WILEY

Quantifying the impact of the modifiable areal unit problem when estimating the health effects of air pollution

Duncan Lee[1] | Chris Robertson[2,3] | Colin Ramsay[3] | Kate Pyper[2]

[1]School of Mathematics and Statistics, University of Glasgow, Glasgow, Scotland

[2]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, Scotland

[3]Health Protection Scotland, Glasgow, Scotland

Correspondence
Duncan Lee, School of Mathematics and Statistics, University Place, University of Glasgow, Glasgow, G12 8SQ, Scotland.
Email: Duncan.Lee@glasgow.ac.uk

Summary

Air pollution is a major public health concern, and large numbers of epidemiological studies have been conducted to quantify its impacts. One study design used to quantify these impacts is a spatial areal unit design, which estimates a population-level association using data on air pollution concentrations and disease incidence that have been spatially aggregated to a set of nonoverlapping areal units. A major criticism of this study design is that the specification of these areal units is arbitrary, and if one changed their boundaries then the aggregated data would change despite the locations of the disease cases and the air pollution surface remaining the same. This is known as the modifiable areal unit problem, and this is the first article to quantify its likely effects in air pollution and health studies. In addition, we derive an aggregate model for these data directly from an idealized individual-level risk model and show that it provides better estimation than the commonly used ecological model. Our work is motivated by a new study of air pollution and health in Scotland, and we find consistent significant associations between air pollution and respiratory disease but not for circulatory disease.

KEYWORDS

epidemiological study, nitrogen dioxide and particulate matter, spatially aggregated disease counts

1 | INTRODUCTION

Air pollution continues to be a major global public health problem, with the World Health Organisation (WHO) linking seven million deaths to it each year worldwide (World Health Organisation, 2016). In the United Kingdom an estimated 40,000 deaths are attributed to air pollution exposure each year (Royal College of Physicians, 2016), and legal limits for individual air pollutants have been stipulated that must not be exceeded (Department for the Environment, Food and Rural Affairs, 2015). The choice of these limits is informed by epidemiological studies, and both individual- and ecological (or group)-level study designs are prevalent in the literature. Ecological-level studies are popular because they utilize available population-level disease incidence data, which makes them comparatively fast and inexpensive to implement. However, all that can be inferred 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* (Wakefield & 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 who exhibit disease and have the highest air pollution exposures.

Spatial ecological studies partition the study region into K nonoverlapping areal units such as electoral wards or census tracts and utilize data on total disease incidence and average air pollution concentrations for the populations living in each unit. Data on air pollution concentrations are available as point-level measurements and/or grid-level modeled concentrations, and average concentrations for each areal unit have been estimated using simple averaging (Lee & Sarran, 2015) or spatial prediction techniques such as block Kriging (Zhu, Carlin, & Gelfand, 2003). A number of different statistical issues have been addressed in the literature in relation to modeling air pollution and its health effects, including different approaches for (i) spatiotemporal pollution prediction (Gilani, Berrocal, & Batterman, 2019; Nicolis, Díaz, Sahu, & Marín, 2019); (ii) estimating individual-level exposures (Clifford et al., 2019); (iii) the impacts of preferential sampling of pollution (Lee, Szpiro, Kim, & Sheppard, 2015); and (iv) allowing for errors in exposure variables (Keller & Peng, 2019; Strand, Sillau, Grunwald, & Rabinovitch, 2015).

However, the specific focus of this article is the spatial structure of the disease data, which are only available as the total number of disease cases for each of the K areal units, and these units have typically been designed to minimize the within-area variation in socioeconomic deprivation, respect physical barriers such as rivers, and have similar population sizes. The specification of these units is crucial for epidemiological air pollution studies, because it is the spatial variation in total disease incidence and average air pollution concentrations between these areal units that drives the estimated pollution–health relationship. In reality however both these variables vary continuously over space, and partitioning the study region into a different set of disjoint areal units, potentially changing the number of units K, will result in a different ecological-level data set for analysis despite the underlying study population remaining unchanged.

This phenomenon is known as the modifiable areal unit problem (MAUP; Openshaw, 1984) and is a major criticism of studies of this type, which reduces the strength of the evidence that such studies provide. The MAUP arises when trying to make inference about spatially continuous processes using data aggregated to a single set of areal units. There has been extensive research into the MAUP in a variety of fields dating back to Openshaw and Taylor (1979), and a review is given by Manley (2014). However, there has been limited research to date on the impact of the MAUP in air pollution and health studies, with Parenteau and Sawada (2011) finding that the estimated impact of NO_2 on respiratory morbidity in Ottawa varied across three different sized spatial scales. This study is rather limited however because it only estimated the impact of the MAUP on a single data set and did not attempt to identify the factors in a data set that lessen or increase the impact of the MAUP. In addition, it focused exclusively on changing the number and sizes of the spatial units, which is known as the *scaling effect* (Openshaw, 1984). It did not assess the *zoning effect*, where the number and average size of the units remains unchanged but the geographical specification of these units does change. We note in passing that the MAUP is not the only spatial rescaling problem that researchers have addressed, with Aregay et al. (2018) developing models for multiple scales of small-area disease data.

This article however is the first to fully tackle the MAUP in air pollution and health studies and is motivated by the following two aims. The first is to quantify the impact that the scaling and zoning effects of the MAUP have on air pollution and health effects estimated from areal unit studies, so that the validity of the effects estimated in existing studies can be better understood. The second is to develop a model that reduces the impact of the MAUP, by deriving an appropriate areal unit-level model directly from an idealized individual-level model. These aims are addressed in the context of both constant and varying air pollution and health effects, the latter varying over areal units. We estimate varying effects using varying coefficient models (Hastie & Tibshirani, 1993), which allows the air pollution and health effect to vary depending on external variables known as *effect modifiers*. These methodological developments are motivated by a new study of air pollution and health in Scotland, and the varying coefficient models are used to quantify whether air pollution disproportionally affects poorer communities, thus widening health inequalities (NHS Health Scotland, 2016) between rich and poor communities. A description of the study data is presented in Section 2, while our methodology is described in Section 3. A simulation study is presented in Section 4, while the result from the motivating study are presented in Section 5. The article concludes with a discussion in Section 6.

2 | MOTIVATING STUDY

Our motivating study is based in Scotland, which while having comparatively low-ambient air pollution concentrations, nonetheless has 38 officially declared Air Quality Management Areas (AQMA, http://www.scottishairquality.co.uk/laqm/aqma), where pollution levels have or are likely to breach European Union (EU, European Parliament, 2008) pollution limits. Partially as a result of the failure to reduce pollution in these AQMA, the Scottish Government has undertaken

a review of its 2015 Cleaner Air for Scotland (CAFS) strategy (http://www.gov.scot/Resource/0048/00488493.pdf), with a view to ensuring that its successor strategy is more evidence based. Providing more recent and more comprehensive epidemiological evidence about the impact of air pollution in Scotland specifically is therefore a necessity to support this policy review. However, there is a lack of recent Scottish-specific research study evidence, with only Lee (2018) and Lee, Robertson, Ramsay, Gillespie, and Napier (2019) presenting studies using data later than 2011 (up to 2016). However, these studies use data over short temporal durations (10 months and 2 years, respectively) and focus exclusively on physical rather than mental health. The study reported here provides a more comprehensive and robust picture of air pollution health effects, by using data on both physical and cognitive health (dementia) for the 7-year period from 2011 to 2017. The study region of mainland Scotland is partitioned into $K = 1,249$ intermediate zones (IZ), which are a Scottish Government created geography used for the allocation, display and analysis of small-area-level statistics, having an average population of around 4,000 people.

2.1 | Disease outcomes

In this study we quantify the impact of air pollution on both physical and cognitive health. While the links between physical health and air pollution are well known (see Bellini, Baccini, Biggeri, & Terracini, 2007; Lee et al., 2019), the association with cognitive health is much less well researched (see Chen et al., 2017; Oudin et al., 2016). For physical health we focus on respiratory (ICD-10 codes J00-J99) and circulatory (ICD-10 codes I00-I99) hospitalizations and deaths, because these diseases have been most strongly linked to air pollution by existing studies, and because they are among Scotland's leading causes of deaths (see http://www.gov.scot/Topics/Statistics/Browse/Health/TrendMortalityRates). For cognitive health we focus on dementia (ICD-10 codes F01 and F03) deaths, because the studies listed above have shown evidence of associations with air pollution in other countries. Finally, we also consider total nonaccidental mortality, as this is a common endpoint used in existing epidemiological air pollution studies.

For each disease outcome we have data on the total numbers of disease events from the populations living within each IZ over the 7-year study duration, yielding a set of spatially referenced counts. In the case of the hospitalization data they relate to the total number of events (not just first events for an individual) where the primary diagnosis was the disease type in question. The total counts for each health outcome grouping are summarized in the top panel of Table 1. This identifies that hospitalizations are more common than deaths as would be expected, and that dementia is the rarest cause of death.

The observed numbers of disease events (Y) will depend on the population size and its age–sex structure in each IZ. This is accounted for by computing the expected numbers of disease events (e) using indirect standardization. Specifically, the number of people living in each IZ in each sex-specific 5-year age strata is multiplied by the Scotland-wide disease rate for that strata, before summing over strata to compute the expected number of disease events. Based on these data the exploratory measure of disease risk is the standardized mortality/morbidity ratio (SMR = Y/e), which is summarized in the middle panel of Table 1. The table shows a wide variation in SMRs across the IZs in mainland Scotland, with the IZs that exhibit the maximum SMRs for most diseases being between two and three times higher than the national average (an SMR of 1). The exception to this is dementia mortality that has a maximum SMR of 6.26, which is due to the uneven distribution of cases across Scotland as well as the small numbers of deaths attributed to this disease leading to an unstable SMR. A map of the SMR for respiratory hospitalizations is displayed in the left panel (a) of Figure 1. This shows that areas with elevated risks are mainly in and around the city of Glasgow, a city known to have some of the worst health in the UK, with marked levels of health inequality.

