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. Within countries, although the measures taken by governments are similar, 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 5.58% (95% credible interval [CI]: 3.35%, 7.86%). This result is consistent across various health models. The analyses can be reproduced and updated routinely using public data sources and shared 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

COVID-19, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is currently widespread and much more dangerous than seasonal flu. COVID-19 is more infectious than seasonal flu and has a higher death rate. Up to 17th September, 2020, it has led to worldwide over 30.3 million cases and 950 thousand deaths. In Germany, the total confirmed cases up to 17th September, 2020 have risen to 269 thousand, with 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.

1.1 Literature review on epidemiology

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 contracting 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 D. 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 sand 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 adverse health effects. Since the particulate matter can be inhaled into the thoracic region of the respiratory tract, there is a plausible reason the relationship could be causal. Examples include D. Lee et al. (2009) and D. 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 adverse health effects because they can travel deeper into lungs.

1.2 Literature review on statistical models

A spatial ecological design can be used to estimate the impacts of air pollution on health by comparing 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; D. 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 (D. Lee et al. 2020). The simulation study from D. 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.

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; D. Lee et al. 2017; Blangiardo et al. 2016; Gryparis et al. 2009).

In this study, we investigate whether long-term average exposure to air pollution increases the risk of COVID-19 infection in Germany using a spatial ecological design. Specifically, in order to reduce the potential ecological bias, we better estimate the real areal pollution concentrations by first using Kriging to pollution monitoring data to obtain predictions on fine 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. This will likely enhance the estimation of people's real exposure for those counties where they generally live at rural areas while their urban pollution are much worse compared to the rural areas. Given that the study from D. 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 Study region

The study region is Germany which has a population of around 83 million people and $K = 401$ counties (administrative districts), among which 294 are rural and 107 are urban. A map of these counties is shown in Figure 1, showing boundaries obtained from Germany's Federal Agency for

Cartography and Geodesy (BKG 2020).

2.1 Data description

The data set used in this study include COVID-19 cases, pollution concentrations, temperature and population data. The accumulated COVID-19 cases used in this study are collected up to 2020-09-13 at the county-level. Both pollution and temperature data are the most recent available few years (2016-2018) average concentrations (representing long-term exposure) from monitoring sites, which are converted to county-level by applying 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).

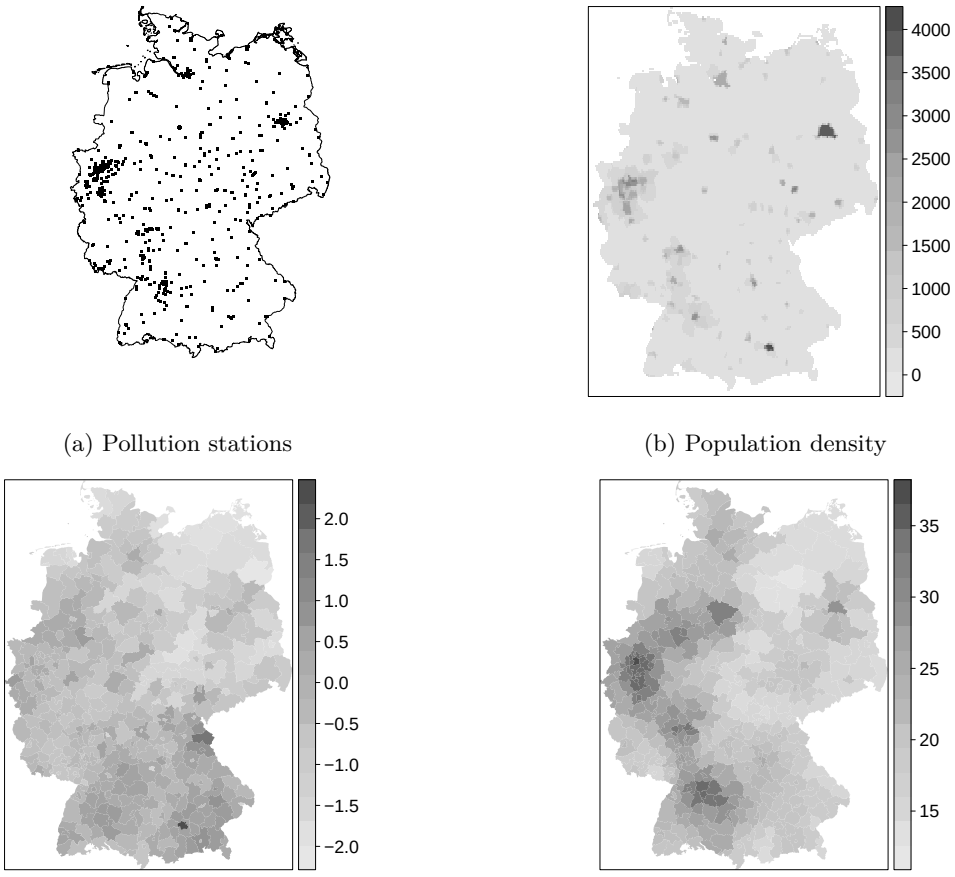
The population data contain fine gridded population density data, the population by sex and age on the federal state level and also the county level population data. The fine gridded population density data are used for calculating population-weighted county level exposure (see Section 3 for details), while the latter two are used to calculate the expected number of cases in each county. Specifically, we denote Y_k as the reported numbers of COVID-19 cases for county k, and 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 $s(k)$. Then we use standardized incidence ratio (SIR) given by $SIR_k = Y_k/E_k$, 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 2020-09-13 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.

