Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Population-weighted exposure to air pollution and COVID-19 incidence in Germany

Guowen Huang*, Patrick E. Brown

Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada
Centre for Global Health Research, St Michael’s Hospital, Toronto, ON, Canada

A R T I C L E   I N F O

Article history:
Received 28 May 2020
Received in revised form 23 October 2020
Accepted 26 October 2020
Available online 3 November 2020

Keywords:
COVID-19
Air pollution
Health impacts
Kriging
INLA

A B S T R A C T

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 worsened 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 (NO₂) is significantly associated with COVID-19 incidence, with a 1 µg m⁻³ increase in long-term exposure to NO₂ increasing the COVID-19 incidence rate by 5.58% (95% credible interval [CI]: 3.35%, 7.86%). This result is consistent across various models. The analyses can be reproduced and updated routinely using public data sources and shared R code.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

COVID-19, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is currently widespread. It is much more dangerous than seasonal flu due to its high infection and death rates. Up to 20th October, 2020, it has led to over 40.5 million cases and 1,120,000 deaths worldwide. In Germany, the total confirmed cases up to 20th October, 2020 have risen to 377,000, with deaths being more than 9800. A recent study by Wu et al. (2020) investigated the impact of

* Corresponding author at: Centre for Global Health Research, St Michael's Hospital and Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada.
E-mail address: hwg0610209@gmail.com (G. Huang).
long-term average exposure to fine particulate matter (PM\textsubscript{2.5}) on the risk of COVID-19 deaths in the United States and found that an increase of 1 $\mu$g/m\textsuperscript{-3} in PM\textsubscript{2.5} was associated with an 8% (95% confidence interval, 2%, 15%) increase in the COVID-19 death rate. Ogen (2020) reported that most COVID-19 fatal cases occurred in those regions with the highest NO\textsubscript{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 pollution impacts

The existing body of research on the impacts of air pollution on human health has linked PM\textsubscript{2.5} and NO\textsubscript{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\textsubscript{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\textsubscript{2} can cause inflammation of the airways. Long-term exposure may affect lung function and respiratory symptoms. For example, research by Bowatte et al. (2017) indicates that long-term exposure to NO\textsubscript{2} was associated with increased risk of respiratory diseases, while Lee et al. (2009) show that long-term exposure to NO\textsubscript{2} was significantly associated with respiratory hospital admissions in Edinburgh and Glasgow, UK. Similarly, Schikowski et al. (2005) suggest that long-term exposure to air pollution from NO\textsubscript{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\textsubscript{10} and PM\textsubscript{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). Exposure to particulate matter is associated with respiratory and cardiovascular illness and mortality as well as other adverse health effects. Since 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 Lee et al. (2009) and Lee (2012), where the authors found that long-term exposure to PM\textsubscript{10} was significantly associated with respiratory hospital admissions. Recent reviews by the Committee on the Medical Effects of Air Pollutants (COMÉAP, 2010) have suggested exposure to PM\textsubscript{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 outcome data are counts of disease cases occurring in each areal unit while the pollution concentrations in each areal unit are typically estimated by applying Kriging (see 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.

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 uncertain as they are only estimates of the true 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 when modelling health outcomes (e.g., Huang et al., 2018; 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 true areal pollution concentrations by first applying Kriging to pollution monitoring data to obtain predictions on a fine grid 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 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 \( \text{PM}_{10} \), 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

2.1. Data description

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 Fig. 1, showing boundaries obtained from Germany’s Federal Agency for Cartography and Geodesy (BKG, 2020).