2.2 | Air pollution concentrations

Measured air pollution concentrations are available from http://www.scottishairquality.scot, but the network of monitors is highly dispersed and is poorly represented at the small area IZ scale used for this study. Therefore in common with existing studies (e.g., Lee & Sarran, 2015) we utilize annual average modeled concentrations from an atmospheric dispersion model. Specifically, we use concentrations estimated by the pollution climate mapping (PCM) model developed for the Department for the Environment, Food and Rural Affairs (DEFRA), which are freely available from https://uk-air. defra.gov.uk/data/pcm-data. This model estimates concentrations on a 1-km^2 square grid, which are spatially misaligned with the irregularly shaped intermediate zones that the disease data relate to. Concentrations of nitrogen dioxide (NO_2) and coarse (PM_{10}) and fine ($PM_{2.5}$) particulate matter are available (measured in μg m^{-3}). These pollutants were chosen

Variable	Percentiles of the distribution				
	0%	25%	50%	75%	100%
Disease incidents (total counts per year)					
Circulatory hospitalization	91	371	465	583	1,174
Circulatory mortality	5	59	80	105	284
Respiratory hospitalization	129	363	489	649	1,607
Respiratory mortality	1	28	39	53	156
Dementia mortality	0	5	12	27	147
Total nonaccidental mortality	22	216	286	377	916
Disease risk (SMR)					
Circulatory hospitalization	0.47	0.84	0.98	1.15	2.03
Circulatory mortality	0.40	0.84	1.00	1.17	2.40
Respiratory hospitalization	0.35	0.72	0.95	1.23	2.32
Respiratory mortality	0.12	0.74	0.97	1.26	2.90
Dementia mortality	0.00	0.42	0.68	1.20	6.26
Total nonaccidental mortality	0.40	0.82	0.98	1.18	2.28
Air pollutants (in $\mu g\ m^{-3}$)					
NO_2	1.4	1.9	2.6	3.5	32.9
$PM_{2.5}$	4.0	4.4	5.0	5.9	11.1
PM_{10}	5.9	6.7	7.6	8.7	17.4

TABLE 1 Summary of the spatial distribution of the disease and pollution data across the 1,249 intermediate zones for 2011 to 2017

because they are the ones specified as exceeding existing limit values in all of Scotland's air quality management areas except one. These yearly concentrations are averaged over the 7-year study period, and the distributions of the resulting concentrations across the IZs are summarized in the bottom panel of Table 1. The table shows that the spatial distributions of the pollutants are skewed to the right, which is because most of the 1-km² grid squares in Scotland are rural and hence have a very low population density. The spatial distribution of PM_{10} is displayed in panel (b) of Figure 1, which shows it is highest in the cities of Glasgow and Edinburgh as well as around the east and southeast coasts, the latter due to transboundary pollution from continental Europe and England, respectively. The three pollutants are highly correlated as expected, with correlations ranging between 0.671 (between NO_2 and PM_{10}) and 0.961 (between $PM_{2.5}$ and PM_{10}).

2.3 | Covariate data

The main confounding factor in spatial studies of disease incidence is socioeconomic deprivation, in part because of its links to smoking prevalence (Kleinschmidt, Hills, & Elliott, 1995). Areas with more deprived populations exhibit worse health on average than more affluent ones. These inequalities in population health are well recognized in Scotland (NHS Health Scotland, 2016). Socioeconomic deprivation is measured in Scotland by the Scottish Index of Multiple Deprivation (SIMD, http://www.gov.scot/Topics/Statistics/SIMD), which is a composite index consisting of deprivation indicators in the domains of access to services, crime, education, employment, health, housing, and income. These are then weighted and combined to create the final index. We have access, at the IZ level, to the seven domain-specific indicators in 2016 as covariates for possible inclusion in our models, although obviously we will not use the health domain as our response variable is health related. The crime indicator has a single IZ with a missing value, which is imputed by computing the average value from geographically neighboring IZs (those sharing a common border). These domain-specific indicators of deprivation are correlated, with the highest pairwise correlations of 0.87 to 0.98 being observed between the income, employment, and education domains. In addition, we also have access to data on the average property price in each IZ, which is used as the effect modifier to assess whether the pollution–health effect depends on the relative affluence of an IZ, thus potentially widening health inequalities.
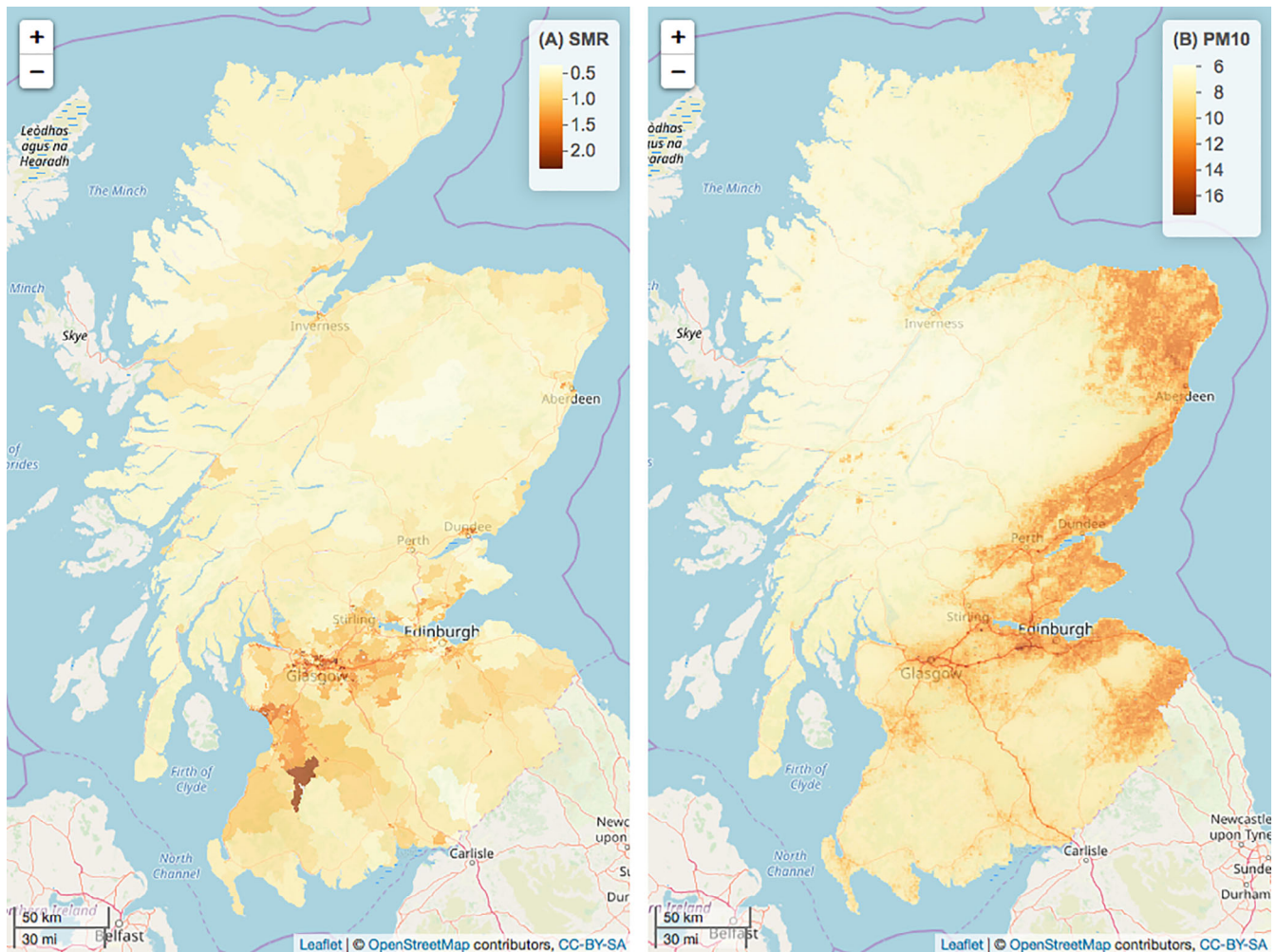
FIGURE 1 Display of the data. (A) The standardized morbidity ratio (SMR) for respiratory hospitalizations and (B) the average concentrations of PM$_{10}$

2.4 | Exploratory analysis

To assess the extent of the residual spatial autocorrelation and overdispersion in the disease data after covariate adjustment, quasi-Poisson log-linear models were fitted to each disease outcome separately. The expected numbers of disease events was included as an offset term, and for the purposes of this exploratory analysis PM$_{10}$ was the only pollutant included in the model. As the grid-level pollution data are spatially misaligned with the IZ-level disease counts, simple averaging was used to spatially realign the pollution data to the IZ level as used by Lee and Sarran (2015). Specifically, each 1-km^2 grid square has an associated centroid (central point), and the estimated pollution concentration for an IZ is computed as the mean of the grid square concentrations whose centroids lie within the IZ. Any IZ that does not contain a grid square centroid is assigned the pollution concentration from the nearest grid square. As the income, employment and education domains of SIMD are highly correlated we only use one of these as a covariate to avoid collinearity problems. Based on the AIC the income domain provided the best fit for five of the six disease outcomes, so for consistency in our modeling we do not consider the education and employment domains any further. The remaining domain-specific SIMD indicators (access to services, crime, housing, and income) were used in each disease-specific model, and the residuals showed substantial overdispersion, with overdispersion parameters ranging between 2.4 (circulatory mortalities) and 16.9 (respiratory hospitalizations). The residuals also showed significant spatial autocorrelation, with p-values based on a Moran's I permutation test (Moran, 1950) all being less than .00001 except for Dementia mortality (p-value of .07).

