


# Sensitivity analysis on the ecological bias for Seoul tuberculosis data

Eunjung Song<sup>1</sup> · Soeun Kim<sup>2</sup> ·  
Seungsik Hwang<sup>3</sup> · Woojoo Lee<sup>1</sup> 

Received: 18 December 2017 / Revised: 2 May 2018 / Published online: 29 May 2018  
© Springer Science+Business Media, LLC, part of Springer Nature 2018

**Abstract** In ecological studies, researchers often try to convey the analysis results to individual level based on aggregate data. In order to do this correctly, the possibility of ecological bias should be studied and addressed. One of the key ideas used to address the ecological bias issue is to derive the ecological model from the individual model and to check whether the parameter of interest in the individual model is identifiable in the ecological model. However, the procedure depends on unverifiable assumptions, and we recommend checking how sensitive the results are to these unverifiable assumptions. We analyzed the tuberculosis data that was collected in Seoul in 2005 using a spatial ecological regression model for the aggregate count data with spatial correlation, and found that the deprivation index is likely to have a small positive effect on the occurrence risk of tuberculosis in individual level in Seoul. We considered this

---

Handling Editor: Pierre Dutilleul.

---

✉ Woojoo Lee  
lwj221@gmail.com

Eunjung Song  
yue0327@inha.edu

Soeun Kim  
Soeun.Kim@uth.tmc.edu

Seungsik Hwang  
cyberdoc@snu.ac.kr

- <sup>1</sup> Department of Statistics, Inha University, 235 Yonghyun-Dong, Nam-Gu, Incheon 402-751, Korea
- <sup>2</sup> Department of Biostatistics and Data Science, University of Texas Health Science Center, Houston, TX, USA
- <sup>3</sup> Department of Public Health Sciences, Seoul National University Graduate School of Public Health, Seoul, Korea

finding in various aspects by performing in depth sensitivity analyses. In particular, our findings are shown to be robust to the distribution assumptions for the individual exposure and missing binary covariate across various scenarios.

**Keywords** Ecological bias · Robustness · Sensitivity analysis · Spatial model

## 1 Introduction

In ecological studies researchers are often faced with aggregate count data with spatial correlation. Spatial ecological regression can be used to obtain results based on aggregate data, which can then be used to try to convey the results to an individual level. Ecological bias becomes an issue in these situations as the aggregate level associations can fail to properly reflect the results from the individual level (Greenland and Morgenstern 1987). The issues related to ecological bias should be studied and addressed in order to correctly convey the results. One of the key ideas that can be used to address these issues is to derive the ecological model from the individual level model and check whether the parameter of interest in the individual model is identifiable in the ecological model. However, the individual model often depends on unverifiable assumptions and requires sensitivity analyses to test the assumptions (Wakefield 2003).

Tuberculosis control is still a major challenge in South Korea. Tuberculosis incidence in South Korea was reported to be seven times higher than the average incidence of countries belonging to the Organization for Economic Co-operation and Development in 2013 (Kim and Yim 2015). Among various risk factors, the relationship of socio-economic deprivation and tuberculosis risk has been an interesting topic for epidemiologic research, because of the possibility of missed opportunities for prevention depending on socio-economic status (Lopez De Fede et al. 2008). There is also a possibility of treatment delay for the deprived population (French et al. 2009), and several studies were conducted to find any association between deprivation and tuberculosis occurrence. In this paper we analyze the tuberculosis data that was collected in Seoul in 2005, using spatial ecological regression to analyze the aggregate count data with spatial correlation. In this application we focus on how to address the ecological bias issue in the Seoul tuberculosis data, and explain in what settings the ecological analysis result is robust against the unverifiable assumptions. In particular, we show how to conduct appropriate sensitivity analyses on the distribution of hypothetical individual level exposures when the contextual effect is considered and when an unmeasured important binary covariate exists. The analysis of Seoul tuberculosis data is presented as a case study that deals with issues related to ecological bias.

In Seoul tuberculosis data, the unit of analysis is “dong” which corresponds to a geographic area comparable to a district. It is desirable to include in the analyses spatial correlation associated with the geographical distances between dongs. This was considered to be a difficult problem 20 years ago as it requires addressing spatial correlation in regression analysis, however, it can now be easily implemented in many software packages such as Winbugs and R. In particular, the Besag–York–Mollié (BYM) model by Besag et al. (1991) is commonly used in the areas of spatial epidemiology and medical sciences (Besag and Kooperberg 1995; Deguen 2010). In

many situations, we are interested in individual level associations, and therefore spatial ecological modeling is not sufficient for the purpose of the analyses. With the possibility of ecological bias, spatial modeling is often a secondary issue in analyzing aggregate spatial data (Wakefield 2003).

