Population-weighted exposure to air pollution and COVID-19 infection in Germany

Guowen Huang[1,2,*], Patrick E Brown[1,2]

Abstract

It is well known that COVID-19, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is to spread mainly from person to person, mainly through respiratory droplets produced when an infected person coughs or sneezes. Therefore, many countries have enforced social distancing to stop the spread of COVID-19. However, within a specific country, although the measures taken by government are almost the same, the incidence rate varies among areas (e.g., counties, cities). One potential explanation is that people in some areas are more vulnerable to the coronavirus disease because of their worsen health conditions caused by long-term exposure to poor air quality. In this study, we investigate whether long-term exposure to air pollution increases the risk of COVID-19 infection in Germany. The results show that nitrogen dioxide (NO_2) is significantly associated with COVID-19 incidence rate, with 1 μgm^{-3} increase of long-term exposure to NO_2, the COVID-19 incidence rate is likely to increase 5.61% (95% credible interval [CI]: 3.36%, 7.91%). This inference is robust across various health models. The analyses can be reproduced and updated routinely by sharing the public data sources and R code.

Keywords

COVID-19, air pollution, health impacts, Kriging, INLA

Author information

[1] Dept. of Statistical Sciences, University of Toronto, Toronto, ON, Canada

[2] Centre for Global Health Research, St Michael's Hospital, Toronto, ON, Canada

[*] Corresponding author: E-mail: hgw0610209@gmail.com

1 Introduction

Currently, COVID-19, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is widespread and much more dangerous than the seasonal flu. COVID-19 is more infectious than the flu and has a higher death rate. Up to 12th September, 2020, it has led to worldwide over 28.5 million cases and 916 thousand deaths. In Germany, the total confirmed cases up to 12th September, 2020 rise to 259 thousand, with the deaths being more than 9.4 thousand. A recent study by Wu et al. (2020) investigated the impact of long-term average exposure to fine particulate matter ($PM_{2.5}$) on the risk of COVID-19 deaths in the United States and found that an increase of only 1 $\mu g/m^{-3}$ in $PM_{2.5}$ was associated with a 8% (95% confidence interval, 2%, 15%) increase in the COVID-19 death rate. Ogen (2020) reported that most of COVID-19 fatality cases occurred in those regions with the highest NO_2 concentrations while studying 66 administrative regions in Italy, Spain, France and Germany. These results suggest that high levels of air pollution may be an important contributor to COVID-19 infections or deaths.

The existing body of research on the impacts of air pollution on human health, has linked $PM_{2.5}$ and NO_2 exposure to health damage, particularly respiratory and lung diseases, which could make

people more vulnerable to contract COVID-19. The main source of NO_2 resulting from human activities is the combustion of fossil fuels (coal, gas and oil), especially fuel used in cars. Exposure to high levels of NO_2 can cause inflammation of the airways. Long term exposure may affect lung function and respiratory symptoms. For example, the research results from Bowatte et al. (2017) indicate that long-term exposure to NO_2 was associated with increased risk of respiratory diseases, while Lee et al. (2009) show that long-term exposure to NO_2 was significantly associated with respiratory hospital admissions in Edinburgh and Glasgow, UK. Similarly, Schikowski et al. (2005) suggests that long-term exposure to air pollution from NO_2 and living near a major road might increase the risk of developing chronic obstructive pulmonary disease (COPD) and can have a detrimental effect on lung function.

On the other hand, particulate matter (both PM_{10} and $PM_{2.5}$) is made up of a wide range of materials and arises from both human-made (such as stationary fuel combustion and transport) and natural sources (such as sea spray and Saharan dust). Concentrations of particulate matter comprises primary particles emitted directly into the atmosphere from combustion sources and secondary particles formed by chemical reactions in the air. Exposure to particulate matter is associated with respiratory and cardiovascular illness and mortality as well as other ill-health effects. Examples include Lee et al. (2009) and Lee (2012), where the authors found that long-term exposure to PM_{10} was significantly associated with respiratory hospital admissions. Recent reviews by Committee on the Medical Effects of Air Pollutants (2010) have suggested exposure to $PM_{2.5}$ had a stronger association with the observed ill-health effects because they can travel deeper into lungs.

In this study, we investigate whether long-term average exposure to air pollution increases the risk of COVID-19 infection in Germany. We use a spatial ecological design to estimate the impacts of air pollution on COVID-19 infection by utilizing geographical contrasts in air pollution and infection risk across K contiguous small areas (Huang et al. (2018), Napier et al. (2018), Rushworth et al.

(2014)).

In such studies, the disease data are counts of disease cases occurring in each areal unit while the spatially representative pollution concentrations in each areal unit are typically estimated by applying Kriging, a spatial model (Diggle and Ribeiro (2007)), to data from a sparse monitoring network, or by computing averages over modelled concentrations (grid level) from an atmospheric dispersion model (Wu et al. (2020), Maheswaran et al. (2006), Lee et al. (2009), Warren et al. (2012)), or by combining both to obtain a better prediction (Huang et al. (2018), Vinikoor-Imler et al. (2014), Sacks et al. (2014)). The downside of these studies is that the inference is a population level association rather than an individual-level causal relationship, and wrongly assuming the two are the same is known as ecological bias (Arbia (1988), Wakefield and Salway (2001)). Such bias is due in part to within-population variation in pollution exposures and disease incidence, because one does not know whether, within a population, it is the same individuals that exhibit disease and have the highest air pollution exposures (Lee et al. (2020)). The simulation study from Lee et al. (2020) also suggests that the estimates of the aggregated model from individual levels almost always exhibit less variation than those from the ecological model.

Therefore, in order to better estimate the real areal pollution concentrations, in this study we first apply Kriging to monitoring pollution data to obtain predictions on a raster grids where population density data are available, then estimate the areal pollution concentrations by taking spatially population-weighted average of the gridded predictions lying within a specific county. In addition, the pollutants considered in this study include common pollutants: particulate matter (both $PM_{2.5}$ and PM_{10}), NO_2, sulfur dioxide (SO_2); and also four poisonous pollutants: benzene, arsenic in PM_{10}, cadmium in PM_{10} and nickel in PM_{10}, given their adverse impacts on human health being reported (e.g., Yu et al. (2003), Smith (2010), Järup et al. (1998), Das et al. (2008)).

