International Journal of Applied Earth Observation and Geoinformation 22 (2013) 65–74

Contents lists available at SciVerse ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Ecological bias in studies of the short-term effects of air pollution on health

Gavin Shaddick[a,*], Duncan Lee[b], Jonathan Wakefield[c]

[a] Department of Mathematical Sciences, University of Bath, UK
[b] Department of Statistics, University of Glasgow, UK
[c] Departments of Statistics and Biostatistics, University of Washington, USA

ARTICLE INFO

Article history:
Received 11 July 2011
Accepted 26 March 2012

Keywords:
Air pollution
Health effects
Particulate matter ecological bias
Exposure modelling

ABSTRACT

There has been a great deal of research into the short-term effects of air pollution on health with a large number of studies modelling the association between aggregate disease counts and environmental exposures measured at point locations, for example via air pollution monitors. In such cases, the standard approach is to average the observed measurements from the individual monitors and use this in a log-linear health model. Hence such studies are ecological in nature being based on spatially aggregated health and exposure data. Here we investigate the potential for bias in the estimates of the effects on health when estimating the short-term effects of air pollution on health. Such ecological bias may occur if a simple summary measure, such as a daily mean, is not a suitable summary of a spatially variable pollution surface. We assess the performance of commonly used models when confronted with such issues using simulation studies and compare their performance with a model specifically designed to acknowledge the effects of exposure aggregation. In addition to simulation studies, we apply the models to a case study of the short-term effects of particulate matter on respiratory mortality using data from Greater London for the period 2002–2005. We found a significant increased risk of 3% (95% CI 1–5%) associated with the average of the previous three days exposure to particulate matter (per 10 $\mu g\,m^{-3}$ PM_{10}).

1. Introduction

The relationship between air pollution exposure and ill health came to public prominence in the mid 1900s, as a result of high air pollution episodes in both Europe (Firket, 1936) and America (Ciocco and Thompson, 1961). Since then a large number of epidemiological studies have consistently reported associations between a variety of pollutants at comparatively low levels and health effects, including particulate matter (Laden et al., 2000), sulphur dioxide (Schwartz, 1991), nitrogen dioxide (Zmirou et al., 1998), carbon monoxide (Conceicao et al., 2001) and ozone (Verhoeff et al., 1996). Associations have also been shown within different sub-groups of the population, such as the elderly (Dominici et al., 2000) and children (Lin et al., 2002) for a range of health outcomes, such as asthma (Yu et al., 2000) and respiratory and circulatory illnesses (Gwynn et al., 2000). More recently, large scale studies have investigated health effects in a large number of cities following to a common protocol, such as the NMMAPS studies in the U.S. (Dominici et al., 2002) and the APHEA and APHEA II studies in Europe (Katsouyanni et al., 1997, 2001). Estimates in the effects of specific pollutants on health have varied over different locations and such variation may be in part to differing temporal patterns and correlation at different locations (Ito et al., 2004) and the characteristics of the locations in which monitoring sites are located (Sarnat et al., 2009).

Whilst a number of studies have examined the longer-term effects of air pollution, the vast majority have investigated associations between short-term changes in air pollution and health. These studies relate changes in exposure with subsequent changes in a specified health outcome using daily health counts and measurements of exposure, the latter often coming from a number of monitoring sites located within an urban area. The majority of studies have estimated pollution exposure on a particular day by averaging the spatial observations, either because of lack of access to the raw data or due to the simplicity of the approach. In addition, routinely available covariate information, such as temperature and humidity, is used. Less easily obtainable information on variables that might be expected to have a relationship with pollution (and health), such as traffic density, are often represented by surrogate variables. For example, 'day of the week' effects are often used in place of traffic density based on the logical assumptions that there will be less traffic at weekends in urban areas.