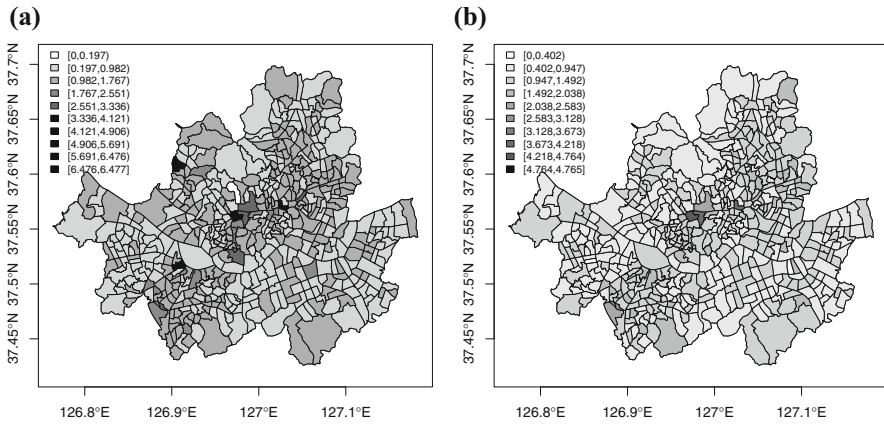
Ecological bias has long been studied in the literature as it involves important issues and is relevant in various applications. Greenland and Morgenstern (1987) discussed the main attributes such as confounding and effect modification for ecological bias, and Richardson et al. (1987) described sources of ecological bias. Wakefield (2007) pointed out that ecological bias occurs due to within-area variability in exposures and confounders, and as consequences of the within-area variability, ecological bias can have different aspects such as pure specification bias and confounding (Wakefield 2007). For dealing with ecological bias, the existing literature emphasizes that the individual-level data should be used together with the ecological data (Jackson et al. 2005). However, this is possible only when the individual level data is available, and the researchers are often left with aggregate level data without access to individual level data. Without the availability of individual level data, sensitivity analysis can be considered. The effects of unmeasured confounders and the problem of pure specification bias can be addressed by sensitivity analysis (Wakefield 2003). As usual in ecological data, individual level data is not available for Seoul tuberculosis data. Therefore, the unverifiable assumptions used in our analysis require us to perform sensitivity analysis. By investigating the results of sensitivity analyses from several aspects, if the results are not very sensitive and do not change its qualitative meaning, it is possible to give a more confident statement of the results that can be conveyed to individual level.

We begin by investigating the relationship between ecological quantities by applying BYM model to Seoul tuberculosis data, and we also address the possibility of ecological bias issues. We tackle the issue by considering a reasonable individual level model and deriving the ecological model from this model as suggested in Wakefield (2007) and Wakefield (2003). The correspondence between the ecological model and derived ecological model from the individual model is examined in depth. We also study whether the ecological analysis results are robust to misspecification of the distributional assumption for within-area exposures and to a missing covariate. This will be exemplified in our analysis in Sects. 6 and 7, followed by concluding remarks.

## 2 Seoul tuberculosis data

Let  $Y_i$  denote the number of tuberculosis patients in  $i$ th dong and  $e_i$  be the expected number of tuberculosis patients in the general population to correct for age structure. In Seoul in 2005, the average dong population for male is 9677.26 and its standard deviation is 3984.82. For female case, the average dong population is 9718.59 and its standard deviation is 4061.62.

Standardized mortality ratio (SMR) is often a quantity of interest for spatial epidemiologists, and is defined to be  $Y_i/e_i$ . We look into SMRs by gender on the map of Seoul in Fig. 1a, b, to take into account that tuberculosis occurrence pattern is different between male and female. The  $x$  and  $y$  axes in the figures denote latitude and



**Fig. 1** Standardized mortality ratio. **a** SMR of Male. **b** SMR of Female

longitude values, respectively. Higher SMR is shaded with darker color. Since gender differences in SMRs are observed, ecological analyses will be performed separately by gender.

