Statistical Internship

Air Pollution and Cardiovascular Disease Hospitalizations in Mashhad, Iran

Department of Statistics
Ludwig-Maximilians-Universität München

Cosima Fröhner, Omaima Mossadeq, Michael Speckbacher, Shmuel Strey

Munich, May 7th, 2024

A report for
Ali Hadianfar, Visiting Scholar at Ludwig-Maximilians-Universität München

Supervised by
Johannes Piller and Prof. Dr. Helmut Küchenhoff

Abstract

This report investigates the association between air pollution, specifically particulate matter measuring less than 2.5 micrometers (PM2.5), and cardiovascular disease related hospitalizations (CVD) in Mashhad, Iran. Data from 22 air quality measurement stations and district-level records spanning the years 2017 to 2020 were used. The study employs random forest imputation to address missing PM2.5 values and subsequent spatial aggregation of the PM2.5 values. For the analysis, the study employs generalized additive models with Poisson regression to investigate both the immediate and delayed effects of PM2.5 exposure on CVD. The findings indicate no significant association between PM2.5 levels and CVD, regardless of whether the exposure is immediate or lagged by up to five days. This conclusion holds across robustness checks, including changes to the flexibility of time-splines, examination for non-linear effect of PM2.5, and capping extreme PM2.5 values. Additionally, the study reveals spatial variation, which, while observed, does not reach statistical significance. Furthermore, subgroup analyses does not demonstrate significant impacts of PM2.5 on CVD of different age groups. Limitations include potential dependencies of the results on the chosen approach for imputation and aggregation of PM2.5. The need for more detailed toxicological data, and for external environmental variables support further data collection and analysis building on the results of this report.

Contents

1 Introduction

PM2.5, short for particulate matter with a diameter of less than 2.5 micrometers, is a leading cause for cardiovascular diseases, which themselves are major drivers for mortality and hospitalizations worldwide (Thangavel et al., 2022; World Health Organization, 2021). With PM2.5 stemming from natural sources, such as windblown dust or wildfires, as well as human-activities, including household heating and traffic, especially densely populated cities in the country of Iran have repeatedly measured high levels of PM2.5 which exceeded recommended air pollution limits (Faridi et al., 2022). Against this background, different studies have been concerned with the investigation of the association of PM2.5 and cardiovascular diseases in Iran (e.g., Hadei et al., 2017; Khaniabadi et al., 2018; Leili et al., 2021).

In the following analysis, we build on this research, focusing on the investigation of a possible short-term effect of PM2.5 on cardiovascular disease related hospitalizations (CVD) in the city of Mashhad, Iran's second-most populous city with 3,372,660 inhabitants according to the census from 2016 (Statistical Centre of Iran, 2016). Mashhad is split into 13 districts which are taken into account by investigating the spatial variation in the relationship of PM2.5 and CVD. Additionally, the different assumptions are tweaked to check whether the results are robust. Specifically, this yields the following research questions - as posed by our project partner and visiting scholar at the Institute of Statistics of LMU Munich, Ali Hadianfar:

1. Is there an association between PM2.5 and cardiovascular disease hospitalizations in Mashhad, Iran?

2. Is there a spatial variation in the relationship?

3. Is the relationship robust?

To this end, the available data is first introduced, followed by a description of the two steps of data preprocessing, namely missing value imputation and spatial aggregation. This is followed by a justification of the modeling approach and the results regarding the main research questions. Afterwards, a robustness analysis is conducted, and the spatial variation assessed, ending with an overview of the results regarding the research questions and the limitations of this study.

2 Data

The following section begins with an introduction to the available data, details the preprocessing of the data, comprising missing value imputation and spatial aggregation of PM2.5 values, and ends with a short description of the resulting PM2.5 data.

2.1 Available Datasets

The following analysis makes use of two datasets provided by our project partner. The first dataset refers to the 22 Air Quality Measurement Stations (AQM-Stations) in Mashhad. Their locations of the stations are indicated on the map in Figure 1 with red dots. The respective dataset contains the X- and Y-coordinates of each station, along with daily average PM2.5 values from January 1, 2017, to December 31, 2020. Due to operational disruptions at the AQM-Stations, approximately one third of the daily mean PM2.5 values are missing. Therefore the data was preprocessed via missing value imputation as described in detail in Chapter 2.2.1.

The second dataset refers to the 13 districts of Mashhad which are outlined in Figure 1. For each district, daily data for the same period (excluding March 20, 2020, which is missing) include: date, official ID of district, number of cardiovascular disease hospitalizations per district per day, age-specific cardiovascular disease hospitalizations (number of hospitalized patients of the age 65 and older and number of hospitalized patients younger than 65), X- and Y-coordinates of the district center, day of the week, public holiday indicator, a covid-19 pandemic indicator, and a screening indicator for a state-mandated cardiovascular disease health screening program for the asymptomatic population. Additional variables include population, percentage of population 65 and older, and percentage of illiterate and unemployed residents per district, with these demographics remaining constant over time.

To explore potential delayed effects of PM2.5 exposure, variables for single lag and cumulative lag (ranging from lag 0 to 5 days) were calculated. Single lag variables capture the effect of PM2.5 on individual days, while cumulative lag variables represent the average exposure over the current and several preceding days. A visualization of the computation of single lag variables can be found in Appendix A. Cumulative lags are defined from cumulative lag 1 to 5, with each number indicating the final day included in the average. For example, cumulative lag 2 is calculated as follows:

$$\text{PM2.5\_cum2} = \frac{\text{PM2.5\_lag0} + \text{PM2.5\_lag1} + \text{PM2.5\_lag2}}{3}. \tag{1}$$

Given the high percentage of missing data in PM2.5 values, the following steps of preprocessing and subsequent data analysis were taken to investigate the research questions from above:

1. Imputation of missing PM2.5 values
2. Spatial aggregation of PM2.5 values on district level
3. Modeling of CVD
4. Robustness of results.

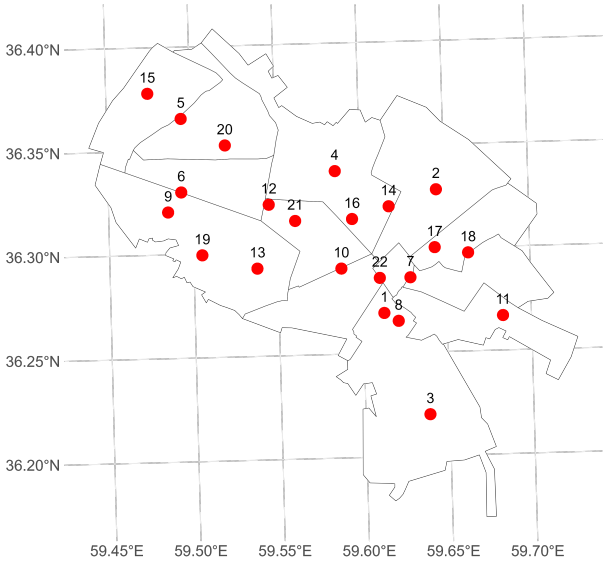


Figure 1: Map of Mashhad with AQM-Stations and District Borders

2.2 Data Preprocessing

Against the background of missing values in PM2.5, the respective data was preprocessed via missing value imputation of PM2.5 station values and subsequent aggregation to district level.

2.2.1 Missing Value Imputation of PM2.5

The following paragraph motivates the missing value imputation for the PM2.5 values of the stations, describes the chosen two-step approach for imputation, evaluates this approach against mean imputation and finally describes the full PM2.5 values of the stations which is used in the following aggregation and modeling.

The extent of missing data is visualized in Figure 2 where each column corresponds to an AQM-station and each row to a specific date. Missing values are highlighted in black. Across stations and dates the PM2.5 data from AQM-stations contain 30.9% of missing values. As this data does not contain any fully complete column (i.e., station) nor complete row (i.e., date), a full-case-analysis is unsuitable for the given PM2.5 data and thus imputation is required.

To this end, imputation based on Random Forest (RF) prediction was conducted. Employing RF for imputation allows to account for relevant challenges in the data. Firstly, it can be assumed that the missing PM2.5 values are non-linear. Therefore, simple imputation methods such as mean imputation or linear interpolation on their own may not be sufficient (Shah et al., 2014). The RF is able to account for this non-linearity. Secondly, there may be temporal and inter-station dependencies that should both be taken into account for imputation. This is especially important as the the missing data gaps are relatively large (Wu et al., 2022). Regarding inter-station similarities, it is important to consider that geographical proximity may not necessarily imply similar PM2.5 levels.