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 13th, September, 2020 at the county level. Both pollution and temperature data are average concentrations for the years 2016–2018 (representing long-term exposure) from monitoring sites, which are converted into 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 \( \text{PM}_{2.5}, \text{PM}_{10}, \text{NO}_2 \) and \( \text{SO}_2 \); and also four poisonous pollutants benzene, arsenic, cadmium and nickel. 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 densities, 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 \( E_k = \frac{P_k}{\bar{P}_{s(k)}} \sum_j r_j \bar{P}_{s(k),j} \), where \( P_k \) is the population in county \( k \), \( r_j \) is the national incidence rate in sex–age group \( j \), and \( \bar{P}_{s(k)} = \sum_j \bar{P}_{s(k),j} \) denotes the population of the state which contains county \( k \). The latter part in the equation \( \sum_j r_j \bar{P}_{s(k),j} \) is the expected number of cases in state \( s(k) \). The standardized incidence ratio (SIR) given by \( \text{SIR}_k = Y_k/E_k \), measures 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 13th September, 2020 can be seen in Fig. 1(c), showing a wide variation in SIRs across the counties in Germany and the majority of the high-risk counties are at the southern and northwestern parts 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 from Kaggle (Heads or Tails, 2020). The COVID-19 cases and deaths are updated daily, with the earliest recorded cases are from 24th January, 2020. The COVID-19 data are originally collected by the Robert Koch Institute, with more details can be found in Heads or Tails (2020). The county level population data are freely available from The City Population (2019). Both 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 (2020), and is shown in Fig. 1(b).

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

3. Method

The observed and expected case counts for each areal unit are used to calculate the standardized incidence ratio, with $\text{SIR}_k = Y_k/E_k$, where $\text{SIR}_k > 1$ represents areas with elevated levels of disease risk, while $\text{SIR}_k < 1$ corresponds to comparatively healthy areas. 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, 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 variance of the estimated risk and the likelihood of excesses estimated 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 (see Rue and Held, 2005). The spatial correlation between the random effects is determined by a binary $K \times K$ neighbourhood matrix $W$. Based on this neighbourhood matrix, the most common models for the random effects include intrinsic autoregressive model (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 fairly large (709) to produce predictions with modest standard errors. We treat the underlying pollution levels in Germany as a spatial Gaussian process $\{S(s) : s \in \mathbb{R}^2\}$ with mean $\mu$, variance $\sigma^2 = \text{Var}(S(s))$ and correlation function $\rho(u) = \text{Corr}(S(s_i), S(s_j)) = \exp(-u/\nu)$, where $u = \|s_i - s_j\|$ denotes the Euclidean distance between $s_i$ and $s_j$. Denote the observed pollution data as $Z = \{Z(s) ; s = s_1, \ldots, s_n\}$, and write $S = \{S(s) ; s = s_1, \ldots, s_n\}$ for the unobserved values of the signal at the sampling locations $s_1, \ldots, s_n$, the pollution model is assumed as

$$Z = S + \epsilon, \quad (1)$$

where $\epsilon(s) \sim N(0, \tau^2 I)$ is uncorrelated with $S$, and $I$ is the identity matrix of size $n$. $S$ is multivariate Gaussian with mean vector $\mu 1$, where $1$ 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 $r_{ij} = \rho(\|s_i - s_j\|)$. Similarly, $Z$ is multivariate Gaussian

$$Z \sim N(\mu 1, \sigma^2 V)$$

$$V = R + \nu^2 I, \quad (2)$$

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

The log-likelihood corresponding to (1) is

$$L(\mu, \tau^2, \sigma^2, \eta) = -0.5(n \log(2\pi) + \log(\sigma^2 R(\eta) + \tau^2 I) + \langle Z - \mu 1, (\sigma^2 R(\eta) + \tau^2 I)^{-1}(Z - \mu 1) \rangle) \quad (3)$$