2.2 Data sources

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), where the COVID-19 cases and deaths will be updated daily, with the earliest recorded cases are from 2020-01-24. More details on the original sources of COVID-19 and state level population data can be found in Heads or Tails (2020). The county level population data are freely obtained from the City Population (Citypopulation 2019). Both two population data sets reflect the (most recent available) estimates on 2018-12-31. The fine gridded population density data are freely available on DIVA-GIS (DIVA-GIS 2020), which is shown in Figure 1b.

Pollution data are available from the Air Quality e-Reporting provided by European Environment Agency (EEA 2020), where the monitoring stations are shown in Figure 1a which tend to be dense where the population density is high (see Figure 1b). The monitoring temperature data can be freely downloaded from European Climate Assessment & Dataset (ECAD 2020).



(a) Pollution stations

(b) Population density

(c) log(SIR) in counties

(d) NO_2 in counties

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

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 (D. Lee 2011). To overcome this problem, the Poisson log-linear spatial models are typically used for the analysis (Elliott et al. 2000; Banerjee et al. 2004; Lawson 2008), where 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.

These 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 $K \times K$ neighbourhood matrix $\mathbf{W}$. Based on this neighbourhood matrix, the most common models for the random effects include intrinsic autoregressive (Besag et al. 1991), convolution 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

For simplicity, in this study we use a univariate model for each pollutant, since the number of monitoring stations is 709 which is a large sample size and produces predictions with modest standard errors. 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 $\mathbf{Z} = \{Z(\mathbf{s}); \mathbf{s} = \mathbf{s}_1, \ldots, \mathbf{s}_n\}$, and write

$\mathbf{S} = \{S(\mathbf{s}); \mathbf{s} = \mathbf{s}_1, \ldots, \mathbf{s}_n\}$ for the unobserved values of the signal at the sampling locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, 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. $\mathbf{S}$ 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 of (1) 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 optimized 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 Population-weighted 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 have the inverse variance for the prediction, $\zeta(\mathbf{s}_0) = \frac{1}{\mathrm{Var}(\hat{S}(\mathbf{s}_0) - S(\mathbf{s}_0))}$. The higher $\zeta(\mathbf{s}_0)$ is, the better quality the prediction has, and we give more weight to the most reliably pollution values while aggregating them (Sanchez-Meca and Marín-Martínez 1998; C. H. Lee et al. 2016).