3 | METHODOLOGY

3.1 | Study design

The study region is partitioned into K nonoverlapping areal units $\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$, and data on disease incidence and socioeconomic confounders are available for each unit over a fixed time interval. The disease incidence data for a single outcome (i.e., hospitalization or mortality) consist of $\mathbf{Y}(\mathcal{A}) = (Y(\mathcal{A}_1), \ldots, Y(\mathcal{A}_K))$ and $\mathbf{e}(\mathcal{A}) = (e(\mathcal{A}_1), \ldots, e(\mathcal{A}_K))$, which respectively denote the observed and expected numbers of disease cases in each areal unit. The latter is computed by indirect standardization and is given by $e(\mathcal{A}_k) = \sum_{j=1}^{J} n_j(\mathcal{A}_k) r_j$, where $n_j(\mathcal{A}_k)$ is the number of people living in area $\mathcal{A}_k$ who are in the jth age–sex strata, while r_j is the nationwide risk of the single disease outcome under study for strata j for $j = 1, \ldots, J$. The matrix of socioeconomic deprivation confounding variables is denoted by $\mathbf{Z}(\mathcal{A}) = (\mathbf{z}^\top(\mathcal{A}_1), \ldots, \mathbf{z}^\top(\mathcal{A}_K))^\top$, where the vector of p values for areal unit $\mathcal{A}_k$ is given by $\mathbf{z}^\top(\mathcal{A}_k) = (z_1(\mathcal{A}_k), \ldots, z_p(\mathcal{A}_k))$. Modeled pollution data are not available at this areal unit level and are instead available on a much finer 1-km^2 square grid $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_M)$. These data are denoted by $\mathbf{x}(\mathcal{G}) = (x(\mathcal{G}_1), \ldots, x(\mathcal{G}_M))$ and are thus spatially misaligned with the aggregated disease and confounder data. Finally, we have modeled population size estimates $\mathbf{P}(\mathcal{G}) = (P(\mathcal{G}_1), \ldots, P(\mathcal{G}_M))$ from Reis et al. (2017) at the same 1-km^2 grid square resolution. In Section 3.2 we outline the standard ecological model that is commonly applied to these data, while in Section 3.3 we outline a more realistic model that should be less prone to the MAUP and ecological bias because it is aggregated from an idealized individual-level risk model.

3.2 | The standard ecological model

Analysis of these data is typically conducted at the areal unit level $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$, because this is the smallest spatial scale at which disease incidence data are available. Thus, one typically aims to estimate the average pollution concentration in each areal unit, namely, $x(\mathcal{A}_k) = \int_{\mathcal{A}_k} x(\mathbf{s}) d\mathbf{s}$, where $x(\mathbf{s})$ denotes the pollution concentration at a single-point location $\mathbf{s}$. From the observed grid-level data this is most easily estimated by averaging the grid-level values $x(\mathcal{G}_i)$ whose centroids $\mathbf{g}_i$ lie within $\mathcal{A}_k$, that is, $\hat{x}(\mathcal{A}_k) = \frac{1}{q(\mathcal{A}_k)} \sum_{i \mid \mathbf{g}_i \in \mathcal{A}_k} x(\mathcal{G}_i)$, where $q(\mathcal{A}_k)$ denotes the number of grid square centroids that lie within $\mathcal{A}_k$. Then using this estimate the commonly used *ecological* regression model is given by

$$
\begin{aligned}
Y(\mathcal{A}_k) &\sim \text{Poisson}(e(\mathcal{A}_k)\theta(\mathcal{A}_k)) \\
\ln[\theta(\mathcal{A}_k)] &= \mathbf{z}(\mathcal{A}_k)^\top \boldsymbol{\alpha} + \hat{x}(\mathcal{A}_k)\beta(\mathcal{A}_k) + \phi(\mathcal{A}_k), \\
\alpha_r &\sim \text{N}(0, 100,000) \quad \text{for} \quad r = 1, \ldots, p.
\end{aligned}
\tag{1}
$$

The confounder regression parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ are assigned independent zero-mean weakly informative Gaussian prior distributions with a large variance, to ensure the data play the dominant role in determining their values. The pollution–health effect parameter $\beta(\mathcal{A}_k)$ is allowed to vary over areal units $\mathcal{A}_k$, and its prior specification is described in Section 3.4. The remaining term in the linear predictor $\boldsymbol{\phi}(\mathcal{A}) = (\phi(\mathcal{A}_1), \ldots, \phi(\mathcal{A}_K))$ is a vector of random effects, which account for any residual spatial autocorrelation in the disease data after adjusting for the covariates. Random effects models for spatial areal unit data are a well-researched topic, and conditional autoregressive (CAR) models, a special case of Gaussian Markov Random Fields (GMRF), are the most popular specification due to their computational efficiency resulting from their sparse precision matrices. These models are given by $\boldsymbol{\phi}(\mathcal{A}) \sim \text{N}(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W})^{-1})$, where the precision matrix $\mathbf{Q}(\mathbf{W})$ is a function of a $K \times K$ neighborhood or adjacency matrix $\mathbf{W} = (w_{kr})$. This matrix determines the spatial autocorrelation structure of the random effects, and if $w_{kr} > 0$ then $(\phi(\mathcal{A}_k), \phi(\mathcal{A}_r))$ are conditionally correlated, whereas if $w_{kr} = 0$ then $(\phi(\mathcal{A}_k), \phi(\mathcal{A}_r))$ are conditionally independent. A binary specification is most often adopted, where $w_{kr} = 1$ if areas $(\mathcal{A}_k, \mathcal{A}_r)$ are spatially close and share a common border, and $w_{kr} = 0$ otherwise. The models proposed by Besag, York, and Mollié (1991) and Leroux, Lei, and Breslow (2000) are the most popular in practice, and here we use the latter because a study by Lee (2011) showed it was preferable to the commonly used BYM model of Besag et al. (1991). This model is most often written in the form of the following univariate full conditional distributions $\phi(\mathcal{A}_k) \mid \boldsymbol{\phi}(-\mathcal{A}_k)$ (where $\boldsymbol{\phi}(-\mathcal{A}_k)$ denotes all the random effects except $\phi(\mathcal{A}_k)$):

$$
\phi(\mathcal{A}_k) \mid \boldsymbol{\phi}(-\mathcal{A}_k) \sim \text{N}\left(\frac{\rho \sum_{r=1}^{K} w_{kr} \phi(\mathcal{A}_r)}{\rho \sum_{r=1}^{K} w_{kr} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{r=1}^{K} w_{kr} + 1 - \rho} \right)
$$

$$\rho \sim \text{Uniform}(0, 1)$$

$$\tau^2 \sim \text{Inverse gamma}(1, 0.01). \tag{2}$$

Here ρ is a spatial dependence parameter defined on the interval [0,1], with $\rho = 1$ corresponding to strong spatial autocorrelation (the intrinsic CAR model proposed by Besag et al., 1991), while $\rho = 0$ corresponds to spatial independence (i.e., $\phi(\mathcal{A}_k) | \phi(-\mathcal{A}_k) \sim \text{N}(0, \tau^2)$). A noninformative uniform prior is assigned to ρ on the unit interval, while a weakly informative conjugate inverse-gamma prior is assigned to τ^2.

3.3 | Aggregation of an individual-level risk model

The *ecological* regression model outlined above is ubiquitously used for modeling areal unit count data, but Wakefield and Shaddick (2006) have considered what an appropriate areal unit-level model would be if one began with an idealized individual-level risk model and aggregated it to the areal unit level. As this aggregated model is based on an individual-level model not susceptible to the MAUP, it should be less susceptible to the MAUP than the naive ecological model outlined above. Thus, we now derive such an aggregate model, and note that it is not an approximation to the naive ecological model (1), but instead an alternative to it that should be less susceptible to the MAUP.

Following Wakefield and Salway (2001), let $Y_{ij}(\mathcal{A}_k)$ denote the outcome of a single disease (e.g., respiratory hospitalization) for the ith individual in the jth age and sex strata (i.e., females, 0-4, females 5-9, etc.) in areal unit $\mathcal{A}_k$. The hospitalization and mortality outcomes from the motivating study are modeled separately here, just as they are in the standard ecological model outlined above. When modeling mortality outcomes $Y_{ij}(\mathcal{A}_k)$ is binary, with $Y_{ij}(\mathcal{A}_k) = 1$ meaning the individual died during the study period while if $Y_{ij}(\mathcal{A}_k) = 0$ they did not. Conversely, when we are separately modeling hospitalization data then $Y_{ij}(\mathcal{A}_k)$ denotes the (probably small) number of times that the individual was admitted to hospital during the study period. Then as the response variable has different sample spaces for these two outcome types, appropriate individual-level risk models for $Y_{ij}(\mathcal{A}_k)$ will be slightly different. Specifically, appropriate individual-level risk models are:

Mortality—$Y_{ij}(\mathcal{A}_k) \sim \text{Bernoulli}[\tilde{\theta}_{ij}(\mathcal{A}_k)]$ and $\tilde{\theta}_{ij}(\mathcal{A}_k) = r_j \exp[\mathbf{z}_{ij}(\mathcal{A}_k)^\top \tilde{\alpha} + x_{ij}(\mathcal{A}_k)\tilde{\beta}(\mathcal{A}_k) + \tilde{\phi}(\mathcal{A}_k)]$.

Hospitalization—$Y_{ij}(\mathcal{A}_k) \sim \text{Poisson}[\tilde{\theta}_{ij}(\mathcal{A}_k)]$ and $\tilde{\theta}_{ij}(\mathcal{A}_k) = r_j \exp[\mathbf{z}_{ij}(\mathcal{A}_k)^\top \tilde{\alpha} + x_{ij}(\mathcal{A}_k)\tilde{\beta}(\mathcal{A}_k) + \tilde{\phi}(\mathcal{A}_k)]. \tag{3}$

A Bernoulli model is used for mortality outcomes because it is a binary event, while a Poisson model is used for hospitalization outcomes because one could be admitted to hospital multiple times during the study period. In the mortality model $\tilde{\theta}_{ij}(\mathcal{A}_k)$ is the probability of death and is modeled with a log-link function because mortality is a rare event, while in the hospitalization model $\tilde{\theta}_{ij}(\mathcal{A}_k)$ represents the estimated number of hospitalizations.