Given $V$, the maximum likelihood estimate (MLE) of $\mu$ and $\sigma^2$ is given by,

$$\hat{\mu}(V) = (1V^{-1} 1)^{-1} 1V^{-1}Z \quad (4)$$

$$\hat{\sigma}^2(V) = n^{-1}(Z - \hat{\mu} 1)V^{-1}(Z - \hat{\mu} 1).$$

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

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

which can be optimized numerically with respect to $\eta$ and $\nu$, 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 by the precision of the predictions. For a new location \( s_0 \), the Kriging formula of \( \hat{S}(s_0) \) (Diggle and Ribeiro, 2007) is used to obtain its prediction by plugging-in the resulting estimates \( \hat{\mu}, \hat{\sigma}^2, \hat{\tau}^2, \hat{\eta} \), which is

\[
\hat{S}(s_0) = \hat{\mu} + C_{s_0} (\hat{\sigma}^2 \hat{V})^{-1} (Z - \hat{\mu} 1),
\]

where \( C_{s_0} = \hat{\sigma}^2 (\exp(-||s_1-s_0||/\hat{\eta}), \ldots, \exp(-||s_n-s_0||/\hat{\eta})) \). The corresponding prediction variance is \( \text{Var}(\hat{S}(s_0) - S(s_0)) = \hat{\sigma}^2 - C_{s_0} (\hat{\sigma}^2 \hat{V})^{-1} C_{s_0} \), based on which we have the inverse variance for the prediction, \( \zeta(s_0) = \frac{1}{\text{Var}(\hat{S}(s_0) - S(s_0))} \). The higher \( \zeta(s_0) \) is, the better quality the prediction has, and we give more weight to the most reliable pollution values while aggregating them (see Sanchez-Meca and Marín-Martínez, 1998).

After obtaining pollution predictions at the centre of all grids where the population density data are available (see Fig. 1(b)) using (6), and denoting the population density at location \( s_i \) as \( G(s_i) \), for a specific county \( k \), the spatially representative pollution concentration is estimated by

\[
z_k = \sum_{s_i \in A_k} \frac{\hat{S}(s_i) G(s_i) \zeta(s_i)}{\sum_{s_j \in A_k} G(s_j) \zeta(s_j)},
\]

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. COVID-19 incidence model

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

\[
Y_k \sim \text{Poisson}(E_k \lambda_k), \quad k = 1, \ldots, K,
\]

\[
\log(\lambda_k) = X_k^T \beta + \phi_k
\]

where the relative risk of disease in county \( k \) is denoted by \( \lambda_k \), and is modelled on the log scale by covariates \( X_k \) and a spatial random effect \( \phi_k \). The covariates are comprised of an intercept, pollutants, temperature and areal population density which is the areal population divided by its area (referred to as popDensity). The regression parameters \( \beta \) are assigned weakly informative zero-mean Gaussian priors with diagonal variance matrix \( \beta \sim N(0, 10^2I) \).

The spatial random effect, \( \phi_k \), 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,

\[
\phi \sim N(0, \kappa \times Q(\xi W)^{-1}),
\]

where \( \phi = \{ \phi_k, k = 1, \ldots, K \} \). Spatial autocorrelation is induced into the random effects by the precision matrix \( Q(\xi W) = \xi (\text{diag}(W_1) - W) + (1-\xi) I \), 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 \( W \), whose \( ij \)th element equals 1 if areas \( i \) and \( j \) share a common border and is zero otherwise. The level of spatial autocorrelation in the random effects is controlled by \( \xi \). Finally, weakly informative hyperpriors are specified for the parameters \( (\kappa, \xi) \) by

\[
\kappa \sim \text{Exp}[\log(2)],
\]

\[
\log \left( \frac{\xi}{1-\xi} \right) \sim N(0, 1.8).
\]
Table 1
Parameter estimation from the spatial pollution model.

| Pollutant | \( \mu \) | \( \sigma^2 \) | \( \tau^2 \) | \( \eta \) | 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  |
| Arsenic    | 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 \( \mu gm^{-3} \) for NO\(_2\), PM\(_{2.5}\), PM\(_{10}\), SO\(_2\), Benzene; ng m\(^{-3}\) for Arsenic, Cadmium, Nickel; and \(^\circ\)C for temperature.

| Pollutant | 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 |
| Arsenic    | 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|