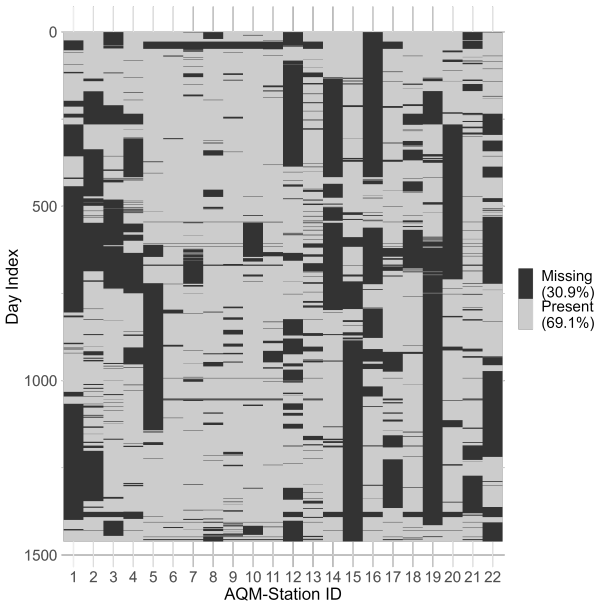


Figure 2: Missing PM2.5 Values

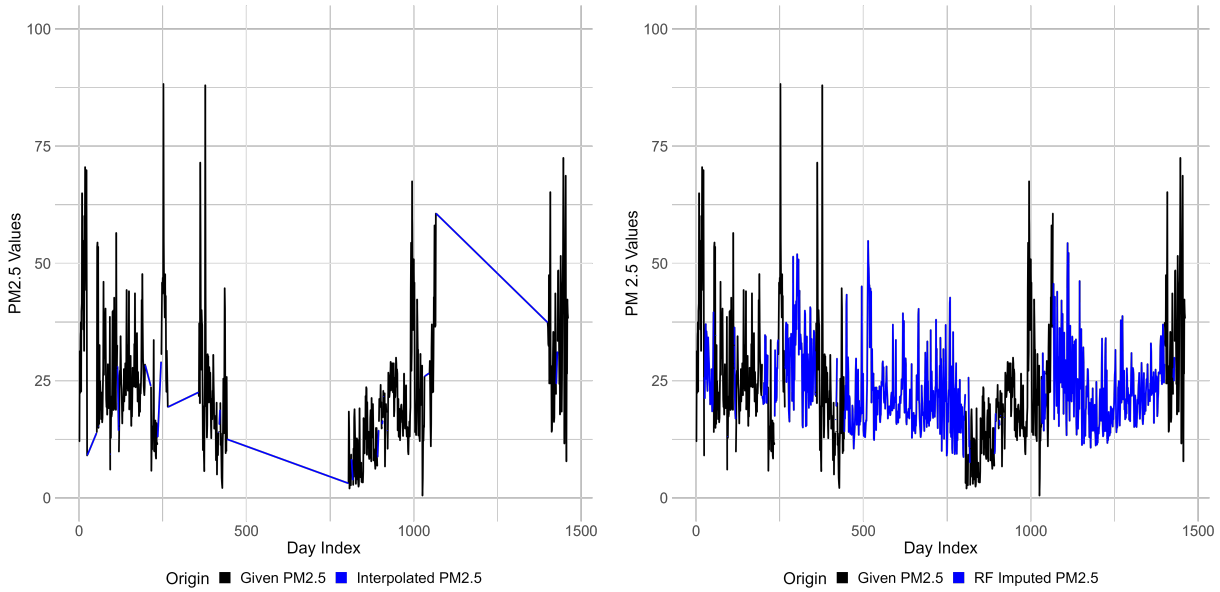
For instance, environmental barriers such as hills between relatively proximate stations may lead to different PM2.5 values of these stations, while two relatively distant stations may be exposed to similar traffic density, leading to similar PM2.5 values. The RF is able to learn and thus account for these different dependencies. In total, RF seems like a reasonable choice. The algorithm was employed in adherence to Jing et al., 2022 and White et al., 2011 as follows:

Step 1. Pre-impute the missing values using linear-interpolated values of the stations acting as a placeholder;
Step 2. Fill missing values with the placeholder data;
Step 3. Train the RF using date and PM2.5 values from other stations as inputs and predict the target stations' missing values;
Step 4. Employ the predicted values of the RF to fill the target stations' missing values;
Step 5. Select the date, fully-filled station values and other stations as the input for the next target station to train a RF;
Step 6. Repeat Step 3 to Step 5 until all variables are fully filled.

As an example, a visualization of the imputation for Station 1 after Step 2 (Figure 3a) and after Step 4 (Figure 3b) indicate suitability of the approach. To evaluate the accuracy, the evaluation criterion of root mean square error (RMSE; Smith et al., 2003) was computed and afterwards compared against the results for mean imputation in a 80/20 train test split per station. The RMSE can be calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{2}$$

where y_i is the true PM2.5 value, $\hat{y}_i$ is the imputed PM2.5 value and n is the number of cases used for testing, which here is 20% of all available cases of the respective station. Comparing our approach to simple mean imputation, in our analysis, the RF-based method demonstrates a lower mean RMSE compared to the mean imputation method. We also note that the difference in accuracy varies between stations, which may be due to the order in which imputation is conducted, or due to differences in variance in PM2.5 values per station (Jing et al., 2022). Overall, the RF method is appropriate for the given use case and effectively addresses the challenges within the data. The visualizations of the imputed data using this method appear reasonable, and it achieves higher accuracy, as indicated by the lower RMSE, compared to the simpler method of mean imputation. Looking at the resulting PM2.5 values, Figure 6a shows the map of Mashhad with its different AQM-stations. The colour of the stations represents the mean PM2.5 values per station after imputation. Some spatial variation is observable: for example, stations in the south-east exhibit comparably higher mean PM2.5 values.



(a) Linear Interpolation (b) Random Forest Imputation

Figure 3: Time-Series of PM2.5 Values at Station 1

2.2.2 Spatial Aggregation of PM2.5

As the target variable CVD is aggregated on district level, PM2.5 also needs to be aggregated on district level. As the stations are not uniformly distributed across the districts it is inappropriate to only consider values of the stations located inside the borders of each district as aggregated PM2.5 value of a district. Also taking into account the information from stations outside the district, aggregation was conducted in three steps. Firstly, a grid was placed over Mashhad, as illustrated in Figure 4. In the real calculations a fine grid of 100 times 100 meters was used in order to have a sufficient number of grid points in each district. Secondly, the value of each grid point was calculated for every day based on the PM2.5 values of each station on that day. For this calculation, the assumption

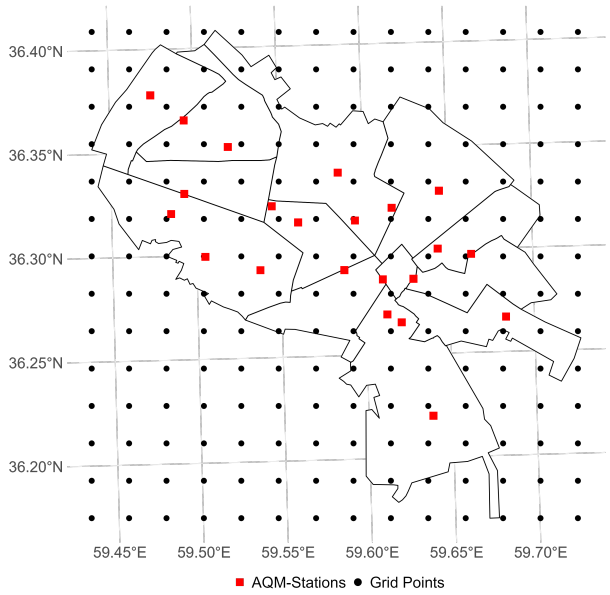


Figure 4: Map of Mashhad with AQM-Stations and Grid Points (2000m x 2000m)

was made that the influence of stations decreases exponentially with distance. Hence, the weight for each station can be calculated using this formula:

$$weight = \exp(-\text{decay\_parameter} \times \text{distance}). \tag{3}$$

The results heavily dependant on the decay parameter, thus the parameter was estimated. To do this, one station was removed at a time, and a loss function based on Mean Squared Error (MSE) was optimized. This process yielded the decay parameters with the lowest MSE for each station and day combination. The median of these estimated decay parameters was calculated, resulting in a decay parameter of approximately 0.00074, which was then used for the computation of the final weights. The weights for each grid point and station are depicted in Figure 5. The x-axis illustrates the distance from a grid point to each station, while the line represents the corresponding weight of the respective station. These weights needed to be normalized across all 22 stations to ensure a total weight of 1. Afterwards, the PM2.5 values of the stations were multiplied by these weights. This process returned daily PM2.5 values for each grid point. Thirdly, the grid points located within a district were determined and the average PM2.5 value of the grid points was computed for each district. This results in PM2.5 values for each day and district as required for modeling CVD. Figure 6b visualizes the mean aggregated PM2.5-value for each district. As expected, when comparing the aggregated district values with the imputed station values, we observe a general correspondence; for example, in both cases, PM2.5 values are higher in the south-east (see Figure 6). Figure 7 visualizes the distribution of the final PM2.5 values. This histograms shows a right-skewed distribution with the majority of day-district combinations in Mashhad exceeding the World Health Organization's recommended maximum of 15 micrograms per cubic meter for PM2.5 (World Health Organization, 2021). This supports that PM2.5 levels are of concern in Mashhad, as safe air quality thresholds are often surpassed.

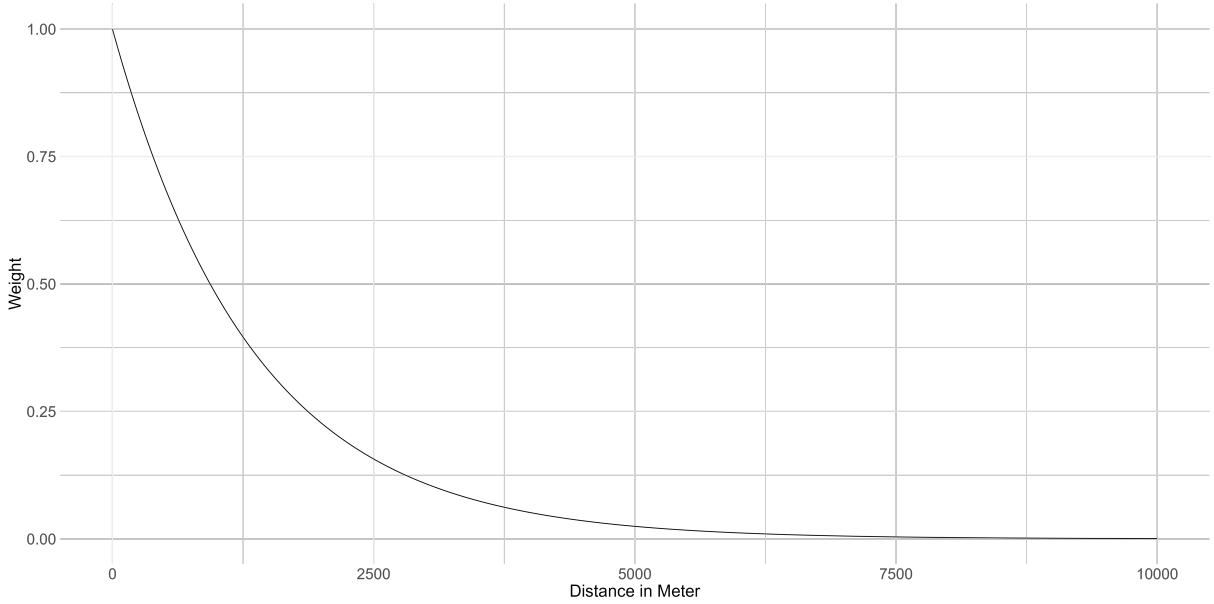


Figure 5: Visualization of Station Weights with Distance

3 Model

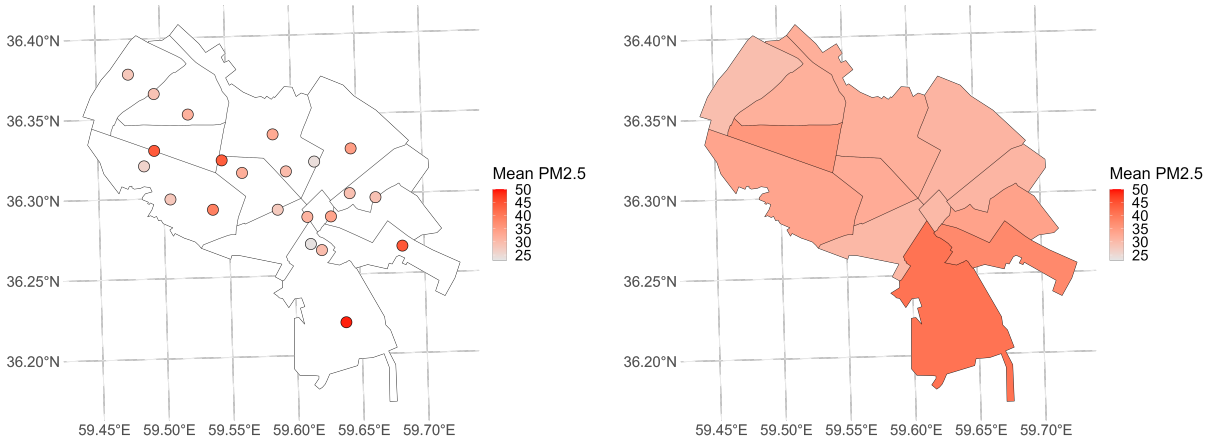
With PM2.5 values on the district level at hand, the following section introduces the model that will be used to investigate the research questions of interest.

