scientific reports

Check for updates

OPEN

A novel spatial heteroscedastic generalized additive distributed lag model for the spatiotemporal relation between $PM_{2.5}$ and cardiovascular hospitalization

Ali Hadianfar[1,2], Helmut Küchenhoff[2], Shahab MohammadEbrahimi[3] & Azadeh Saki[1]✉

Many studies have examined the impact of air pollution on cardiovascular hospitalization (CVH), but few have looked at the delayed effects of air pollution on CVH. Additionally, there has been no research on the spatial and temporal differences in how environmental pollutants affect CVH. This study seeks to identify spatial heteroscedasticity in the relation between PM2.5 and CVH by developing a Generalized Additive Distributed Lag (GADL) model. Data on hospitalization due to cardiovascular disease were collected from the Hospital Information System (HIS) of Mashhad University of Medical Science from 2017 to 2020. Air pollution data from 22 air quality monitoring (AQM) stations were obtained from the Environmental Pollution Monitoring Center of Mashhad administrates. Markov Random Field (MRF) smoother was utilized in the GADL model to account for spatial heteroscedasticity in the observations. This developed model is a Spatial Heteroscedastic Generalized Additive Distributed Lag (SHGADL) model. Our use of GADL allowed us to discover a significant relationship between PM2.5 exposures and the risk of CVH at lags 0 and 1 in all districts. Our results reveal heteroscedasticity in the Relative Risks (RR) of PM2.5 on CVH across different districts. After accounting for this spatial heteroscedasticity, we found that the RR of PM2.5 on CVH at lags 0 and 1 were 1.0102 (95% CI: 1.0034, 1.0170) and 1.0043 (95% CI: 1.0009, 1.0078) respectively. The central and southeastern districts showed higher RR for CVH. The developed SHGADL model provides evidence of a significant lagged effect of PM2.5 exposures on CVH, and identifies low- and high-risk districts for CVH in Mashhad. This finding can assist decision-makers in allocating resources and planning strategically, with a focus on local interventions to manage ambient air pollution and providing emergency care for CVH.

Keywords Cardiovascular hospitalization, $PM_{2.5}$, Generalized additive distributed lag model, Spatial Heteroscedastic generalized Additive distributed lag model, Exposure-lag-response association, Markov Random Field

Cardiovascular diseases (CVDs), such as heart disease, stroke, and arrhythmias, are a significant contributor to illness and death, placing substantial financial strain on countries globally[1]. In Iran, CVDs are the primary cause of death, leading to millions of disability-adjusted life years (DALYs)[2]. Additionally, CVDs are responsible for 46% of all fatalities and contribute to 20–23% of the overall disease burden in Iran[3]. Predictions suggest that the DALYs linked to CVDs in Iranians aged 30 years and older will double by 2025 compared to the statistics from 2005[2].

Air pollution presents a widespread global health concern that heightens the likelihood of cardiovascular disease. Particulate matter (PM), specifically those with a diameter smaller than 2.5 micrometers ($PM_{2.5}$), is a significant contributor to the detrimental impact on cardiovascular health[4]. Recent research, including epidemiological and clinical studies, has indicated that exposure to air pollution, particularly $PM_{2.5}$, is linked to a rise in both mortality and morbidity related to cardiovascular disease[5–9].

[1]Department of Epidemiology and Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran. [2]Statistical Consulting Unit StaBLab, Department of Statistics, Ludwig-Maximilians- Universität, Munich, Germany. [3]Department of Medical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran. ✉email: Sakia@mums.ac.ir; azadehsaki@yahoo.com

Various models and methodologies have been used to study the impact of air pollution on health outcomes. Time series methods often use moving averages of air pollution concentrations for exposure assessment, but this may underestimate the overall effect due to the failure to capture the complete lag structure[10]. Distributed Lag Models (DLMs) have gained prominence in epidemiological and environmental health studies, allowing for the estimation of associations between environmental exposures and health outcomes at various lagged periods. Recent attempts have aimed to move beyond linear relationships by exploring non-linear effects, leading to the development of the Generalized Additive Distributed Lag (GADL) model. This modeling approach has gained popularity in environmental health studies, particularly in the fields of climate change and air pollution research[11–21]. Despite significant progress in environmental exposure models, there is limited research on spatial heteroscedasticity and spatial dependency in assessing the delayed effects of environmental exposure on health outcomes. Studies have shown that there is significant spatial heteroscedasticity in air pollution concentrations at a small scale within urban areas. However, most existing studies on cardiovascular disease have only looked at average ambient pollutant concentrations across entire cities, ignoring the within-city spatial variability in air pollutant levels. By considering spatial heteroscedasticity, GADLs can better capture geographical relationships and provide more accurate estimations, especially in scenarios where spatial factors are important[22,23]. Therefore, incorporating the complex spatial structure aspects into GADL is crucial for improving the accuracy and precision of the investigated relationships.

A previous study reported significant variations in $PM_{2.5}$ levels across Mashhad districts, but existing studies on air pollution and cardiovascular disease use the average of $PM_{2.5}$for all districts, neglecting spatial heteroscedasticity[24]. The goal of present study is to incorporating this spatial heteroscedasticity in air pollution concentrations and its effect on the risk of CVH by a Spatial Heteroscedastic Generalized Additive Distributed Lag (SHGADL) model.

Materials and methods

Study area

This research was carried out in Mashhad, which is the capital of Razavi Khorasan province in the northeastern region of Iran. With a population estimated at nearly 3.8 million, it is the second most populous city in Iran[25]. As shown in Fig. 1 the city is comprised of 13 administrative districts.

Data source

The Hospital Information System (HIS) of Mashhad University of Medical Science provided hospitalization data for cardiovascular disease from 2017 to 2020, as depicted in Fig. 2. For these patients all methods were performed in accordance with the relevant guidelines and regulations. The selection of cardiovascular diseases was based on the International Classification of Diseases, 10th Revision (ICD-10), using codes I00-I99. Patients' residential addresses were geocoded to latitude and longitude formats using Google Maps, and only patients residing within Mashhad city boundaries were included in the study. The average 24-hour concentration of $PM_{2.5}$ was gathered from 22 AQM stations located across various parts of Mashhad, as shown in Figs. 1 and 2. These stations are managed by the Environmental Pollution Monitoring Center of Mashhad.

Missing data imputation

In our original AQM station data, we found that there was an average of 30% missing data in the $PM_{2.5}$measurements (Fig. 1 in Appendix S1). Machine learning methods are best to impute missing data to improve accuracy[26]. Recent research has indicated that machine learning techniques, particularly the Random Forest (RF) and k-Nearest Neighbor (KNN) algorithms, are effective for filling in missing data[27,28]. Random Forest (RF) is a powerful machine learning algorithm that constructs an ensemble of decision trees to make predictions. One of its key advantages is its robustness to overfitting, as it averages the results from multiple trees to improve accuracy. RF is particularly effective at capturing complex nonlinear relationships in the data[29,30]. KNN is a non-parametric method that imputes missing values based on the values of the k nearest neighbors. KNN is simple to implement and requires minimal parameter tuning, primarily the choice of k, the number of neighbors to consider. A significant advantage of KNN is that it makes no assumptions about the underlying data distribution, making it flexible and broadly applicable[30]. We used both models to fill in the missing $PM_{2.5}$ data for each AQM station, and then compared their performance using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) (refer to Table 1 in Appendix S1). The results showed that the random forest method outperformed the k-Nearest Neighbor algorithm in filling in missing $PM_{2.5}$ data, with the exception of station 6. Therefore, we chose the random forest method as the preferred model for filling in missing data. Figure 2 in Appendix S1, illustrates how the imputed $PM_{2.5}$ data aligns with the actual information at each station.

District-level exposurse assessment.