The prior distribution of \( \xi \) is fairly non-informative as it is roughly uniformly distributed within [0,1]. The prior distribution of \( \kappa \) allows small values, which are what we expect for the variation of the log scale of relative risk. The COVID-19 incidence models are implemented in INLA (Rue et al., 2009) which uses a computationally effective and extremely powerful method for fitting Bayesian models, and has 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 obtained by applying the model in (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 are all well below those from a non-spatial model that does not incorporate the component \( R \) in (2). This suggests that spatial structure is an important component of the pollution models. The main outputs from the fitted pollution models are the population-weighted county level exposure for each pollutant. The population-weighted NO\(_2\) exposures are shown in Fig. 1(d), suggesting that the western part of Germany is much more exposed to NO\(_2\). A summary of the estimated population-weighted county level exposure is presented in Table 2, while scatterplots of the natural logarithm of COVID-19 SIR against the population-weighted NO\(_2\) and PM\(_{2.5}\) are displayed in the upper part of Fig. 2. The latter seems to indicate a linear relationship between NO\(_2\) and log COVID-19 risk.

4.2. Model validation

Before presenting the estimated effects of environmental factors on COVID-19, we assess the necessity of including spatial autocorrelation via the random effects model (9). In order to do so, we fitted a simplified version of model (8) without the spatial random effects term \( \phi_k \). The
residuals from this model show substantial spatial autocorrelation, with significant Moran’s I statistics (Moran, 1950) shown as Fig. 2(c). The empirical semi-variogram of the residuals, illustrated in Fig. 2(d), shows several points are lying outside the 95% Monte Carlo simulation envelopes. This suggests that some spatial autocorrelation remains in the residuals, and thus that including the spatial random effect model (9) is necessary.

4.3. Effects of pollution on COVID-19

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, and the results from other commonly used CAR models: the “Besag” intrinsic autoregressive model proposed by Besag et al. (1991), the “BYM” convolution model (also proposed by Besag et al., 1991), and the non-spatial model (referred to as “IID”). In addition, as PM_{10} and PM_{2.5} are highly correlated (with correlation coefficient 0.75), we run two health models separately, with each model including either PM_{10} or PM_{2.5} to avoid collinearity. The results from having PM_{2.5} in the model are presented in Table 3, while those from having PM_{10} are presented in the Appendix in Table 4.
The results from the Leroux model show that NO\textsubscript{2} exposure is significantly associated with COVID-19 incidence in Germany, with a 1 \( \mu \text{g m}^{-3} \) increase in long-term exposure to NO\textsubscript{2} increasing the COVID-19 incidence rate by 5.58\% (95\% CI: 3.35\%, 7.86\%). This statistically significant association between NO\textsubscript{2} and COVID-19 incidence is consistent across various health models, including the Leroux model with PM\textsubscript{2.5}, and commonly used BYM, Besag, IID models (see Table 3). SO\textsubscript{2} is just at the border of having a significant association with COVID-19 incidence, since the posterior probabilities of its increasing relative risk is 0.96 (see Table 3). SO\textsubscript{2} is significantly associated with COVID-19 incidence in the model having PM\textsubscript{10} rather than PM\textsubscript{2.5} (see the Leroux model results from Table 4).

### 5. Discussion

Ogen (2020) states that poisoning our environment means poisoning our own body, and when it experiences chronic respiratory stress its ability to defend itself from infections is limited. Existing research has linked pollutants (e.g., PM\textsubscript{2.5} and NO\textsubscript{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 individual's air pollution exposure. The results show that long-term exposure to NO\textsubscript{2} is significantly associated with COVID-19 incidence rate in Germany, with a 1 \( \mu \text{g m}^{-3} \) increase in long-term exposure to NO\textsubscript{2}.
Table 4
Posterior medians and 95% CI for the percentage increase in relative risk from one-unit increase in each covariate, and the WAIC from fitting various health models (having PM$_{10}$), including the employed Leroux model, and the commonly used BYM, Besag, IID models. ‘Pr’ is the posterior probabilities that covariate increases relative risk.

