Spatial and Spatio-temporal Epidemiology 18 (2016) 44–52

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Mitigating the effects of preferentially selected monitoring sites for environmental policy and health risk analysis

Gavin Shaddick[a,*], James V. Zidek[b], Yi Liu[a]

[a] Department of Mathematical Sciences, University of Bath, UK
[b] Department of Statistics, University of British Columbia, Canada

ARTICLE INFO

Article history:
Received 20 December 2015
Revised 14 February 2016
Accepted 22 March 2016
Available online 16 April 2016

Keywords:
Spatio–temporal models
Bayesian hierarchical modelling
Air pollution
Preferential sampling

ABSTRACT

The potential effects of air pollution are a major concern both in terms of the environment and in relation to human health. In order to support both environmental and health policy there is a need for accurate estimates of the exposures that populations might experience. The information for this typically comes from environmental monitoring networks but often the locations of monitoring sites are preferentially located in order to detect high levels of pollution. Using the information from such networks has the potential to seriously affect the estimates of pollution that are obtained and that might be used in health risk analyses. In this context, we explore the topic of preferential sampling within a long–standing network in the UK that monitored black smoke due to concerns about its effect on public health, the extent of which came to prominence during the famous London fog of 1952. Abatement measures led to a decline in the levels of black smoke and a subsequent reduction in the number of monitoring locations that were thought necessary to provide the information required for policy support. There is evidence of selection bias during this process with sites being kept in the most polluted areas. We assess the potential for this to affect the estimates of risk associated air pollution and show how using Bayesian spatio–temporal exposure models may be used to attempt to mitigate the effects of preferential sampling in this case.

1. Introduction

Air pollution has been a concern for many centuries: during the middle ages, monarchs in several countries tried to reduce air pollution by banning practices such as burning coal, and travellers in the seventeenth centuries commented on the poor air quality in many cities. Following the industrial revolution, problems associated with air pollution worsened in many areas of Europe. During the first half of the twentieth century major pollution episodes occurred in London, notably in 1952 an episode of fog, in which levels of black smoke exceeded 4500 μgm^{-3}, was associated with 4000 excess deaths (Ministry of Health, 1954). Other early episodes, which were caused by a combination of industrial pollution sources and adverse weather conditions, resulted in large numbers of deaths in the Meuse valley (Firket, 1936) and the US (Ciocco and Thompson, 1961).

Attempts to measure levels of air pollution in a regular and systematic way largely arose as a result of these episodes. Early air pollution control legislations were focussed on setting restrictions on the use of smoke–producing fuels and smoke–producing equipment (Garner and Crow, 1969; Stern, 1973) and in 1961 the world's first co–ordinated national air pollution monitoring network was established in the UK, the 'National Survey', which

* Corresponding author.
E-mail addresses: g.shaddick@bath.ac.uk (G. Shaddick), jim@stat.ubc.ca (J.V. Zidek), yl238@bath.ac.uk (Y. Liu).

http://dx.doi.org/10.1016/j.sste.2016.03.005
1877-5845/

was used to monitor black smoke and sulphur dioxide at around 1000 sites (Clifton, 1964). Since then all European countries have begun to establish monitoring networks, some of them run at the national level, others by local authorities or municipalities. Because of the different ways in which these have developed, and the different purposes for which they have been established, many of the networks vary in terms of which pollutants they measure, how they measure them, where monitoring sites are located, and how results are reported. In addition, over time many of the networks have changed; some growing, others shrinking, as attention has shifted to new pollutants and geographical areas. During much of the twentieth century, for example, the main concern was soot (or black smoke) and sulphur dioxide from industry and domestic fires. Most networks thus focussed on measuring these pollutants, especially in industrial areas where concentrations were likely to be high.

Following legislation at both the national and international level and the WHO air quality guidelines (AQGs) the monitoring of air pollution has significantly increased. The AQGs were designed to offer global guidance on reducing the health impacts of air pollution. Yet despite this, the information that is available to support air pollution policy and management is far from sufficient and three specific problems conspire to limit its utility; (i) monitoring is expensive and so monitoring networks are typically sparse, (ii) concentrations may vary greatly over small distances, epically in urban areas and (iii) networks are often designed to monitor compliance with standards and may not give a true representation of levels over an area.