It is worth to note that another challenge in air pollution health effect studies is how to allow for the uncertainty in the estimated pollution concentrations when estimating their health effects

(Huang et al. (2018), Blair et al. (2007)). Specifically, the areal level pollution predictions produced from the pollution data are with uncertainty as they are only the estimates of the true spatially-varying concentrations. The disadvantage of using a point estimate is that one may overstate the certainty about the connection between the outcome and the covariate. A number of approaches have been proposed to incorporate pollution uncertainties and measurement errors into the health model (e.g., Huang et al. (2018), Lee et al. (2017), Blangiardo et al. (2016), Gryparis et al. (2008)). Given that the study from Lee et al. (2017) showed that treating the posterior predictive pollution distribution as a prior in the disease model has produced similar results to ignoring the uncertainty except for PM_{10}, and Blangiardo et al. (2016) also found that incorporating uncertainty in pollution by making multiple sets of estimated exposure and then fitting the disease model separately for each set before combining the estimated health effects, did not change the substantive conclusions, we do not address exposure uncertainty in this study. Instead, we incorporate the reliability of gridded pollution predictions while aggregating them spatially, with details can be found in Section 3.2.

The remainder of this paper is organized as follows. The data and its exploratory analysis are presented in Section 2, while the statistical methodology is outlined in Section 3. The results of the study are reported in Section 4, and the key conclusions are presented in Section 5.

2 Motivating study

The study region is Germany which has a population of around 83 million people. It consists of $K = 401$ counties (administrative districts), among which 294 are rural and 107 are urban. A map of these counties is shown in Figure 1, with the shapefile publicly available from Germany's Federal Agency for Cartography and Geodesy (see Table 4). The data set used in this study are county-level aggregated, including COVID-19 cases, pollution concentrations, temperature and

population data. Note that the lockdown measures in Germany issued on March 22, 2020 have led to reduced vehicle traffic and air pollution. However, long-term exposure to polluted air before the pandemic may be more important than current levels of pollution. Therefore, we use the most recent available few years (2016-2018) average concentrations of pollutants for estimating the long-term exposure for each county. Full description of the data set is introduced as follows.

2.1 Data description

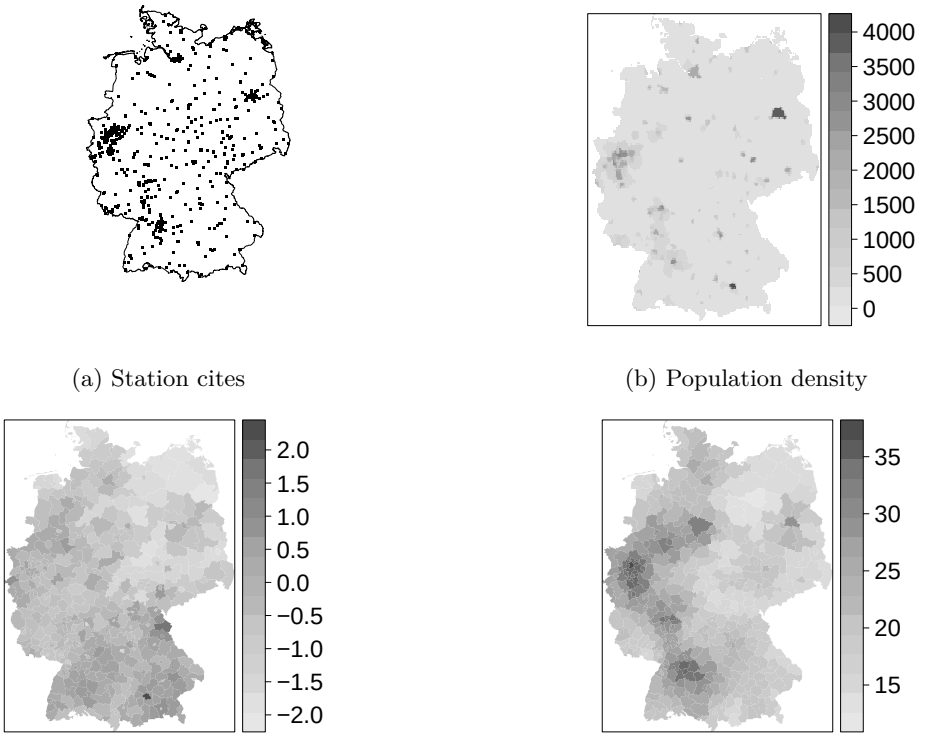
The COVID-19 cases by county, and the population by sex and age on the federal state level in Germany are publicly available on Kaggle (Heads or Tails (2020)), contributed by one with a user name 'Heads or Tails', where the COVID-19 cases and deaths will be updated daily, with the earliest recorded cases are from 2020-01-24. The accumulated COVID-19 cases used in this study are up to 2020-09-09. According to Heads or Tails (2020), the original data are being collected by Germany's Robert Koch Institute (Table 4) and can be downloaded through the National Platform for Geographic Data (Table 4). The population data on the federal state level could be downloaded from Germany's Federal Office for Statistics (Statistisches Bundesamt, (Table 4)) through their Open Data platform GENESIS (Table 4). In addition, the county level population data are freely obtained from the City Population (Table 4). Both these two population data sets reflect the (most recent available) estimates on 2018-12-31. The raster gridded population density data are freely available on DIVA-GIS (Table 4), which is shown in Figure 1b.

We denote Y_k as the reported numbers of COVID-19 cases for county k. As the number of case in a county depends on its population, we calculate the expected number of cases in each county by $E_k = \frac{P_k}{P_{s(k)}} \sum_j r_j P_{s(k),j}$, where P_k is the population in county k, r_j is the national incidence rate in sex-age group j, and $P_{s(k)} = \sum_j P_{s(k),j}$ denotes the population of the state which contains county k. The latter part in the equation $\sum_j r_j P_{s(k),j}$ is the expected cases in state $s(k)$. We use

standardized incidence ratio (SIR) given by $\text{SIR}_k = Y_k/E_k$, to measure the risk of disease, and an SIR of 1.1 indicates a 10% increased risk of disease compared to that expected. A spatial map of the natural logarithm of SIR for COVID-19 (the scale will be modelled on) as of 2020-09-09 can be seen in Figure 1c, showing a wide variation in SIRs across the counties in Germany and the majority of the high-risk counties are at the east and south part of Germany.

Pollution data are available from the Air Quality e-Reporting provided by European Environment Agency (EEA) (Table 4), where the monitoring stations are shown in Figure 1a which tend to be dense where the population density is high (Figure 1b). We first calculate the average concentration for each station through 2016-2018. Then we apply the spatial modelling and prediction method described in Section 3 to obtain the spatially population-weighted representative concentrations for each county. The pollutants considered in this study include common pollutants: $PM_{2.5}$ and PM_{10}, NO_2, SO_2; and also four poisonous pollutants: benzene, arsenic in PM_{10}, cadmium in PM_{10} and nickel in PM_{10}. These pollutants could have potential harmful health effects, such as damage to the lungs and nasal cavity, reducing lung function, causing chronic bronchitis and cancers of the bladder and lungs (Yu et al. (2003), Smith (2010), Järup et al. (1998), Das et al. (2008)). A summary of the estimated population-weighted county level exposure is presented in Table 1. An exploratory of the scatterplots of the natural logarithm of COVID-19 SIR against NO_2 and $PM_{2.5}$ are displayed in Figure 2, which would appear to indicate a linear correlation between NO_2 and log COVID-19 SIR.