|          | NO$_2$   | SO$_2$  | Temperature | Benzene | Arsenic | Cadmium | Nickel | popDensity | WAIC |
|----------|----------|---------|-------------|---------|---------|---------|--------|------------|-------|
| Est      | 6.06     | 18.82   | -10.89      | -0.80   | -13.74  | 13.52   | -0.70  | -2.08      | 3814.55 |
| CI       | (3.50, 8.67) | (-12.49, 7.63) | (-20.34, -0.24) | (-18.69, 20.39) | (-29.20, 3.35) | (-8.03, 41.20) | (-12.51, 12.67) | (-7.30, 3.43) | 3814.55 |
| Pr       | [1.00]   | [0.98]  | [0.02]      | [0.47]  | [0.07]  | [0.88]  | [0.46] | [0.22]     | 3814.55 |
| Est      | 5.51     | 6.40    | -8.09       | -0.80   | -9.18   | 13.52   | -0.70  | -1.92      | 3815.64 |
| CI       | (2.91, 8.20) | (-9.64, 25.21) | (-18.09, 3.13) | (-23.67, 20.39) | (-26.99, 12.85) | (-8.03, 41.20) | (-12.51, 12.67) | (-7.09, 3.52) | 3815.64 |
| Pr       | [1.00]   | [0.77]  | [0.07]      | [0.47]  | [0.19]  | [0.87]  | [0.46] | [0.24]     | 3815.64 |
| Est      | 5.57     | 6.18    | -8.05       | -0.80   | -9.42   | 13.52   | -0.70  | -1.92      | 3816.13 |
| CI       | (2.94, 8.26) | (-9.94, 25.18) | (-18.12, 3.24) | (-24.07, 24.39) | (-27.34, 12.85) | (-8.03, 41.20) | (-12.51, 12.67) | (-7.09, 3.52) | 3816.13 |
| Pr       | [1.00]   | [0.76]  | [0.08]      | [0.39]  | [0.19]  | [0.87]  | [0.46] | [0.25]     | 3816.13 |
| WAIC     | 3814.55  | 50.66   | 16.21       | 8.46    | -12.82  | 25.46   | 13.52  | -5.89      | 3814.58 |

Increasing the COVID-19 incidence rate by 5.58% (95% CI: 3.35%, 7.86%). No substantial associations were found between the COVID-19 incidence rate and the other pollutants, including PM$_{2.5}$, PM$_{10}$, SO$_2$, Benzene, Arsenic, Cadmium and Nickel. Temperature and population density are adjusted for in the model, and spatial random effects are also included to capture the residual spatial autocorrelation after the covariate effects have been accounted for.

For comparison purposes, we compared our results to other commonly used CAR models, including the intrinsic autoregressive model (Besag et al., 1991), convolution model (Besag et al., 1991), and non-spatial model. In addition, as PM$_{10}$ and PM$_{2.5}$ are highly correlated, we ran two health models separately, with each model including either PM$_{10}$ or PM$_{2.5}$ to avoid collinearity. We found that the statistically significant associations between NO$_2$ and COVID-19 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 adjusted for a set of other pollutants, temperature and population density. Another limitation is lacking COVID-19 testing numbers, since the confirmed cases (positive testing numbers) in a county rely on the total number of tests being conducted in that county. The limitation in regard to estimating pollution exposures is that the current model is univariate, which is potentially losing some power by not borrowing strength over correlated pollutants as would occur in a multivariate pollution model.

Fig. 3. Posterior means of relative risk $E(\lambda_k | Y)$ and probabilities of 50% excess risk $Pr(\exp(\phi) > 1.5 | Y)$.
Finally, COVID-19 deaths, not only infections, should be focused when (or if) more deaths occur in the future. Such studies will help us better understand COVID-19, and also help the global communities and health organizations stay informed and make data driven decisions.