For both disease outcome models $[\mathbf{z}_{ij}(\mathcal{A}_k), x_{ij}(\mathcal{A}_k)]$ are respectively individual-level covariates (confounders) and air pollution exposures for the ith individual from the jth age–sex strata living in areal unit $\mathcal{A}_k$. In addition, r_j denotes the nationwide risk of disease for age–sex strata j to adjust for individual demographics and is used in the computation of $e(\mathcal{A}_k)$ described above for the ecological model. The parameters $[\tilde{\alpha}, \tilde{\phi}(\mathcal{A}_k)]$ are respectively confounder regression parameters and spatial random effects (to control for areal unit-level unmeasured spatial confounding), which are not necessarily equal to the corresponding parameters $[\alpha, \phi(\mathcal{A}_k)]$ from the ecological model (1). Similarly, the regression parameter $\tilde{\beta}(\mathcal{A}_k)$ quantifying the effect of air pollution on disease risk is not necessarily equal to the value from the ecological model, that is, $\beta(\mathcal{A}_k) \neq \tilde{\beta}(\mathcal{A}_k)$. Furthermore, as the areal units have been designed to be homogeneous in terms of socioeconomic deprivation, we do not allow $\tilde{\beta}(\mathcal{A}_k)$ to further vary across individuals within an areal unit.

These individual-level models cannot be fitted because the only available disease data for each outcome are the areal unit-level total disease counts $Y(\mathcal{A}_k)$, which sum the individual disease outcomes over all $n_j(\mathcal{A}_k)$ individuals in strata j and across all J strata in areal unit $\mathcal{A}_k$. Thus while this enforced aggregation returns us to the MAUP, its effects should be reduced if we build a model by aggregating (separately by outcome) these individual-level risk models, instead of using the *ecological* model outlined in Section 3.2. Therefore as the aggregated disease count for areal unit $\mathcal{A}_k$ can be written as $Y(\mathcal{A}_k) = \sum_{j=1}^{J} \sum_{i \in S_j(\mathcal{A}_k)} Y_{ij}(\mathcal{A}_k)$, its expectation based on either of the individual-level models in (3) (which is the same) is given by

$$\mathbb{E}[Y(\mathcal{A}_k)] = \sum_{j=1}^{J} \sum_{i \in S_j(\mathcal{A}_k)} r_j \exp[\mathbf{z}_{ij}(\mathcal{A}_k)^\top \tilde{\alpha} + x_{ij}(\mathcal{A}_k)\tilde{\beta}(\mathcal{A}_k) + \tilde{\phi}(\mathcal{A}_k)], \tag{4}$$

where $S_j(\mathcal{A}_k)$ denotes the set of individuals in strata j and areal unit $\mathcal{A}_k$. Thus regardless of whether one models mortalities or hospitalizations, $\mathbb{E}[Y(\mathcal{A}_k)]$ from (4) is the same. However, the only data available for model fitting is $[Y(\mathcal{A}_k), e(\mathcal{A}_k), \mathbf{z}(\mathcal{A}_k), x(\mathcal{G}_i), P(\mathcal{G}_i)]$, and thus we need to make some assumptions to be able to fit the above model. The first of these is that $\mathbf{z}_{ij}(\mathcal{A}_k) = \mathbf{z}(\mathcal{A}_k)$ for all individuals in $\mathcal{A}_k$, because we only have areal unit-level confounders and not individual-level ones. This assumption simplifies the expected disease count to

$$
\begin{aligned}
\mathbb{E}[Y(\mathcal{A}_k)] &= \exp[\mathbf{z}(\mathcal{A}_k)^\top \tilde{\alpha} + \tilde{\phi}(\mathcal{A}_k)] \sum_{j=1}^{J} r_j \sum_{i \in S_j(\mathcal{A}_k)} \exp[x_{ij}(\mathcal{A}_k)\tilde{\beta}(\mathcal{A}_k)] \\
&= \exp[\mathbf{z}(\mathcal{A}_k)^\top \tilde{\alpha} + \tilde{\phi}(\mathcal{A}_k)] \sum_{j=1}^{J} n_j(\mathcal{A}_k) r_j \frac{1}{n_j(\mathcal{A}_k)} \sum_{i \in S_j(\mathcal{A}_k)} \exp[x_{ij}(\mathcal{A}_k)\tilde{\beta}(\mathcal{A}_k)].
\end{aligned}
\tag{5}
$$

However, this model still cannot be fitted in its current form, because we do not know the set of individual pollution exposures $x_{ij}(\mathcal{A}_k)$ for the ith individual in strata S_j in areal unit $\mathcal{A}_k$. Therefore we need to approximate $\frac{1}{n_j(\mathcal{A}_k)} \sum_{i \in S_j(\mathcal{A}_k)} \exp[x_{ij}(\mathcal{A}_k)\tilde{\beta}(\mathcal{A}_k)]$, the population averaged risk of pollution for people in strata j, using the gridded pollution and population data we do have. To do this we first assume that this population-level average risk does not differ by strata, that is,

$$
\frac{1}{n_j(\mathcal{A}_k)} \sum_{i \in S_j(\mathcal{A}_k)} \exp[x_{ij}(\mathcal{A}_k)\tilde{\beta}(\mathcal{A}_k)] = \frac{1}{n_h(\mathcal{A}_k)} \sum_{i \in S_h(\mathcal{A}_k)} \exp[x_{ih}(\mathcal{A}_k)\tilde{\beta}(\mathcal{A}_k)],
\tag{6}
$$

for all pairs of strata (j,h). Then we approximate this unknown population-level average risk by the population weighted average risk based on the known grid-level concentrations $x(\mathcal{G}_i)$. This approach gives a larger weight to a grid square's concentrations if more people live in that square. Specifically, we make the approximation

$$
\frac{1}{n_j(\mathcal{A}_k)} \sum_{i \in S_j(\mathcal{A}_k)} \exp[x_{ij}(\mathcal{A}_k)\tilde{\beta}(\mathcal{A}_k)] \approx \frac{\sum_{\mathcal{G}_l \in \mathcal{A}_k} P(\mathcal{A}_k \cap \mathcal{G}_l) \exp[x_l(\mathcal{G}_l)\tilde{\beta}(\mathcal{A}_k)]}{\sum_{\mathcal{G}_h \in \mathcal{A}_k} P(\mathcal{A}_k \cap \mathcal{G}_h)} \quad \text{for } j = 1, \dots, J,
\tag{7}
$$

where $P(\mathcal{A}_k \cap \mathcal{G}_l)$ is the estimated population in the intersection area $\mathcal{A}_k \cap \mathcal{G}_l$, and $\mathcal{G}_l \in \mathcal{A}_k$ denotes that grid square $\mathcal{G}_l$ has a nonzero intersection with areal unit $\mathcal{A}_k$. We note that in the simplifying case that all the populations $P(\mathcal{A}_k \cap \mathcal{G}_l)$ are equal for all l, we return to the model proposed by Wakefield and Shaddick (2006) albeit using modelled rather than monitored concentrations.

The observed population data $P(\mathcal{G}_l)$ are at the grid level, and we estimate the population in a grid square and areal unit intersection by $P(\mathcal{A}_k \cap \mathcal{G}_l) = P(\mathcal{G}_l) \frac{a(\mathcal{A}_k \cap \mathcal{G}_l)}{a(\mathcal{G}_l)}$, where $a(\mathcal{A}_k \cap \mathcal{G}_l)$ and $a(\mathcal{G}_l)$ denote the area of $\mathcal{A}_k \cap \mathcal{G}_l$ and $\mathcal{G}_l$, respectively. Thus, the intersection population $P(\mathcal{A}_k \cap \mathcal{G}_l)$ is estimated as the proportion of the population in grid square $\mathcal{G}_l$ that lies within $\mathcal{A}_k$. This assumes that population density is constant within a grid square, and these assumptions lead to the modified risk model

$$
\begin{aligned}
\mathbb{E}[Y(\mathcal{A}_k)] &= \exp[\mathbf{z}(\mathcal{A}_k)^\top \tilde{\alpha} + \tilde{\phi}(\mathcal{A}_k)] \sum_{j=1}^{J} n_j(\mathcal{A}_k) r_j \frac{\sum_{\mathcal{G}_l \in \mathcal{A}_k} P(\mathcal{A}_k \cap \mathcal{G}_l) \exp[x_l(\mathcal{G}_l)\tilde{\beta}(\mathcal{A}_k)]}{\sum_{\mathcal{G}_l \in \mathcal{A}_k} P(\mathcal{A}_k \cap \mathcal{G}_l)} \\
&= \exp[\mathbf{z}(\mathcal{A}_k)^\top \tilde{\alpha} + \tilde{\phi}(\mathcal{A}_k)] \frac{\sum_{\mathcal{G}_l \in \mathcal{A}_k} P(\mathcal{A}_k \cap \mathcal{G}_l) \exp[x_l(\mathcal{G}_l)\tilde{\beta}(\mathcal{A}_k)]}{\sum_{\mathcal{G}_h \in \mathcal{A}_k} P(\mathcal{A}_k \cap \mathcal{G}_h)} \sum_{j=1}^{J} n_j(\mathcal{A}_k) r_j \\
&= \exp[\mathbf{z}(\mathcal{A}_k)^\top \tilde{\alpha} + \tilde{\phi}(\mathcal{A}_k)] \frac{\sum_{\mathcal{G}_l \in \mathcal{A}_k} P(\mathcal{G}_l) \frac{a(\mathcal{A}_k \cap \mathcal{G}_l)}{a(\mathcal{G}_l)} \exp[x_l(\mathcal{G}_l)\tilde{\beta}(\mathcal{A}_k)]}{\sum_{\mathcal{G}_h \in \mathcal{A}_k} P(\mathcal{G}_h) \frac{a(\mathcal{A}_k \cap \mathcal{G}_h)}{a(\mathcal{G}_h)}} e(\mathcal{A}_k) \\
&= e(\mathcal{A}_k) \exp[\mathbf{z}(\mathcal{A}_k)^\top \tilde{\alpha} + \tilde{\phi}(\mathcal{A}_k)] \frac{\sum_{\mathcal{G}_l \in \mathcal{A}_k} P(\mathcal{G}_l) a(\mathcal{A}_k \cap \mathcal{G}_l) \exp[x_l(\mathcal{G}_l)\tilde{\beta}(\mathcal{A}_k)]}{\sum_{\mathcal{G}_h \in \mathcal{A}_k} P(\mathcal{G}_h) a(\mathcal{A}_k \cap \mathcal{G}_h)}.
\end{aligned}
\tag{8}
$$