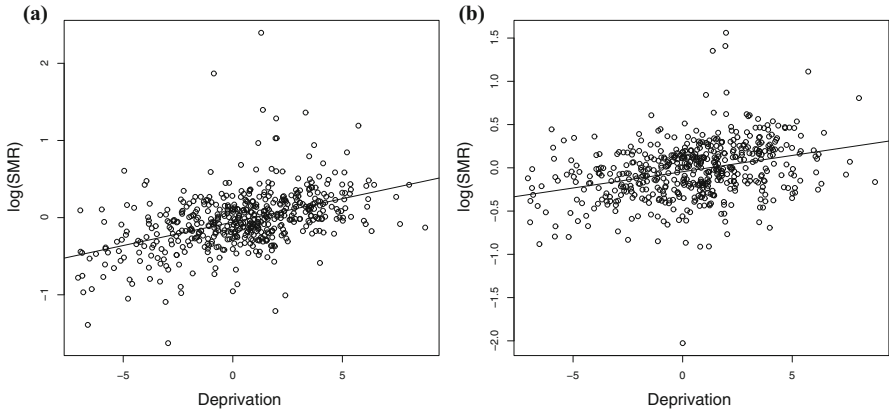
In this study, the covariate of interest is the deprivation index of each dong (Townsend 1987). The deprivation index was developed by the British Department for Communities and Local Government as a measure to identify how deprived different parts of England are. The index consists of seven different domains (McLennan et al. 2011): (1) Income, (2) Employment, (3) Health and Disability, (4) Education, Skills and Training, (5) Barriers to Housing and Services, (6) Living environment, and (7) Crime. These domains are considered and combined to represent an overall measure that shows the level of deprivation at a small area level. Figure 2 are scatter plots of  $\log(Y_i/e_i)$  versus the deprivation index for male and female, respectively. Both plots show linear increasing trends, particularly for males. The relationship of socio-economic deprivation and tuberculosis risk in Seoul will be examined in depth as in the following sections.

### 3 Spatial ecological model

We use the BYM model for our tuberculosis data to account for spatial correlation in aggregate level data (Besag et al. 1991). For the aggregate count data  $Y_i$  of the  $i$ th dong ( $i = 1, \dots, n$ ), the BYM model can be written by

$$\log E(Y_i|x_i, u_i, v_i) = \log e_i + \beta_0 + \beta_1 x_i + u_i + v_i \tag{1}$$

where  $u_i|u_{-i} \sim N(\bar{u}_i, \sigma_u^2/q_i)$  and  $v_i \sim N(0, \sigma_v^2)$ .  $u_i$  is introduced to explain the spatial correlation and intrinsic conditional autoregressive model (ICAR) (Besag 1974) is employed. ICAR has gained its popularity for analysis of aggregated spatial data in spatial epidemiology, disease mapping, agricultural experiments and image analysis (Besag 1974; Besag and Kooperberg 1995; Besag and Higdon 1999).  $\bar{u}_i$  is the mean of



**Fig. 2** The relationship between log SMR and deprivation index. **a** Male. **b** Female

the  $u_i$  for the neighborhoods of  $i$ th dong that share administrative borderline.  $q_i$  is the number of neighborhoods of  $i$ th dong. Since  $u_i$  is conditionally specified,  $\sigma_u^2$  should be interpreted not as the marginal variance but as the conditional variance (Wakefield 2007). Therefore it is not reasonable to compare  $\sigma_u^2$  and  $\sigma_v^2$  directly, because  $\sigma_v^2$  is specified as the marginal variance.  $v_i$  is introduced to explain the area-specific heterogeneity, and can capture additional variability beyond what is captured by Poisson distribution.  $x_i$  denotes the deprivation index of  $i$ th dong.

We consider two different priors in order to check whether the results from our ecological model are robust against the prior specification. Prior 1 is specified following the recommendation in Wakefield (2007), so that the prior is specified for total variability instead of specifying priors for each of the variance components, because the total variability is a quantity with available prior knowledge. Let  $\sigma_v^2 = (1 - p)\tau^{-1}$  and  $\sigma_u^2 = p\tau^{-1}$  where  $\tau = 1/(\sigma_v^2 + \sigma_u^2)$  and  $p = \sigma_u^2/(\sigma_u^2 + \sigma_v^2)$ . Then  $p \sim \text{Beta}(1, 1)$ ,  $\tau \sim \text{Gamma}(1, 0.0260)$ , and the improper uniform prior,  $dflat()$ , was used for the regression coefficients  $\beta_0$  and  $\beta_1$ . Prior 2 is very commonly used and obtained from GeoBUGS user manual ver 1.2 (Thomas and Best 2004), with the assumption that  $\beta_0 \sim dflat()$ ,  $\beta_1 \sim N(0, 10^5)$ . For the inverse of variance components  $1/\sigma_u^2$  and  $1/\sigma_v^2$ ,  $\text{Gamma}(0.5, 0.0005)$  and  $\text{Gamma}(0.5, 0.0005)$  were used, respectively.