These studies are ecological in nature, being based on spatially aggregated health and exposure data modelled at the same resolution. As such, there is the potential for ecological bias; assuming that associations observed at the level of the area hold for the individuals

* Corresponding author.
E-mail address: g.shaddick@bath.ac.uk (G. Shaddick).

0303-2434/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jag.2012.03.011

within the areas can lead to the so-called ecological fallacy. Ecological bias can manifest itself in a variety of ways. For a review of the problems of ecological bias and possible approaches for corrections, see Wakefield (2008).

Whilst the vast majority of studies have opted for the simple approach of using a daily average of measurements from a set of monitors, a number of studies have incorporated spatial modelling within health studies, see for example Zidek et al. (1998), Zhu et al. (2003) and Fuentes et al. (2006) for earlier examples. To a large extent, this has been because the health and exposure data were recorded at different geographical locations or scales, an issue termed the 'change of support problem' by Gelfand et al. (2001). More recently, Szpiro et al. (2011) and Chang et al. (2011) have formulated the difference between measurements from individual monitoring sites and the underlying level of pollution within a measurement error framework and provide methods for correcting for the fact that the underlying exposures are predictions from a spatial model. Other examples of acknowledging the uncertainty that will be present when using modelled predictions can be found in Lee and Shaddick (2010), who sample from the posterior predictive distribution within a Bayesian spatial–temporal model and Peng and Bell (2010), who adopt a regression calibration approach. These approaches consider the correction of estimates of risk calculated on a summary measure of exposure obtained from a spatial model rather than the specific effect of ecological bias induced by aggregation. Approaches which directly address the issues of ecological bias include hybrid case–control and ecological designs (Haneuse and Wakefield, 2007; Haneuse et al., 2008) and two-phase designs (Wakefield and Sebastien, 2008) but these require individual level outcome data which is not usually available in studies of air pollution and health.

In this paper we investigate the possibility of ecological bias being induced by aggregation within short-term epidemiological studies. Results of using the standard ecological model are compared with those from models which acknowledge such bias. The remainder of this paper is organised as follows. Section 2 describes the 'standard' modelling approach used in time-series air pollution and health studies. Section 3 describes the true underlying model at the individual level but for which data are unlikely to be available and compares its aggregated form with the 'standard' Poisson or quasi-likelihood model. This section also describes alternative modelling approaches that may alleviate such problems. Section 4 presents a simulation study that assesses the biases that may arise from using the different modelling approaches. Section 5 provides a case study comprising of an epidemiological case study investigating the association between respiratory mortality and particulate matter concentrations in Greater London for the period 2002–2005. Finally, Section 6 provides a concluding discussion.

2. Time series studies of air pollution and health

The majority of short-term air pollution and mortality studies are based on an ecological time series design, that use mortality, pollution and meteorological data that relate to a geographical region $\mathcal{R}$ (such as a city or extended urban area) for n consecutive days. Only daily counts of mortality or morbidity events from the population living within the study region are available, and are denoted here by $\mathbf{y} = (y_1, \ldots, y_n)$. These data are regressed against ambient (background) air pollution concentrations and a vector of q covariates, the latter of which are denoted by the $n \times q$ matrix $Z = (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_n^{\mathrm{T}})^{\mathrm{T}}$ where $\mathbf{z}_t^{\mathrm{T}} = (z_{t1}, \ldots, z_{tq})$ representing the realisations for day t. The covariates remove the influence of unmeasured risk factors that induce long-term trends, seasonal variation, overdispersion and temporal correlation into the daily health counts. The influence of such factors are typically modelled by smooth functions of time (i.e. day of the study) and meteorological covariates, as well as indicator variables for 'day of the week' effects and influenza epidemics.