3.1 Theoretical Background

To address the main research questions, generalized additive models (GAMs) with Poisson regression were employed. According to Biggeri (2004), the use of GAMs is widely established in analyzing short-term effects of air pollution on health. The modeling approach requires the dependent variable to consist of non-negative count values. Another assumption of the Poisson model is that the conditioned mean of the distribution equals the conditioned variance, expressed as $\mathbb{E}(Y|x) = Var(Y|x)$, indicating no overdispersion. This is checked by fitting a quasi-Poisson model and checking the overdispersion parameter. A parameter of 1 indicates no overdispersion. Additionally, the independence of observations of the target variable is required. This assumption is checked by examining Partial Autocorrelation Function (PACF) plots, where minimal values indicate a lack of autocorrelation. The visual inspection of quantile residuals against fitted values and quantile residuals against index serves to ensure a lack of patterns. Further checking the Q-Q plot serves to ensure that quantile residuals are normally distributed. The latter diagnostics thus further ensure overall model fit, assessing if the model is appropriate for the data.

Using Poisson regression, the expected values are calculated as follows:

$$\mathbb{E}(Y_i) = \exp\left(\beta_0 + \beta_1 \cdot \text{variable}_{i,1} + \beta_2 \cdot \text{variable}_{i,2} + \ldots + \beta_k \cdot \text{variable}_{i,k}\right), \quad i \in \{1, 2, \ldots, n\}, \quad k \in \{1, 2, \ldots, k\}. \tag{4}$$



(a) Mean PM2.5 Values per AQM-Station (b) Mean PM2.5 Values per District

Figure 6: Mean PM2.5 Values after Imputation and after Aggregation

A β-coefficient of 0 results in no effect on the dependant variable since $e^0 = 1$. A positive β-coefficient indicates an increase in the estimated dependant variable because $e^\beta > 1$ when $\beta > 0$. Conversely, a negative β-coefficient leads to a decrease since $e^\beta < 1$ for $\beta < 0$.

3.2 Formula and Covariables

To assess the association between PM2.5 and CVD, a model is run separately for each lag. The model formula is defined as follows:

$$\begin{aligned}\mathbb{E}(CVD_i) = \exp\big(&\beta_0 + \beta_{1,j}\ \text{PM2.5\_lag}_{i,j} + \beta_2\ \text{factor}(\text{districtID}_i) + \\ &\beta_3 \text{factor}(\text{DOW}_i) + \text{ns}(\text{date}_i, 20) + \beta_4\ \text{covid}_i + \\ &\beta_5\ \text{screening}_i + \beta_6\ \text{holiday}_i + \text{offset}(\log(\text{population}_i))\big),\end{aligned} \tag{5}$$

$$i \in \{0, 1, \ldots, n\}, \quad j \in \{lag0, lag1, \ldots, lag5, cum\_lag1, \ldots, cum\_lag5\},$$

where $\mathbb{E}(CVD)$ denotes the expected daily count of cardiovascular disease hospitalizations in one district; β_0 represents the constant term; $PM2.5\_lag_j$ refers to the different lag terms. To prevent multicollinearity, only one lag term is included per model; *districtID* is a factor with 13 levels; *DOW* is a factor with 7 levels; ns(*date*, df) is a natural cubic spline function to control for potential nonlinear confounding effects of time, with 20 degrees of freedom (df), equivalent to 5 df per year, as used by Wu et al., 2022 and Kloog et al., 2013. Furthermore, *covid*, *screening* and *holiday* are included as dummy variables and log(*population*) is used as an offset to account for population-differences across districts. The coefficients for these variables are represented by the beta coefficients.

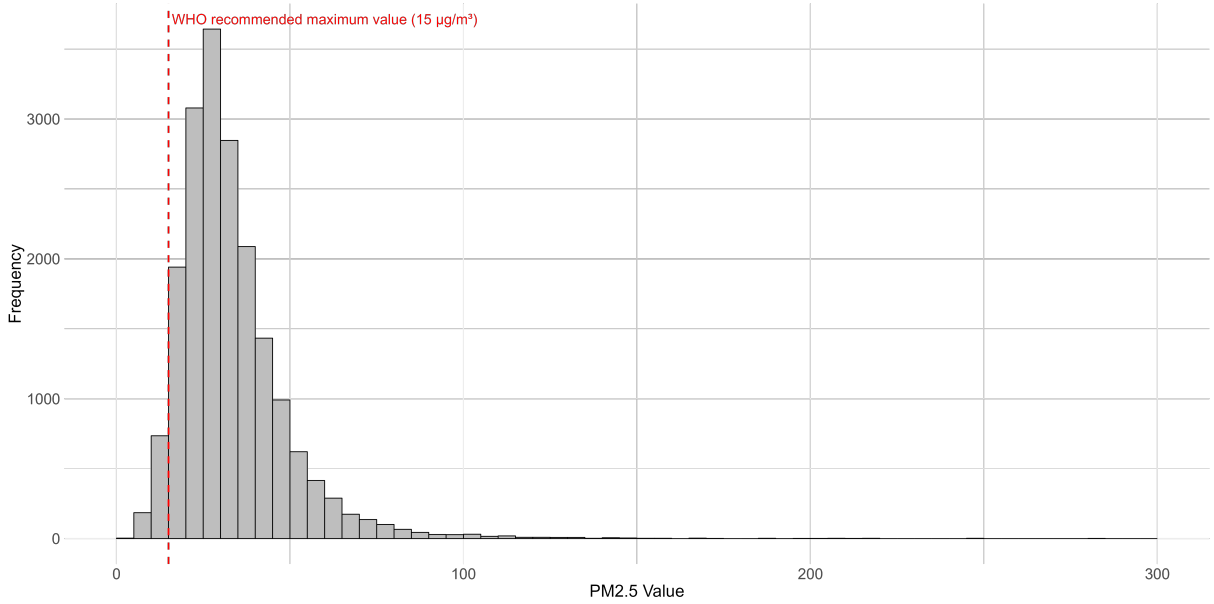


Figure 7: Histogram of imputed and aggregated PM2.5 Values

Before continuing with the model, a short descriptive analysis of the confounders in the data serves to justify their inclusion beyond literature. Appendix B gives an overview of the differences in *CVD* between the different levels of the confounders *districtID*, *DOW*, *covid*, *screening*, and *holiday*. Firstly, *districtID* is included as districts show differences in mean *CVD* value. For example, district 13 has a relatively low mean *CVD* value, whereas district 2 has a comparably high mean *CVD* value, emphasizing the need to include district as factor variable to account for geographical differences in *CVD*. Secondly, *DOW* is included as differences between the weekdays are notable. For example, Friday shows lower mean *CVD* values compared to Saturday. This variability highlights the importance of considering *DOW* as a confounding factor. Thirdly and fourthly, *covid* is included because the number of *CVD* cases is lower for its respective time period and *screening* is included because the number of *CVD* cases cases is higher for its respective time period, motivating to account for their impact on *CVD* in the model. Fifthly, the variable *holiday* is included in the model. The mean *CVD* cases for holidays are lower compared to non-holidays thus supporting the inclusion of *holiday* as a binary variable in the model to control for their impact on *CVD*. Lastly, *date* is included in the model "to control for those unobserved confounders that could have a systematic temporal behavior" (Biggeri, 2004, p. 58). To further justify the inclusion, a time series plot shows the monthly aggregated *CVD* cases over the four-year period of the study (see Appendix C). Events, such as the duration of the screening program and the occurrence of the COVID-19 pandemic, are marked on the timeline. The plot indicates temporal trends for these events, further supporting the inclusion of not only a date variable but also *covid* and *screening* in the model to account for their potential impacts.

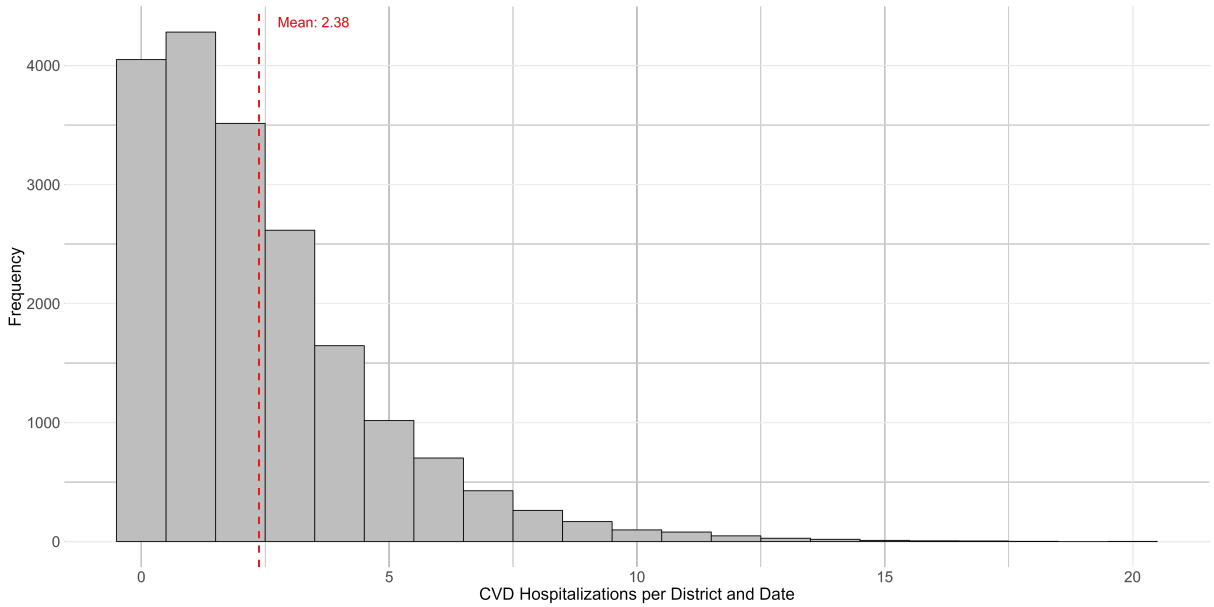


Figure 8: Histogram of *CVD* with Mean Value

3.3 Model Fit

In accordance with the assumptions of Poisson that were described in Section 3.1, *CVD* is a count variable with non-negative values, as visualized in Figure 8. Fitting the quasi-Poisson model gave an overdispersion parameter of 1.06. Therefore, overdispersion is negligible and the according assumption that the conditioned variance equals the conditioned expected value can be considered fulfilled. The PACF plots for each lag-term show negligible violations of the independence assumption. The plots for the model with *PM2.5_lag0* are given in the Appendix D for districts 1 to 6. With very similar results in PACF plots of all districts and across different models (i.e., with different lag-variables), this assumption can also be considered met. Finally, Figure 9 shows diagnostics plots for the model with *PM2.5_lag0*. There are no obvious patterns in the plots (first row) and quantile residuals are approximately normally distributed (second row), indicating a very good model fit for the model with *PM2.5_lag0* overall. Given the minor changes between the different models, this good model fit holds for each lag-model, allowing to continue with the investigation of the main research questions.