To begin with, we perform ecological analysis using Winbugs ver 1.4. In order to check whether both random effects are necessary, we compute the deviance information criterion (DIC) (Spiegelhalter et al. 2002) for four models: (1) the model with only  $u_i$ , (2) only  $v_i$ , (3) both  $u_i$  and  $v_i$  and (4) in addition to  $u_i$  and  $v_i$ , including longitude and latitude as linear covariates.

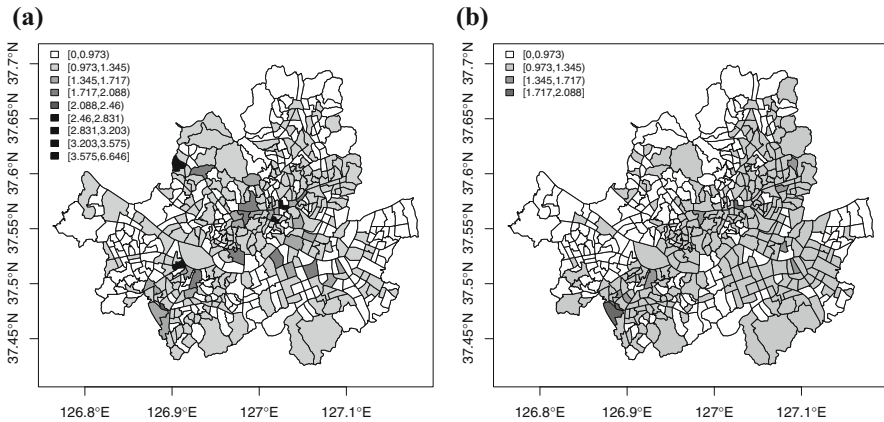
As shown in Tables 1 and 2, the models with both  $u_i$  and  $v_i$  have the smallest DIC, and are selected to be the best models. To see which random effect has a dominant effect, we need to compare  $\sigma_v^2$  and the marginal variance of  $u_i$ , but the latter quantity is difficult to compute. Thus, using similar method to what was done in Haneuse and Wakefield (2004), we use  $E(s_u^2|y_1, \dots, y_n)$  where  $s_u^2 = (n - 1)^{-1} \sum_{i=1}^n (u_i - \bar{u})^2$  and  $\bar{u} = n^{-1} \sum_{i=1}^n u_i$  as an approximate estimate of marginal variance for  $u_i$ . From Table 1, we observe that the area-specific heterogeneity seems more dominant than the spatial

**Table 1** Summary of ecological analysis result: Male

Prior 1	Model	$\beta_1$	SE	DIC	$E(s_u^2 y_1, \dots, y_n)$	$\text{Var}(v_i)$
	$u_i$	0.047	0.005	3722.030	0.075	
	$v_i$	0.058	0.005	3698.210		0.075
	$v_i, u_i$	0.051	0.006	3691.180	0.020	0.054
	Longitude, latitude, $v_i, u_i$	0.052	0.006	3692.020	0.023	0.053
Prior 2	Model	$\beta_1$	SE	DIC	$E(s_u^2 y_1, \dots, y_n)$	$\text{Var}(v_i)$
	$u_i$	0.047	0.005	3722.770	0.075	
	$v_i$	0.059	0.005	3698.560		0.076
	$v_i, u_i$	0.052	0.005	3691.870	0.021	0.053
	Longitude, latitude, $v_i, u_i$	0.052	0.006	3695.190	0.021	0.054

**Table 2** Summary of ecological analysis result: Female

Prior 1	Model	$\beta_1$	SE	DIC	$E(s_u^2 y_1, \dots, y_n)$	$\text{Var}(v_i)$
	$u_i$	0.029	0.005	3330.540	0.034	
	$v_i$	0.036	0.004	3345.730		0.037
	$v_i, u_i$	0.031	0.005	3319.910	0.019	0.017
	Longitude, latitude, $v_i, u_i$	0.031	0.005	3321.070	0.020	0.015
Prior 2	Model	$\beta_1$	SE	DIC	$E(s_u^2 y_1, \dots, y_n)$	$\text{Var}(v_i)$
	$u_i$	0.029	0.005	3330.440	0.034	
	$v_i$	0.035	0.004	3345.980		0.036
	$v_i, u_i$	0.030	0.005	3321.300	0.022	0.010
	Longitude, latitude, $v_i, u_i$	0.030	0.005	3322.620	0.022	0.013