The monitoring temperature data can be freely downloaded from European Climate Assessment & Dataset (Table 4). Similar to pollution exposure estimation, they are firstly used to calculate the averaged temperature from 2016-1018 for each monitor, then spatial modelling is applied to obtain population-weighted county level exposure. A simple summary can be seen in Table 1 as well.



(a) Station cites

(b) Population density

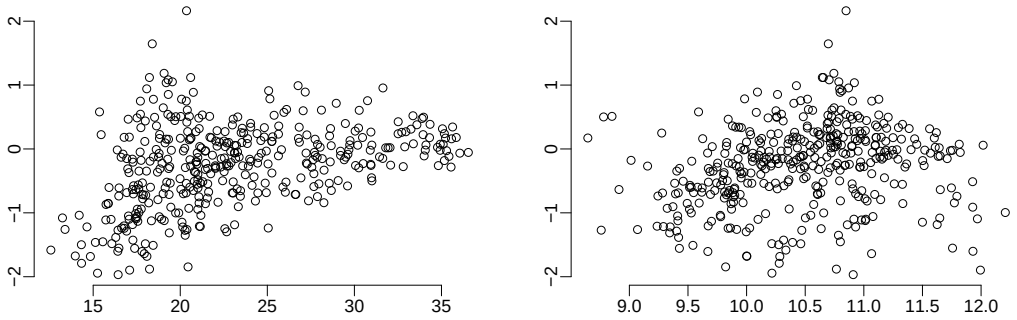
(c) log(SIR) in counties

(d) NO_2 in counties

Figure 1: Pollution stations, population density, log COVID-19 SIR and NO_2 (μgm^{-3}) in Germany.

Table 1: Population-weighted county level exposure summary, with μgm^{-3} unit for NO_2, PM_{25}, PM_{10}, SO_2, Benzene; ngm^{-3} for Aresenic, admium, Nickel; and $^{\circ}C$ for temperature.

Variable	**min**	**quantile25**	**median**	**mean**	**quantile75**	**max**
no2	12.57	18.79	21.58	23.02	26.84	36.54
pm25	8.65	9.98	10.52	10.48	10.93	12.21
pm10	14.48	16.84	17.73	17.74	18.57	21.07
so2	0.69	1.21	1.52	1.65	1.95	4.22
Benzene	0.69	0.81	0.85	0.88	0.96	1.15
Aresenic	0.32	0.40	0.44	0.45	0.50	0.67
Cadmium	0.08	0.10	0.10	0.11	0.13	0.20
Nickel	0.97	1.32	1.44	1.63	1.85	3.22
Temperature	6.69	9.61	10.15	10.09	10.54	11.84



(a) NO_2 (b) $PM_{2.5}$

Figure 2: Scatterplots of log COVID-19 SIR against NO_2 (μgm^{-3}) and $PM_{2.5}$ (μgm^{-3}).

3 Method

Based on the observed and expected counts of disease cases occurring in each areal unit, we calculate $\text{SIR}_k = Y_k/E_k$, to measure the risk of disease. $\text{SIR}_k > 1$ represents areas with elevated levels of disease risk, while $\text{SIR}_k < 1$ corresponds to comparatively healthy areas. The elevated risks are likely to happen by chance if E_k is small, which can occur if the disease in question is rare and/or

the population at risk is small (Lee (2011)). To overcome this problem, a Poisson log-linear models are typically used for the analysis (Elliott et al. (2000), Banerjee et al. (2004), Lawson (2008)). The linear predictor includes pollutant concentrations and potential confounders. These known covariates are augmented by a set of random effects to capture the residual spatial autocorrelation after the covariate effects have been accounted for. The random effects borrow strength from values in neighbouring areas, which reduces the likelihood of excesses in risk occurring by chance.

The random effects are commonly modelled by the class of conditional autoregressive (CAR) prior distributions, which are a type of Markov random field model (Rue and Held (2005)). The spatial correlation between the random effects is determined by a binary $n \times n$ neighbourhood matrix $\mathbf{W}$. Based on this neighbourhood matrix, the most common models for the random effects include intrinsic autoregressive (IAR) (Besag et al. (1991)), convolution or BYM model (Besag et al. (1991)), as well as those proposed by Cressie (1993) and Leroux et al. (1999). These CAR models differ by holding different assumptions about how the random effects depend on each other across space.

3.1 Pollution model

It is known that the relationship between pollutants could improve predictions (e.g., Huang et al. (2018)), especially where one pollutant is missing but the others are not. The prediction of the missing pollutant could borrow strength from the other pollutants. However, for simplicity, in this study we prefer a univariate model for each pollutant, since the number of monitoring stations is 709 which is a fairly large sample size. We treat the underlying pollution levels in Germany as a spatial Gaussian process $\{S(\mathbf{s}) : \mathbf{s} \in \mathbb{R}^2\}$ with mean μ, variance $\sigma^2 = \text{Var}\{S(\mathbf{s})\}$ and correlation function $\rho(u) = \text{Corr}\{S(\mathbf{s}), S(\mathbf{s}')\} = \exp(-u/\eta)$, where $u = \|\mathbf{s}-\mathbf{s}'\|$ denotes the Euclidean distance between $\mathbf{s}$ and $\mathbf{s}'$. Denote the observed pollution data as $\mathbf{Z} = \{Z(\mathbf{s}); \mathbf{s} = \mathbf{s}_1, \dots, \mathbf{s}_n\}$, and write $\mathbf{S} = \{S(\mathbf{s}); \mathbf{s} = \mathbf{s}_1, \dots, \mathbf{s}_n\}$ for the unobserved values of the signal at the sampling locations $\mathbf{s}_1, \dots, \mathbf{s}_n$,

the pollution model is assumed as

$$\mathbf{Z} = \mathbf{S} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon}(\mathbf{s}) \sim \mathrm{N}(\mathbf{0}, \tau^2\mathbf{I})$ is uncorrelated with $\mathbf{S}$, and $\mathbf{I}$ is the identity matrix of order n. Note that, $\mathbf{S}$ is multivariate Gaussian with mean vector $\mu\mathbf{1}$, where $\mathbf{1}$ denotes a vector each of whose elements is 1, and variance matrix $\tau^2 R$, where R is the n by n matrix with elements $r_{ij} = \rho(\|\mathbf{s}_i - \mathbf{s}_j{}'\|)$. Similarly, Z is multivariate Gaussian

$$\begin{aligned} \mathbf{Z} &\sim \mathrm{N}(\mu\mathbf{1}, \sigma^2 V) \\ V &= R + \nu^2\mathbf{I}, \end{aligned} \tag{2}$$

where $\nu^2 = \tau^2/\sigma^2$ is the noise-to-signal variance ratio.