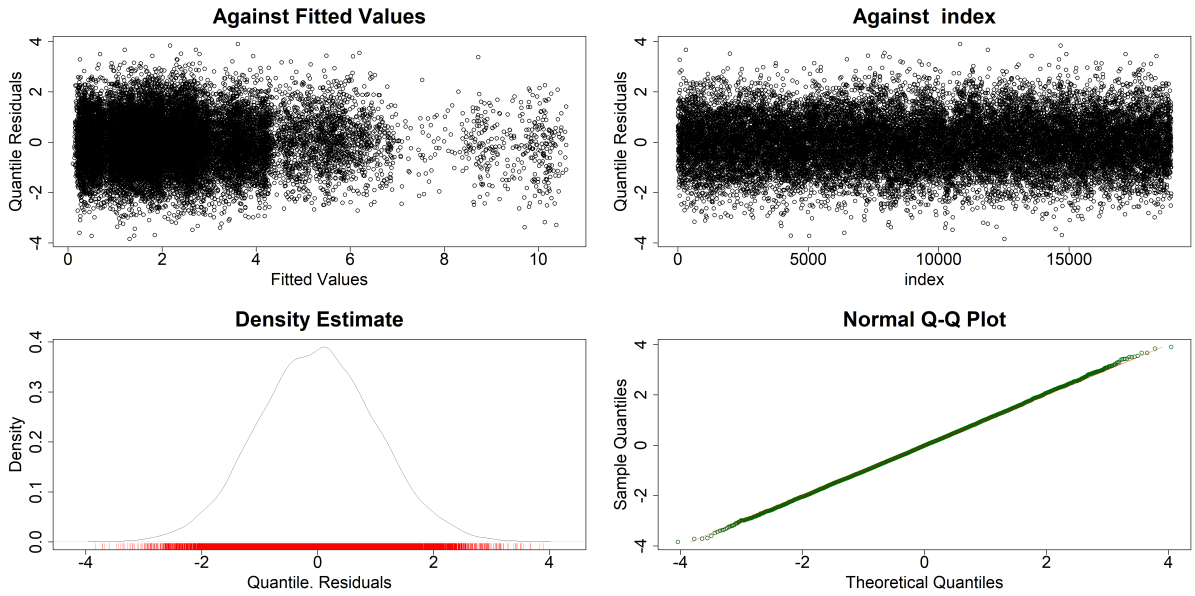


Figure 9: Diagnostics of GAM with Poisson Regression and *PM2.5_lag0*

4 Results

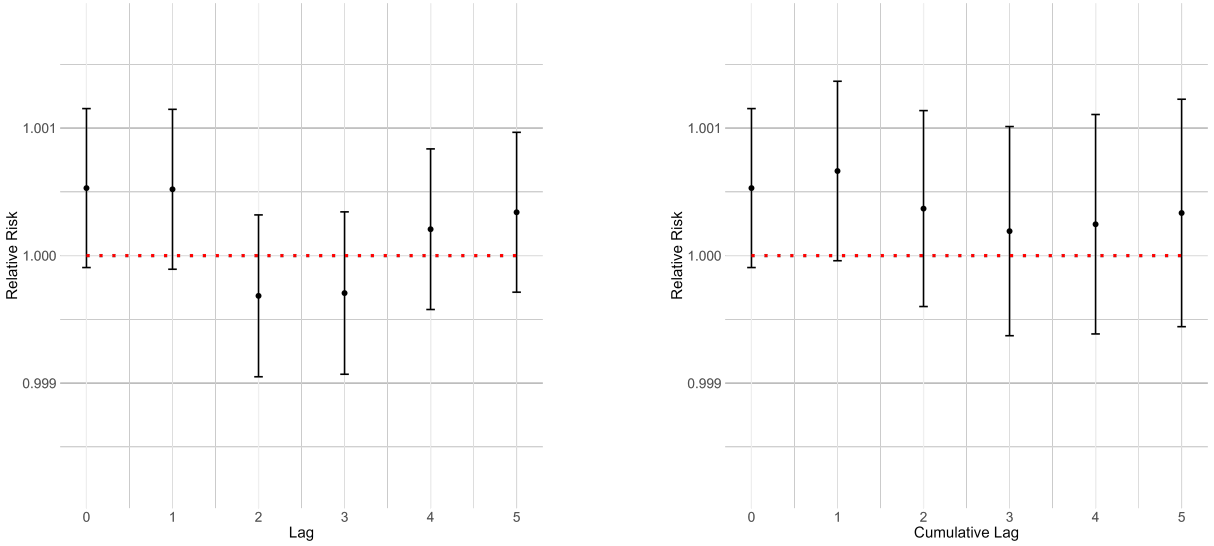
In this section, the models detailed in the previous section are used to address the research questions.

4.1 Association between CVD and PM2.5

The results of the analysis of the association between CVD and PM2.5 are presented below, followed by an examination of the robustness of this association.

4.1.1 Main Models

The main models are used to examine the impact of PM2.5 on CVD, focusing on the relative risk (RR) associated with different lags of PM2.5 exposure. The results are illustrated in Figure 10a, showing the exponential of the coefficients (i.e., RR) across lags 0 through 5, along with their 95% confidence intervals (CIs). The y-axis represents the RR, where lags 0, 1, 4, and 5 exhibit a RR greater than 1, indicating a positive impact of PM2.5 on CVD (i.e., higher PM2.5 are associated with higher expected CVD). Lags 2 and 3 display a RR less than 1, indicating a negative impact of PM2.5 on CVD (i.e. higher PM2.5 are associated with lower expected CVD). None of the lag-terms are statistically significant, as all 95% CIs cover a RR of 1 (y = 1), indicating no significant association between CVD and PM2.5 exposure at any lag within the models. The RR of the cumulative lag-terms of PM2.5 are visualized in Figure 10b. For each model the RR of PM2.5 is higher than 1. The 95% CIs again cover 1 (y = 1), indicating that none of these models show a significant association between PM2.5 and CVD. In conclusion, none of the different PM2.5 lag-terms exhibits a significant association with CVD.



(a) Relative Risk for Single Lags

(b) Relative Risk for Cumulative Lags

Figure 10: Relative Risk of PM2.5

4.1.2 Robustness Analysis

The robustness of these findings was assessed by altering specific model specifications to determine if these impact the results.

Adaption of Degrees of Freedom. The flexibility of the natural cubic spline for *date* was adjusted by changing its df, as done in previous research in the field (e.g., Biggeri, 2004; Kan et al., 2010) and recommended by Biggeri (2004). For the initial model, 20 df, equivalent to five df per year, were used, as a common choice in the literature of this field. This choice resulted in non-significant effects for both the cumulative and single PM2.5 lag terms, as shown in Section 4.1.1. The df for the date spline were altered to 10, 15, 25, and 50, with separate models for each PM2.5 lag term. The results are presented in Table 1 for single lag variables and in Table 2 for cumulative lag variables. For the wide majority of these combinations, the PM2.5 coefficient yields a p-value higher than 0.05. It is important to treat p-values with caution due to the presence of multiple testing. Therefore, it is assumed that the lack of significance in the effect of PM2.5 on CVD is not related to the choice of df in the date component of the model.

Non-Linearity. The possibility of a non-linear relationship between PM2.5 and CVD was investigated. To that end, penalized B-splines for PM2.5 were fit, and the effects visualized through spline effect plots (see Appendix E for plots with single lags). As before, the analysis covers all models with its different lag and cumulative lag terms for PM2.5. The results indicate no significant non-linear effect for most examined lag terms. This lack of evidence for non-linearity supports the decision to employ PM2.5 coefficients without extra flexibility. The linear approach not only simplifies the interpretation, but is also justified by the data, which overall do not suggest a complex relationship between different PM2.5 lags and CVD outcomes.

Table 1: Effects of PM2.5 Single Lags for Different Degrees of Freedom in Natural Cubic Date Spline

df of date-spline	lag-term	PM2.5 coefficient	p-value	AIC
10	PM2.5_lag0	0.00074	0.01887	63343.3
10	PM2.5_lag1	0.00072	0.02244	63343.6
10	PM2.5_lag2	-0.00010	0.75914	63348.6
10	PM2.5_lag3	-0.00008	0.80845	63348.7
10	PM2.5_lag4	0.00042	0.18513	63347.0
10	PM2.5_lag5	0.00055	0.08374	63345.8
15	PM2.5_lag0	0.00055	0.08038	63302.4
15	PM2.5_lag1	0.00054	0.08941	63302.5
15	PM2.5_lag2	-0.00029	0.36476	63304.6
15	PM2.5_lag3	-0.00027	0.39914	63304.7
15	PM2.5_lag4	0.00023	0.47757	63304.9
15	PM2.5_lag5	0.00035	0.26723	63304.2
20	PM2.5_lag0	0.00053	0.09578	63286.0
20	PM2.5_lag1	0.00052	0.10398	63286.2
20	PM2.5_lag2	-0.00032	0.32989	63287.8
20	PM2.5_lag3	-0.00029	0.36585	63288.0
20	PM2.5_lag4	0.00021	0.51941	63288.4
20	PM2.5_lag5	0.00034	0.28782	63287.7
25	PM2.5_lag0	0.00047	0.14072	63236.9
25	PM2.5_lag1	0.00049	0.13104	63236.8
25	PM2.5_lag2	-0.00035	0.28328	63237.9
25	PM2.5_lag3	-0.00032	0.32022	63238.0
25	PM2.5_lag4	0.00019	0.55675	63238.7
25	PM2.5_lag5	0.00033	0.30006	63238.0
50	PM2.5_lag0	0.00046	0.16079	62961.5
50	PM2.5_lag1	0.00050	0.12767	62961.2
50	PM2.5_lag2	-0.00038	0.24824	62962.1
50	PM2.5_lag3	-0.00036	0.27551	62962.3
50	PM2.5_lag4	0.00019	0.57288	62963.1
50	PM2.5_lag5	0.00036	0.26459	62962.2