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

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

where A_k represents county k. Therefore, z_k is a spatial metric of pollution concentrations weighted by population density and also the inverse of their Kriged variances.

3.3 Health model

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

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

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 spatial random effect ϕ_k. The regression parameters $\boldsymbol{\beta}$ are assigned weakly informative zero-mean Gaussian priors with diagonal variance matrix $\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, 10^2\mathbf{I})$.

The spatial random effect, ϕ_k, is 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 CAR model proposed by Leroux et al. (1999). The spatial dependence in the data is captured by an $K \times K$ 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

$$\begin{aligned} \kappa &\sim \text{Exp}[\log(2)], \\ \log(\frac{\xi}{1-\xi}) &\sim \text{N}(0, 1.8). \end{aligned} \tag{10}$$

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. Health models are implemented in INLA which uses the Integrated Nested Laplace Approximation (Rue et al. 2009), 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 1 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 independent errors, without spatial component R in model (3)), 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 pollutant, with an example of NO_2 population-weighted exposure being shown in Figure 1d that the east part of Germany has much higher NO_2 exposure levels. A summary of the estimated population-weighted county level exposure is presented in Table 2, and an exploratory of the scatterplots of the natural

Table 1: Parameter estimation from pollution model (1).

	μ	σ^2	τ^2	η	AIC	Non-spatial AIC
NO_2	20.181	60.532	107.134	0.537	4501.620	4644.213
$PM_{2.5}$	10.395	1.156	1.525	0.736	741.625	771.677
PM_{10}	17.483	4.375	10.265	0.554	2138.916	2178.230
SO_2	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

Table 2: Population-weighted county level exposure summary, with unit μgm^{-3} for NO_2, PM_{25}, PM_{10}, SO_2, Benzene; ngm^{-3} for Aresenic, admium, Nickel; and $^{\circ}C$ for temperature.

	Min	Quantile25	Median	Mean	Quantile75	Max
NO_2	12.57	18.79	21.62	23.03	26.86	36.54
$PM_{2.5}$	8.64	9.98	10.52	10.48	10.94	12.21
PM_{10}	14.48	16.84	17.74	17.74	18.57	21.07
SO_2	0.69	1.21	1.51	1.66	1.95	4.23
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

logarithm of COVID-19 SIR against the population-weighted NO_2 and $PM_{2.5}$ are displayed in the upper of Figure 2, which appears to indicate a linear relationship between NO_2 and log COVID-19 SIR.

4.2 Health model validation

Before presenting the health effects results from health model (8), we assess the necessity of including spatial autocorrelation via the random effects model (9) by fitting a simplified version of model

(8) without spatial random effects term ϕ_k. The residuals from this model show substantial spatial autocorrelation, with significant Moran's I statistics (see the middle of Figure 2) (Moran 1950). The empirical semi-variogram of the residuals in Figure 2d 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.

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 4.

The bottom of Figure 2 shows both the prior and posterior distributions of the spatial dependence parameter ξ and variance parameter κ from fitting the health model (8) (having $PM_{2.5}$), respectively, suggesting that both of them are well estimated from the data. Figure 2e 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 (9). Similarly, Figure 2f shows the estimate of the spatial variance parameter κ is around 0.75. The

predicted COVID-19 SIR presented in Figure 3a 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 the counties likely to have high probabilities of excess risk (see Figure 3b).

The main results are presented in Table 3, including the posterior medians and 95% credible intervals of relative risk from one-unit increase for each covariate, and the widely applicable information criterion (WAIC) (Watanabe 2010) from fitting various health models, including the employed Leroux model, and commonly used BYM, Besag, IID models. Table 3 shows that the WAIC from different CAR 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 5.58% (95% CI: 3.35%, 7.86%). This statistically significant association between NO_2 and COVID-19 SIR is consistent across various health models, including the BYM, Besag, IID models (and also those from Table 4 where health model has PM_{10} rather than $PM_{2.5}$), which enhances the plausibility of the results.