The log-likelihood function is

$$L(\mu, \tau^2, \sigma^2, \eta) = -0.5\{n\log(2\pi) + \log\{\mid \sigma^2 R(\eta) + \tau^2\mathbf{I} \mid\} + (\mathbf{Z} - \mu\mathbf{1})'(\sigma^2 R(\eta) + \tau^2\mathbf{I})^{-1}(\mathbf{Z} - \mu\mathbf{1})\} \tag{3}$$

Given V, the maximum likelihood estimate (mle) of μ and σ^2 is given by,

$$\begin{aligned} \hat{\mu}(V) &= (\mathbf{1}'V^{-1}\mathbf{1})^{-1}\mathbf{1}'V^{-1}\mathbf{Z} \\ \hat{\sigma}^2(V) &= n^{-1}(\mathbf{Z} - \hat{\mu}\mathbf{1})'V^{-1}(\mathbf{Z} - \hat{\mu}\mathbf{1}). \end{aligned} \tag{4}$$

By substituting $\hat{\mu}(V)$, and $\hat{\sigma}^2(V)$ into the log-likelihood function, we have,

$$L_0(\nu^2, \eta) = -0.5\{n\log(2\pi) + n\log\hat{\sigma}^2(V) + \log\mid V \mid + n\}, \tag{5}$$

which can be optimised numerically with respect to η and ν, followed by back substitution to

obtain $\hat{\sigma}^2$ and $\hat{\mu}$. This is achieved by function **likfit()** in **geoR** package by providing initial values for the covariance parameters (Diggle and Ribeiro (2007)).

3.2 Areal pollution exposure

The areal pollution exposure is estimated by aggregating the gridded predictions weighted by population density and also the precision of prediction. For a new location $\mathbf{s}_0$, the kriging formula of $S(\mathbf{s}_0)$ (Diggle and Ribeiro (2007)) is used to obtain its prediction by plugging-in the resulting estimates $\hat{\mu}, \sigma^2, \hat{\tau}^2, \hat{\eta}$, which is

$$\hat{S}(\mathbf{s}_0) = \hat{\mu} + \mathbf{C'}_{\mathbf{s}_0}(\hat{\sigma}^2\hat{V})^{-1}(\mathbf{Z} - \hat{\mu}\mathbf{1}), \tag{6}$$

where $\mathbf{C}_{\mathbf{s}_0} = \hat{\sigma}^2(\exp(-\|\mathbf{s}_1 - \mathbf{s}_0\|/\hat{\eta}), \dots, \exp(-\|\mathbf{s}_n - \mathbf{s}_0\|/\hat{\eta}))'$. The corresponding prediction variance is $\mathrm{Var}(\hat{S}(\mathbf{s}_0) - S(\mathbf{s}_0)) = \hat{\sigma}^2 - \mathbf{C'}_{\mathbf{s}_0}(\hat{\sigma}^2\hat{V})^{-1}\mathbf{C}_{\mathbf{s}_0}$, based on which we define the veracity score for the prediction by $\zeta(\mathbf{s}_0) = \frac{1}{\sqrt{\mathrm{Var}(\hat{S}(\mathbf{s}_0) - S(\mathbf{s}_0))}}$. The higher the veracity score is, the better quality the prediction has, and then it should speak more while being aggregated to spatial areal exposure.

After obtaining pollution predictions at the center of all grids where the population density data are available (see Figure 1b) using (6), by denoting the population density at location $\mathbf{s}_i$ as $G(\mathbf{s}_i)$, for a specific k county, the spatially representative pollution concentration is estimated by

$$z_k = \sum_{\mathbf{s}_i \in A_k} \frac{Z(\mathbf{s}_i)G(\mathbf{s}_i)\zeta(\mathbf{s}_i)}{\sum_{\mathbf{s}_j \in A_k} G(\mathbf{s}_j)\zeta(\mathbf{s}_i)}, \tag{7}$$

where A_k represents county k. Therefore, z_k is a spatial metric weighted by population density and also the veracity score.

3.3 Health model

The disease data are counts of the numbers of cases occurring in each county in Germany, and thus Poisson log-linear models are typically used for the analysis (Shaddick and Zidek (2015)). Recall that the observed and expected number of COVID-19 cases for county k as Y_k, E_k, respectively. The health model is given by

$$\begin{aligned} Y_k &\sim \text{Poisson}(E_k\lambda_k), \quad k = 1, \ldots, K, \\ \ln(\lambda_k) &= \mathbf{X}_k^\top \boldsymbol{\beta} + \phi_k \end{aligned} \tag{8}$$

where the relative risk of disease in country k is denoted by λ_k, and is modelled on the log scale by covariates, $\mathbf{X}_k$ containing an intercept column and covariates (pollutants, temperature and areal population density [population divided by area] referred to as popDensity), and a random effect ϕ_k. The regression parameters $\boldsymbol{\beta}$ are assigned weakly informative zero-mean Gaussian priors with a large diagonal variance matrix $\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, 10^2\mathbf{I})$.

ϕ_k is a random effect included to allow for any residual spatial autocorrelation remaining in the disease counts after the covariate effects have been accounted for, and is modelled by,

$$\boldsymbol{\phi} \sim \text{N}(\mathbf{0}, \kappa^2\mathbf{Q}(\xi, \mathbf{W})^{-1}), \tag{9}$$

where $\boldsymbol{\phi} = \{\phi_k, k = 1, \ldots, K\}$. Spatial autocorrelation is induced into the random effects by the precision matrix $\mathbf{Q}(\xi, \mathbf{W}) = \xi(\text{diag}(\mathbf{W1}) - \mathbf{W}) + (1 - \xi)\mathbf{I}$, which corresponds to the conditional autoregressive (CAR) prior proposed by Leroux et al. (1999). The spatial dependence in the data is captured by an $n \times n$ neighbourhood matrix $\mathbf{W}$, whose ijth element equals 1 if areas (i, j) share a common border and is zero otherwise. The level of spatial autocorrelation in the random effects

is controlled by ξ. Finally, weakly informative hyperpriors are specified for the parameters (κ^2, ξ) by

$$\begin{aligned} \kappa &\sim \exp(ln2), \\ \log(\frac{\xi}{1-\xi}) &\sim \mathrm{N}(0, 1.8). \end{aligned} \tag{10}$$

These priors are presented in Figure 4, by which the disease models are implemented. The prior distribution of ξ is likely non-informative as it is roughly uniformly distributed within [0,1]. The prior distribution of κ allows small values, which are what we expected for the variation of the log scale of relative risk. INLA uses the Integrated Nested Laplace Approximation, a computationally effective and extremely powerful alternative to implement Bayesian models, and is an increasingly popular analysis package in R. For details on how to fit spatial and spatio-temporal models with R-INLA, refer to Blangiardo et al. (2013).