Acknowledgements

We would like to thank the editors, and two anonymous referees for their insightful and constructive comments, which greatly improved the presentation of the article. Professor Brown is funded by the Natural Sciences and Engineering Research Council of Canada.

Appendix

The results from fitting COVID-19 incidence models having PM_10 are shown in Table 4. The data and R code used in this study are shared on Github (https://github.com/hgw0610209/Germany-covid-paper).

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Selected Papers of Hirotugu Akaike. Springer New York, New York, NY, pp. 199–213.
Arbia, G., 1988. Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems. Springer.
Banerjee, S., Carlin, B.P., Gelfand, A.E., 2004. Hierarchical Modeling and Analysis of Spatial Data, first ed. Chapman and Hall/CRC Press.
Besag, J., York, J., Mollie, A., 1991. Bayesian Image restoration with two applications in spatial statistics. Ann. Inst. Stat. Math. 43, 1–59.
BKG, 2020. Federal agency for cartography and geodesy: Digitale geodaten. https://gdz.bkg.bund.de/index.php/default/digitale-geodaten.html.
Blair, A., Stewart, P., Lubin, J.H., Forastiere, F., 2007. Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. Amer. J. Ind. Med. 50 (3), 199–207.
Blangiardo, M., Cameletti, M., Baiocchi, G., Rue, H., 2013. Spatial and spatio-temporal models with R-INLA. Spatial Spatio-temporal Epidemiol. 4, 33–49.
Blangiardo, M., Finazzi, F., Cameletti, M., 2016. Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. Spatial Spatio-temporal Epidemiol. 18, 1–12, Environmental Exposure and Health.
Bowatte, G., Erbas, B., Lodge, C.J., Knibbs, L.D., Gurrin, L.C., Marks, G.B., Thomas, P.S., Johns, D.P., Giles, G.G., Hui, J., Dennekamp, M., Perret, J.L., Abramson, M.J., Walters, E.H., Matheson, M.C., Dharmage, S.C., 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. Eur. Respir. J. 50 (4).
COMEAP, 2010. The Mortality Effects of Long-Term Exposure to Particulate Air Pollution in the United Kingdom. Crown.
Cressie, N., 1993. Statistics for Spatial Data. Wiley, New York.
Das, K., Das, S., Dhumde, S., 2008. Nickel, its adverse health effects & oxidative stress. Indian J. Med. Res. 128 (4), 412–425.
Diggle, P., Ribeiro, P., 2007. Model-Based Geostatistics: In: Springer Series in Statistics, Springer.
DIVA-GIS, 2020. DIVA-GIS: Free spatial data. https://www.diva-gis.org/gdata.
EEA, 2020. The European climate assessment & dataset. https://www.ecad.eu/.
Elliott, P., Wakefield, J., Best, N., Briggs, D., 2000. Spatial Epidemiology: Methods and Applications, first ed. Oxford University Press.
Gryparis, A., Paciorek, C., Zeka, A., Schwartz, J., Coull, B., 2009. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics 10 (2), 258–274.
Heads or Tails, 2020. COVID-19 tracking Germany, version 156. https://www.kaggle.com/headsorfalls/covid19-tracking-germany/version/156.
Huang, G., Lee, D., Scott, E.M., 2018. Multivariate space-time modelling of multiple air pollutants and their health effects accounting for exposure uncertainty. Stat. Med. 37 (7), 1134–1148.
Järup, L., Berglund, M., Elinder, C.G., Nordberg, G., Vanter, M., 1998. Health effects of cadmium exposure – a review of the literature and a risk estimate. Scand. J. Work Environ. Health 24, 1–51.
Lawson, A.B., 2008. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, first ed. Chapman and Hall/CRC Press.
Lee, D., 2011. A comparison of conditional autoregressive models used in Bayesian disease mapping. Spatial Spatio-temporal Epidemiol. 2 (2), 79–89.
Lee, D., 2012. Using spline models to estimate the varying health risks from air pollution across Scotland. Stat. Med. 31 (27), 3366–3378.
Lee, D., Ferguson, C., Mitchell, R., 2009. Air pollution and health in Scotland: a multicity study. Biostatistics 10 (3), 409–423.