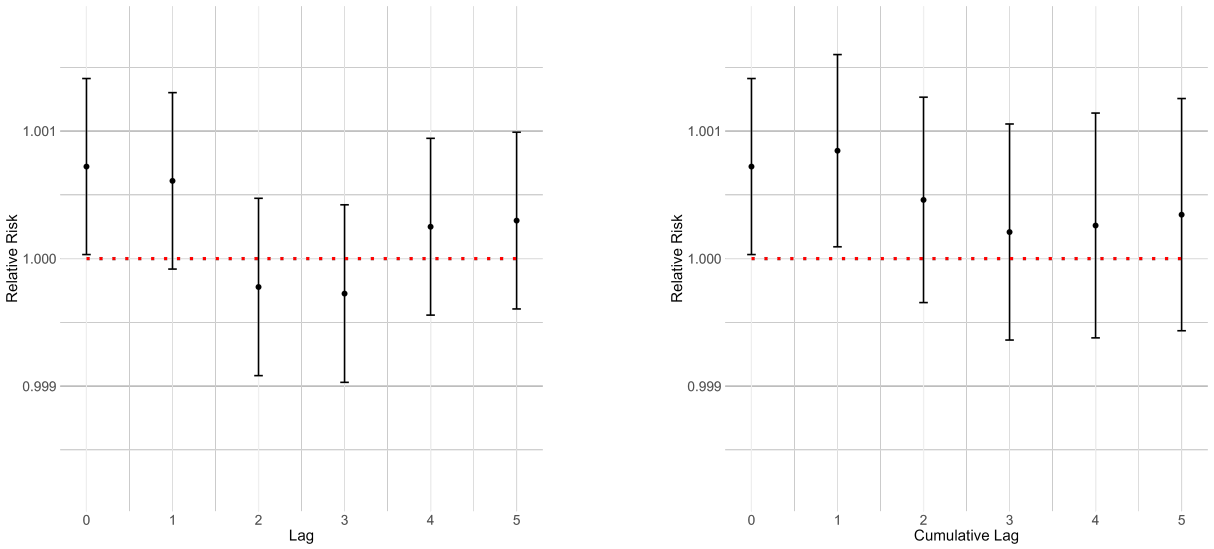
Capping. All PM2.5 values above a threshold of 100 were set to the value of 100, corresponding to the adaption of less than 1% of all PM2.5 values. Spline Plots from the previous paragraph highlight the uncertainty of the effect of PM2.5 for values above 100, raising interest in the impact of adapting these values on the results. Shown in Figure 11, the PM2.5 lag 0 and cumulative lag 1 show a significant positive effect on CVD ($p \leq .05$), $while all the other lags do not show significant effects after capping. It is important to note that the thre$

Table 2: Effects of PM2.5 Cumulative Lags for Different Degrees of Freedom in Natural Cubic Date Spline

df of date-spline	lag-term	PM2.5 coefficient	p-value	AIC
10	PM2.5_lag0	0.00074	0.01887	63343.3
10	PM2.5_cum1	0.00092	0.00947	63342.1
10	PM2.5_cum2	0.00068	0.08004	63345.7
10	PM2.5_cum3	0.00054	0.18770	63347.0
10	PM2.5_cum4	0.00063	0.14362	63346.6
10	PM2.5_cum5	0.00074	0.09565	63346.0
15	PM2.5_lag0	0.00055	0.08038	63302.4
15	PM2.5_cum1	0.00069	0.05304	63301.7
15	PM2.5_cum2	0.00040	0.30262	63304.3
15	PM2.5_cum3	0.00023	0.58227	63305.1
15	PM2.5_cum4	0.00029	0.51276	63305.0
15	PM2.5_cum5	0.00037	0.40861	63304.7
20	PM2.5_lag0	0.00053	0.09578	63286.0
20	PM2.5_cum1	0.00066	0.06457	63285.4
20	PM2.5_cum2	0.00037	0.34684	63287.9
20	PM2.5_cum3	0.00019	0.64674	63288.6
20	PM2.5_cum4	0.00025	0.57540	63288.5
20	PM2.5_cum5	0.00033	0.46257	63288.2
25	PM2.5_lag0	0.00047	0.14072	63236.9
25	PM2.5_cum1	0.00061	0.09342	63236.2
25	PM2.5_cum2	0.00031	0.43668	63238.4
25	PM2.5_cum3	0.00013	0.76525	63238.9
25	PM2.5_cum4	0.00018	0.68063	63238.9
25	PM2.5_cum5	0.00028	0.54750	63238.7
50	PM2.5_lag0	0.00046	0.16079	62961.5
50	PM2.5_cum1	0.00062	0.09771	62960.7
50	PM2.5_cum2	0.00030	0.46725	62962.9
50	PM2.5_cum3	0.00010	0.82449	62963.4
50	PM2.5_cum4	0.00016	0.73071	62963.3
50	PM2.5_cum5	0.00027	0.56973	62963.1

4.2 Spatial Variation in the Effect of PM2.5 on CVD

Separate models were fit for each district to investigate potential spatial variation in the relationship between PM2.5 and CVD, incorporating both single and cumulative lag effects of PM2.5. The analysis began with models using single PM2.5 lag terms, with results depicted in Figure 12. For example, districts 8 and 13 exhibit RR higher than 1, suggesting a positive effect of PM2.5 on CVD across all single lag terms, whereas district 9 shows RR below 1 for certain lags, indicating a potential negative impact on CVD. In some other districts, the models often yield both positive and negative RR, varying



(a) Relative Risk for Single Lags

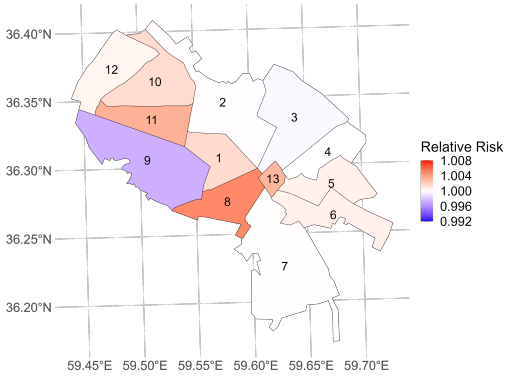
(b) Relative Risk for Cumulative Lags

Figure 11: Relative Risk of Capped PM2.5

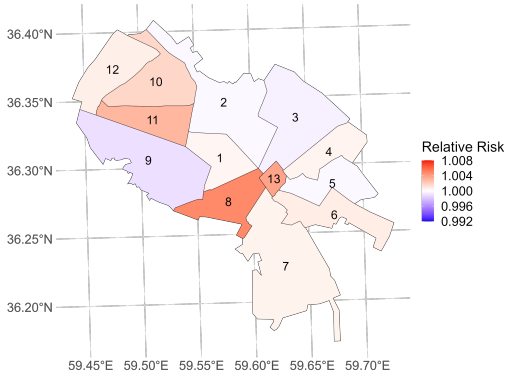
with the specific lag analyzed. For cumulative lag terms, shown in Figure 13, districts 8, 10, 11, 12, and 13 display comparably high RR, pointing to a consistent positive effect of PM2.5 on CVD across different cumulative lags in these districts. In contrast, districts 3 and 9 exhibit lower RR, suggesting a negative impact on CVD. Cumulative lag results generally offer a more consistent perspective on a district's overall RR. Despite these spatial variations, it is crucial to note that the coefficients for PM2.5 lag terms are non-significant in nearly all models ($p > .05$), particularly when accounting for p-value adjustments due to many tests conducted.

4.3 Subgroup Analysis

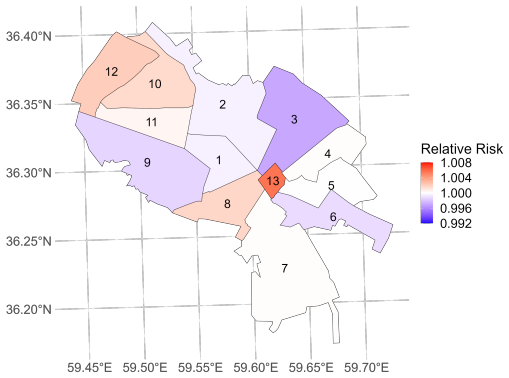
Different epidemiological studies indicate that older adults are particularly vulnerable to the adverse cardiovascular effects of exposure to PM2.5 (Wang et al., 2015). In particular, studies suggest a higher number of cardiovascular-related hospitalizations in individuals aged 64 and older compared to younger adults (e.g., Pope et al., 2008). However, other research has found no increased risk of cardiovascular-related hospitalizations among older adults when exposed to PM2.5 compared to younger age groups (e.g., Metzger et al., 2004). This was investigated in a subgroup analysis in addition to the main analyses. Figure 14 shows the RR for each lag split up for the two different age groups, along with the respective 95% confidence intervals. The results show no significant effect of PM2.5 for any of these single lag or cumulative lag terms in either of the two groups ($p > .05$). The impact of PM2.5 on CVD is very similar in both age groups. In summary, there is no significant effect of PM2.5 on CVD, even when analyzed separately by age group.