The exposure to $PM_{2.5}$at the district level was calculated using inverse distance weighting (IDW) interpolation. IDW is a deterministic interpolation method that uses the inverse distance between points for weighting. IDW is computationally straightforward to implement, which can be advantageous in studies requiring rapid assessments or when computational resources are limited. Various studies used IDW for PM trend evaluation and PM exposure analysis[32–35]. Initially, a 500-meter by 500-meter grid was established within the study area. The values at each grid point were estimated by averaging the weighted sum of values from nearby monitoring stations within the study area. The weights assigned to each neighboring value were determined based on the inverse distance between the monitoring station locations and unmeasured locations, raised to a non-negative power[36]. Finally, daily averages of $PM_{2.5}$for each district were calculated based on the average of the interpolated grid values within that specific district. To validate the IDW method, a leave-one-out cross-validation was used

	Total	Mean	SD	Min	Percentile			Max
					$P_{0.25}$	$P_{0.50}$	$P_{0.75}$	
CVH	45,089	31	12	3	22	30	39	77
Gender								
Male	22,851 (54.67%)	17	7	1	12	17	22	42
Female	18,943 (45.32%)	14	7	1	10	14	18	47
Age(year)								
<65	22,781 (50.52%)	16	7	0	10	15	20	44
≥65	22,307 (49.47%)	15	6	1	11	15	19	42
$PM_{2.5}$(μg/m^3)		32.70	14.04	11.87	24.92	30.56	37.88	90.00

Table 1. Descriptive statistics of daily CVH and $PM_{2.5}$ concentrations in Mashhad, from 2017 to 2020.

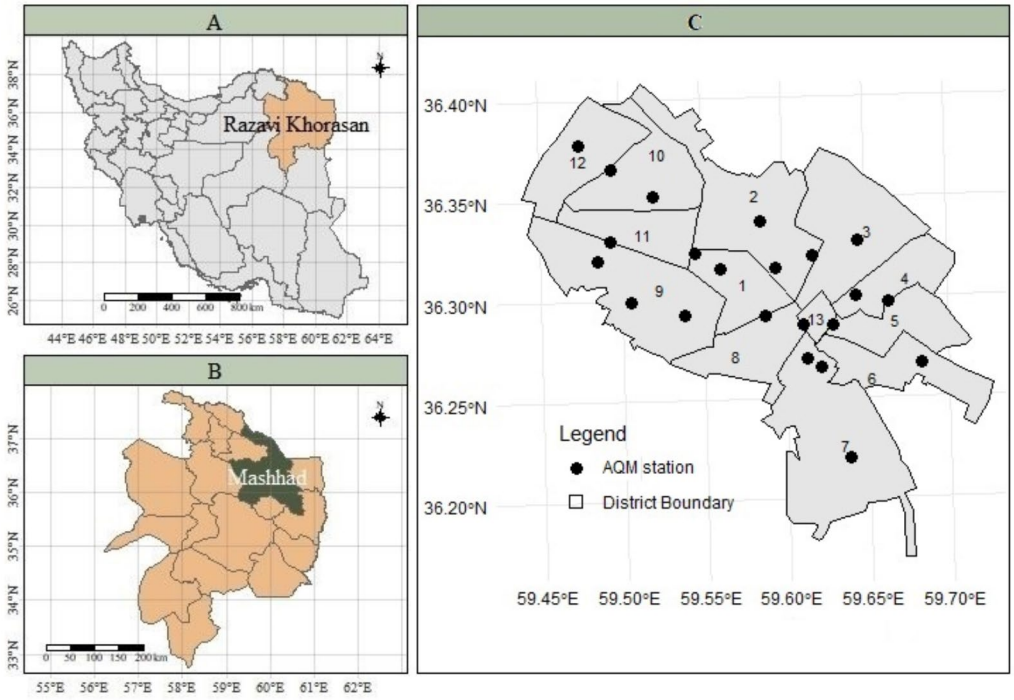

Fig. 1. Study area location (**A** and **B**) with distribution of 22 AQM stations within 13 administrative districts in Mashhad (**C**), Iran.

to determine the optimal power value for distance measurement and the number of nearest neighbors. The parameter set that minimized the prediction root mean square error (RMSE) was selected for each day[37]. The power of distance measurement was tested using five values from 1 to 3, and the number of neighbors varied from 3 to the maximum number of stations.

Dealing with confounding factors

Two Confounding factors influenced the daily number of CVHs during the study period. The first factor was the introduction of a health screening program for Iranians aged 30 and above, which aimed to identify high-risk individuals and refer them to hospitals for further evaluation and treatment. The second factor was the onset of the COVID-19 pandemic, which had a significant impact on hospitals (Fig. 2). The pandemic led to healthcare facilities needing to allocate resources and capacity to treat COVID-19 patients, creating challenges in the delivery of healthcare services. To account for the influence of these two factors, we introduced two dummy variables. Days during the COVID-19 pandemic or days during the screening health program were assigned a value of 1, while other days were assigned a value of 0.

Data analysis

The Generalized Additive Distributed Lag (GADL) model is designed to analyze the relationship between environmental exposures and health outcomes over time, allowing for both nonlinear and lagged effects. Its key functions include:

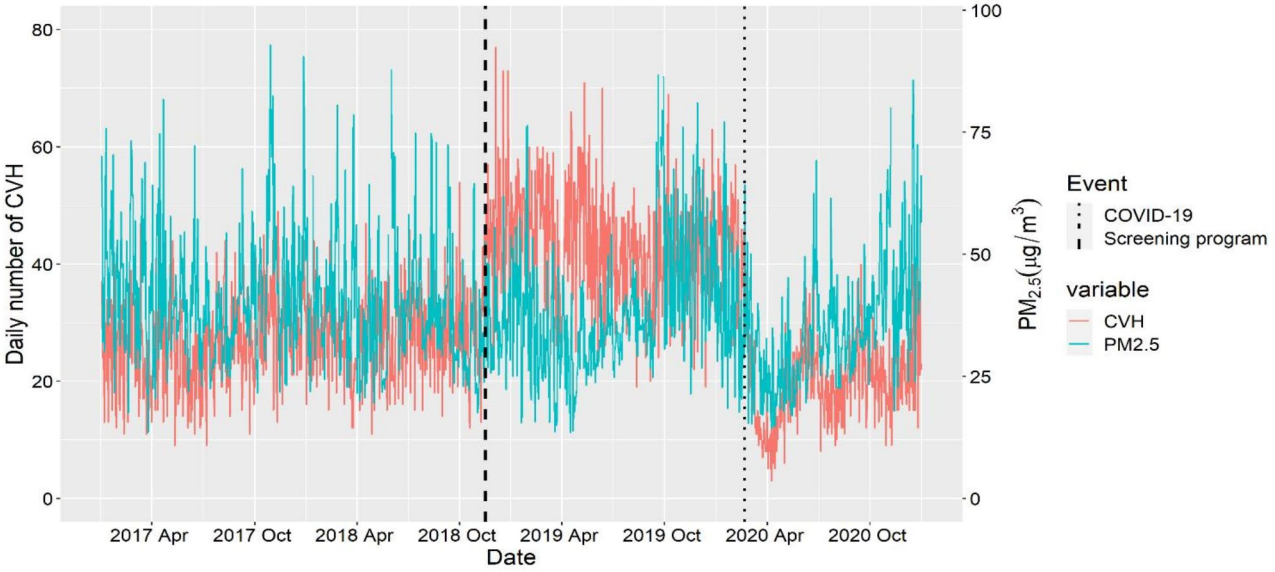

Fig. 2. Time series plot of $PM_{2.5}$ concentrations and CVH from 2017 to 2020, Mashhad, Iran.