**Fig. 3** Residual relative risk:  $\exp(u_i + v_i)$ . **a**  $\exp(u_i + v_i)$  of Male. **b**  $\exp(u_i + v_i)$  of Female

component for male. The situation is reversed for females, as shown in Table 2. The estimate of  $\beta_1$  is small but significant in both genders.  $\beta_1$  from the smallest DIC model for male is 0.051 (Prior 1). This means that given  $u_i$  and  $v_i$ , the expected number of male tuberculosis patients increases by 5% when one unit of deprivation index increases, which is a significant number for the public health agency. The increase in the expected number of female tuberculosis patients is 3%, which is a little smaller than that of the male case. These interpretations do not depend on the choice of priors. The spatial component explains 27.0 and 52.8% of the residual variability for male and female, respectively. In order to see the relative contribution of the deprivation index ( $x_i$ ),  $u_i$  and  $v_i$ , we compute  $\text{Var}(x_i \beta_1)$  and compare it with the estimates for  $E(s_u^2 | y_1, \dots, y_n)$  and  $\text{Var}(v_i)$ .  $\text{Var}(x_i \beta_1)$  is 0.022 and 0.008 for male and female, respectively. The contribution by the deprivation index is similar to that of  $u_i$  for males, but it is less than the half of  $u_i$  for females. From the spatial ecological model using Prior 1, the residual relative risk  $\exp(u_i + v_i)$  is illustrated in Fig. 3a, b, respectively. The two figures show the tuberculosis risk after adjusting for the deprivation index.

From the ecological analysis results above, can we say that as the individual social economic level deteriorates (i.e. the individual deprivation index is higher), the individual risk for tuberculosis increases? In order to be able to fill the gap from ecological analysis to individual interpretation, we need to deal with the ecological bias issue.

An operational procedure to deal with them was given by Diggle and Elliott (1995) and Wakefield (2007). To be complete, we summarize it here:

1. Specify an individual level model.
2. Derive the ecological level model from the individual-level model.
3. Check the sources of ecological bias by comparing the derived ecological model and ordinary ecological model.

In the following section, we take the above steps for Seoul tuberculosis data. While the aggregate model is derived from the individual model, we identify explicitly which assumptions are used. Some assumptions can be justified from the data, but others



cannot be justified because of the lack of individual-level data. In the latter case, sensitivity analysis is desirable.

#### 4 An individual level analysis for Seoul tuberculosis data

Let  $Y_{ij}$  be a Bernoulli random variable that denotes whether  $j$ th person in  $i$ th dong is a tuberculosis patient.  $x_{ij}$  is the deprivation index of *ban* (submunicipal level division in Korea) where  $j$ th person in  $i$ th dong lives, and  $x_i$  is the average deprivation index of  $i$ th dong. Let

$$Y_{ij}|x_{ij}, x_i, u_i, v_i \sim Ber(p_{ij}) \tag{2}$$

where

$$p_{ij} = E(Y_{ij}|x_{ij}, x_i, u_i, v_i) = \exp(\alpha_0 + \alpha_1^{ind} x_{ij} + \alpha_1^{con} x_i + u_i + v_i + \gamma_{kij}). \tag{3}$$

Model (3) is appropriate for rare disease such as tuberculosis. The prevalence of tuberculosis is known to be approximately 159/100,000 according to WHO Global tuberculosis report in 2014 (World Health Organization 2014).  $\exp(\gamma_{kij})$  is used to denote the risk associated with the  $j$ th person who live in  $i$ th dong and belong to  $k$ th age-level.  $\alpha_1^{con}$  denotes the contextual effect, which is often considered in social epidemiology or infectious disease epidemiology (Salway and Wakefield 2005). Greenland (2001) emphasized that even though the primary objective of research is to estimate the contextual effect, the ecological level model should be derived from an individual-level model including  $x_i$  as well as  $x_{ij}$ . Some statistical arguments for the non-separability of  $\alpha_1^{ind}$  and  $\alpha_1^{con}$  in ecological analysis are given in Greenland (2002).