4 Results

4.1 Exposure estimation

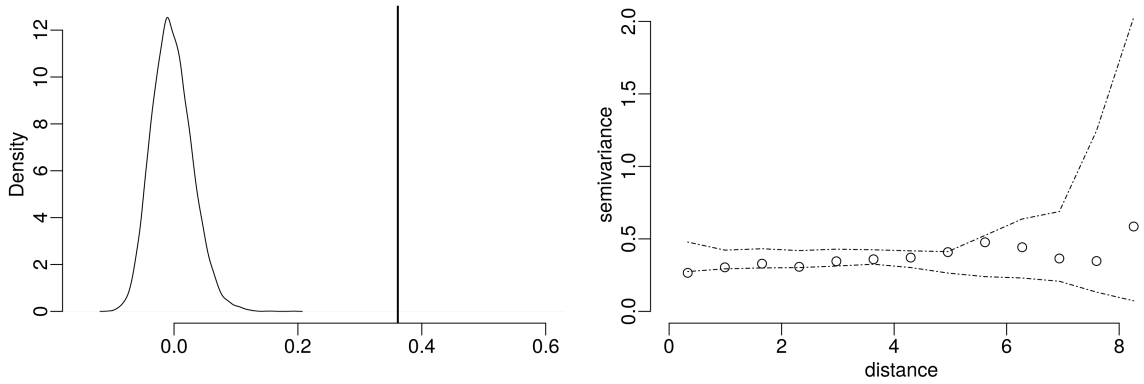
Table 2 presents the estimation of pollution model parameters while applying model (1) to different pollutants separately. The main message from the table is that the Akaike information criterion (AIC, Akaike (1973)) from the proposed spatial pollution model (1) are all well less than those from non-spatial models (an intercept with error model without spatial component, and details can be found in Diggle and Ribeiro (2007)), indicating the necessity of the spatial structure in pollution model. The main results from pollution model are the population-weighted county level exposure for each pollutants, with an example of NO_2 exposure being shown in Figure 1d that the east part of Germany has much higher NO_2 exposure levels.

Table 2: Parameter estimation from pollution model.

	μ	σ^2	τ^2	η	AIC	Non-spatial AIC
no2	20.181	60.532	107.134	0.537	4501.620	4644.213
pm25	10.395	1.156	1.525	0.736	741.625	771.677
pm10	17.483	4.375	10.265	0.554	2138.916	2178.230
so2	1.507	0.751	0.115	0.681	301.668	365.301
Benzene	0.846	0.022	0.093	1.584	96.896	107.443
Aresenic	0.446	0.014	0.031	1.473	-71.610	-54.280
Cadmium	0.110	0.001	0.002	1.159	-532.236	-499.639
Nickel	1.468	0.532	1.239	0.922	603.834	631.438
Temperature	9.810	1.527	0.245	0.857	1783.719	2175.323

4.2 Health model validation

Before presenting the health effects results from health model (8), we assess the necessity of allowing for spatial autocorrelation in the disease data via the random effects (9), by fitting a simplified version of that model without spatial random effects term ϕ_k. The residuals from this model show substantial spatial autocorrelation, with significant Moran's I statistics (see Figure 3) (Moran (1950)). The empirical semi-variogram of the residuals in Figure 3b also shows that few points are lying outside the 95% Monte Carlo simulation envelopes, suggesting strong spatial autocorrelation is remained in the residuals and including the spatial random effect term (9) in health model is appropriate.



(a) Moran's I test, p-value<0.0001 (b) Empirical semi-variogram

Figure 3: The Moran's I test and the empirical semi-variogram of the residuals from the non-spatial geostatistical model (circles), with 95% Monte Carlo simulation envelopes (dashed lines).

4.3 Pollution health effects

In this section, we present the air pollution health effects, which are the main results in this study. For comparison purposes, we show both the results from our employed health model with the Leroux et al. (1999) CAR model to account for spatially correlated residuals (referred to as Leroux), and the results from other commonly used CAR models, including the intrinsic autoregressive proposed by Besag et al. (1991) (referred to as Besag), convolution model also proposed by Besag et al. (1991) (referred to as Bym), and also the non-spatial model (referred to as iid). In addition, as PM_{10} and $PM_{2.5}$ are highly correlated (with correlation coefficient being 0.75 in our study), we run two separated health model to avoid collinearity, with each model including either PM_{10} or $PM_{2.5}$ and all the other covariates. The results from having $PM_{2.5}$ in the model are presented in this section in Table 3, while those from having PM_{10} are presented in the Appendix in Table 5. The health model was implemented in INLA (Rue et al. (2009), *www.r-inla.org*) which uses the Integrated Nested Laplace Approximation to implement Bayesian models (Blangiardo et al. (2013)).

Figure 4a and 4b show both the prior and the posterior distributions of the spatial dependence parameter ξ and variance parameter κ from fitting the health model (having $PM_{2.5}$), respectively, suggesting that both of them are well estimated from the data. Figure 4a shows that the estimate of ξ is around 0.8 which indicates high spatial autocorrelation in the disease data after the covariate effects have been accounted for, validating the use of the spatial random effects model. Similarly, Figure 4b shows the estimate of the spatial variance parameter κ is around 0.75. The estimated COVID-19 SIR from Figure 4c from health model (8) shows that the majority of the high-risk counties are at the east and south part of Germany, that is where likely to have remaining positive random effects (see Figure 4d).

The main results are presented in Table 3, including the posterior medians and 95% credible intervals of relative risk from one-unit increase in each covariate, and the widely applicable information criterion (WAIC) (Watanabe (2010)) from fitting various health model, including the employed Leroux model, and commonly used Bym, Besag, iid models. Table 3 shows that the WAIC from different models are similar, while it is slightly lower (better) from the currently used Leroux model. The results from Leroux model show that NO_2 is significantly (at 0.05 level) associated with the COVID-19 SIR, with 1 μgm^{-3} increase of long-term exposure to NO_2, the COVID-19 incidence rate is likely to increase 5.61% (95% CI: 3.36%, 7.91%). This statistically significant association between NO_2 and COVID-19 SIR is robust across various health models, including the Bym, Besag, iid models (and also those from Table 5 where health model has PM_{10} rather than $PM_{2.5}$), which enhances the plausibility of the inference.