Here $e(\mathcal{A}_k) = \sum_{j=1}^{J} n_j(\mathcal{A}_k) r_j$ as defined in Section 3.1 and $a(\mathcal{G}_l) = a(\mathcal{G}_h)$ as all grid squares are the same size. This model differs from that of Wakefield and Shaddick (2006) in that they did not have access to population data $P(\mathcal{G}_i)$, and hence

did not weight their pollution concentrations by estimated population density. In addition, they only considered constant pollution effects, that is, $\tilde{\beta}(\mathcal{A}_k) = \tilde{\beta}$. This mean model for $Y(\mathcal{A}_k)$ suggests that given the constraints on data availability, an appropriate *aggregate* model is

$$
\begin{aligned}
Y(\mathcal{A}_k) &\sim \text{Poisson}(\mu(\mathcal{A}_k)) \quad k = 1, \dots, K \\
\mu(\mathcal{A}_k) &= e(\mathcal{A}_k) \exp[\mathbf{z}(\mathcal{A}_k)^\top \tilde{\alpha} + \tilde{\phi}(\mathcal{A}_k)] \frac{\sum_{\mathcal{G}_l \in \mathcal{A}_k} P(\mathcal{G}_l) a(\mathcal{A}_k \cap \mathcal{G}_l) \exp[x_l(\mathcal{G}_l)\tilde{\beta}(\mathcal{A}_k)]}{\sum_{\mathcal{G}_h \in \mathcal{A}_k} P(\mathcal{G}_h) a(\mathcal{A}_k \cap \mathcal{G}_h)} \\
\alpha_r &\sim \text{N}(0, 100{,}000) \quad \text{for} \quad r = 1, \dots, p \\
\phi(\mathcal{A}_k)|\boldsymbol{\phi}(-\mathcal{A}_k) &\sim \text{N}\left(\frac{\tilde{\rho} \sum_{r=1}^{K} w_{kr} \tilde{\phi}(\mathcal{A}_r)}{\tilde{\rho} \sum_{r=1}^{K} w_{kr} + 1 - \tilde{\rho}}, \frac{\tilde{\tau}^2}{\tilde{\rho} \sum_{r=1}^{K} w_{kr} + 1 - \tilde{\rho}} \right) \\
\tilde{\rho} &\sim \text{Uniform}(0, 1) \\
\tilde{\tau}^2 &\sim \text{Inverse gamma}(1, 0.01),
\end{aligned}
\tag{9}
$$

where the random effects model is identical to that outlined in Section 3.2. A Poisson data likelihood model is assumed here because once spatially aggregated the disease variable $Y(\mathcal{A}_k)$ is a count of the number of events (hospitalizations or mortalities) from the population living in areal unit $\mathcal{A}_k$. Thus, the main difference between the naive ecological model (1) and the aggregate model (9) is that the former first averages over the gridded pollution concentrations within each areal unit to create a single pollution measure, $\hat{x}(\mathcal{A}_k)$, in the risk model, while the aggregate model averages the risks, $\exp[x_l(\mathcal{G}_l)\tilde{\beta}(\mathcal{A}_k)]$, across all grid squares $\mathcal{G}_l$ within each areal unit. We undertake a simulation study in the next section to illustrate which of these two competing models better estimates the effects of air pollution on human health.

3.4 | Constant and varying coefficients models for $\tilde{\beta}(\mathcal{A}_k)$

We consider both constant and varying coefficients for $\tilde{\beta}(\mathcal{A}_k)$, with the former used because it gives a single overall effect size and is commonly used in existing studies. We use the later to quantify whether air pollution disproportionally affects less affluent communities, thus making it a cause of health inequalities. We describe our varying coefficient model in terms of the aggregate model parameter $\tilde{\beta}(\mathcal{A}_k)$, but note that we also use it for the ecological model parameter $\beta(\mathcal{A}_k)$. The prior distributions for the constant and varying coefficient models are given below, where in both cases weakly informative Gaussian priors are specified for the unknown parameters.

Constant $\tilde{\beta}(\mathcal{A}_k) = \tilde{\beta}$ and $\tilde{\beta} \sim \text{N}(0, 100{,}000)$.
Varying $\tilde{\beta}(\mathcal{A}_k) = \tilde{\delta}_0 + \tilde{\delta}_1 m(\mathcal{A}_k)$ and $\tilde{\delta}_0, \tilde{\delta}_1 \sim \text{N}(0, 100{,}000)$.

For the varying coefficient model $\tilde{\beta}(\mathcal{A}_k)$ is assumed to be linearly related to an areal unit-level measure of socioeconomic deprivation or affluence $m(\mathcal{A}_k)$, called an *effect modifier*. In initial analysis of the study data we investigated nonlinear relationships with affluence via cubic regression splines, but the results showed no evidence against linearity and hence this simpler model is presented here. This linear model is also consistent with the earlier work of Lee (2012).

3.5 | Inference

Inference for both models is conducted in a Bayesian framework using Markov chain Monte Carlo (MCMC) simulation. The *ecological* model (1) and (2) can be implemented using the CARBayes package (Lee, 2013), while software in the form of an R implementation to run the *aggregate* model (9) with and without the varying coefficient effect is available at https://github.com/duncanplee/Aggregate-spatial-models. Also available in that repository is an R code file that runs a sample analysis using all four models (ecological and aggregate models with constant and varying pollution–health effects), as well as an R code file to run the simulation study presented below. Finally, the repository contains a subset of the data analyzed here, but the observed disease counts have been jittered to preserve confidentiality as they are not publicly available.

4 | SIMULATION STUDY

This section presents a novel simulation study that is the first to answer the following key questions: (i) What impact does the scaling and zoning components of the MAUP have on air pollution and health effects? and (ii) Does the aggregate model (9) perform better than the naive ecological model (1)?

4.1 | Data generation

Data are generated on a 71×71 square grid study region $\mathcal{A}$ comprising just over 5,000 grid squares ($M = 5,041$) and representing a pseudocontinuous spatial surface, and the data generated at this scale are subsequently spatially aggregated to observe the effect of the MAUP. For each simulated data set a population of around 3.5 million individuals are generated across this region, where the grid square population counts are denoted by $P(\mathcal{G}_i)$. These population counts are generated on the square root scale from a spatially correlated multivariate normal distribution, which ensures that "cities" and "rural areas" are created that comprise geographically adjacent groups of grid squares that respectively have high and low populations. The square root scale is used to ensure that the distribution of population counts is skewed to the right when back transformed by squaring, because cities have much higher population counts than rural areas.

The disease counts are generated from $Y(\mathcal{G}_i) \sim \text{Poisson}(e(\mathcal{G}_i)\theta(\mathcal{G}_i))$, where the expected counts $e(\mathcal{G}_i) = \xi P(\mathcal{G}_i)$ depend on the population size and the disease prevalence denoted by ξ. Grid-level disease risk $\theta(\mathcal{G}_i)$ is modeled on the log scale by $\ln[\theta(\mathcal{G}_i)] = x(\mathcal{G}_i)\beta + \phi(\mathcal{G}_i)$, where the regression parameter is fixed at $\beta = .1$. The pollution covariate $x(\mathcal{G}_i)$ is generated as either independent or spatially correlated, so that the impact of the MAUP on these two different types of covariates can be observed. Both pollution covariate types are generated from zero-mean multivariate normal distributions, with the independent covariate having an identity correlation matrix $\mathbf{\Sigma} = \mathbf{I}$. The correlated covariate has a correlation matrix defined by the spatial exponential correlation function $\mathbf{\Sigma} = \exp(-\upsilon \mathbf{D})$, where $\mathbf{D}$ is an $M \times M$ distance matrix between all pairs of grid squares and $\upsilon = 0.5$ is the dependence parameter chosen so that the correlation between adjacent grid squares is 0.6. The random effects $\phi(\mathcal{G}_i)$ are included to mimic the presence of spatially correlated unmeasured confounding and are also generated from a zero-mean multivariate normal distribution with an exponential correlation function with $\upsilon = 0.1$ (adjacent grid squares have a correlation of 0.9).

Once a grid-level data set has been generated it is spatially aggregated to K randomly generated nonoverlapping areal units, which is the type of data available in practice. This is achieved by randomly choosing K grid squares as anchor points, and then assigning the remaining grid squares to the areal unit that has the closest anchor point. The grid level observed and expected disease counts $(Y(\mathcal{G}_i), e(\mathcal{G}_i))$ are summed within areal units to obtain the aggregated counts $(Y(\mathcal{A}_k), e(\mathcal{A}_k))$, while the pollution covariate $x(\mathcal{G}_i)$ is averaged across each areal unit to obtain $x(\mathcal{A}_k)$. This spatial aggregation process is repeated $Q = 20$ times with different spatial aggregations, yielding Q differentially aggregated areal unit data sets for analysis from each grid-level data set. The variation in the regression parameter estimates $\{\hat{\beta}_1, \ldots, \hat{\beta}_Q\}$ from these Q spatially aggregated data sets quantifies the size of the zoning effect component of the MAUP. Data are generated under 48 different scenarios based on all possible combinations of the following five factors, which assesses the effect of the MAUP to how the data were generated.

1. *Scaling effect* The number of areal units is set to $K = 100$ (on average 50 grid squares per areal unit), $K = 500$ (on average 10 grid squares per areal unit), or $K = 1,000$ (on average five grid squares per areal unit).
2. *Disease prevalence* The disease is relatively common $e(\mathcal{G}_i) = 0.05P(\mathcal{G}_i)$ (one in 20 people have it) or rarer $e(\mathcal{G}_i) = 0.01P(\mathcal{G}_i)$ (one in 100 people have it).
3. *Pollution covariate correlation* The pollution covariate is independent or correlated in space, which is determined by $\mathbf{\Sigma}$ as described above.
4. *Pollution covariate variation* The pollution covariate has relatively little ($\text{SD}[x(\mathcal{G}_i)] = 1$) or large ($\text{SD}[x(\mathcal{G}_i)] = 5$) variation across space, where $\text{SD}[.]$ denotes standard deviation.
5. *Residual confounding* The unmeasured spatial confounding is relatively small ($\text{SD}[\phi(\mathcal{G}_i)] = 0.04$) or large ($\text{SD}[\phi(\mathcal{G}_i)] = 0.2$).