1. Nonlinear Relationships: The GADL model employs smoothing functions to capture complex, nonlinear associations between exposure variables (e.g., PM2.5 levels) and health outcomes (e.g., cardiovascular hospitalizations). This flexibility is crucial in environmental health studies, where relationships are not always linear.
2. Lagged Effects: The model allows for the examination of delayed effects of exposures, enabling researchers to assess how health outcomes may be influenced by past exposure levels. This is particularly important in understanding the cumulative impact of pollutants like PM2.5 over time.
3. Handling Confounding Factors: The GADL can include additional covariates to control for potential confounders, such as demographic factors or seasonal variations, enhancing the robustness of the findings.

A GADL to analyze the association of a time series of outcomes to covariates is given by

$$g(\mu_t) = \alpha + \sum_{l=1}^{L} s_l(x_{tl}; \boldsymbol{\beta}_l) + \sum_{k=1}^{K} \gamma_k z_{tk}, \qquad t = 1, \ldots, T$$

Where, $\mu_t = E(Y_t)$, g is a monotonic link function, and Y is assumed to follow a distribution belonging to the exponential family. The functions s_l represent smoothed relationships between the variables x_l and the linear predictor, defined by the parameter vectors $\boldsymbol{\beta}_l$. The variables z_k includes additional predictors or confounders with linear effects specified by the related

coefficients γ_k. Here, l and Lrepresent the lag number and the maximum lag, respectively[16]. Initially, we established a fundamental model based on the original GADL to examine the non-linear lag effect of $PM_{2.5}$ on CVH. Suppose Y_t is the number of CVH at day t, following a Poisson distribution $Poisson(\mu_t)$, where μ_t is the expected value of CVH on day *t*. Due to the extra variation in Y_t relative to μ_t, we considered a quasi-Poisson model to deal with the over-dispersion problem. So, The global model for the association between $PM_{2.5}$ on CVH is defined by the following formula:

$$\log(E(CVH_t)) = \alpha + cb(PM_{2.5}, \text{lag}) + ns(time_t, df) + \beta_1 factor(dw_t) + \beta_2 holidy_t + \beta_3 screening_t + \beta_4 covid_t + offset$$

Where log(E(CVH_t)) is the canonical link function for count data with Poisson and quasi-Poisson distribution, α represents the constant term; cb indicates the cross-basis matrix obtained by applying the GADL to the concentration of $PM_{2.5}$; with lag representing the number of lag; *ns ()* indicates the smoother of the nature cubic spline for time trend (Fig. 3 in Appendix S1). The "*dw*" and "*holiday*" are the dummy variables for the day of the week and public holidays. Also two dummy variables; "screening" and "covid" included in the GADL for health screening program and the COVID-19 pandemic confounders. The offset term is the logarithm of population.

This GADL model is applicable in three key areas. First, in the epidemiological domain, it is utilized to investigate the effects of air pollution on health outcomes in urban environments, considering both immediate and delayed impacts. Second, in the context of climate change, this model is effective for assessing how changes in environmental conditions affect public health over time, particularly concerning air pollution. Lastly, in policy evaluation, the GADL model aids in assessing the effectiveness of interventions designed to reduce pollution levels and their subsequent health impacts, offering valuable insights for policymakers.

Spatial heteroscedastic generalized additive distributed lag (SHGADL)

To incorporate spatial heteroscedasticity, we employed a SHGADL by adding the spatial random effect ($s\,(i)$) for the districts in GADL model as follows:

$$\log\left(E(CVH_{i,t})\right) = \alpha + cb(PM_{2.5,i}, \text{lag}) + ns\,(time_t, df) + \beta_1 factor(dw_t) + \beta_2 holidy_t + \beta_3 screening_t + \beta_4 covid_t + s(i) + offset(i)$$

Where $CVH_{i,t}$ and $PM_{2.5,i}$ represent the number of CVH and average $PM_{2.5}$ concentration on day t in district i respectively. The spatial function $s\,(i)$ was estimated using the Markov Random Field (MRF) smoother to account for spatial autocorrelations and spatial heteroscedasticity. It is implemented through a conditional autoregressive prior following a Normal distribution:

$$s\left(i'\right)|i \neq i' \sim N\left(\sum_{i' \in \Omega_i} \frac{s\,(i')}{N_i},\ \frac{\sigma_i^2}{N_i}\right)$$

where Ω_i represents the neighborhood set containing all adjacent districts sharing boundaries with district i, and N_i denotes the total number of districts in Ω_i. Additionally, σ_i^2 is an unknown variance parameter in district i that needs to be estimated[38]. Therefore, by including a MRF smoother in the GADL, it can be developed into a special type of conditional autoregressive model. By using log link function, the exponential of estimated parameters can be interpreted as the relative risk for CVH. Moreover, the spatial function can identify each district as a low or high-risk area for CVH after controlling other factors, as the 95% confidence interval (CI) of the spatial estimate is significantly greater than zero.

We used the Akaike information criterion for the quasi-Poisson model (Q-AIC) as a measure of goodness-of-fit[39] to find the appropriate exposure response association. Based on the minimum Q-AIC, our approach involved utilizing a linear function exposure response function. For the lag-response association, a B-spline function with 2 degrees of freedom was employed, incorporating a maximum lag of 7 days in the creation of the cross-basis.

All statistical analyses were conducted using R software (version 4.3.2) with the "caret" package for missing imputation, "gstat" package for spatial interpolation of $PM_{2.5}$, "dlnm"and "mgcv" packages for modeling, and "ggplot2" for creating figures. The R codes for this study are presented in supplementary file.

Results

Table 1 presents the descriptive statistics for the daily Cardiovascular Hospitalization (CVH) and $PM_{2.5}$ concentrations in Mashhad, covering the period from 2017 to 2020, in addition of the descriptive statistics for daily $PM_{2.5}$ concentrations. A total of 45,089 patients within Mashhad addresses with CVH were selected for study, 22,851 (54.67%) of these patients were male and 22,307 (49.47%) were older than 65 years. During the 4-year study period, the daily mean of CVH and $PM_{2.5}$ concentrations was 31.0 ± 12.0 and 32.70 ± 14.04 μg/m^3, respectively.

Figure 3 presents an overview of the impact of $PM_{2.5}$ on CVH based on the GADL results. The graph in 3-**a** illustrates the RR across various levels of $PM_{2.5}$ concentration and time lags, showing a significant immediate effect of $PM_{2.5}$ on CVH, with a delayed effect for higher $PM_{2.5}$ values. In plot 3-**b**, there is a clear demonstration of a substantial and progressively increasing association between cumulative $PM_{2.5}$ exposures and CVH, highlighting the cumulative impact of $PM_{2.5}$ on CVH and indicating a significant rise in risk with higher $PM_{2.5}$ exposures. The plot 3-**c** shows the lag-response association at a specific $PM_{2.5}$ exposure value of 10, indicating how the relationship between $PM_{2.5}$ exposures and CVH varies with different time lags. The plot 3-**d** displays the overall cumulative association, indicating a significant cumulative effect of a 10-unit increase in $PM_{2.5}$ on CVH for both the current day and a following day.

The result of GADL fitting reveals a significant immediate effect of $PM_{2.5}$ on CVH (RR = 1.019, $CI_{95\%}$: (1.005, 1.033)).

Results of SHGADL model

Table 2. displays district-specific average daily $PM_{2.5}$ concentrations and the rate of CVH per 100,000 of population. The average daily $PM_{2.5}$ concentration varied between 29.8 ± 10.1 and 34.6 ± 13.3, and the highest level was in District 7. Additionally, the average daily CVH varied across districts, ranging from 0.8 to 3.6, and the highest value was in District 13.

Figure 4 illustrates the variation in CVH rates among the 13 districts of Mashhad. Over the course of our four-year study, five districts surpassed the average CVH rate. Three of these districts had particularly high rates, with crude rates ranging from 1000 to 5200 cases per 100,000 of population.

