data and Pixel Information Expert Engine (PIE-Engine)." *GIScience & Remote Sensing* 60 (1): 2286746. https://doi.org/10.1080/15481603.2023.2286746.

Lundberg, S. M., and S. I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, 4768– 4777. https://dl.acm.org/doi/10.5555/3295222.3295230.

Ma, C., Y. Xie, S. B. Duan, W. Qin, Z. Guo, G. Xi, X. Zhang, Q. Bie, H. Duan, and L. He. 2022. "Characterization of Spatiotemporal Patterns of Grassland Utilization Intensity in the Selinco Watershed of the Qinghai-Tibetan Plateau from 2001 to 2019 Based on Multisource Remote Sensing and Artificial Intelligence Algorithms." *GIScience & Remote Sensing* 59 (1): 2217–2246.

Mahakalkar, A. U., L. Gianquintieri, L. Amici, M. A. Brovelli, and E. G. Caiani. 2024. "Geospatial Analysis of Short-Term Exposure to Air Pollution and Risk of Cardiovascular Diseases and Mortality–A Systematic Review." *Chemosphere* 353:141495. https://doi.org/10.1016/j.chemosphere.2024.141495.

Manne, R. 2021. "Machine Learning Techniques in Drug Discovery and Development." *International Journal of Applied Research* 7 (4): 21–28. https://doi.org/10.22271/allresearch.2021.v7.i4a.8455.

Morgenstern, H. 1995. "Ecologic Studies in Epidemiology: Concepts, Principles, and Methods." *Annual Review of Public Health* 16 (1): 61–81. https://doi.org/10.1146/annurev.pu.16.050195.000425.

Nakashima, T., S. Ogata, T. Noguchi, Y. Tahara, D. Onozuka, S. Kato, Y. Yamagata, et al. 2021. "Machine Learning Model for Predicting Out-Of-Hospital Cardiac Arrests Using Meteorological and Chronological Data." *Heart* 107 (13): 1084–1091.

Odebiri, O., O. Mutanga, J. Odindi, K. Peerbhay, and S. Dovey. 2020. "Predicting Soil Organic Carbon Stocks Under Commercial Forest Plantations in KwaZulu-Natal Province, South Africa Using Remotely Sensed Data." *GIScience &*

Remote Sensing 57 (4): 450–463. https://doi.org/10.1080/15481603.2020.1731108.

Ogata, S., M. Takegami, T. Ozaki, T. Nakashima, D. Onozuka, S. Murata, Y. Nakaoku, et al. 2021. "Heatstroke Predictions by Machine Learning, Weather Information, and an All-Population Registry for 12-Hour Heatstroke Alerts." *Nature Communications* 12 (1): 4575.

Papadogeorgou, G., and F. Dominici. 2020. "A Causal Exposure Response Function with Local Adjustment for Confounding: Estimating Health Effects of Exposure to Low Levels of Ambient Fine Particulate Matter." *The Annals of Applied Statistics* 14 (2): 850. https://doi.org/10.1214/20-AOAS1330.

Parikh, R. B., C. Manz, C. Chivers, S. H. Regli, J. Braun, M. E. Draugelis, L. M. Schuchter, et al. 2019. "Machine Learning Approaches to Predict 6-Month Mortality Among Patients with Cancer." *JAMA Network Open* 2 (10): e1915997–e1915997. https://doi.org/10.1001/jamanetworkopen.2019.15997.

Park, J., H. J. Kim, C. H. Lee, C. H. Lee, and H. W. Lee. 2021. "Impact of Long-Term Exposure to Ambient Air Pollution on the Incidence of Chronic Obstructive Pulmonary Disease: A Systematic Review and Meta-Analysis." *Environmental Research* 194:110703. https://doi.org/10.1016/j.envres.2020.110703.

Peng, J., K. Zou, M. Zhou, Y. Teng, X. Zhu, F. Zhang, and J. Xu. 2021. "An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients." *Journal of Medical Systems* 45 (5): 1–9. https://doi.org/10.1007/s10916-021-01736-5.

Pranata, R., R. Vania, A. E. Tondas, B. Setianto, and A. Santoso. 2020. "A Time-To-Event Analysis on Air Pollutants with the Risk of Cardiovascular Disease and Mortality: A Systematic Review and Meta-Analysis of 84 Cohort Studies." *Journal of Evidence-Based Medicine* 13 (2): 102–115.