It is vital that the information obtained from these networks is accurate and reflects not only levels recorded at the locations of monitoring sites but can be used to accurately reflect the levels of exposures that may be experienced by populations. This is unlikely to be the case if monitors are intentionally placed in locations where pollution might be expected to be high; a practice known as *preferential sampling*. In the context of air pollution and health in epidemiological analyses, Guttorp and Sampson (2010) state that air pollution monitoring sites may be preferentially located for a number of reasons, including to measure: (i) background levels outside of urban areas; (ii) levels in residential areas; and (iii) levels near pollutant sources. Standard geostatistical methods which assume sampling is non–preferential are often employed despite the presence of a preferential sampling scheme. Ignoring preferential sampling may lead to incorrect inferences and biased estimates of pollution concentrations and thus any subsequent estimation of health risks.

The implications of preferential sampling on the estimation of health risk have not received much attention, although Szpiro and Sheppard (2010) demonstrated by simulation that preferential sampling does induce bias and uncertainty in estimates of the health effects of air pollution. In this paper we assess the potential effects of preferential sampling on the estimation of health risks associated with air pollution. The remainder of the paper is organised as follows; in Section 2 we describe preferential sampling and present the results of simulation studies that show the effects it may have on the estimation of health risks. In Section 3, we describe the use of exposure models for predicting exposures at preferentially sampled locations based on data that may not be subject to the same biases and in Section 4 we apply this to a case study of the health effects of black smoke in the UK. Section 5 contains a concluding discussion and describes possible avenues for future research in this area.

Throughout the paper, models are presented for both health counts and exposures. To avoid ambiguity between the two, we use $Y^{(1)}$, $X^{(1)}$, $Z^{(1)}$, $\theta^{(1)}$ for the health models and $Y^{(2)}$, $X^{(2)}$, $Z^{(2)}$, $\theta^{(2)}$ for the exposure models. In both cases, Y denotes the response, X covariates and Z an underlying latent process, or the underlying true level. It is noted that although the health counts, $Y^{(1)}$, can be considered to be measurements from an underlying true level with differences occurring, for example due to misclassification or data anomalies, here we consider them to be an accurate reflection of the truth, i.e. $Y^{(1)} = Z^{(1)}$ and so in the health models this distinction is dropped.

2. Preferential sampling

Preferential sampling is a common phenomenon in environmental studies, as the monitoring locations in a spatial network are often subjectively chosen for objectives such as accommodating a change in government policies or monitoring high levels of pollution. For example, if monitors are positioned close to known pollution sources, such as on a roadside, near an industrial polluter, or within a city centre, then the estimated pollution surface is likely to be overestimated. Both the number and locations of the pollution monitors will affect the accuracy of estimates of the true exposure surface. However, it is often implicitly assumed that the true exposure surface is based on the random sampling of the complete temporal–spatial pollution field. When, as is likely this is not the case, the exposure measurements obtained from preferentially sampled networks may lead to inaccurate estimation of exposure to air pollution and consequently to the estimation of relative risks in epidemiological studies.

Recently there have been a small number of papers published on the subject of preferential sampling in an environmental setting, which occurs when the process that determines the locations of the monitoring sites and the process being modelled (air pollution concentrations) are in some ways dependent. Diggle et al. (2010) extend the classical geostatistical model in two ways; (i), the monitoring locations are treated as random quantities of a log–Gaussian Cox process rather than being fixed; (ii) the exposures are modelled conditionally on the locations assuming a Gaussian spatial process. Through simulation examples they show that ignoring preferential sampling can lead to misleading inferences, especially with spatial predictions. Pati et al. (2011) adapt this approach within a Bayesian framework and demonstrate it's use in a case study of ozone data over eastern U.S.A which shows significant evidence of preferential sampling. Other examples of the application of this approach include (Lee et al., 2011) who implement it when constructing air quality indicators for a case study set in Greater London.