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 1c, the COVID-19 SIRs are generally higher in the south where actually has a lower long-term temperature surface compared to the north (see Maps of World 2020). This explains the negative association 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 of having a significant association with COVID-19 SIR, since the posterior probabilities of its increasing relative risk is 0.96 (see Table 3). And SO_2 is significantly associated with COVID-19 SIR from the model having PM_{10} rather than $PM_{2.5}$ (see the Leroux model results from Table 4).

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 models (having $PM_{2.5}$), including the employed Leroux model, and commonly used BYM, Besag, IID models. Pr is the posterior probabilities that covariate increases relative risk.

	Leroux			**BYM**			**Besag**			**IID**		
	Est	CI	Pr	Est	CI	Pr	Est	CI	Pr	Est	CI	Pr
NO_2	5.58	(3.35, 7.86)	[1.00]	5.21	(2.98, 7.50)	[1.00]	5.36	(3.06, 7.71)	[1.00]	5.60	(3.94, 7.29)	[1.00]
$PM_{2.5}$	4.59	(-12.57, 24.79)	[0.69]	1.62	(-15.76, 22.51)	[0.57]	0.45	(-17.49, 22.27)	[0.52]	8.04	(-2.70, 19.94)	[0.93]
SO_2	15.83	(-1.42, 35.45)	[0.96]	6.29	(-9.20, 24.36)	[0.78]	5.45	(-10.56, 24.31)	[0.74]	39.51	(26.44, 53.94)	[1.00]
Temperature	-11.72	(-20.84, -1.46)	[0.01]	-8.52	(-18.01, 2.05)	[0.05]	-8.48	(-18.23, 2.41)	[0.06]	-18.12	(-24.55, -11.16)	[0.00]
Benzene	-1.21	(-19.21, 20.12)	[0.45]	-2.75	(-23.00, 22.75)	[0.41]	-3.45	(-24.59, 23.58)	[0.39]	9.32	(-1.58, 21.41)	[0.95]
Aresenic	-16.72	(-31.45, 1.67)	[0.04]	-9.34	(-26.32, 11.47)	[0.17]	-10.27	(-27.89, 11.64)	[0.16]	-22.38	(-30.83, -12.92)	[0.00]
Cadmium	16.44	(-5.67, 44.53)	[0.92]	23.93	(-1.12, 55.49)	[0.97]	27.08	(0.21, 61.15)	[0.98]	6.86	(-5.49, 20.80)	[0.86]
Nickel	-1.35	(-13.13, 12.03)	[0.42]	-1.41	(-13.44, 12.26)	[0.41]	-1.56	(-14.22, 12.95)	[0.41]	-1.47	(-8.90, 6.57)	[0.35]
popDensity	-2.12	(-7.34, 3.39)	[0.22]	-2.23	(-7.37, 3.19)	[0.20]	-1.83	(-6.98, 3.61)	[0.25]	-6.15	(-11.35, -0.65)	[0.01]
WAIC		3814.64			3814.8			3816.18			3815.05	