The areal population density dose not have a significant association with COVID-19 SIR, while temperature displays a negative association (at 0.05 level) with COVID-19 incidence rate. As shown in Figure 1, the COVID-19 SIRs are generally higher in the south where actually has a lower long-term temperature surface compared to the north (https://www.mapsofworld.com/germany/thematic-maps/germany-temperature-map.html). This explain the negative associations

between temperature and COVID-19 incidence rate. No substantial associations (at 0.05 level) were found between COVID-19 incidence rate and the other pollutants, including $PM_{2.5}$, SO_2, Benzene, Aresenic, Cadmium and Nickel. Note that SO_2 is just at the border to have a significant association with COVID-19 SIR, since the posterior probabilities of its increasing relative risk is 0.96 (see Table 3). And it is significant associated from the model having PM_{10} (see Table 3).

Table 3: Posterior medians and 95% CI for the relative risk (%) from one-unit increase in each covariate, and the WAIC from fitting various health model (having $PM_{2.5}$), including the employed Leroux model, and commonly used Bym, Besag, iid models. The result format is: point estimate of relative risk (lower estimate, upper estimate), [the posterior probabilities that covariate increase relative risk].

	Leroux	**Bym**	**Besag**	**iid**
no2	5.61 (3.36, 7.91), [1.00]	5.37 (3.06, 7.74), [1.00]	5.40 (3.08, 7.78), [1.00]	5.54 (3.86, 7.24), [1.00]
pm25	4.91 (-12.41, 25.34), [0.70]	1.01 (-17.09, 22.99), [0.54]	0.78 (-17.37, 22.90), [0.53]	8.16 (-2.66, 20.18), [0.93]
so2	16.02 (-1.57, 36.07), [0.96]	5.47 (-10.77, 24.62), [0.73]	5.33 (-10.98, 24.60), [0.73]	40.07 (26.78, 54.75), [1.00]
Temperature	-11.96 (-21.10, -1.66), [0.01]	-8.77 (-18.52, 2.14), [0.05]	-8.77 (-18.55, 2.18), [0.06]	-18.00 (-24.47, -10.98), [0.00]
Benzene	-0.92 (-19.06, 20.60), [0.46]	-3.04 (-24.26, 24.09), [0.40]	-3.17 (-24.51, 24.20), [0.40]	9.58 (-1.41, 21.79), [0.96]
Aresenic	-16.74 (-31.54, 1.77), [0.04]	-10.08 (-27.75, 11.86), [0.17]	-10.23 (-27.99, 11.88), [0.17]	-22.55 (-31.03, -13.03), [0.00]
Cadmium	16.57 (-5.67, 44.85), [0.92]	26.70 (-0.10, 60.76), [0.97]	27.24 (0.15, 61.64), [0.98]	7.40 (-5.08, 21.51), [0.87]
Nickel	-1.67 (-13.48, 11.75), [0.40]	-1.89 (-14.50, 12.56), [0.39]	-1.92 (-14.63, 12.67), [0.39]	-1.65 (-9.12, 6.43), [0.34]
popDensity	-2.74 (-7.98, 2.78), [0.16]	-2.52 (-7.69, 2.94), [0.18]	-2.46 (-7.62, 2.99), [0.18]	-6.62 (-11.83, -1.12), [0.01]
WAIC	3803.84	3805.05	3805.39	3804.41

5 Discussion

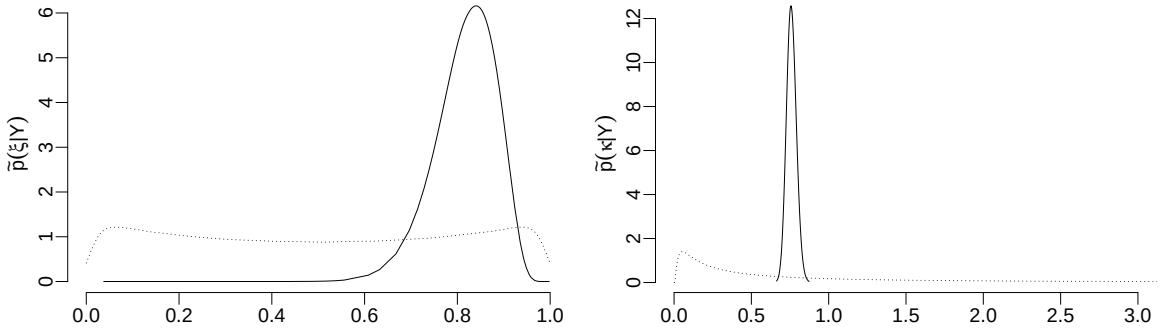
"Poisoning our environment means poisoning our own body, and when it experiences chronic respiratory stress its ability to defend itself from infections is limited" (Ogen (2020)). Given that the existing research has linked pollutants (e.g., $PM_{2.5}$ and NO_2) exposure to health damage, particularly respiratory and lung diseases, which could make people more vulnerable to contract COVID-19. This study uses a spatial ecological design to estimate the impacts of air pollution on COVID-19 infection by utilizing geographical contrasts in air pollution and infection risk across K contiguous small areas, and uses population-weighted method to better estimate the real areal pollution concentrations. The results show that long-term exposure to NO_2 is significantly associated with COVID-19 incidence rate, with 1 μgm^{-3} increase of long-term exposure to NO_2, the COVID-19 incidence rate is likely to increase 5.61% (95% CI: 3.36%, 7.91%). No substantial associations were found between COVID-19 incidence rate and the other pollutants, including $PM_{2.5}$, PM_{10}, SO_2, Benzene, Aresenic, Cadmium and Nickel. Our results are based on population-weighted average exposure, which better estimates the real areal pollution exposure. In addition, temperature and population density are adjusted in the model, and a set of random effects are also included to capture the residual spatial autocorrelation after the covariate effects have been accounted for.

For comparison purposes, we also run the health models with other commonly used CAR models, including the intrinsic autoregressive proposed by Besag et al. (1991), convolution model also proposed by Besag et al. (1991), and also the non-spatial model. In addition, as PM_{10} and $PM_{2.5}$ are highly correlated, we run two separated health model to avoid collinearity, with each model including either PM_{10} or $PM_{2.5}$ and all the other covariates. We found that the statistically significant associations between NO_2 and COVID-19 SIR are robust across these various health models, which enhances the plausibility of the inference.