In each scenario $C = 50$ grid-level data sets are generated, each of which is then spatially aggregated in $Q = 20$ different ways, yielding 1,000 spatially aggregated data sets under each scenario. We fit two separate models to each of these 1,000 aggregated data sets in each scenario, the first of which is the naive ecological model (denoted Ecological) given by (1)

that is applied to the aggregated data $\{Y(\mathcal{A}_k), e(\mathcal{A}_k), x(\mathcal{A}_k)\}$. The second is the aggregate model (denoted Aggregate) given by (9), which is applied to the aggregated disease data $\{Y(\mathcal{A}_k), e(\mathcal{A}_k)\}$ and the grid-level pollution data $\{x(\mathcal{G}_i)\}$, the latter mimicking the available grid-level pollution data from the motivating study. Inference for both models is based on 2,000 MCMC samples, which were obtained by running a Markov chain for 300,000 samples, removing the first 100,000 as the burnin period, and thinning the remaining samples by 100 to remove their autocorrelation.

4.2 | Results

We present the results in two stages: the first quantifies the impact of the MAUP, while the second quantifies the accuracy with which the true pollution–health effect is estimated.

4.2.1 | The impact of the MAUP

The impacts of the scaling and zoning components of the MAUP are presented in Table 2, which quantifies the variation in $\{\hat{\beta}_1, \ldots, \hat{\beta}_Q\}$ from the $Q = 20$ spatially aggregated data sets relating to each grid-level data set. Specifically, the interquartile range (IQR) is computed for these $Q = 20$ estimates, and then the median of these IQRs is computed for the $C = 50$ grid-level data sets generated under each scenario. These average IQRs are presented in the table as percentages of the true value of $\beta = .1$ to make the results interpretable, and the table shows a number of key findings. First, the median IQR is relatively low if you have at least 500 areal units and the pollution covariate is spatially correlated, with median IQRs less than 10% of the true value of β in all cases. This suggests that the zoning component of the MAUP should have relatively little impact on the estimates from the motivating study, because there the pollution covariate is correlated and the number of IZs is over 1,000. Second, the impact of the zoning effect reduces as the number of areal units K increases (the scaling effect), which is not surprising because one has more data points with which to estimate β and is getting closer to the grid (individual) level. Third, the estimates from the aggregate model almost always exhibit less variation than those from the ecological model, with percentage differences around 1% in most scenarios.

Fourth, the remaining factors considered here all influence the magnitude of the zoning effect within the MAUP, with reductions in the zoning effect occurring if (i) the pollution covariate is spatially correlated with a high variation; (ii) the disease is common, and (iii) the level of residual spatial confounding is low. If the pollution covariate is spatially correlated then neighboring grid squares will have similar covariate values, hence different spatial aggregations of these grid squares will have similar sets of areal-level covariate values across the K areal units, resulting in similar estimated risks and thus reducing the zoning effect. As the variation in the pollution covariate increases there are greater contrasts in its areal-level values across the K areal units, leading to more precise estimation of β and thus reducing the variation between estimates from different spatial aggregations. Similarly, as the disease prevalence increases there are more disease cases across the study region. This increases the size of, and the variation between, the areal unit-level disease counts, which provides more precise estimates of covariate effects and hence a reduction in the zoning effect between different spatial aggregations. Finally, as the amount of residual spatial confounding reduces the grid-level risks are more homogeneous between neighboring grid squares, meaning that the choice of spatial aggregation has less effect on the set of K areal-level average risks, resulting in more consistent estimation of β (reduction of the zoning effect).

4.2.2 | The accuracy of β

The accuracy with which β can be estimated is presented in Table 3, which displays root mean square errors (RMSE) of $\hat{\beta}$ from all 1,000 aggregated data sets for each model and scenario. As before, the results are presented as percentages of the true value of $\beta = .1$. The table shows that as long as the pollution covariate is spatially correlated and there are more than 500 areal units, as is the case in the motivating study, then estimation accuracy is good, with RMSEs less than 12% of the true value in all cases. Estimates from the aggregate model are almost always better than those from the ecological model, with average reductions in RMSE of around 1%. Increasing the number of areal units increases estimation accuracy (reduces the RMSE) markedly, as does having: (i) a pollution covariate that is spatially correlated with a high variation; (ii) a common disease with larger numbers of disease events, and (iii) a low level of residual spatial confounding.

TABLE 2 Estimated impacts of the scaling and zoning effects in the MAUP from the simulation study for different scenarios

Scenario	Scaling effect K	Correlated covariate		Independent covariate	
		Ecological	Aggregate	Ecological	Aggregate
Prevalence—Rare	$K = 100$	11.0	10.2	37.7	34.0
Covariate variation—Low	$K = 500$	5.0	4.9	14.6	14.4
Residual confounding—Low	$K = 1,000$	4.4	4.3	10.4	9.9
Prevalence—Rare	$K = 100$	26.3	25.1	81.7	73.9
Covariate variation—Low	$K = 500$	9.8	9.2	19.4	20.1
Residual confounding—High	$K = 1,000$	7.1	7.0	12.3	11.6
Prevalence—Rare	$K = 100$	3.3	2.0	9.6	5.6
Covariate variation—High	$K = 500$	1.3	0.9	3.6	2.0
Residual confounding—Low	$K = 1,000$	1.0	0.8	2.1	1.6
Prevalence—Rare	$K = 100$	6.2	4.9	16.9	11.9
Covariate variation—High	$K = 500$	2.2	1.9	4.2	2.9
Residual confounding—High	$K = 1,000$	1.5	1.3	2.5	2.0
Prevalence—Common	$K = 100$	7.2	6.2	23.5	21.3
Covariate variation—Low	$K = 500$	3.0	2.9	7.4	7.2
Residual confounding—Low	$K = 1,000$	2.3	2.2	4.2	4.3
Prevalence—Common	$K = 100$	27.6	25.3	68.5	65.8
Covariate variation—Low	$K = 500$	8.5	7.8	14.2	13.0
Residual confounding—High	$K = 1,000$	5.4	5.2	7.2	6.9
Prevalence—Common	$K = 100$	3.3	1.3	7.2	2.9
Covariate variation—High	$K = 500$	1.0	0.5	2.9	1.1
Residual confounding—Low	$K = 1,000$	0.7	0.4	1.4	0.7
Prevalence—Common	$K = 100$	6.7	5.1	16.4	14.6
Covariate variation—High	$K = 500$	2.4	1.6	4.0	2.3
Residual confounding—High	$K = 1,000$	1.4	0.9	2.0	1.2

Note: First, the interquartile range of $\{\hat{\beta}_1, \ldots, \hat{\beta}_Q\}$ is computed that arise from the $Q = 20$ spatially aggregated areal unit data sets from each grid-level data set. Then the average (median) of these IQRs from the $C = 50$ grid-level data sets is presented in the table above to quantify the zoning effect. In all cases this median value is presented as a percentage of the true value of $\beta = .1$.

5 | RESULTS FROM THE MOTIVATING STUDY

5.1 | Model description

Both the ecological and aggregate models are fitted to each disease outcome separately, to ensure that the estimated associations are not influenced by any cross correlations between diseases. Only a single pollutant is included in each model run because the three pollutants are highly correlated, which results in $18 = 6 \times 3$ disease and pollutant combinations. In each case the observed disease count variable, $Y(\mathcal{A}_k)$, is modeled by the expected number of disease cases, $e(\mathcal{A}_k)$, as an offset in the model, one of the three pollutants, a vector of socioeconomic confounders, and the spatially correlated random effects. The socioeconomic confounders include the access to services, crime, housing, and income domains of the SIMD, with the other domains being removed due to collinearity as discussed in Section 2.4.

The ecological and aggregate models are fitted first assuming a constant pollution–health effect to provide a comparison with existing studies, before relaxing that assumption and allowing the estimated effect size to vary across the IZs. In the latter case our aim is to estimate the extent to which the pollution–health effect is modified by the level of

TABLE 3 Root mean square errors (RMSE) of the regression parameter estimates $\{\hat{\beta}\}$ for each scenario

Scenario	Scaling effect K	Correlated covariate		Independent covariate	
		Ecological	Aggregate	Ecological	Aggregate
Prevalence—Rare	$K = 100$	12.8	12.7	40.8	36.3
Covariate variation—Low	$K = 500$	8.1	8.0	16.2	15.8
Residual confounding—Low	$K = 1,000$	6.3	6.4	11.5	11.3
Prevalence—Rare	$K = 100$	32.5	33.0	79.2	77.0
Covariate variation—Low	$K = 500$	11.9	11.6	20.0	19.8
Residual confounding—High	$K = 1,000$	9.5	9.5	12.4	11.6
Prevalence—Rare	$K = 100$	3.8	2.6	8.9	5.3
Covariate variation—High	$K = 500$	1.6	1.2	3.3	2.2
Residual confounding—Low	$K = 1,000$	1.4	1.3	2.4	1.7
Prevalence—Rare	$K = 100$	7.1	5.7	17.8	11.7
Covariate variation—High	$K = 500$	2.7	2.3	4.5	3.2
Residual confounding—High	$K = 1,000$	2.1	2.0	2.7	2.1
Prevalence—Common	$K = 100$	8.6	8.3	22.9	21.4
Covariate variation—Low	$K = 500$	4.5	4.5	7.4	7.0
Residual confounding—Low	$K = 1,000$	3.0	3.0	4.9	4.7
Prevalence—Common	$K = 100$	37.5	40.0	70.8	69.3
Covariate variation—Low	$K = 500$	9.7	9.3	13.0	12.2
Residual confounding—High	$K = 1,000$	6.8	6.7	7.5	7.2
Prevalence—Common	$K = 100$	3.5	1.4	6.8	3.3
Covariate variation—High	$K = 500$	1.3	0.8	2.6	1.1
Residual confounding—Low	$K = 1,000$	1.0	0.6	1.7	0.7
Prevalence—Common	$K = 100$	8.7	8.6	18.4	22.7
Covariate variation—High	$K = 500$	2.2	1.7	3.8	2.2
Residual confounding—High	$K = 1,000$	1.7	1.4	2.0	1.2

Note: All values are presented as a percentage of the true value of $\beta = .1$.

socioeconomic deprivation in an IZ. Our hypothesis is that there may be larger pollution–health effects in less afflu-ent communities, which will increase the magnitude of health inequalities. To assess this hypothesis we allow the pollution–health effect $\tilde{\beta}(\mathcal{A}_k)$ to depend on the natural log of the average property price in an IZ. Property price was cho-sen because it was a measure of affluence that was not in the SIMD and hence not in the covariate part of the model. A natural log transformation was taken so that the variable was symmetric and did not contain large outliers that may affect the results. We assume a linear relationship $\tilde{\beta}(\mathcal{A}_k) = \tilde{\delta}_0 + \tilde{\delta}_1 m(\mathcal{A}_k)$ between the pollution–health effect in areal unit $\mathcal{A}_k$, $\tilde{\beta}(\mathcal{A}_k)$, and the log of the average property price in area $\mathcal{A}_k$, $m(\mathcal{A}_k)$, because exploratory analyses of these data with nonlinear spline based relationships showed no evidence of nonlinearity.