The spatially adjusted effect of $PM_{2.5}$on CVH is depicted in Fig. 5 using the SHGADL model. The exposure-lag-response surface in the plot 5-**a** of this figure shows the lag response at different exposure intensities and the exposure-response at various lags. By integrating spatial heteroscedasticity into the GADL model, the results indicate a significant spatially adjusted association between $PM_{2.5}$ exposures and CVH. The plot 5-**b** the figure illustrates a direct association between cumulative $PM_{2.5}$ exposures and CVH. The plot 5-**c** demonstrates the lag response association at a specific $PM_{2.5}$exposure value of 10 across the lag space. It was found that a 10-unit increase in $PM_{2.5}$ exposures significantly raised the risk of CVH on the same day as well as in the subsequent two days. The plot 5-**d** in the figure shows a significant cumulative effect of a 10-unit increase in $PM_{2.5}$ on CVH for both the current day and the following day. The covariate estimations of SHGADL model are summarized in Table 2 of Appendix S1. As shown in this table the main effects of COVID-19 and PM2.5 were statistically

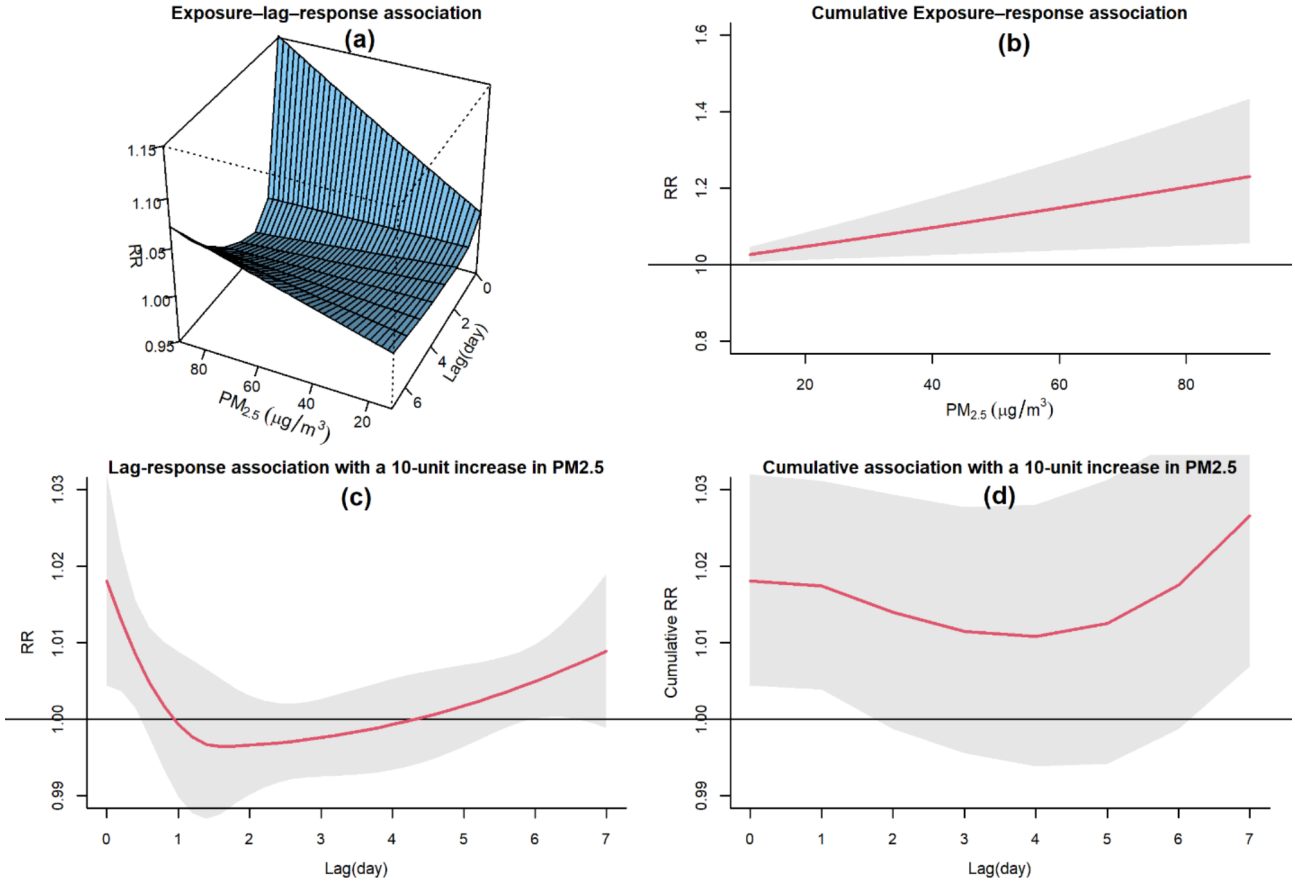

Fig. 3. The association between air pollution exposure and CVH in Mashhad, Iran, from 2017 to 2020. **(a)**: Exposure–lag–response surface risk. **(b)**: cumulative exposure-response association. **(c)**: lag- response association at 10-unit exposure. **(d)**: overall cumulative association.

Districts	$PM_{2.5}$ (μg/m³)		Daily CVH per 100,000 population	
	Mean	SD	Mean	SD
1	31.5	11.6	1.5	1.1
2	31.1	11.7	1.2	0.6
3	30.9	11.6	0.9	0.6
4	30.9	11.6	1.3	0.9
5	32.3	11.9	0.8	0.8
6	33.8	12.5	0.8	0.7
7	33.9	10.7	1.4	1.0
8	30.2	10.3	1.8	1.5
9	31.0	10.9	0.8	0.6
10	30.6	11.4	0.8	0.6
11	31.9	10.7	0.9	0.7
12	29.7	11.5	0.9	1.2
13	30.3	10.4	3.6	6.0

Table 2. Descriptive statistics of daily $PM_{2.5}$ concentration and daily rate of CVH in districts of Mashhad, Iran, during 2017 to 2020.