5 Discussion

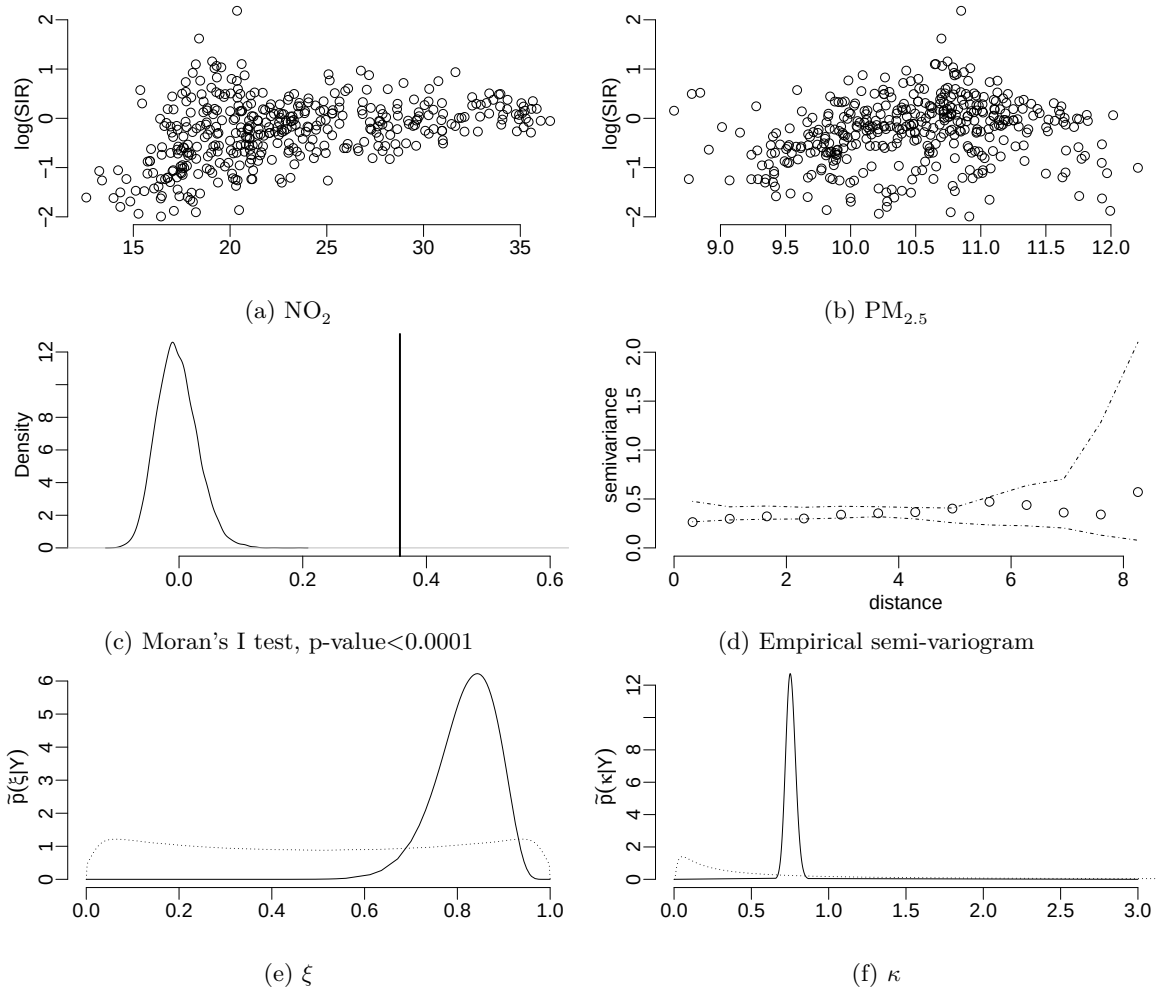
"Poisoning our environment means poisoning our own body, and when it experiences chronic respiratory stress its ability to defend itself from infections is limited" (Ogen 2020). Given that the existing research has linked pollutants (e.g., $PM_{2.5}$ and NO_2) exposure to health damage, particularly respiratory and lung diseases, which could make people more vulnerable to contracting COVID-19. This study uses a spatial ecological design to estimate the impacts of air pollution on COVID-19 infection in Germany by comparing geographical contrasts in air pollution and infection risk across K contiguous small areas, where we use 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 in Germany, with 1 μgm^{-3} increase of long-term exposure to NO_2, the COVID-19 incidence rate is likely to increase 5.58% (95% CI: 3.35%, 7.86%). 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. 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 IID model. In addition, as $PM_{2.5}$ and PM_{10} are highly correlated, we run two separated health models to avoid collinearity, with each model including either $PM_{2.5}$ or PM_{10} and all the other covariates. We found that the statistically significant associations between NO_2 and COVID-19 SIR are consistent across these various health models, which enhances the plausibility of the results.