Consider the individuals in  $k$ th age group only:

$$p_{ij}^k = \exp(\alpha_0 + \alpha_1^{ind} x_{ij} + \alpha_1^{con} x_i + u_i + v_i + \gamma_k)$$

Then,

$$\begin{aligned} p_i^k &= \sum_j p_{ij}^k / N_{ik} \approx E(\exp(\alpha_0 + \alpha_1^{ind} x_{ij} + \alpha_1^{con} x_i + u_i + v_i + \gamma_k) | x_i, u_i, v_i) \\ &= \exp\left(\alpha_0 + \alpha_1 x_i + \frac{1}{2} s_i^2 (\alpha_1^{ind})^2 + u_i + v_i + \gamma_k\right) \end{aligned} \tag{4}$$

where  $N_{ik}$  is the population of  $k$ th age category in  $i$ th dong and  $\alpha_1 = \alpha_1^{ind} + \alpha_1^{con}$ . The expectation is taken with respect to  $x_{ij}$ , and we assume that the exposure follows a normal distribution:

$$x_{ij} \sim N(x_i, s_i^2). \tag{5}$$

The term  $\frac{1}{2}s_i^2(\alpha_1^{ind})^2$  in the Eq. (4) is obtained from the moment generating function of this normal distribution. The normal assumption will be examined in the sensitivity analysis later. Then, the mean number of the tuberculosis patients in the  $i$ th dong is given by

$$\mu_i = \sum_k N_{ik} p_i^k \quad (6)$$

$$\begin{aligned} &= \sum_k N_{ik} \exp\left(\alpha_0 + \alpha_1 x_i + \frac{1}{2}s_i^2(\alpha_1^{ind})^2 + u_i + v_i + \gamma_k\right) \\ &= \left(\sum_k N_{ik} \exp(\gamma_k)\right) \exp\left(\alpha_0 + \alpha_1 x_i + \frac{1}{2}s_i^2(\alpha_1^{ind})^2 + u_i + v_i\right) \\ &= e_i \exp\left(\alpha_0 + \alpha_1 x_i + \frac{1}{2}s_i^2(\alpha_1^{ind})^2 + u_i + v_i\right). \end{aligned} \quad (7)$$

Equations (6) and (7) use the large sample approximation (4) in each dong. Note that  $e_i = \sum_k N_{ik} \exp(\gamma_k)$  is the same quantity that appeared in (1) and is treated as the offset variable.

The derived ecological model is given by

$$Y_i | x_i, u_i, v_i \sim \text{Poisson}\left(e_i \exp\left(\alpha_0 + \alpha_1 x_i + \frac{1}{2}s_i^2(\alpha_1^{ind})^2 + u_i + v_i\right)\right). \quad (8)$$

This Poisson approximation is valid for the sum of independent, non-identical Bernoulli variables under the rare disease assumption. The rigorous technical condition was first given in Le Cam (1960), and a simple explanation can be found in Steele (1994). This approximation has been used in ecological studies, for example, in Wakefield (2007). If we compare the model (8) with the ecological model (1), the potentially problematic term is  $\frac{1}{2}s_i^2(\alpha_1^{ind})^2$ . The relationship between  $\alpha_1$  and  $\beta_1$  depends on the relationship between  $s_i^2$  and  $x_i$ , but  $s_i^2$  is unknown to us, so we cannot check the relationship between mean and variance based on the data. Sensitivity analysis with respect to the change in  $s_i^2$  is desirable, which will be considered in the following section.

In the derivation of the ecological model from the individual model, no missing covariate is assumed in the individual level model. However, some important variables that relate to lifestyles, such as smoking, are not available in our dataset. Since smoking status is deemed to be important in analyzing lung-related disease, we need to consider this variable as a missing binary covariate in the individual-level model. This will be discussed in Sect. 7.

### 5 Sensitivity to the functional relationship between $s_i^2$ and $x_i$

In this section, we consider pure specification bias issue. Various scenarios can be considered for the relationship between  $s_i^2$  and  $x_i$ .

- Scenario (1) If the mean level of the deprivation index ( $x_{ij}$ ) is high, its variability can be high, and if the mean level of the deprivation index is low, its variability can be low. This implies that there exist wealthy towns in which only the rich live (low variability), but that the rich and the poor can coexist in areas with high poverty rates.
- Scenario (2) If the mean level of the deprivation index ( $x_{ij}$ ) is high, its variability can be low, however if the mean level of the deprivation index is low, its variability can be high. This implies that there are poor towns that consist only of poor population, but the rich and the poor can coexist in some wealthy towns.
- Scenario (3) The rich and the poor coexist regardless of the deprivation index.