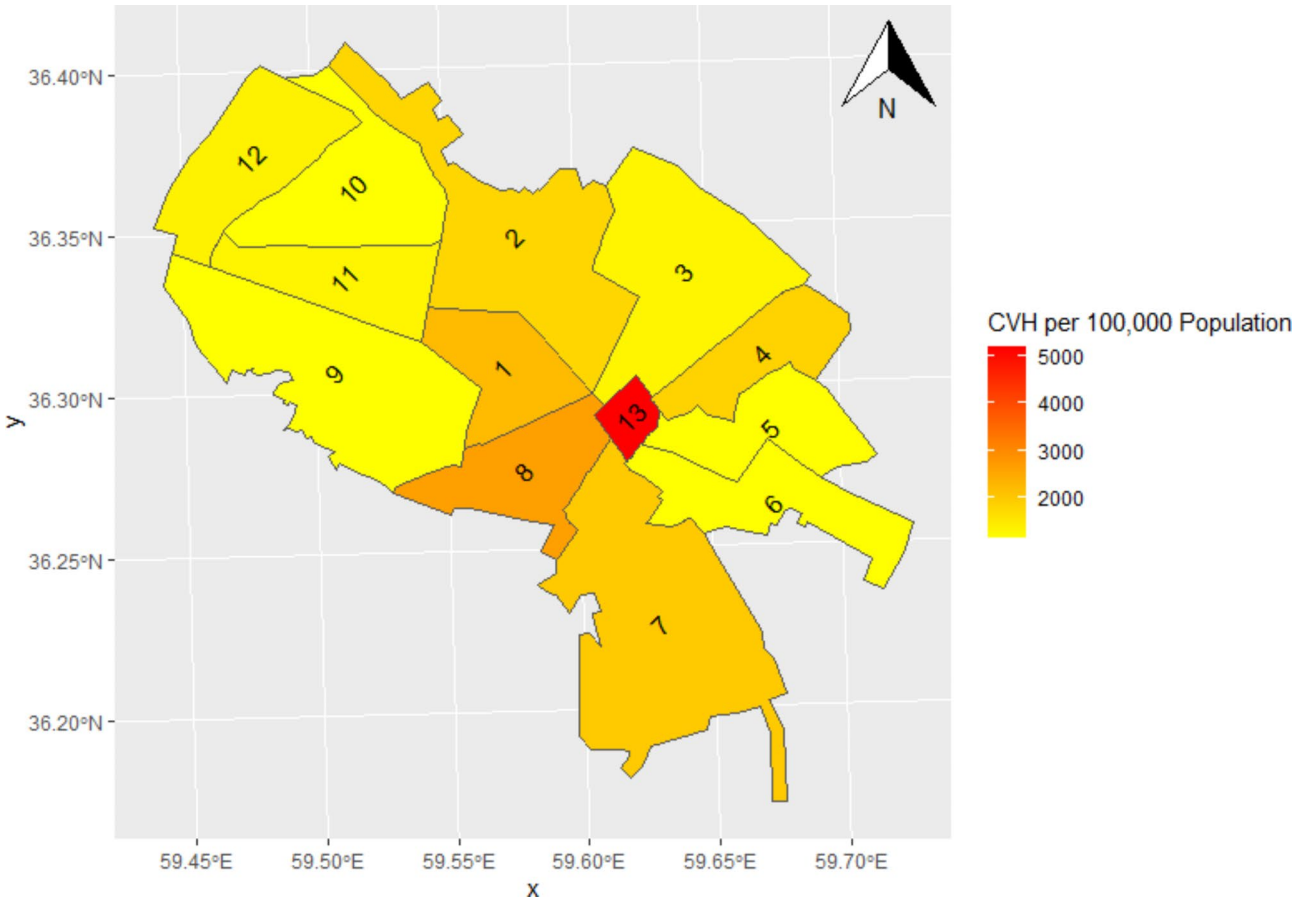

Fig. 4. Geographic distribution of total CVH per 100,000 of population from 2017 to 2020, Mashhad, Iran.

significant, indicating that each had a notable impact on CVH independently. However, the interaction term between COVID-19 and PM2.5 was not statistically significant, demonstrating that the relationship between PM2.5 and CVH did not significantly change during the COVID-19 pandemic.

Table 3 displays the RRs of lagged and cumulative exposure to $PM_{2.5}$ on CVH using the SHGADL model. The results indicate significant positive associations at lag 0 and 1 for the lagged effect. A 10 μg/m^3 increase in $PM_{2.5}$ is shown to significantly elevate the risk of CVH at lag 0 (RR = 1.0102, CI95%= (1.0034, 1.0170)) and lag 1 (RR = 1.0043, CI95% =(1.0009, 1.0078)). Moreover, cumulative exposure to $PM_{2.5}$ is found to significantly impact the risk of CVH at lag 0–3 days post-exposure.

Figure 6 depicts the estimated spatial function, which illustrates the spatial heteroscedasticity of the relative risk of CVH. The map reveals a distinct and disparate pattern within our study area. Specifically, districts in the central and southeastern regions exhibit higher relative risks for CVH than other areas. Among the 13 districts, six have RRs greater than 1. So, when the $PM_{2.5}$ concentrations are elevated, these districts are at a heightened risk for CVH in Mashhad.

Discussion

Air pollutants, particularly $PM_{2.5}$, have been suggested as a potential factor affecting human health and have been identified as a major global public health threat according to the WHO's report[40]. This study utilized a novel SHGADL model to investigate the spatial correlation between the time series of $PM_{2.5}$ concentrations and cardiovascular health in Mashhad from 2017 to 2021. Our approach goes beyond previous research by incorporating spatial variability into the analysis, addressing an important aspect that had not been explored in earlier studies. Initially, we employed a non-spatial GADL model, which revealed a significant link between $PM_{2.5}$ exposures and an increased risk of hospitalizations related to cardiovascular disease on the same day. Specifically, for every 10 μg/m^3 increase in $PM_{2.5}$ concentrations, the relative risk of daily hospitalizations for cardiovascular disease increased by RR = 1.019, CI 95%= (1.005, 1.033).

Our results are consistent with a study conducted by Xu et al., in which they noted a significant increase in CVD hospital admissions (0.30, CI 95%= (0.20, 0.39) for every 10 μg/m^3 rise in $PM_{2.5}$concentration from the previous day to the current (lag 0–1)[41]. Similarly, a study carried out in Wuhan, China, found that a 10 μg/m^3 increase in $PM_{2.5}$(lag 0–2 days) was linked to a 1.23% (CI 95% = (1.01–1.45)) increase in the risk of hospital admissions for CVD[42]. In Iran, a systematic review and meta-analysis showed that for every 10 μg/

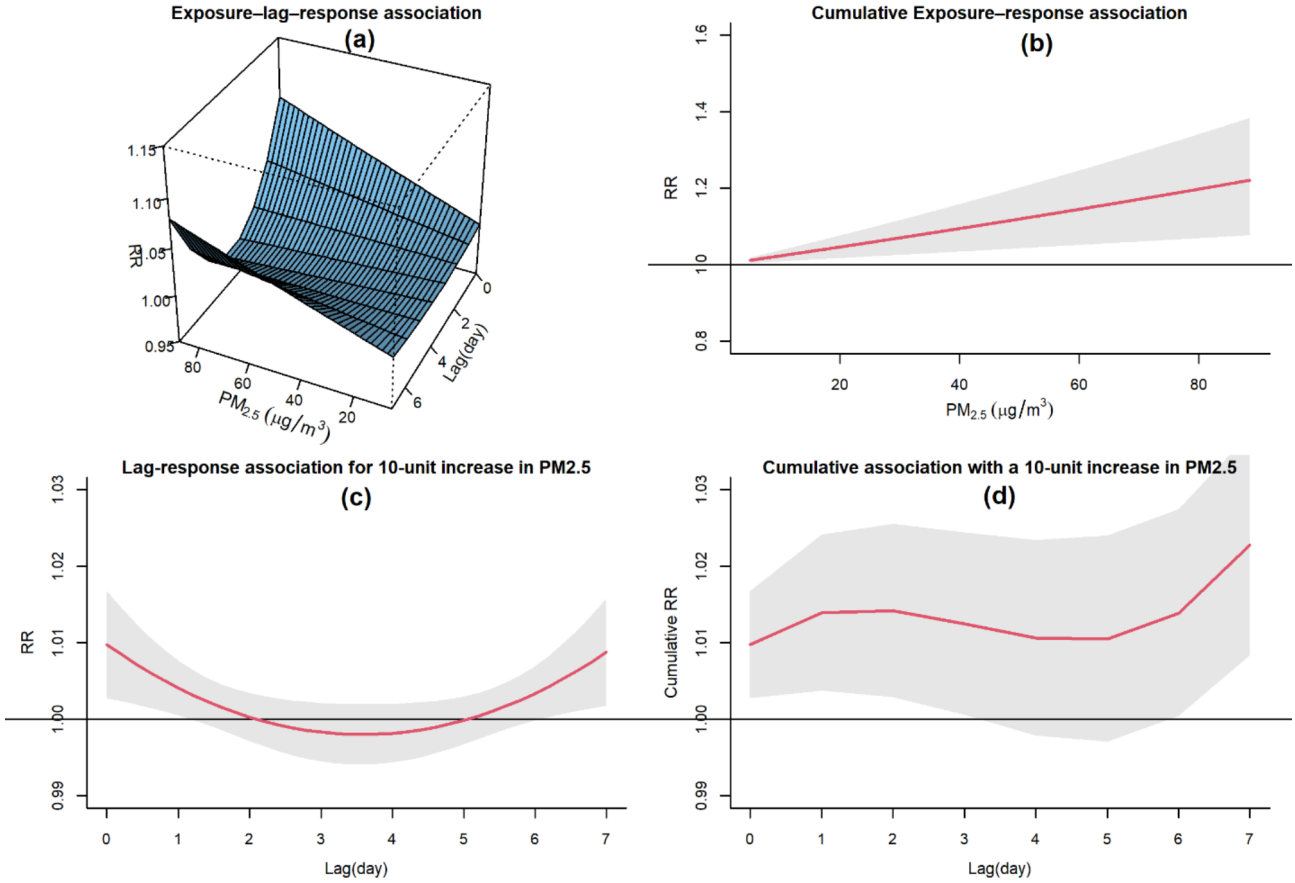

Fig. 5. The spatial adjusted association between air pollution exposure and CVH in Mashhad, Iran, from 2017 to 2020. **(a)**: Exposure–lag–response surface risk. **(b)**: cumulative exposure-response association. **(c)**: lag-response association at 10-unit exposure. **(d)**: overall cumulative association.