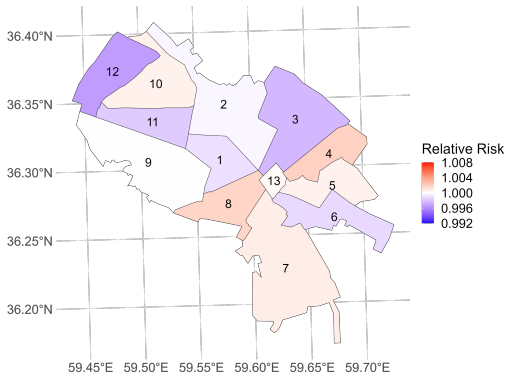
(a) Relative Risk for PM2.5_lag0



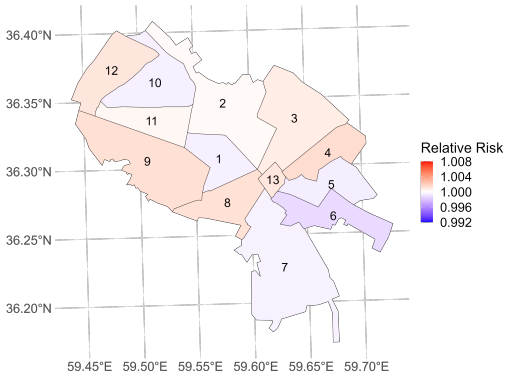
(b) Relative Risk for PM2.5_lag1



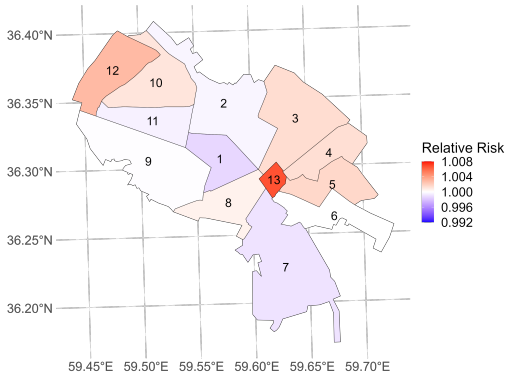
(c) Relative Risk for PM2.5_lag2



(d) Relative Risk for PM2.5_lag3

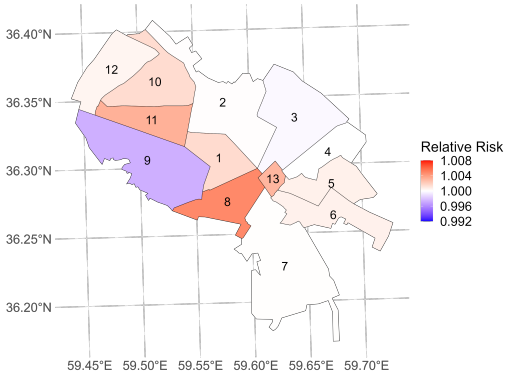


(e) Relative Risk for PM2.5_lag4

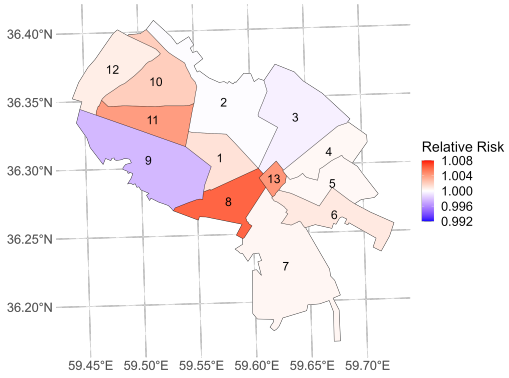


(f) Relative Risk for PM2.5_lag5

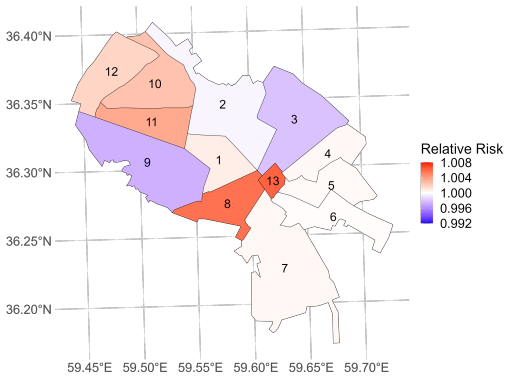
Figure 12: Relative Risk for Single PM2.5 Lags and each District



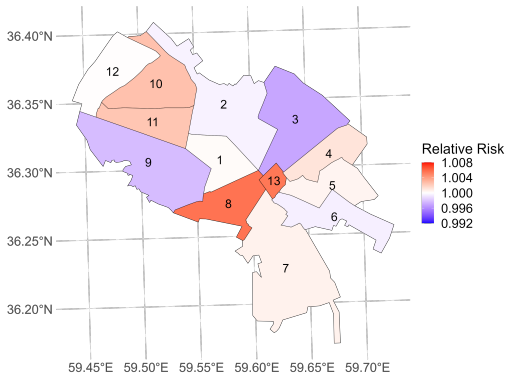
(a) Relative Risk for PM2.5_lag0



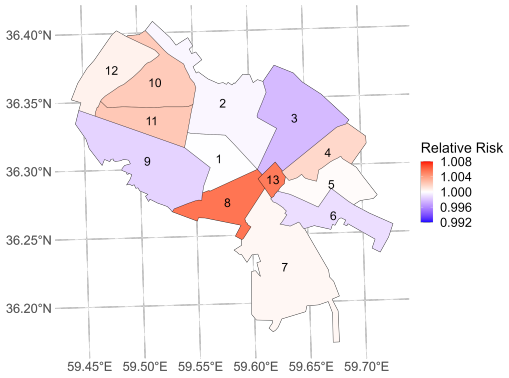
(b) Relative Risk for PM2.5_cum1



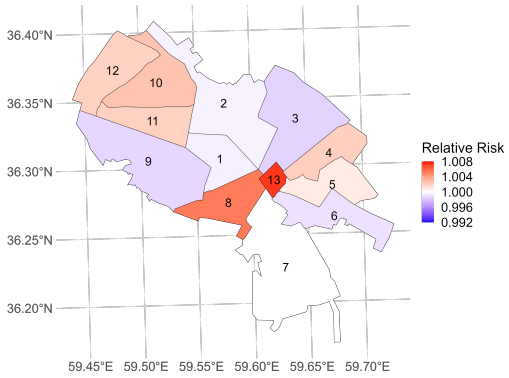
(c) Relative Risk for PM2.5_cum2



(d) Relative Risk for PM2.5_cum3

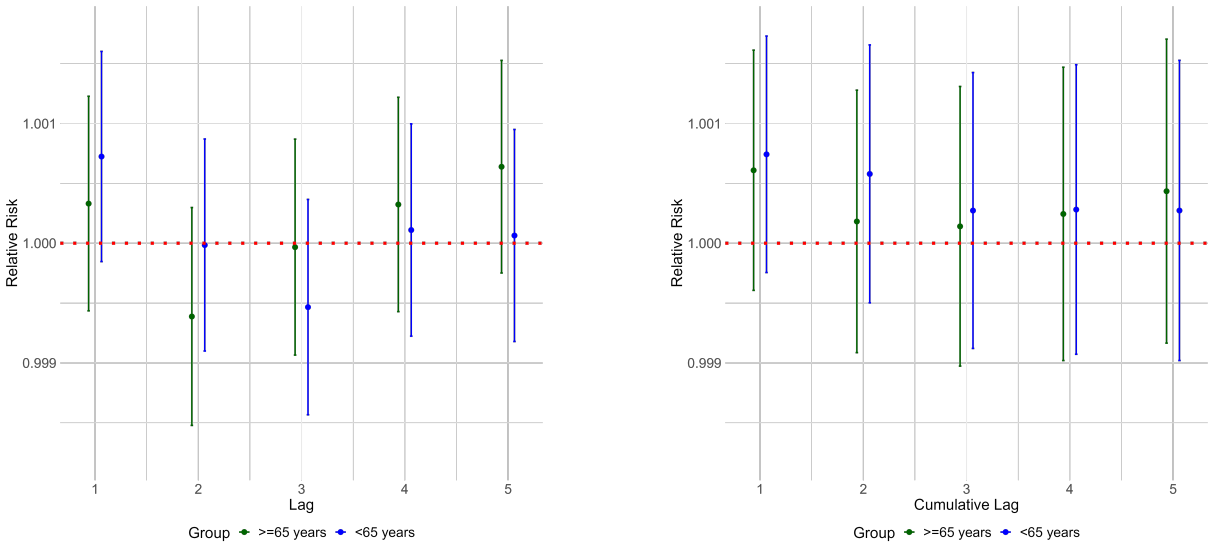


(e) Relative Risk for PM2.5_cum4



(f) Relative Risk for PM2.5_cum5

Figure 13: Relative Risk for Cumulative PM2.5 Lags and each District



(a) Relative Risk for Single Lags

(b) Relative Risk for Cumulative Lags

Figure 14: Relative Risk of PM2.5 Split by Age

5 Summary

We can advance the following conclusions with respect to the initial research questions. The data do not show a significant association between cardiovascular-disease related hospitalizations and PM2.5 in Mashhad, using RF for imputation and spatially aggregating under the assumption of an exponentially decreasing influence of stations. This observation holds for the consideration of delayed effects of PM2.5 on hospitalizations, more precisely of isolated, delayed effects of PM2.5 at each of 5 previous days as well as delayed effects of PM2.5 from previous days combined (i.e. cumulative). Furthermore, these findings are robust against re-specifications of df for time. There is also no clear indication of a non-linear effect of PM2.5 on CVD nor support for the assumption that capping very high PM2.5 values leads to different results. With respect to the second research question, the data show spatial variation in the effect of PM2.5 on cardiovascular disease related hospitalizations between districts, with for instance districts 8 and 13 exhibiting comparably high RR. Finally, an additional analysis within subgroups of age below 65 and 65-and-above reveals no effect of PM2.5 on CVD in either of these groups.

6 Limitations

This study does not remain without limitations. One relevant aspect to consider is the presence of many missing PM2.5 values, which required imputation and subsequent aggregation. The results are thus dependent on the choice of methods for these preprocessing steps. Additionally, cardiovascular disease related hospitalizations in the data are assigned based on the district of residence rather than the district of most exposure. This geographical attribution could potentially distort the results if individuals spend a lot of time outside their residential districts. Another critical issue is the assumption made in the study that all PM2.5 particles possess the same level of toxicity. However, existing literature highlights that the source and composition of PM2.5 can vary, thus affecting its toxicity (Li et al., 2019; Thurston & Balmes, 2017; Thurston et al., 2022). This variation is relevant, emphasizing that the risk associated with PM2.5 exposure possibly cannot be uniformly assessed due to differing toxicity levels, which complicates the accurate estimation of its health impacts. Furthermore, the relationship between PM2.5 exposure and CVD is also thought to be influenced by other environmental variables, such as temperature (e.g., Cui et al., 2019; Mohammadi et al., 2021; Xiong et al., 2017) and humidity (e.g., Klompmaker et al., 2021), suggesting that additional environmental factors should be considered when modeling CVD to better understand the health impacts of PM2.5. These limitations must be kept in mind when drawing conclusions from the results. They can justify the need for extended data collection and motivate further research in this field.

References

Biggeri, A. (2004). *Statistical modelling: Proceedings of the 19th international workshop on statistical modelling, florence (italy), 4-8 july, 2004*. Firenze university press.

Cui, L., Geng, X., Ding, T., Tang, J., Xu, J., & Zhai, J. (2019). Impact of ambient temperature on hospital admissions for cardiovascular disease in hefei city, china. *International journal of biometeorology*, *63*, 723–734.