The pollution data are obtained from k fixed site monitors located across $\mathcal{R}$ and measure ambient pollution concentrations continuously throughout the day. A daily average is typically calculated at each monitoring location, which for day t and spatial location $\mathbf{s}_l$ is denoted by $w_t(\mathbf{s}_l)$. The set of pollution locations are collectively denoted by $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_k\}$ (where $\mathbf{s}_l = (a_l, b_l) \in \mathcal{R}$), and for day t the pollution levels are summarised by the $k \times 1$ vector $\mathbf{w}_t(\mathcal{S}) = (w_t(\mathbf{s}_1), \ldots, w_t(\mathbf{s}_k))^{\mathrm{T}}$. The pollution data for all n days are collected into an $n \times k$ matrix $W(\mathcal{S}) = (\mathbf{w}_1(\mathcal{S})^{\mathrm{T}}, \ldots, \mathbf{w}_n(\mathcal{S})^{\mathrm{T}})^{\mathrm{T}}$, which is likely to contain a small proportion (typically less than 10%) of missing values. From these data the vector of daily pollution exposures are almost exclusively estimated by $\mathbf{w} = (w_1, \ldots, w_n)$, where

$$w_t = \frac{1}{k}\sum_{l=1}^{k} w_t(\mathbf{s}_l) \quad \text{for } t = 1, \ldots, n, \tag{1}$$

the average value across the k monitors on day t (missing values are typically ignored).

The relationship between ($\mathbf{y}$, $\mathbf{w}$, Z) is estimated using quasi-Poisson log-linear or additive models, in which only the mean and variance of y_t are specified using a quasi-likelihood approach. The moments resemble those from a Poisson distribution, except that the variance is allowed to be a multiple of the mean. The quasi-Poisson model has expectation $\mathbb{E}[y_t|\mathrm{w}_t, \mathbf{z}_t] = \mu_t$ and variance $\mathrm{Var}[y_t|\mathrm{w}_t, \mathbf{z}_t] = \kappa\mu_t$, where κ is the over-dispersion parameter. In addition the vector $y_1, \ldots, y_n$ are assumed to be independent, which may not be true as the number of events on successive days are likely to be correlated. Pollution concentrations at a single or multiple lags can be included into the model, with the specification above incorporating exposures $\tilde{\mathbf{w}}_t = (w_t, w_{t-1}, \ldots, w_{t-l})$ from the same day up to a maximum lag of l days, where l will typically range from between zero and five (Dominici et al., 2000). The mean log-linear function is given by

$$\mu_t = \exp\left(\alpha_0 + \sum_{j=1}^{p} z_{tj}\alpha_j^E + \sum_{j=p+1}^{q} f(z_{tj}|\alpha_j^E)\right) \exp\left(\tilde{\mathbf{w}}_t^{\mathrm{T}}\boldsymbol{\beta}^E\right) \tag{2}$$

allowing the covariates to have log-linear (e.g. $z_{tj}\alpha_j$) or log non-linear (e.g. $f(z_{tj}|\alpha_j)$) relationships with the health data.

A commonly used outcome measure in epidemiology is the relative risk (RR), which is the rate of risks of an event (or of developing a disease) with the denominator typically a baseline level of exposure. From the above model, the estimate of $\boldsymbol{\beta}^E$ gives us the relationship between pollution and health and the relative risk is RR = $\exp(\boldsymbol{\beta}^E)$ with interest lying primarily in whether this is significantly greater than one.

3. Statistical modelling

The 'standard' ecological model described by (1) and (2) may be deficient in a number of ways, and here we focus on two: (i) the form of the mean function and (ii) the exposure measure. To illustrate these deficiencies we begin by describing the desired individual level model, and then aggregate it to the ecological level.

3.1. Individual level model

The desired individual level model is based on data (y_{it}, x_{it}, $\mathbf{z}_{it}$) for the entire population of $i = 1, \ldots, N$ individuals living in the study region $\mathcal{R}$ over all $t = 1, \ldots, n$ days of the study. Here y_{it} is the Bernoulli indicator variable equalling one if individual i has a mortality or morbidity event on day t and zero otherwise, while x_{it} is the