Lag	Lagged effect	Cumulative effect
	RR($CI_{95\%}$)	RR($CI_{95\%}$)
0	**1.0102 (1.0034, 1.0170)**	**1.0102 (1.0034, 1.0171)**
1	**1.0043 (1.0009, 1.0078)**	**1.0146 (1.0047, 1.0246)**
2	1.0004 (0.9973, 1.0034)	**1.0149 (1.0040, 1.0260)**
3	0.9982 (0.9945, 1.0020)	**1.0132 (1.0017, 1.0247)**
4	0.9978 (0.9942, 1.0017)	1.0111 (0.9989, 1.0234)
5	0.9996 (0.9965, 1.0026)	1.0107 (0.9979, 1.0236)
6	1.0030 (0.9996, 1.0064)	**1.0137 (1.0010, 1.0266)**
7	1.0083 (1.0015, 1.0152)	**1.0222 (1.0088, 1.0358)**

Table 3. The RRs of exposure to $PM_{2.5}$ on CVH by fitting the SHGADL model.

m^3increase in the combined concentrations of all air pollutants, there is a 0.7% (CI 95% = (0.6–0.9%)) increase in hospitalizations due to cardiovascular problems[43].

However, many studies have presented evidence connecting air pollutants to adverse health effects by assuming uniform exposure within each region or city. This assumption may result in inaccuracies in estimating exposure[44]. By integrating a spatially varying MRF process to account for spatial heteroscedasticity among districts, we were also able to address the spatial autocorrelation between observations. This approach is likely to generate more precise estimates compared to using citywide average concentrations. Our analysis showed that a 10 μg/m^3 increase in $PM_{2.5}$exposures was significantly linked to an increase in CVH at lag 0 and lag 1. Consistent with prior research[45,46], our findings indicate that the effect estimates derived from district-level $PM_{2.5}$ (using SHGADL) were higher than those obtained from averaged $PM_{2.5}$ values across multiple stations (using GADL).

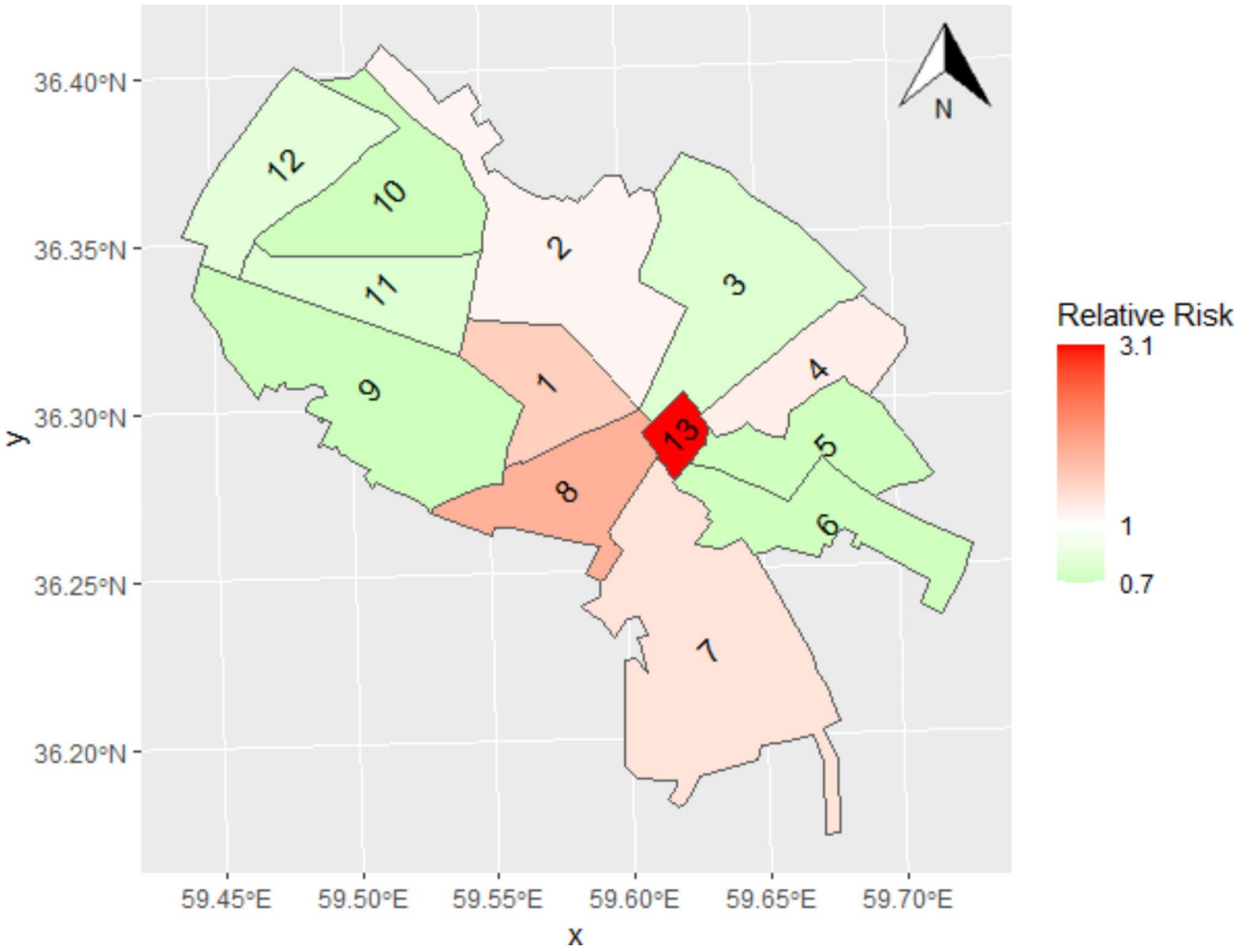

Fig. 6. Spatial heteroscedasticity of the relative risk of CVH in Mashhad, Iran.

This suggests that previous time-series studies relying on average levels may have underestimated the impact of $PM_{2.5}$[45].

Our research reveals that there is a heteroscedasticity in the estimated risk of cardiovascular health issues across the 13 districts of Mashhad, with higher relative risk values observed in the central and southeastern regions compared to other areas. District 13 of Mashhad is the oldest district of the city, which due to the presence of a pilgrimage destination (the Holy shrine of Imam Reza) is a crowded and heavily trafficked area. During certain times of each year, 1 to 3 million pilgrims from different regions of Iran come to Mashhad and stay in the numerous accommodations in District 13. Most of these pilgrims are elderly and are among the cardiovascular hospitalization cases in this district, although they are not part of the district's population. During these special periods, the air pollution in District 13 increases significantly due to heavy vehicle traffic. As a result, the highest risk of cardiovascular hospitalization is observed in this district.