Faridi, S., Bayat, R., Cohen, A. J., Sharafkhani, E., Brook, J. R., Niazi, S., Shamsipour, M., Amini, H., Naddafi, K., & Hassanvand, M. S. (2022). Health burden and economic loss attributable to ambient pm2.5 in iran based on the ground and satellite data. *Scientific reports*, *12*(1), 14386. https://doi.org/10.1038/s41598-022-18613-x

Hadei, M., Nazari, S. S. H., Eslami, A., Khosravi, A., Yarahmadi, M., Naghdali, Z., & Shahsavani, A. (2017). Distribution and number of ischemic heart disease (ihd) and stroke deaths due to chronic exposure to pm2. 5 in 10 cities of iran (2013-2015); an airq+ modelling. *Journal of air pollution and health*, *2*(3), 129–136.

Jing, X., Luo, J., Wang, J., Zuo, G., & Wei, N. (2022). A multi-imputation method to deal with hydro-meteorological missing values by integrating chain equations and random forest. *Water Resources Management*, *36*(4), 1159–1173. https://doi.org/10.1007/s11269-021-03037-5

Kan, H., Chen, B., Zhao, N., London, S. J., Song, G., Chen, G., Zhang, Y., & Jiang, L. (2010). Part 1. a time-series study of ambient air pollution and daily mortality in shanghai, china. *Research report (Health Effects Institute)*, (154), 17–78.

Khaniabadi, Y. O., Sicard, P., Khaniabadi, A. O., Mohammadinejad, S., Keishams, F., Takdastan, A., Najafi, A., De Marco, A., & Daryanoosh, M. (2018). Air quality modeling for health risk assessment of ambient pm10, pm2. 5 and so2 in iran. *Human and Ecological Risk Assessment: An International Journal*.

Klompmaker, J. O., Hart, J. E., James, P., Sabath, M. B., Wu, X., Zanobetti, A., Dominici, F., & Laden, F. (2021). Air pollution and cardiovascular disease hospitalization – are associations modified by greenness, temperature and humidity? *Environment International*, *156*, 106715. https://doi.org/https://doi.org/10.1016/j.envint.2021.106715

Kloog, I., Ridgway, B., Koutrakis, P., Coull, B. A., & Schwartz, J. D. (2013). Long- and short-term exposure to pm2. 5 and mortality: Using novel exposure models. *Epidemiology*, *24*(4), 555–561.

Leili, M., Nadali, A., Karami, M., Bahrami, A., & Afkhami, A. (2021). Short-term effect of multi-pollutant air quality indexes and pm2. 5 on cardiovascular hospitalization in hamadan, iran: A time-series analysis. *Environmental Science and Pollution Research*, *28*(38), 53653–53667.

Li, X., Jin, L., & Kan, H. (2019). Air pollution: A global problem needs local fixes.

Metzger, K. B., Tolbert, P. E., Klein, M., Peel, J. L., Flanders, W. D., Todd, K., Mulholland, J. A., Ryan, P. B., & Frumkin, H. (2004). Ambient air pollution and cardiovascular emergency department visits. *Epidemiology (Cambridge, Mass.)*, *15*(1), 46–56. https://doi.org/10.1097/01.EDE.0000101748.28283.97

Mohammadi, D., Zare Zadeh, M., & Zare Sakhvidi, M. J. (2021). Short-term exposure to extreme temperature and risk of hospital admission due to cardiovascular diseases.

International journal of environmental health research, *31*(3), 344–354. https://doi.org/10.1080/09603123.2019.1663496

Pope, C. A., Renlund, D. G., Kfoury, A. G., May, H. T., & Horne, B. D. (2008). Relation of heart failure hospitalization to exposure to fine particulate air pollution. *The American journal of cardiology*, *102*(9), 1230–1234. https://doi.org/10.1016/j.amjcard.2008.06.044

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study. *American journal of epidemiology*, *179*(6), 764–774. https://doi.org/10.1093/aje/kwt312

Smith, B. L., Scherer, W. T., & Conklin, J. H. (2003). Exploring imputation techniques for missing data in transportation management systems. *Transportation Research Record: Journal of the Transportation Research Board*, *1836*(1), 132–142. https://doi.org/10.3141/1836-17

Statistical Centre of Iran. (2016). Census 2016 of mashhad, iran. https://www.citypopulation.de/en/iran/admin/khor%C4%81s%C4%81n_e_razavi/0916__mashhad/

Thangavel, P., Park, D., & Lee, Y.-C. (2022). Recent insights into particulate matter (pm2.5)-mediated toxicity in humans: An overview. *International journal of environmental research and public health*, *19*(12). https://doi.org/10.3390/ijerph19127511

Thurston, G. D., & Balmes, J. R. (2017). We need to "think different" about particulate matter.

Thurston, G. D., Chen, L. C., & Campen, M. (2022). Particle toxicity's role in air pollution. *Science*, *375*(6580), 506–506.

Wang, C., Tu, Y., Yu, Z., & Lu, R. (2015). Pm2.5 and cardiovascular diseases in the elderly: An overview. *International Journal of Environmental Research and Public Health*, *12*(7), 8187–8197. https://doi.org/10.3390/ijerph120708187

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

World Health Organization. (2021). *Who global air quality guidelines: Particulate matter (pm2,5 and pm10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide.*

Wu, R., Hamshaw, S. D., Yang, L., Kincaid, D. W., Etheridge, R., & Ghasemkhani, A. (2022). Data imputation for multivariate time series sensor data with large gaps of missing data. *IEEE Sensors Journal*, *22*(11), 10671–10683. https://doi.org/10.1109/JSEN.2022.3166643

Xiong, J., Lan, L., Lian, Z., & Lin, Y. (2017). Effect of different temperatures on hospital admissions for cardiovascular and cerebrovascular diseases: A case study. *Indoor and Built Environment*, *26*(1), 69–77. https://doi.org/10.1177/1420326X15604492

Code Availability:

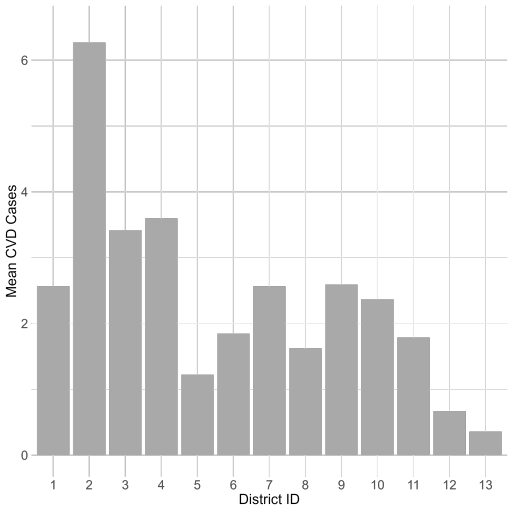
Code for the analysis is available online at github.com/CFroehner/StatPrak-CVD.git.

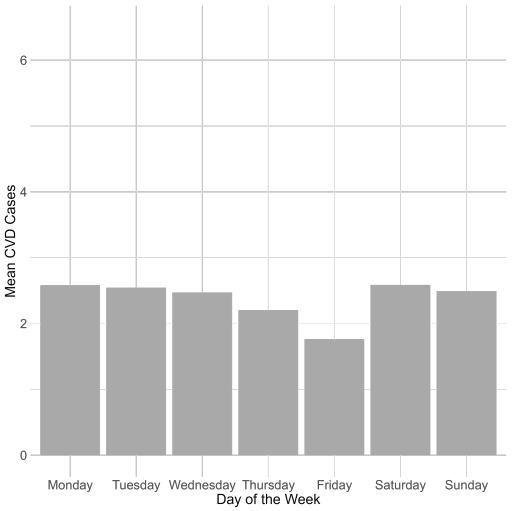
Appendix

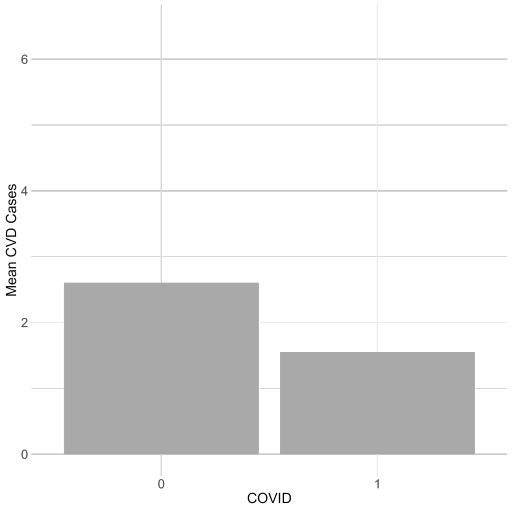
A PM2.5 Lag Variables Visualization

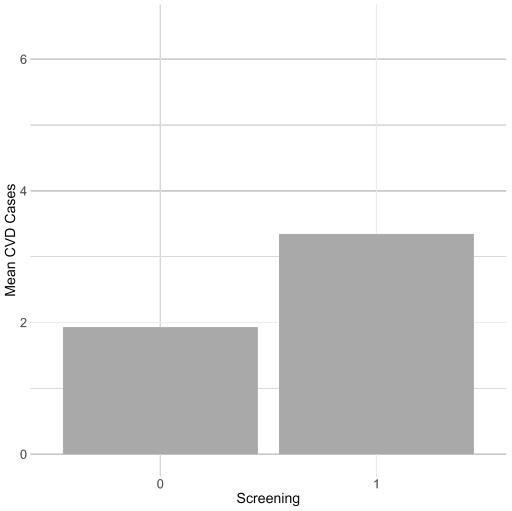
Date	PM2.5_lag$_0$	PM2.5_lag$_1$	PM2.5_lag$_2$	PM2.5_lag$_3$	PM2.5_lag$_4$	PM2.5_lag$_5$
2017-01-01		NA	NA	NA	NA	NA
2017-01-02			NA	NA	NA	NA
2017-01-03				NA	NA	NA
2017-01-04					NA	NA
2017-01-05						NA
2017-01-06						
2017-01-07						