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 for 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 and temperature. Another limitation is lacking COVID-19 testing numbers, since the confirmed cases (positive testing numbers) in a county mainly rely on the total testing numbers being conducted in that county, 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. The limitation in regard to pollution model is that the current one is a univariate model, which is potentially losing some power by not borrowing strength over correlated pollutants compared to a multivariate pollution model.

Finally, besides COVID-19 infection rate, its death rate and multi-country studies should also be focused when (or 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.

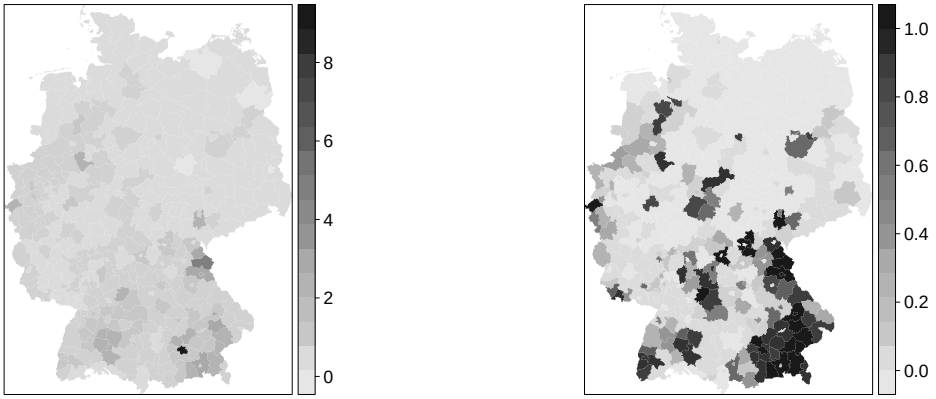


(a) NO_2 (b) $PM_{2.5}$

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

(e) ξ (f) κ

Figure 2: Upper: scatterplots of log COVID-19 SIR against NO_2 (μgm^{-3}) and $PM_{2.5}$ (μgm^{-3}); Middle: the Moran's I test and the empirical semi-variogram of the residuals from the non-spatial health model (circles), with 95% Monte Carlo simulation envelopes (dashed lines); Bottom: posterior (solid line) and prior (dashed line) plots for ξ and κ from health model (8).



(a) Predicted risk

(b) Exceedance probabilities

Figure 3: Posterior means of relative risk $\mathrm{E}(\lambda_k|Y)$ and probabilities of excess risk $\Pr(\exp(\phi_i) > 1.5 \mid Y)$.

Appendix

The results from fitting health models having PM_{10} are shown in Table 4. The raw data and R code used in this study will be shared on Github shortly.

Table 4: Posterior medians and 95% CI for the relative risk (%) from one-unit increase in each covariate, and the WAIC from fitting various health models (having $PM_{2.5}$), including the employed Leroux model, and commonly used BYM, Besag, IID models. Pr is the posterior probabilities that covariate increases relative risk.