It is important to take into account the spatial heteroscedasticity of ambient air pollutants when identifying populations at high risk. Similar study in Beijing, China investigated the spatial heteroscedasticity in the relationship between particulate matter less than 10 μm (PM_{10}) concentrations and cardiovascular mortality. The results indicated that a 2.46% (CI95%= (1.22, 3.72)) increase in daily cardiovascular mortality was associated with an interquartile rise in PM_{10} concentration. These findings emphasize that the risk of mortality related to air pollutants differs across geographic districts.

To the best of our knowledge, this is the first study to investigate the spatiotemporal association between CVH and air pollution that simultaneously accounts for spatial heteroscedasticity and nonlinear lag effects. Indeed, nonlinear lag analysis and spatial analysis have been treated as separate entities in previous research studies. However, it is worth noting that one study has demonstrated the importance of incorporating a spatial function into the GADL[46]. By doing so, this approach acknowledges and accounts for spatial dependencies, offering a more comprehensive and accurate representation of the complex relationships between variables. The integration of complex spatial structure aspects into the GADL can increase the robustness of the analysis, and capturing additional variances that may be missed when these aspects are not considered.

There are some limitations to this study. First, data for PM10 and other air pollutants were not available, so it was not possible to adjust for these variables. Second, our study only considered exposure within residential

areas and did not take into account important factors influencing individual exposure, such as daily commuting and working in different areas. Third, some factors including personal and socioeconomics information that could affect the effects were not controlled for in this study.

Conclusion

In conclusion, this study shows the complex relationship between $PM_{2.5}$ exposures and CVH. The results show a significant immediate and delayed effect of $PM_{2.5}$ on CVH. Spatial analysis explored distinct patterns, indicating higher relative risks of $PM_{2.5}$ for CVH in the southern and central regions of Mashhad. These significant spatiotemporal relationships provide evidence about the causal effect of $PM_{2.5}$ on CVH. Our findings underscore the need for targeted interventions in specific high-risk areas to mitigate the health impact of air pollution.

The developed SHGADL model contributes to the understanding of the spatial and temporal dynamics of air pollution's impact on cardiovascular health, providing crucial information for public health planning and policy implementation to reduce the burden of CVH in Mashhad and similar urban settings. Moreover, our analysis reveals analytically that considering spatial heteroscedasticity improves the precision of estimating the delayed effect of $PM_{2.5}$ on CVH.

Data availability

The datasets generated and analyzed during the current study are available upon request to the corresponding author upon permission from the deputy of research of Mashhad University of Medical Sciences.

Received: 5 June 2024; Accepted: 25 November 2024
Published online: 26 November 2024

References

1. Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V. & Roth, G. A. The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health. J Am Coll Cardiol [Internet]. ;80(25):2361–71. (2022). https://www.sciencedirect.com/science/article/pii/S0735109722073120
2. Saki, N. et al. Prevalence of cardiovascular diseases and associated factors among adults from southwest Iran: baseline data from Hoveyzeh Cohort Study. *BMC Cardiovasc. Disord.* **22** (1), 309 (2022).
3. Sarrafzadegan, N. & Mohammmadifard, N. Cardiovascular disease in Iran in the last 40 years: prevalence, mortality, morbidity, challenges and strategies for cardiovascular prevention. *Arch. Iran. Med.* **22** (4), 204–210 (2019).
4. Basith, S. et al. The Impact of Fine Particulate Matter 2.5 on the Cardiovascular System: A Review of the Invisible Killer. Vol. 12, Nanomaterials. (2022).
5. Jalali, S. et al. Long-term exposure to PM2.5 and cardiovascular disease incidence and mortality in an Eastern Mediterranean country: findings based on a 15-year cohort study. Environ Heal [Internet]. ;20(1):112. (2021). https://doi.org/10.1186/s12940-021-00797-w
6. Ren, Z. et al. Effect of ambient fine particulates (PM2.5) on hospital admissions for respiratory and cardiovascular diseases in Wuhan, China. Respir Res [Internet]. ;22(1):128. (2021). https://doi.org/10.1186/s12931-021-01731-x
7. Wu, T. et al. Acute effects of fine particulate matter (PM2.5) on hospital admissions for cardiovascular diseases in Lanzhou, China: a time-series study. Environ Sci Eur [Internet]. ;34(1):55. (2022). https://doi.org/10.1186/s12302-022-00634-y
8. Krittanawong, C. et al. PM2.5 and cardiovascular diseases: State-of-the-Art review. Int J Cardiol Cardiovasc Risk Prev [Internet]. ;19:200217. (2023). https://www.sciencedirect.com/science/article/pii/S2772487523000508
9. Krittanawong, C. et al. PM2.5 and Cardiovascular Health Risks. Curr Probl Cardiol [Internet]. ;48(6):101670. (2023). https://www.sciencedirect.com/science/article/pii/S0146280623000877
10. Zanobetti, A. & Schwartz, J. Mortality displacement in the association of ozone with mortality: an analysis of 48 cities in the United States. Am J Respir Crit Care Med. ;177(2):184-9. doi: (2008). https://doi.org/10.1164/rccm.200706-823OC. Epub 2007 Oct 11. PMID: 1793237.
11. Wang, Z. et al. Association between short-term exposure to air pollution and ischemic stroke onset: a time-stratified case-crossover analysis using a distributed lag nonlinear model in Shenzhen, China. Environ Heal [Internet]. ;19(1):1. (2020). https://doi.org/10.1186/s12940-019-0557-4
12. Qiu, X. et al. Inverse probability weighted distributed lag effects of short-term exposure to PM2.5 and ozone on CVD hospitalizations in New England Medicare participants - Exploring the causal effects. Environ Res [Internet]. ;182:109095. (2020). https://www.sciencedirect.com/science/article/pii/S0013935119308916
13. Lubczyńska, M. J., Christophi, C. A. & Lelieveld, J. Heat-related cardiovascular mortality risk in Cyprus: a case-crossover study using a distributed lag non-linear model. *Environ. Health.* **14**, 39. https://doi.org/10.1186/s12940-015-0025-8 (2015). PMID: 25930213; PMCID: PMC4.
14. Hadianfar, A., Rastaghi, S., Tabesh, H. & Saki, A. Application of distributed lag models and spatial analysis for comparing the performance of the COVID-19 control decisions in European countries. *Sci. Rep.* **13** (1), 17466. https://doi.org/10.1038/s41598-023-44830-z (2023).
15. Aßenmacher, M., Kaiser, J. C., Zaballa, I., Gasparrini, A. & Küchenhoff, H. Exposure-lag-response associations between lung cancer mortality and radon exposure in German uranium miners. *Radiat. Environ. Biophys.* **58** (3), 321–336. https://doi.org/10.1007/s00411-019-00800-6 (2019).
16. Gasparrini, A., Armstrong, B. & Kenward, M. G. Distributed lag non-linear models. *Stat. Med.* **29** (21), 2224–2234. https://doi.org/10.1002/sim.3940 (2010). PMID: 20812303; PMCID: PMC2998707.
17. Kim, Y. M., Park, J. W. & Cheong, H. K. Estimated effect of climatic variables on the transmission of Plasmodium Vivax malaria in the Republic of Korea. *Environ. Health Perspect.* **120** (9), 1314–1319. https://doi.org/10.1289/ehp.1104577 (2012). Epub 2012 Jun 18. PMID: 22711788; PMCID.
18. Stanišić Stojić, S., Stanišić, N. & Stojić, A. Temperature-related mortality estimates after accounting for the cumulative effects of air pollution in an urban area. *Environ. Health.* **15**, 73. https://doi.org/10.1186/s12940-016-0164-6 (2016).
19. Yang, J. et al. Daily temperature and mortality: a study of distributed lag non-linear effect and effect modification in Guangzhou. *Environ. Health.* **11**, 63. https://doi.org/10.1186/1476-069X-11-63 (2012).
20. Tian, Y. et al. Association between ambient fine particulate pollution and hospital admissions for cause specific cardiovascular disease: time series study in 184 major Chinese cities. BMJ [Internet]. ;367:l6572. (2019). http://www.bmj.com/content/367/bmj.l6572.abstract
21. Wright, N. et al. Long-term ambient air pollution exposure and cardio-respiratory disease in China: findings from a prospective cohort study. Environ Heal [Internet]. ;22(1):30. (2023). https://doi.org/10.1186/s12940-023-00978-9