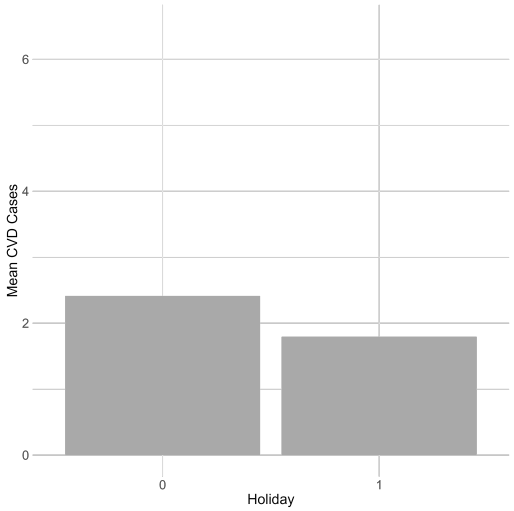
B Barplots of Confounders



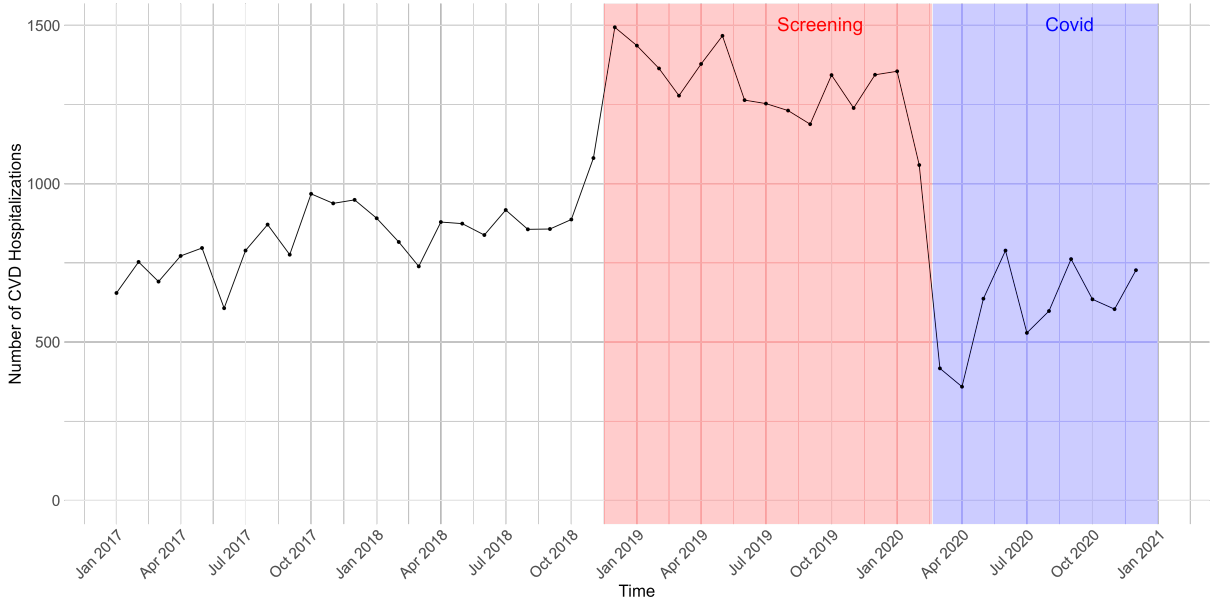




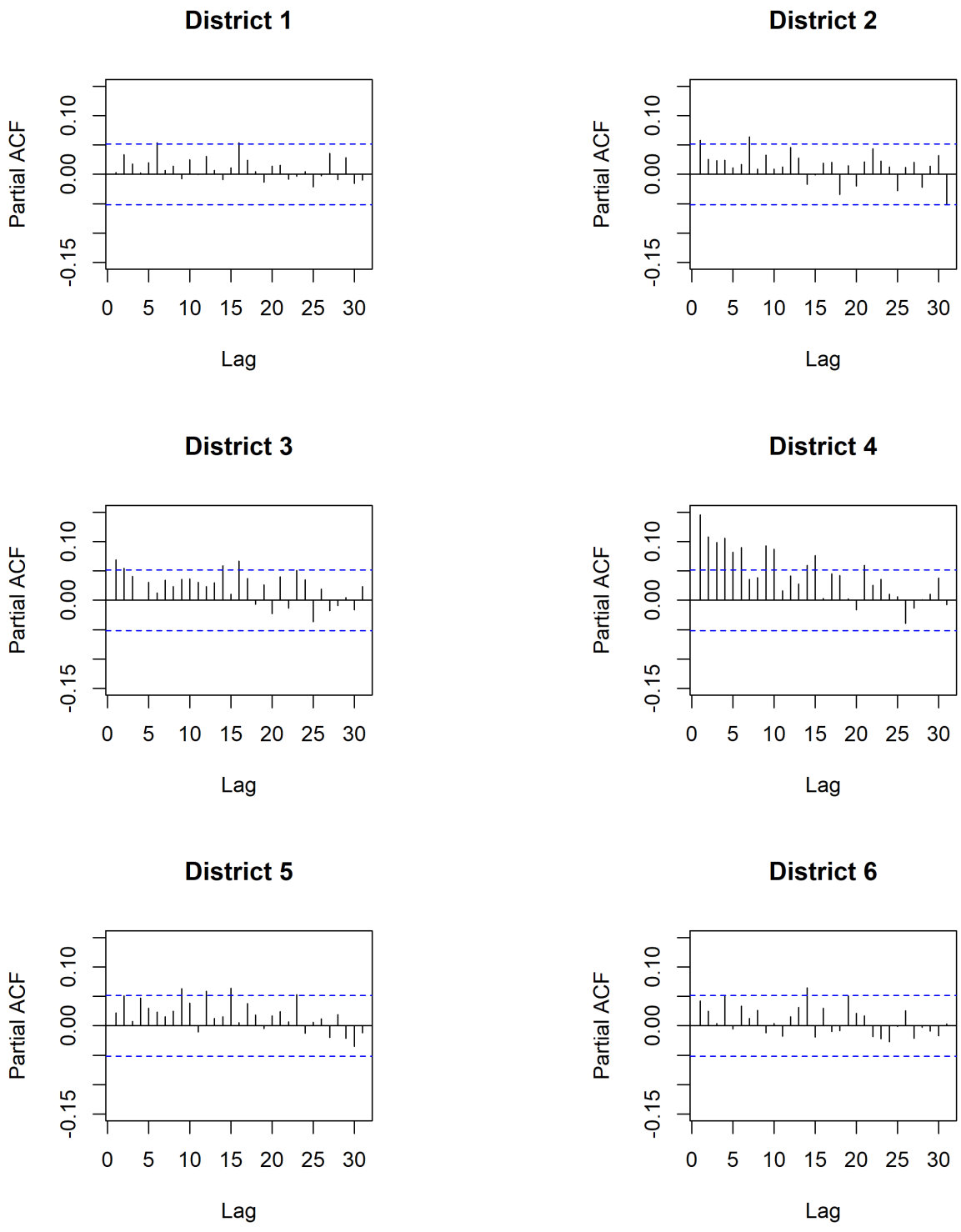




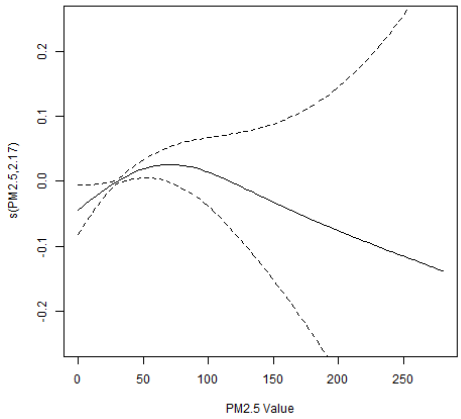
C Timeseries CVD with Screening and Covid Period



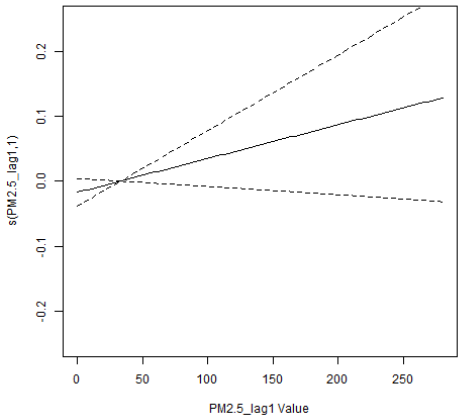
D PACF for PM2.5_lag0 and District 1-6



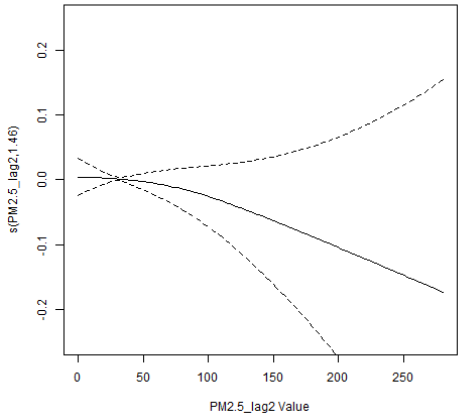
E P-Splines for PM2.5 Single Lags



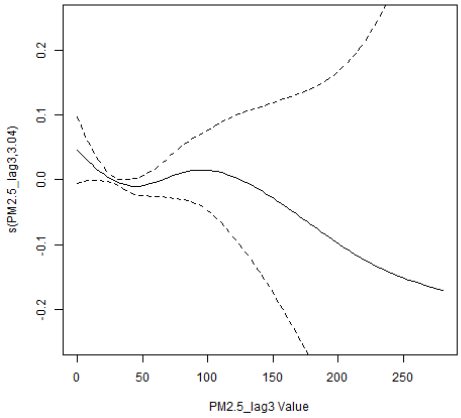
(a) PM2.5_lag0



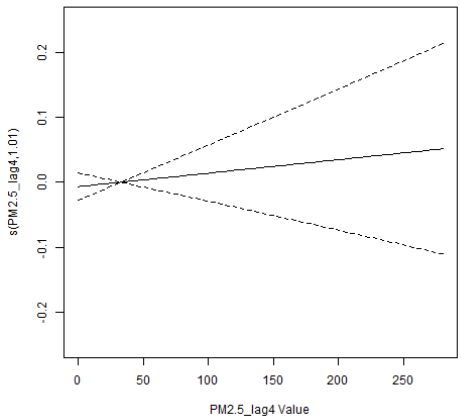
(b) PM2.5_lag1



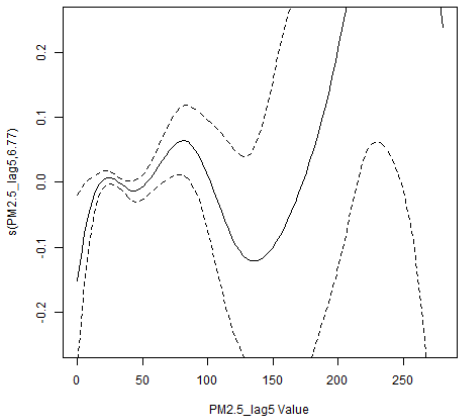
(c) PM2.5_lag2



(d) PM2.5_lag3



(e) PM2.5_lag4



(f) PM2.5_lag5