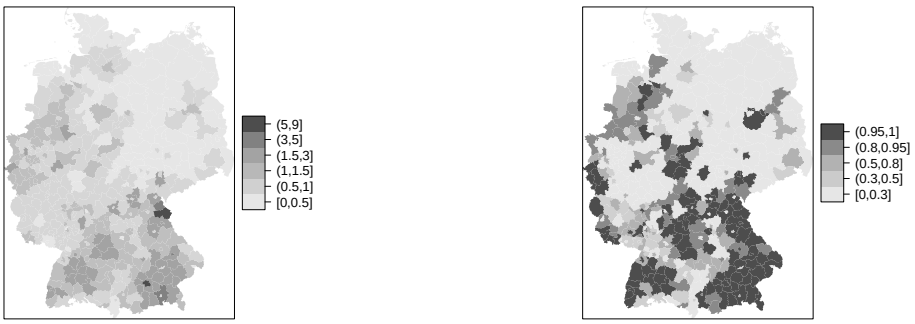
Several limitations to this pilot study need to be acknowledged. First, due to data availability,

no socioeconomic or health care related covariates were included in the health model which, if included, would provide the possibility of sensitivity analyses and help testing the robustness of the findings. However, in our health model, we do include a spatial random effects term to allow any spatial autocorrelation residuals after accounting for the known covariates, and the main findings of NO_2 are actually adjusted for a set of other pollutants. Another limitation is lacking COVID-19 testing numbers. The confirmed cases (positive testing numbers) mainly rely on the total testing numbers being conducted, without which the infection rates in some counties could be higher or lower estimated compared to others'. Therefore, we put an effort to better estimate the expected cases in each county by utilizing national sex-age standardized infection rate.

Finally, besides COVID-19 infection rate, its death rate and multi-country studies should also be focused when (if) more deaths occur in the future. Such studies will help us better understanding COVID-19, and also help the global community and health organizations stay informed and make data driven decisions.



(a) ξ (b) κ



(c) $\hat{R}_k$ (d) $\Pr(\hat{\phi}_k > 0)$

Figure 4: Posterior density plots for ξ and κ from health model (dashed line is the prior), and the estimated COVID-19 SIR and the posterior probability of positive random effects.

Appendix

The publicly available data sources used in the analysis can be seen in Table 4, while the results from fitting health model having PM_{10} are shown in Table 5. The raw data and R code used in this study will be shared on github shortly.

Table 4: Publicly available data sources used in the analysis.

Source	**Link**
Federal Agency for Cartography and Geodesy	https://www.bkg.bund.de/EN/Home/home.html
Robert Koch Institute	https://www.rki.de/EN/Home/homepage_node.html
National Platform for Geographic Data	https://npgeo-corona-npgeo-de.hub.arcgis.com/
Federal Office for Statistics: Statistisches Bundesamt	https://www.destatis.de/EN/Home/_node.html
Open Data platform GENESIS	https://www-genesis.destatis.de/genesis/online
City Population	https://www.citypopulation.de/en/germany/admin/
DIVA-GIS	https://www.diva-gis.org/gdata
European Environment Agency: Air Quality e-Reporting	https://www.eea.europa.eu/

Table 5: Posterior medians and 95% CI for the relative risk (%) from one-unit increase in each covariate, and the WAIC from fitting various health model (having PM_{10}), including the employed Leroux model, and commonly used Bym, Besag, iid models. The result format is: point estimate of relative risk (lower estimate, upper estimate), [the posterior probabilities that covariate increase relative risk].

	Leroux	**Bym**	**Besag**	**iid**
no2	6.06 (3.47, 8.69), [1.00]	5.39 (2.80, 8.05), [1.00]	5.57 (2.92, 8.30), [1.00]	6.81 (4.89, 8.78), [1.00]
pm10	-2.66 (-12.27, 8.08), [0.30]	-0.68 (-10.74, 10.47), [0.45]	-1.25 (-11.55, 10.23), [0.41]	-7.55 (-13.87, -0.77), [0.01]
so2	18.97 (0.71, 39.62), [0.98]	6.65 (-9.33, 25.40), [0.78]	6.00 (-10.42, 25.40), [0.75]	51.12 (37.44, 66.16), [1.00]
Temperature	-11.19 (-20.66, -0.51), [0.02]	-8.50 (-18.37, 2.57), [0.06]	-8.41 (-18.50, 2.93), [0.07]	-16.17 (-22.98, -8.77), [0.00]
Benzene	-0.44 (-18.49, 20.96), [0.48]	-2.12 (-22.48, 23.53), [0.43]	-2.91 (-23.95, 23.91), [0.40]	8.84 (-2.07, 20.94), [0.94]
Aresenic	-13.85 (-29.39, 5.36), [0.07]	-8.81 (-26.32, 12.79), [0.20]	-9.48 (-27.51, 13.01), [0.19]	-13.25 (-23.34, -1.84), [0.01]
Cadmium	13.72 (-7.97, 41.63), [0.88]	24.10 (-1.47, 56.43), [0.97]	26.55 (-0.46, 60.88), [0.97]	-0.71 (-12.03, 12.05), [0.45]
Nickel	-1.03 (-12.87, 12.39), [0.43]	-1.58 (-13.76, 12.30), [0.40]	-1.73 (-14.41, 12.81), [0.40]	-1.35 (-8.80, 6.71), [0.37]
popDensity	-2.70 (-7.94, 2.82), [0.16]	-2.74 (-7.89, 2.70), [0.16]	-2.45 (-7.62, 3.00), [0.18]	-6.37 (-11.56, -0.87), [0.01]
WAIC	3803.76	3804.22	3805.34	3803.93

References

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Selected Papers of Hirotugu Akaike*. New York, NY: Springer New York, pp. 199–213.

Arbia, G. (1988). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Springer.

Banerjee, S., B. Carlin, and A. Gelfand (2004). *Hierarchical modelling and analysis for spatial data (1st ed)*. Chapman and Hall.

Besag, J., J. York, and A. Mollie (1991). "Bayesian image restoration with two applications in spatial statistics". In: *Annals of the Institute of Statistics and Mathematics* 43, pp. 1–59.

Blair, A., P. Stewart, J. H. Lubin, and F. Forastiere (2007). "Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures". In: *American Journal of Industrial Medicine* 50.3, pp. 199–207.

Blangiardo, M., M. Cameletti, G. Baio, and H. Rue (2013). "Spatial and spatio-temporal models with R-INLA". In: *Spatial and Spatio-temporal Epidemiology* 4, pp. 33–49.

Blangiardo, M., F. Finazzi, and M. Cameletti (2016). "Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions". In: *Spatial and Spatio-temporal Epidemiology* 18. Environmental Exposure and Health, pp. 1–12. URL: http://www.sciencedirect.com/science/article/pii/S1877584515300460.

Bowatte, G., B. Erbas, C. J. Lodge, L. D. Knibbs, L. C. Gurrin, G. B. Marks, P. S. Thomas, D. P. Johns, G. G. Giles, J. Hui, M. Dennekamp, J. L. Perret, M. J. Abramson, E. H. Walters, M. C. Matheson, and S. C. Dharmage (2017). "Traffic-related air pollution exposure over a 5-year period is associated with increased risk of asthma and poor lung function in middle age". In: *European Respiratory Journal* 50.4. eprint: https://erj.ersjournals.com/content/50/4/1602357.full.pdf. URL: https://erj.ersjournals.com/content/50/4/1602357.