Inference for all models fitted to the study data is based on 6,000 MCMC samples generated from three parallel Markov chains, each of which was burnt in for 100,000 samples (by which point convergence was assessed to have been reached), before being run for a further 400,000 samples which were then thinned by 200 to remove their autocorrelation. The results for the constant pollution–health effects are presented in the next section, while the varying effects are presented in Section 5.3.

TABLE 4 Estimated relative risks between each pollutant and disease outcome

Disease outcome	Model	NO$_2$	PM$_{2.5}$	PM$_{10}$
Circulatory hospitalization	Ecological	1.000 (0.991, 1.009)	1.007 (0.992, 1.023)	1.008 (0.992, 1.026)
	Aggregate	0.999 (0.988, 1.010)	1.008 (0.987, 1.028)	1.010 (0.986, 1.034)
Circulatory mortality	Ecological	**0.989 (0.981, 0.997)**	0.996 (0.983, 1.010)	1.005 (0.990, 1.021)
	Aggregate	**0.990 (0.982, 0.998)**	0.999 (0.985, 1.015)	1.009 (0.992, 1.027)
Respiratory hospitalization	Ecological	1.010 (0.998, 1.022)	**1.041 (1.020, 1.062)**	**1.049 (1.025, 1.072)**
	Aggregate	**1.020 (1.006, 1.034)**	**1.058 (1.032, 1.086)**	**1.052 (1.026, 1.088)**
Respiratory mortality	Ecological	**1.022 (1.010, 1.033)**	**1.040 (1.019, 1.060)**	**1.034 (1.011, 1.058)**
	Aggregate	**1.028 (1.015, 1.040)**	**1.041 (1.017, 1.064)**	**1.035 (1.008, 1.061)**
Dementia mortality	Ecological	1.009 (0.975, 1.042)	1.010 (0.954, 1.070)	1.002 (0.936, 1.070)
	Aggregate	1.028 (0.992, 1.069)	1.044 (0.972, 1.117)	1.018 (0.939, 1.101)
Total nonaccidental mortality	Ecological	1.001 (0.994, 1.007)	1.009 (0.997, 1.021)	**1.013 (1.000, 1.026)**
	Aggregate	1.004 (0.996, 1.011)	1.012 (0.999, 1.025)	**1.017 (1.001, 1.032)**

Note: The relative risks relate to a 3 μg m^{-3} increase for NO$_2$, a 1 μg m^{-3} increase for PM$_{2.5}$, and a 2 μg m^{-3} for PM$_{10}$, which are close to their spatial standard deviations. Significant effects are in bold.

5.2 | Constant pollution–health effects

The estimated pollution–health effects from each model and pollutant–disease combination are presented in Table 4. All results are presented as relative risks for the following increases in each pollutant: NO$_2$—3 μg m^{-3}, PM$_{2.5}$—1 μg m^{-3}, and PM$_{10}$—2 μg m^{-3}. These increases are chosen for the pollutants because they are close to their respective spatial standard deviations, thus making them realistic increases in concentrations that might occur. The main message from Table 4 is that air pollution has consistent significant effects on respiratory disease (hospitalizations and mortality), while no effects are observed for any of the other disease outcomes. For respiratory disease 11 out of the 12 relative risks (2 disease outcomes × 2 models × 3 pollutants) have 95% credible intervals wholly greater than 1, with point estimates ranging between a 2% and a 5.8% increased risk for the given increase in the pollutant. By contrast, there is consistently no effect observed for circulatory disease, with the one significantly negative association likely being due to the multiple testing inherent when estimating the 12 pollutant–disease effects. Similarly, there is no consistent association for dementia mortality although some of the point estimates are relatively large, and the lack of significance may be due to the wide uncertainty intervals that arise from having very small disease counts for dementia, a disease which is a relatively rare cause of death compared with respiratory disease. In addition, there is no consistent association with total nonaccidental mortality either across the six pollution–health estimates, with only PM$_{10}$ showing significant associations. Finally, there are some differences between the estimates from the aggregate and ecological models, for example, the associations estimated between respiratory hospitalizations and PM$_{2.5}$, and the simulation study suggests that the aggregate model is likely to provide the more accurate results.

5.3 | Varying pollution–health effects

Varying pollution–health effects were only estimated for the respiratory disease outcomes (hospitalization and mortality), because the remaining outcomes showed no evidence of significant relationships with air pollution from Table 4. The results are displayed in Figure 2, where separate effects are shown for each pollutant (in each row of the figure), disease outcome (hospitalization and mortality) and model (ecological and aggregate). Each figure presents the linear relationship between the relative risk of air pollution and affluence, where the latter is represented by the natural log of the average property price in each IZ. The solid line depicts the posterior mean, while the dashed lines represent 95% credible intervals. The relative risk scales relate to the same increases in each pollutant that were presented in Table 4, namely, NO$_2$—3 μg m^{-3}; PM$_{2.5}$—1 μg m^{-3}; and PM$_{10}$—2μg m^{-3}. The main finding is that the relative risk of air pollution decreases
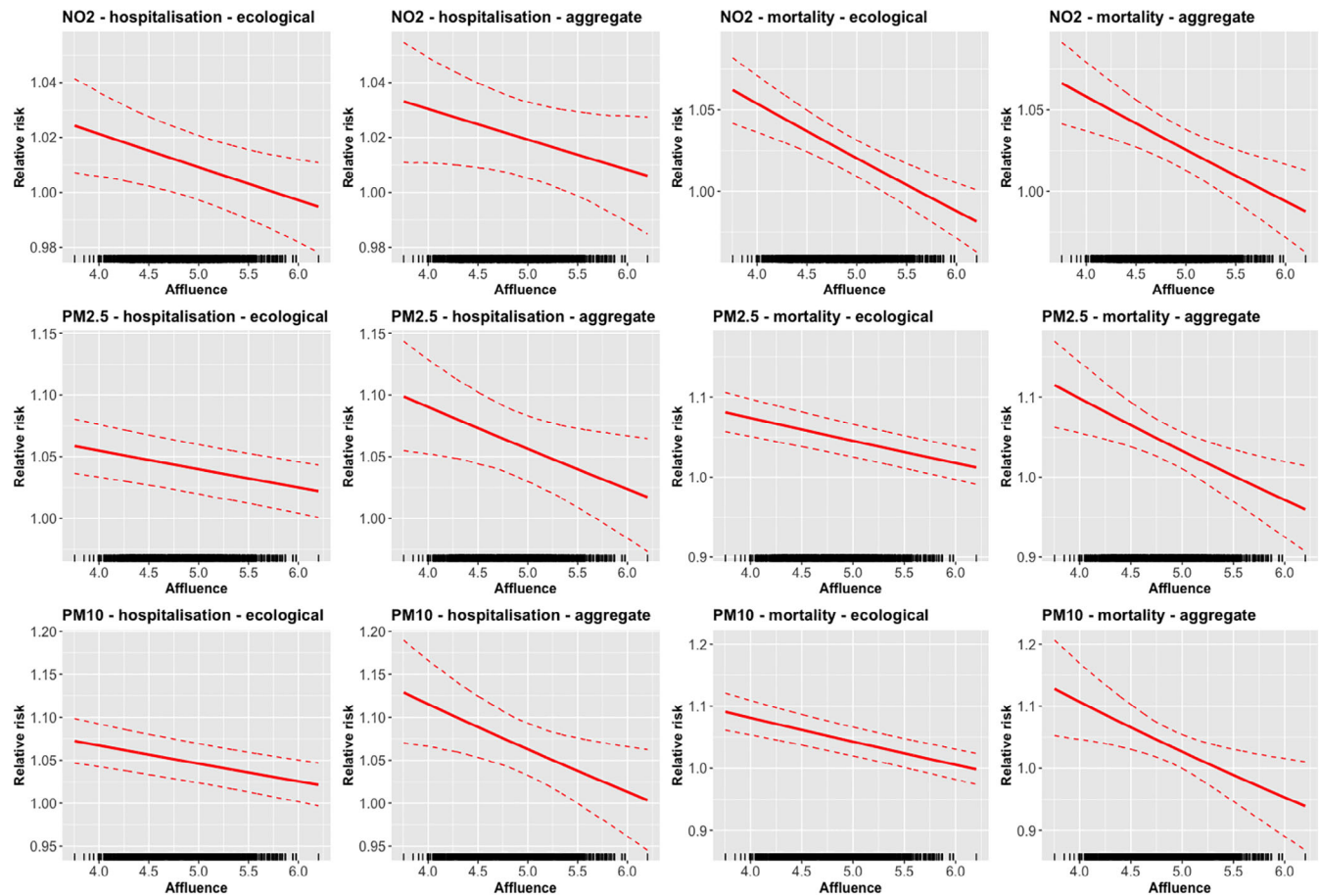
FIGURE 2 The estimated relationship between the pollution–disease relative risks for each disease outcome, pollutant and model (ecological and aggregate), and affluence. Affluence is measured by the natural log of the average property price in each IZ

as the level of affluence increases, which is consistently observed for all pollutants, disease outcomes, and models. This suggests that air pollution does widen health inequalities, because it has a greater impact on less affluent communities compared with more affluent ones. These differences in relative risk are significant for the mortality outcome for all pollutants, as well as for the PM_{10}–hospitalization association, because one cannot draw a horizontal line between the 95% credible intervals.