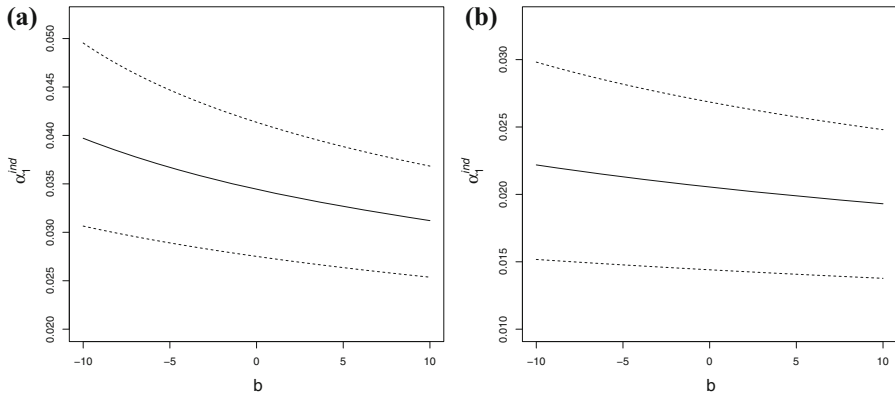
These scenarios can be captured approximately by the following linear model:

$$s_i^2 = a + bx_i$$

The same linear model was used for the sensitivity analysis in Wakefield (2003). The first, second and third scenarios correspond to the case with  $b > 0$ ,  $b < 0$  and  $b = 0$ , respectively. Then the derived ecological model becomes

$$\begin{aligned} Y_i|u_i, v_i &\sim \text{Poisson} \left( e_i \exp \left( \alpha_0 + \alpha_1 x_i + \frac{1}{2} (\alpha_1^{ind})^2 (a + bx_i) + u_i + v_i \right) \right) \\ &= \text{Poisson} \left( e_i \exp \left( \alpha_0 + \frac{1}{2} (\alpha_1^{ind})^2 a + \left( \alpha_1 + \frac{1}{2} (\alpha_1^{ind})^2 b \right) x_i \right. \right. \\ &\quad \left. \left. + u_i + v_i \right) \right) \end{aligned}$$

The correspondence between the ecological model and the derived model from the individual model is given by  $\beta_0 = \alpha_0 + \frac{1}{2} a (\alpha_1^{ind})^2$  and  $\beta_1 = \alpha_1 + \frac{1}{2} b (\alpha_1^{ind})^2$ . The role of  $\alpha_1^{ind}$  is important because it determines the difference between  $\beta_0$  and  $\alpha_0$ , and  $\beta_1$  and  $\alpha_1$ . Note that under Scenario 3,  $\beta_1$  is equal to  $\alpha_1 (= \alpha_1^{ind} + \alpha_1^{con})$ . In practice, using ecological data, we can obtain an estimate for  $\beta_1$ , which can be used in the equation to solve for  $\alpha_1^{ind}$ . Let  $\alpha_1^{con} = \kappa \alpha_1^{ind}$  for some known value  $\kappa > 0$ . Since  $\beta_1 = \alpha_1 + \frac{1}{2} b (\alpha_1^{ind})^2 = (1 + \kappa) \alpha_1^{ind} + \frac{1}{2} b (\alpha_1^{ind})^2$  is a quadratic equation with respect to  $\alpha_1^{ind}$ , it can give two solutions for  $\alpha_1^{ind}$ . Since  $\alpha_1^{ind}$  is believed to be a small nonzero value in our application (Jee et al. 2009), we take  $\alpha_1^{ind}$  with the same sign as that of  $\beta_1$ . This is reasonable because  $\beta_1 = (1 + \kappa) \alpha_1^{ind}$  when  $b = 0$  and  $\alpha_1^{ind}$  varies continuously as  $b$  changes near the origin. In making a choice of  $\alpha_1^{ind}$  we need to be cautious and take into account subject knowledge about the size of  $\alpha_1^{ind}$  if available. Note that small  $\beta_1$  does not necessarily imply that  $\alpha_1^{ind}$  is small. For example,  $\alpha_1^{ind} = 1$ ,  $\kappa = 0.1$  and  $b = -2$ ,  $\beta_1$  becomes 0.1. In other words,  $\beta_1$  can be small even when  $\alpha_1^{ind}$  is not so small according to the values of  $b$ . Therefore it is important to note that small  $\beta_1$  does not directly imply that  $\alpha_1^{ind}$  is small.