Lee, D., Mukhopadhyay, S., Rushworth, A., Sahu, S.K., 2017. A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health. Biostatistics 18 (2), 370–385.

Lee, D., Robertson, C., Ramsay, C., Pyper, K., 2020. Quantifying the impact of themodifiable areal unit problem when estimating the health effects of air pollution. Environmetrics.

Leroux, B., Lei, X., Breslow, N., 1999. Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence. Springer-Verlag, New York, pp. 125–178.

Maheswaran, R., Haining, R., Pearson, T., Law, J., Brindley, P., Best, N., 2006. Outdoor NOx and stroke mortality adjusting for small area level smoking prevalence using a Bayesian approach. Stat. Methods Med. Res. 15, 499–516.

Moran, P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

MOW, 2020. Germany Weather map. https://www.mapsofworld.com/germany/thematic-maps/germany-temperature-map.html.

Napier, G., Lee, D., Robertson, C., Lawson, A., 2018. A Bayesian space-time model for clustering areal units based on their disease trends. Biostatistics.

Ogen, Y., 2020. Assessing nitrogen dioxide (NO2) levels as a contributing factor to coronavirus (COVID-19) fatality. Sci. Total Environ. 726, 138605.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman and Hall/CRC, New York.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2), 319–392.

Rushworth, A., Lee, D., Mitchell, R., 2014. A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. Spatial Spatio-temporal Epidemiol. 10, 29–38.

Sacks, J.D., Rappold, A.G., Allen Davis Jr., J., Richardson, D.B., Waller, A.E., Luben, T.J., 2014. Influence of urbanicity and county characteristics on the association between ozone and asthma emergency department visits in North Carolina. Environ. Health Perspect 122 (5), 506–512.

Sanchez-Meca, J., Marin-Martinez, F., 1998. Weighting by inverse variance or by sample size in meta-analysis: A simulation study. Educat. Psychol. Measur. 58 (2), 211–220.

Schikowski, T., Sugiri, D., Ranft, U., Gehring, U., Heinrich, J., Wichmann, H.-E., Kraemer, U., 2005. Long-term air pollution exposure and living close to busy roads are associated with COPD in women. Respir. Res. 6, 152.

Shaddick, G., Zidek, J., 2015. Spatio-Temporal Methods in Environmental Epidemiology. In: CRC Texts in Statistical Science, Chapman & Hall, UK United Kingdom.

Smith, M.T., 2010. Advances in understanding benzene health effects and susceptibility. Annu. Rev. Public Health 31 (1), 133–148.

The City Population, 2019. GERMANY: Administrative division: States and counties. https://www.citypopulation.de/en/germany/admin/.

Vinikoor-Imler, L.C., Davis, J.A., Meyer, R.E., Messer, L.C., Luben, T.J., 2014. Associations between prenatal exposure to air pollution, small for gestational age, and term low birthweight in a state-wide birth cohort. Environ. Res. 132, 132–139.

Wakefield, J., Salway, R., 2001. A statistical framework for ecological and aggregate studies. J. R. Stat. Soc. A 164 (1), 119–137.

Warren, J., Fuentes, M., Herring, A., Langlois, P., 2012. Bayesian Spatial-temporal model for cardiac congenital anomalies and ambient air pollution risk assessment. Environmetrics 23 (8), 673–684.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.

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

Yu, W.H., Harvey, C.M., Harvey, C.F., 2003. Arsenic in groundwater in Bangladesh: A geostatistical and epidemiological framework for evaluating health effects and potential remedies. Water Resour. Res. 39 (6).