Committee on the Medical Effects of Air Pollutants (2010). *The Mortality Effects of Long-Term Exposure to Particulate Air Pollution in the United Kingdom*. Crown.

Cressie, N. (1993). *Statistics for Spatial Data*. revised. New York: Wiley.

Das, K., S. Das, and S. Dhundasi (Oct. 2008). "Nickel, its adverse health effects & oxidative stress". In: *The Indian journal of medical research* 128.4, pp. 412–425.

Diggle, P. and P. Ribeiro (2007). *Model-based Geostatistics*. Springer Series in Statistics, Springer.

Elliott, P., J. Wakefield, N. Best, and D. Briggs (2000). *Spatial epidemiology: methods and applications (1st ed)*. Oxford University Press.

Gryparis, A., C. J. Paciorek, A. Zeka, J. Schwartz, and B. A. Coull (Oct. 2008). "Measurement error caused by spatial misalignment in environmental epidemiology". In: *Biostatistics* 10.2, pp. 258–274.

Heads or Tails (Sept. 2020). *COVID-19 Tracking Germany, version 150*. URL: https://www.kaggle.com/headsortails/covid19-tracking-germany/version/150.

Huang, G., D. Lee, and E. M. Scott (2018). "Multivariate space-time modelling of multiple air pollutants and their health effects accounting for exposure uncertainty". In: *Statistics in Medicine* 37.7, pp. 1134–1148.

Järup, L., M. Berglund, C. G. Elinder, G. Nordberg, and M. Vanter (1998). "Health effects of cadmium exposure – a review of the literature and a risk estimate". In: *Scandinavian Journal of Work, Environment & Health* 24, pp. 1–51.

Lawson, A. (2008). *Bayesian disease mapping: hierarchical modelling in spatial epidemiology (1st ed)*. Chapman and Hall.

Lee, D. (2011). "A comparison of conditional autoregressive models used in Bayesian disease mapping". In: *Spatial and Spatio-temporal Epidemiology* 2.2, pp. 79–89.

– (2012). "Using spline models to estimate the varying health risks from air pollution across Scotland". In: *Statistics in Medicine* 31.27, pp. 3366–3378.

Lee, D., C. Ferguson, and R. Mitchell (2009). "Air pollution and health in Scotland: a multicity study". In: *Biostatistics* 10.3, pp. 409–423.

Lee, D., S. Mukhopadhyay, A. Rushworth, and S. K. Sahu (Apr. 2017). "A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health". In: *Biostatistics (Oxford, England)* 18.2, pp. 370–385.

Lee, D., C. Robertson, C. Ramsay, and K. Pyper (2020). "Quantifying the impact of the modifiable areal unit problem when estimating the health effects of air pollution". In: *Environmetrics*.

Leroux, B., X. Lei, and N. Breslow (1999). "Estimation of disease rates in small areas: A new mixed model for spatial dependence". In: Springer-Verlag, New York. Chap. Statistical Models in Epidemiology, the Environment and Clinical Trials, Halloran, M and Berry, D (eds), pp. 135–178.

Maheswaran, R., R. Haining, T. Pearson, J. Law, P. Brindley, and N. Best (2006). "Outdoor NOx and stroke mortality adjusting for small area level smoking prevalence using a Bayesian approach". In: *Statistical Methods in Medical Research* 15, pp. 499–516.

Moran, P. (1950). "Notes on continuous stochastic phenomena". In: *Biometrika* 37, pp. 17–23.

Napier, G., D. Lee, C. Robertson, and A. Lawson (June 2018). "A Bayesian space–time model for clustering areal units based on their disease trends". In: *Biostatistics* 20.4, pp. 681–697.

Ogen, Y. (2020). "Assessing nitrogen dioxide (NO2) levels as a contributing factor to coronavirus (COVID-19) fatality". In: *Science of The Total Environment* 726, p. 138605. URL: http://www.sciencedirect.com/science/article/pii/S0048969720321215.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory And Applications (Monographs on Statistics and Applied Probability)*. Chapman & Hall/CRC.

Rue, H., S. Martino, and N. Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 319–392.

Rushworth, A., D. Lee, and R. Mitchell (2014). "A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London". In: *Spatial and Spatio-temporal Epidemiology* 10, pp. 29–38.

Sacks, J. D., A. G. Rappold, J. Allen Davis Jr., D. B. Richardson, A. E. Waller, and T. J. Luben (2014). "Influence of urbanicity and county characteristics on the association between ozone and asthma emergency department visits in North Carolina". In: *Environ Health Perspect* 122.5, pp. 506–512.

Schikowski, T., D. Sugiri, U. Ranft, U. Gehring, J. Heinrich, H.-E. Wichmann, and U. Kraemer (2005). "Long-term air pollution exposure and living close to busy roads are associated with COPD in women". In: *Respiratory Research* 6, pp. 152–152.

Shaddick, G. and J. Zidek (June 2015). "Spatio-Temporal Methods in Environmental Epidemiology". English. In: CRC Texts in Statistical Science.

Smith, M. T. (2010). "Advances in Understanding Benzene Health Effects and Susceptibility". In: *Annual Review of Public Health* 31.1, pp. 133–148.

Vinikoor-Imler, L. C., J. A. Davis, R. E. Meyer, L. C. Messer, and T. J. Luben (2014). "Associations between prenatal exposure to air pollution, small for gestational age, and term low birthweight in a state-wide birth cohort". In: *Environmental Research* 132, pp. 132–139.

Wakefield, J. and R. Salway (2001). "A Statistical Framework for Ecological and Aggregate Studies". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 164.1, pp. 119–137. URL: http://www.jstor.org/stable/2680539.

Warren, J., M. Fuentes, A. Herring, and P. Langlois (2012). "Bayesian spatial-temporal model for cardiac congenital anomalies and ambient air pollution risk assessment". In: *Environmetrics* 23.8, pp. 673–684.

Watanabe, S. (Dec. 2010). "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory". In: *J. Mach. Learn. Res.* 11, pp. 3571–3594.

Wu, X., R. C. Nethery, B. M. Sabath, D. Braun, and F. Dominici (2020). "Exposure to air pollution and COVID-19 mortality in the United States". In: *medRxiv*. URL: https://www.medrxiv.org/content/10.1101/2020.04.05.20054502v2.

Yu, W. H., C. M. Harvey, and C. F. Harvey (2003). "Arsenic in groundwater in Bangladesh: A geostatistical and epidemiological framework for evaluating health effects and potential remedies". In: *Water Resources Research* 39.6. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002WR001327. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002WR001327.