Rauch, S., H. Taubenböck, C. Knopp, and J. Rauh. 2021. "Risk and Space: Modelling the Accessibility of Stroke Centers Using Day- & Nighttime Population Distribution and Different Transportation Scenarios." *International Journal of Health Geographics* 20 (1): 1–15. https://doi.org/10.1186/s12942-021-00284-y.

Razavi-Termeh, S. V., A. Sadeghi-Niaraki, and S. M. Choi. 2022. "Coronavirus Disease Vulnerability Map Using a Geographic Information System (GIS) from 16 April to 16 May 2020." *Physics and Chemistry of the Earth, Parts A/B/C* 126:103043. https://doi.org/10.1016/j.pce.2021.103043.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier."*Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco, California, USA, 1135–1144. August.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2018. "Anchors: High-Precision Model-Agnostic Explanations." *Proceedings of the AAAI conference on artificial intelligence*, New Orleans, Louisiana, USA (Vol. 32 , April, No. 1).

Rivera-González, L. O., Z. Zhang, B. N. Sánchez, K. Zhang, D. G. Brown, L. Rojas-Bracho, A. Osornio-Vargas, F. Vadillo-Ortega, and M. S. O'Neill. 2015. "An Assessment of Air Pollutant Exposure Methods in Mexico City, Mexico."

Journal of the Air & Waste Management Association 65 (5): 581–591. https://doi.org/10.1080/10962247.2015.1020974.

Rosenberg, P., M. Kano, I. Ludford, A. Prasad, and H. Thomson. 2016. Global Report on Urban Health: Equitable, Healthier Cities for Sustainable Development. World Health Organization.

Ross, E. G., N. H. Shah, R. L. Dalman, K. T. Nead, J. P. Cooke, and N. J. Leeper. 2016. "The Use of Machine Learning for the Identification of Peripheral Artery Disease and Future Mortality Risk." *Journal of Vascular Surgery* 64 (5): 1515–1522. https://doi.org/10.1016/j.jvs.2016.04.026.

Roth, G. A., G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, N. C. Barengo, et al. 2020. "Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update from the GBD 2019 Study." *Journal of the American College of Cardiology* 76 (25): 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. "Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization." *Proceedings of the IEEE international conference on computer vision*, Venice, Italy, 618–626.

Seto, K. C., S. Parnell, and T. Elmqvist. 2013. *Urbanization, Biodiversity and Ecosystem Services: Challenges and Opportunities: A Global Assessment*, 1–12. https://doi.org/10.1007/978-94-007-7088-1_1.

Shimada-Sammori, K., T. Shimada, R. E. Miura, R. Kawaguchi, Y. Yamao, T. Oshima, T. Oami, K. Tomita, K. Shinozaki, and T. A. Nakada. 2023. "Machine Learning Algorithms for Predicting Days of High Incidence for Out-Of-Hospital Cardiac Arrest." *Scientific Reports* 13 (1): 9950. https://doi.org/10.1038/s41598-023-36270-6.

Shtiliyanova, A., G. Bellocchi, D. Borras, U. Eza, R. Martin, and P. Carrère. 2017. "Kriging-Based Approach to Predict Missing Air Temperature Data." *Computers and Electronics in Agriculture* 142:440–449. https://doi.org/10.1016/j.compag.2017.09.033.

Son, J. Y., M. L. Bell, and J. T. Lee. 2014. "The Impact of Heat, Cold, and Heat Waves on Hospital Admissions in Eight Cities in Korea." *International Journal of Biometeorology* 58 (9): 1893–1903. https://doi.org/10.1007/s00484-014-0791-y.

Son, J. Y., J. T. Lee, G. B. Anderson, and M. L. Bell. 2011. "Vulnerability to Temperature-Related Mortality in Seoul, Korea." *Environmental Research Letters* 6 (3): 034027. https://doi.org/10.1088/1748-9326/6/3/034027.

Song, Y. J. 2009. "The South Korean Health Care System." *Jmaj* 52 (3): 206–209.

Stewart, S., A. K. Keates, A. Redfern, and J. J. McMurray. 2017. "Seasonal Variations in Cardiovascular Disease." *Nature Reviews Cardiology* 14 (11): 654–664.

Sun, C., J. Li, Y. Liu, L. Cao, J. Zheng, Z. Yang, J. Ye, and Y. Li. 2022. "Ecological Quality Assessment and Monitoring Using a Time-Series Remote Sensing-Based Ecological Index (Ts-RSEI)." *GIScience & Remote Sensing* 59 (1): 1793–1816.

Tan, X., L. Han, X. Zhang, W. Zhou, W. Li, and Y. Qian. 2021. "A Review of Current Air Quality Indexes and Improvements Under the Multi-Contaminant Air Pollution Exposure."