6 | DISCUSSION

This article has investigated the extent to which the modifiable areal unit problem is likely to affect the population-level air pollution and health associations estimated from spatial areal unit studies, which is a commonly made criticism of this study design. Our main finding is that as long as the pollution concentrations are spatially correlated and that the number of areal units is large, both of which occur in the motivating study, then the MAUP is unlikely to have a sizeable impact on the results. For example, the simulation study showed that the average IQR in the set of effect estimates resulting from different spatial aggregations of the same data (the zoning effect) was less than 10% of the true effect size. This suggests that the estimated associations from the motivating study are of the correct order of magnitude and not affected by the MAUP to any great extent. This would not be the case if the covariate of interest was independent and/or the number of spatial units was low, in which case the simulation study showed that the zoning effect could cause the variation (as measured by the median IQR) in the effects estimated from different spatial aggregations of the same data to be up to 80% of the true effect size, hence making the estimates unreliable.

Our second main finding is that the aggregate model derived in Section 3 generally outperforms the ecological model commonly used in studies of this type, both in terms of the impact of the MAUP and estimation accuracy of the pollution–health association. However, the difference in estimation performance between the aggregate and ecological

models is not large, with improvements in RMSE of around 1% of the size of the true association in most scenarios considered. Thus, while the aggregate model does not remove the impact of the MAUP it does mitigate it slightly, and is thus the model of choice for spatial ecological association studies.

The main finding from our motivating study is that increased levels of air pollution were consistently associated with increased risks of respiratory disease (both hospitalization and mortality), whereas no consistent associations were observed for circulatory disease. This agrees with the existing epidemiological literature in Scotland such as Willocks et al. (2012), Yap et al. (2012), and Lee et al. (2019), which also shows associations for respiratory but not for circulatory disease using a variety of study designs. Thus, the Scottish evidence is consistent with the majority of the international evidence on the existence of an association between air pollution and respiratory disease. However, the consistent failure to identify evidence of a significant association between air pollution and cardiovascular disease in Scottish studies is at odds with most of the international literature, with a review of the latter provided by World Health Organisation (2013). Therefore, future work is required to understand the reasons for these differences, which could be due to numerous factors including lower average pollutant concentrations in Scotland, or differences in the underlying causes and cofactors determining the Scottish population's vulnerability to circulatory disease. No associations were also observed between any pollutant and mortality attributed to dementia, although as some of the point estimates were relatively large (e.g., with $PM_{2.5}$) this might be due to the wide uncertainty intervals that result from a lack of power caused by their being small numbers of disease events over the 7-year study period. Our other main finding is that air pollution appears to consistently have a greater impact on less affluent communities compared with more affluent ones, so widening the level of health inequalities, the reduction of which is a current priority for Scottish Government (NHS Health Scotland, 2016).

The major limitation of this and all other epidemiological air pollution studies is that the air pollution exposure for each areal unit has been estimated using population-weighted outdoor residential concentrations. As the derivation of the aggregate model in Section 3 shows, using individual pollution exposures from the populations living in each areal unit would make the model more realistic, likely resulting in more accurate pollution–health associations. However, personal exposure estimates need to allow for the fact that people spend a large proportion of their time indoors and are hence exposed to indoor sources of pollution, as well as accounting for movement patterns away from their residence such as traveling to work, the shops, and so forth on a daily basis. Such personal exposure profiles are not currently available, and their estimation constitutes the next significant challenge in air pollution research.

DATA AVAILABILITY STATEMENT
The data and code that support the findings of this study are available at https://github.com/duncanplee/Aggregate-spatial-models, although the disease count data have been jittered as they are not publicly available

ORCID
Duncan Lee 🄳 https://orcid.org/0000-0002-6175-6800

REFERENCES
Aregay, M., Lawson, A., Faes, C., Kirby, R., Carroll, R., & Watjou, K. (2018). Zero-inflated multiscale models for aggregated small area health data. *Environmetrics*, *29*, e2477.

Bellini, P., Baccini, M., Biggeri, A., & Terracini, B. (2007). The meta-analysis of the Italian studies on short-term effects of air pollution (misa): Old and new issues on the interpretation of the statistical evidences. *Environmetrics*, *18*, 219–229.

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

Chen, H., Kwong, J., Copes, R., Hystad, P., Donkelaar, A., Tu, K., … Burnett, R. (2017). Exposure to ambient air pollution and the incidence of dementia: A population-based cohort study. *Environment International*, *108*, 271–277.

Clifford, S., Low-Choy, S., Mazaheri, M., Salimi, F., Morawska, L., & Mengersen, K. (2019). A bayesian spatiotemporal model of panel design data: Airborne particle number concentration in Brisbane, Australia. *Environmetrics*, *30*(7), e2597.

Department for the Environment, Food and Rural Affairs (2015). National air quality objectives and european directive limit and target values for the protection of human health. Retrieved from https://uk-air.defra.gov.uk/assets/documents/Air_Quality_Objectives_Update.pdf.

European Parliament. (2008). Directive 2008/50/EC of the European parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union*, *152*, 1–44.

Gilani, O., Berrocal, V., & Batterman, S. (2019). Nonstationary spatiotemporal bayesian data fusion for pollutants in the near-road environment. *Environmetrics*, *30*(7), e2581.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B*, *55*, 757–796.

NHS Health Scotland (2016). Health inequalities - what are they and how do we reduce them? Retrieved from http://www.healthscotland.scot/media/1086/health-inequalities-what-are-they-how-do-we-reduce-them-mar16.pdf.

Keller, J., & Peng, R. (2019). Error in estimating area-level air pollution exposures for epidemiology. *Environmetrics*, *30*(8), e2573.

Kleinschmidt, I., Hills, M., & Elliott, P. (1995). Smoking behaviour can be predicted by neighbourhood deprivation measures. *Journal of Epidemiology and Community Health*, *49*, S71–S77.

Lee, A., Szpiro, A., Kim, S., & Sheppard, L. (2015). Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics*, *26*(4), 255–267.

Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, *2*, 79–89.

Lee, D. (2012). Using spline models to estimate the varying health risks from air pollution across Scotland. *Statistics in Medicine*, *31*, 3366–3378.

Lee, D. (2013). CARBayes: an R package for Bayesian spatial modelling with conditional autoregressive priors. *Journal of Statistical Software*, *55*(13).

Lee, D. (2018). A locally adaptive process-convolution model for estimating the health impact of air pollution. *Annals of Applied Statistics*, *12*, 2540–2558.

Lee, D., Robertson, C., Ramsay, C., Gillespie, C., & Napier, G. (2019). Estimating the health impact of air pollution in Scotland, and the resulting benefits of reducing concentrations in city centres. *Spatial and Spatio-Temporal Epidemiology*, *29*, 85–96.

Lee, D., & Sarran, C. (2015). Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies. *Environmetrics*, *26*(7), 477–487.

Leroux, B., Lei, X., & Breslow, N. (2000). *Statistical models in epidemiology, the environment and clinical trials*. In M. Halloran & D. Berry (Eds.), *Estimation of disease rates in small areas: A new mixed model for spatial dependence* (pp. 135–178). New York, NY: Springer-Verlag.

Manley, D. (2014). *Handbook of regional science*. In M. Fischer, P. Nijkamp, & D. Berry (Eds.), *Scale, aggregation and the modifiable areal unit problem* (pp. 1157–1171). Heidelberg/Germany: Springer-Verlag.

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

Nicolis, O., Díaz, M., Sahu, S., & Marín, J. (2019). Bayesian spatiotemporal modeling for estimating short-term exposure to air pollution in Santiago de Chile. *Environmetrics*, *30*(7), e2574.

Openshaw, S. (1984). *The modifiable areal unit problem (concepts and techniques in modern geography)*. Norwich, England: GeoBooks.

Openshaw, S., & Taylor, P. (1979). *A million or so correlation coefficients: Three experiments on the modifiable areal unit problem* (pp. 127–144). London, Engand: Pion.

Oudin, A., Forsberg, B., Nordin Adolfsson, A., Lind, N., Modig, L., Nordin, M., … Nilsson, L. (2016). Traffic-related air pollution and dementia incidence in northern Sweden: A longitudinal study. *Environmental Health Perspectives*, *124*, 306–312.

Parenteau, M., & Sawada, M. (2011). The modifiable areal unit problem (MAUP) in the relationship between exposure to NO_2 and respiratory health. *International Journal of Health Geographics*, *10*, 58.

Reis, S., Liska, T., Steinle, S., Carnell, E., Leaver, D., Roberts, E., …, Dragosits, U. (2017). UK gridded population 2011 based on census 2011 and land cover map 2015.

Royal College of Physicians (2016). Every breath we take: The lifelong impact of air pollution. Retrieved from https://www.rcplondon.ac.uk/projects/outputs/every-breath-we-take-lifelong-impact-air-pollution.

Strand, M., Sillau, S., Grunwald, G., & Rabinovitch, N. (2015). Regression calibration with instrumental variables for longitudinal models with interaction terms, and application to air pollution studies. *Environmetrics*, *26*(6), 393–405.

Wakefield, J., & Salway, R. (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society Series A*, *164*, 119–137.

Wakefield, J., & Shaddick, G. (2006). Health-exposure modeling and the ecological fallacy. *Biostatistics*, *7*, 438–455.

Willocks, L., Bhaskar, A., Ramsay, C., Lee, D., Brewster, D., Fischbacher, C., … Scott, M. (2012). Cardiovascular disease and air pollution in Scotland: No association or insufficient data and study design? *BMC Public Health*, *12*, 227.

World Health Organisation (2013). *Review of evidence on health aspects of air pollution – REVIHAAP project*.

World Health Organisation (2016). Ambient air pollution: A global assessment of exposure and burden of disease. Retrieved from http://apps.who.int/iris/bitstream/10665/250141/1/9789241511353-eng.pdf.

Yap, C., Beverland, I., Heal, M., Cohen, G., Robertson, C., Henderson, D., … Agius, R. (2012). Association between long-term exposure to air pollution and specific causes of mortality in Scotland. *Occupational and Environmental Medicine*, *69*, 916–924.

Zhu, L., Carlin, B., & Gelfand, A. (2003). Hierarchical regression with misaligned spatial data: Relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics*, *14*, 537–557.