22. Song, R. et al. Spatial variations in urban air pollution: impacts of diesel bus traffic and restaurant cooking at small scales. Air Qual Atmos Heal [Internet]. ;14(12):2059–72. (2021). https://doi.org/10.1007/s11869-021-01078-8
23. Dias, D. & Tchepel, O. Spatial and Temporal Dynamics in Air Pollution exposure Assessment. *Int. J. Environ. Res. Public Health*. **15** (3), 558. https://doi.org/10.3390/ijerph15030558 (2018).
24. Esmaili, R. & Amini, F. L. Identification of hot spots PM2. 5 in Mashhad air pollution. *J. Clim. Res.* **1399** (44), 63–78 (2021).
25. Statistical Center of Iran. Official report of statistical survey of population in Mashhad city archived by the Statistical Center of Iran. 2018–2019. https://www.amar.org.ir/english/Iran-Statistical-Yearbook/Statistical-Yearbook-2018-2019. Accessed 19 Ju.
26. Rizvi, S. T. H., Latif, M. Y., Amin, M. S., Telmoudi, A. J. & Shah, N. A. Analysis of Machine Learning Based Imputation of Missing Data. Cybern Syst [Internet].:1–15. https://doi.org/10.1080/01969722.2023.2247257
27. Li, J. et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. BMC Med Res Methodol [Internet]. ;24(1):41. (2024). https://doi.org/10.1186/s12874-024-02173-x
28. Emmanuel, T. et al. A survey on missing data in machine learning. J Big Data [Internet]. ;8(1):140. (2021). https://doi.org/10.1186/s40537-021-00516-9
29. Ou, H., Yao, Y. & He, Y. Missing Data Imputation Method Combining Random Forest and Generative Adversarial Imputation Network. *Sens. (Basel)* ;**24**(4):1112. doi: https://doi.org/10.3390/s24041112. (2024). PMID: 38400270; PMCID: PMC10893362.
30. Holloway-Brown, J., Helmstedt, K. J. & Mengersen, K. L. Stochastic spatial random forest (SS-RF) for interpolating probabilities of missing land cover data. J Big Data [Internet]. ;7(1):55. (2020). https://doi.org/10.1186/s40537-020-00331-8
31. Lu, G. Y. & Wong, D. W. An adaptive inverse-distance weighting spatial interpolation technique. *Comput. Geosci.* **34** (9), 1044–1055 (2008).
32. Halek, F. & Kavousi-rahim, A. GIS assessment of the PM10, PM2.5 and PM10 concentrations in urban area of Tehran in warm and cold seasons. *Int. Arch. Photogramm Remote Sens. Spat. Inf. Sci.* **XL-2/W3**, 141–149 (2014).
33. Li, L., Losser, T., Yorke, C. & Piltner, R. Fast Inverse Distance weighting-based spatiotemporal interpolation: a web-based application of Interpolating Daily Fine Particulate Matter PM2.5 in the contiguous U.S. using parallel programming and k-d tree. *Int. J. Environ. Res. Public. Heal.* **11**, 9101–9141 (2014).
34. Kan, H. et al. Traffic exposure and lung function in adults: the atherosclerosis risk in communities study. *Thorax* **62** (10), 873–879 (2007).
35. Lipsett, M. J. et al. Long-term exposure to air pollution and cardiorespiratory disease in the California teachers study cohort. *Am. J. Respir Crit. Care Med.* **184** (7), 828–835 (2011).
36. Wong, D. W., Yuan, L. & Perlin, S. A. Comparison of spatial interpolation methods for the estimation of air quality data. J Expo Sci Environ Epidemiol. ; 14(5):404–15. (2004). https://doi.org/10.1038/sj.jea.7500338 PMID: 15361900.
37. Li, Z. & Wang, P. Intelligent Optimization on Power Values for Inverse Distance Weighting. In: 2013 International Conference on Information Science and Cloud Computing Companion. Guangzhou: IEEE; 370–5. (2013). p https://doi.org/10.1109/ISCC-C.2013.81
38. Fahrmeir, L., Lang, S. & Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. J R Stat Soc Ser C (Applied Stat [Internet]. ;50(2):201–20. (2001). http://www.jstor.org/stable/2680887
39. Peng, R. D., Dominici, F. & Louis, T. A. Model choice in time series studies of air pollution and mortality. *J. R Stat. Soc. Ser. Stat. Soc.* **169** (2), 179–203 (2006).
40. WHO. WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide: Global Update 2005. Summary of Risk Assessment. Available online: (2006). http://apps.who.int/iris/bitstream/10665/69477/1/WHO_SDE_PHE_OEH_06.02_eng.pdf (ac.
41. Amsalu, E. et al. Acute effects of fine particulate matter (PM2.5) on hospital admissions for cardiovascular disease in Beijing, China: a time-series study. Environ Heal [Internet]. ;18(1):70. (2019). https://doi.org/10.1186/s12940-019-0506-2
42. Ren, Z. et al. Effect of ambient fine particulates (PM2.5) on hospital admissions for respiratory and cardiovascular diseases in Wuhan, China. *Respir Res.* **22** (1), 128. https://doi.org/10.1186/s1293 (2021).
43. Karimi, B., shokrinezhad, B. & Samadi, S. Mortality and hospitalizations due to cardiovascular and respiratory diseases associated with air pollution in Iran: A systematic review and meta-analysis. Atmos Environ [Internet]. ;198:438–47. (2019). https://www.sciencedirect.com/science/article/pii/S1352231018307611
44. Li, W. et al. The spatial variation in the effects of air pollution on cardiovascular mortality in Beijing, China. J Expo Sci Environ Epidemiol. ;28(3):297–304. doi: (2018). https://doi.org/10.1038/jes.2016.21. PMID: 29666509.
45. Xu, M. et al. Spatiotemporal analysis of particulate air pollution and ischemic heart disease mortality in Beijing, China. *Environ. Health*. **13**, 109. https://doi.org/10.1186/1476-069X-13-109 (2014).
46. Zhang, Y. et al. The spatial characteristics of ambient particulate matter and daily mortality in the urban area of Beijing, China. *Sci. Total Environ.* **435–436**, 14–20. https://doi.org/10.1016/j.scitotenv.2012.06.092 (2012).

Acknowledgements

This paper is extracted from Ali Hadianfar's Ph.D. thesis (research code: 4001789) supervised by Dr. Azadeh Saki at the Department of Epidemiology and Biostatistics of Mashhad University of Medical Sciences. The authors wish to acknowledge Omaima Mossadeq, Cosima Froehner, Shmuel Leibel Strey, and Michael Speckbacher, statistics students at LMU, for their contributions to missing data imputation.

Author contributions

H. A: Conducted statistical analysis and manuscript writing. K. H: Provided scientific advice on data analysis, contributed to manuscript development. M.S: Participated in data collection, data pre-processing, and execution of geocoding tasks. S.A: Conceptualized the research, supervised the project, and scientifically revised and editing the manuscript. All authors read and approved the final manuscript.

Funding

None.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

The study protocol was approved by the Ethics Committee of Mashhad University of Medical Science with

number: IR.MUMS.FHMPM.REC.1401.026 and the informed consent was obtained from all subjects and/or their legal guardian(s).

Consent for publication

All authors approved the manuscript and this submission.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-81036-3.

Correspondence and requests for materials should be addressed to A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.