Journal of Environmental Management 279:111681. https://doi.org/10.1016/j.jenvman.2020.111681.

Tayal, A., J. Gupta, A. Solanki, K. Bisht, A. Nayyar, and M. Masud. 2021. "DL-CNN-based Approach with Image Processing Techniques for Diagnosis of Retinal Diseases." *Multimedia Systems* 28 (4): 1–22. https://doi.org/10.1007/s00530-021-00769-7.

Wang, C., L. Feng, and Y. Qi. 2021. "Explainable Deep Learning Predictions for Illness Risk of Mental Disorders in Nanjing, China." *Environmental Research* 202:111740. https://doi.org/10.1016/j.envres.2021.111740.

Wen, X., Y. Wang, and Z. Shao. 2024. "The Spatiotemporal Trend of Human Brucellosis in China and Driving Factors Using Interpretability Analysis." *Scientific Reports* 14 (1): 4880. https://doi.org/10.1038/s41598-024-55034-4.

Wertis, L., M. M. Sugg, J. D. Runkle, and D. Rao. 2023. "Socio-Environmental Determinants of Mental and Behavioral Disorders in Youth: A Machine Learning Approach." *GeoHealth* 7 (9): e2023GH000839. https://doi.org/10.1029/2023GH000839.

Willmott, C. J. 1981. "On the Validation of Models." *Physical Geography* 2 (2): 184–194. https://doi.org/10.1080/02723646.1981.10642213.

Wu, Y., B. Wen, A. Gasparrini, B. Armstrong, F. Sera, E. Lavigne, S. Li, et al. 2024. "Temperature Frequency and Mortality: Assessing Adaptation to Local Temperature." *Environment International* 187:108691. https://doi.org/10.1016/j.envint.2024.108691.

Xiong, J., C. Ye, T. Zhou, and W. Cheng. 2019. "Health Risk and Resilience Assessment with Respect to the Main Air Pollutants in Sichuan." *International Journal of Environmental Research and Public Health* 16 (15): 2796. https://doi.org/10.3390/ijerph16152796.

Yang, J. S., Y. Q. Wang, and P. V. August. 2004. "Estimation of Land Surface Temperature Using Spatial Interpolation and Satellite-Derived Surface Emissivity." *Journal of*

Environmental Informatics 4 (1): 37–44. https://doi.org/10.3808/jei.200400035.

Yao, Y., H. Yin, C. Xu, D. Chen, L. Shao, Q. Guan, and R. Wang. 2022. "Assessing Myocardial Infarction Severity from the Urban Environment Perspective in Wuhan, China." *Journal of Environmental Management* 317:115438. https://doi.org/10.1016/j.jenvman.2022.115438.

Yeager, R. A., T. R. Smith, and A. Bhatnagar. 2020. "Green Environments and Cardiovascular Health." *Trends in Cardiovascular Medicine* 30 (4): 241–246. https://doi.org/10.1016/j.tcm.2019.06.005.

Yitshak-Sade, M., I. Kloog, and V. Novack. 2017. "Do Air Pollution and Neighborhood Greenness Exposures Improve the Predicted Cardiovascular Risk?" *Environment International* 107:147–153. https://doi.org/10.1016/j.envint.2017.07.011.

Yoo, C., J. Im, Q. Weng, D. Cho, E. Kang, and Y. Shin. 2023. "Diurnal Urban Heat Risk Assessment Using Extreme Air Temperatures and Real-Time Population Data in Seoul." *Iscience* 26 (11): 108123. https://doi.org/10.1016/j.isci.2023.108123.

Zhang, L., Z. Xin, Q. Guan, L. Feng, C. Hu, C. Zhang, and H. Zhou. 2024. "Monitoring and Understanding Chlorophyll-A Concentration Changes in Lakes in Northeastern China Using MERIS and OLCI Satellite Data." *GIScience & Remote Sensing* 61 (1): 2285166.

Zhou, H., J. A. Wang, J. Wan, and H. Jia. 2010. "Resilience to Natural Hazards: A Geographic Perspective." *Natural Hazards* 53 (1): 21–41. https://doi.org/10.1007/s11069-009-9407-y.

Zhu, Y., C. Geiß, and E. So. 2024. "Simulating Urban Expansion with Interpretable Cycle Recurrent Neural Networks." *GIScience & Remote Sensing* 61 (1): 2363576. https://doi.org/10.1080/15481603.2024.2363576.

Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society Series B: Statistical Methodology, O* 67 (2): 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.