	Leroux			**BYM**			**Besag**			**IID**		
	Est	CI	Pr	Est	CI	Pr	Est	CI	Pr	Est	CI	Pr
NO_2	6.06	(3.50, 8.67)	[1.00]	5.51	(2.91, 8.20)	[1.00]	5.57	(2.94, 8.26)	[1.00]	6.93	(5.02, 8.88)	[1.00]
PM_{10}	-2.98	(-12.49, 7.63)	[0.28]	-1.43	(-11.56, 9.81)	[0.40]	-1.61	(-11.78, 9.72)	[0.38]	-7.88	(-14.13, -1.17)	[0.01]
SO_2	18.82	(0.90, 39.05)	[0.98]	6.40	(-9.64, 25.21)	[0.77]	6.18	(-9.94, 25.18)	[0.76]	50.66	(37.18, 65.45)	[1.00]
Temperature	-10.89	(-20.34, -0.24)	[0.02]	-8.09	(-18.09, 3.13)	[0.07]	-8.05	(-18.12, 3.24)	[0.08]	-16.21	(-22.97, -8.87)	[0.00]
Benzene	-0.80	(-18.69, 20.39)	[0.47]	-3.00	(-23.67, 23.12)	[0.40]	-3.27	(-24.07, 23.22)	[0.39]	8.46	(-2.34, 20.43)	[0.94]
Aresenic	-13.74	(-29.20, 5.35)	[0.07]	-9.18	(-26.99, 12.85)	[0.19]	-9.42	(-27.34, 12.88)	[0.19]	-12.82	(-22.89, -1.44)	[0.01]
Cadmium	13.52	(-8.03, 41.21)	[0.88]	25.46	(-0.87, 59.05)	[0.97]	26.31	(-0.47, 60.27)	[0.97]	-1.37	(-12.55, 11.21)	[0.41]
Nickel	-0.70	(-12.51, 12.67)	[0.46]	-1.32	(-13.82, 12.95)	[0.42]	-1.37	(-14.00, 13.11)	[0.42]	-1.17	(-8.59, 6.85)	[0.38]
popDensity	-2.08	(-7.30, 3.43)	[0.22]	-1.92	(-7.09, 3.52)	[0.24]	-1.82	(-6.98, 3.61)	[0.25]	-5.89	(-11.08, -0.41)	[0.02]
WAIC		3814.55			3815.65			3816.13			3814.58	

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. P. Carlin, and A. E. Gelfand (2004). *Hierarchical Modeling and Analysis of Spatial Data (1st ed)*. Chapman and Hall/CRC Press.

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.

BKG (2020). "Federal Agency for Cartoraphy and Geodesy: Digitale Geodaten". In: URL: https://gdz.bkg.bund.de/index.php/default/digitale-geodaten.html.

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.

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.

Citypopulation (2019). "GERMANY: Administrative Division: States and Counties". In: URL: https://www.citypopulation.de/en/germany/admin/.

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.

DIVA-GIS (2020). "DIVA-GIS: Free Spatial Data". In: URL: https://www.diva-gis.org/gdata.

ECAD (2020). "The European Climate Assessment & Dataset". In: URL: https://www.ecad.eu/.

EEA (2020). "The European Environment Agency: Air Quality e-Reporting". In: URL: https://www.eea.europa.eu/.

Elliott, P., J. Wakefield, N. Best, and D. Briggs (2000). *Spatial Epidemiology: Methods and Applications (1st ed)*. Oxford University Press.

Gryparis, A., C. Paciorek, A. Zeka, J. Schwartz, and B. Coull (2009). "Measurement error caused by spatial misalignment in environmental epidemiology". In: *Biostatistics* 10.2, pp. 258–274.

Heads or Tails (2020). "COVID-19 Tracking Germany, version 156". In: URL: https://www.kaggle.com/headsortails/covid19-tracking-germany/version/156.

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. B. (2008). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology (1st ed)*. Chapman and Hall/CRC Press.

Lee, C. H., S. Cook, J. S. Lee, and B. Han (Dec. 2016). "Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores". In: *Genomics & Informatics* 14, p. 173.

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* 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.

Maps of World (2020). "Germany Weather Map". In: URL: https://www.mapsofworld.com/germany/thematic-maps/germany-temperature-map.html.

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 spaceâ time model for clustering areal units based on their disease trends". In: *Biostatistics*.

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.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. New York: Chapman and 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.

Sanchez-Meca, J. and F. Marín-Martínez (1998). "Weighting by Inverse Variance or by Sample Size in Meta-Analysis: A Simulation Study". In: *Educational and Psychological Measurement* 58.2, pp. 211–220.

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 (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.

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. (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.