**Fig. 4** Sensitivity analysis. **a** Sensitivity analysis (male). **b** Sensitivity analysis (female)

Note that  $a$ ,  $b$  and  $\kappa$  are not identifiable from the aggregate data. Their values should be based purely on subject matter knowledge. We perform sensitivity analysis of the effect on  $\alpha_1^{ind}$  of changing  $b$  and  $\kappa$  over some reasonable range. We plot the estimate of  $\alpha_1^{ind}$  with its pointwise 95% credible interval in Fig. 4 by moving  $b$  over  $-10$  to  $10$  for different  $\kappa = 0, 0.1, 0.5$ . To select a reasonable range of  $b$ , we need the within-area variances  $s_i^2$ . For example, Wakefield (2003) derived plausible values of  $b$  by using the interquartile range of  $s_i^2$ . However, in our situation,  $s_i^2$  are not available. Therefore, we decide to make a conservative choice for the range of  $b$ . Considering the total range of  $x_i (-7.061, 8.776)$ ,  $(-10, 10)$  seems sufficiently wide to include plausible values of  $b$ . The three values for  $\kappa$  reflect that  $\alpha_1^{ind}$  has a dominant effect compared to  $\alpha_1^{con}$ . As a conservative choice for  $\kappa$ , we consider  $\kappa$  up to  $0.5$  which implies that the contextual effect by  $x_i$  on the tuberculosis risk corresponds to  $50\%$  of the effect by  $x_{ij}$  when  $x_{ij}$  and  $x_i$  are defined on the same scale. For each fixed  $b$  and  $\kappa$ , the pointwise credible interval is obtained by solving  $\beta_1^* = (1 + \kappa)\alpha_1^{ind} + \frac{1}{2}b(\alpha_1^{ind})^2$  with respect to  $\alpha_1^{ind}$  where  $\beta_1^*$  correspond to the leftmost and rightmost points of 95% credible set for  $\beta_1$ . Since  $\beta_1$  itself is a small value, the change of  $\alpha_1^{ind}$  is also small and the credible set does not touch  $0$  over  $b \in [-10, 10]$  for different  $\kappa$ . For example, consider  $\kappa = 0.5$ . For male,  $\alpha_1^{ind}$  lies within  $(0.031, 0.040)$  (Fig. 4a) and for female,  $\alpha_1^{ind}$  lies within  $(0.019, 0.022)$  (Fig. 4b). Thus, in Seoul tuberculosis data, when the ecological result is conveyed to the individual level, it can be argued that the functional relationship between  $x_i$  and  $s_i^2$  is not likely to have a big influence on the effect of deprivation index.

## 6 Sensitivity to the distribution assumption on $x_{ij}$

Since the normal distribution for  $x_{ij}$  is a convenient choice rather than a theoretically supported choice, it is meaningful to check whether our analyses results have robustness property against the distribution misspecification. In line with this, we first

consider some flat shape or leptokurtic distributions for  $x_{ij}$ . These distributions are possible forms that can occur in practice for various exposure variables.

If  $\alpha_1^{ind}$  is small and the moment generating function of  $x_{ij}$  is continuously differentiable, we can use the following approximation:

$$\begin{aligned} M(\alpha_1^{ind}) &= E[\exp(\alpha_1^{ind} x_{ij})] \\ &= M(0) + M'(0)\alpha_1^{ind} + O((\alpha_1^{ind})^2) \\ &\approx \exp(\alpha_1^{ind} E(x_{ij})) \\ &= \exp(\alpha_1^{ind} x_i). \end{aligned} \tag{9}$$

We will see that this small  $\alpha_1^{ind}$  approximation leads to small pure specification bias through some examples. Wakefield and Salway (2001) also noted that for small  $\alpha_1$  the pure specification bias is expected to be small when  $x_{ij}$  follows normal and gamma distributions. Some specific examples are considered below.

### 6.1 Uniform distribution

Suppose that

$$x_{ij} \sim \text{Uniform}(x_i - \delta, x_i + \delta). \tag{10}$$

where  $E(x_{ij}) = x_i$ , and  $2\delta$  denotes the range of the uniform distribution. This uniform within-area exposure distribution was discussed in Greenland (1992). The derived ecological model becomes

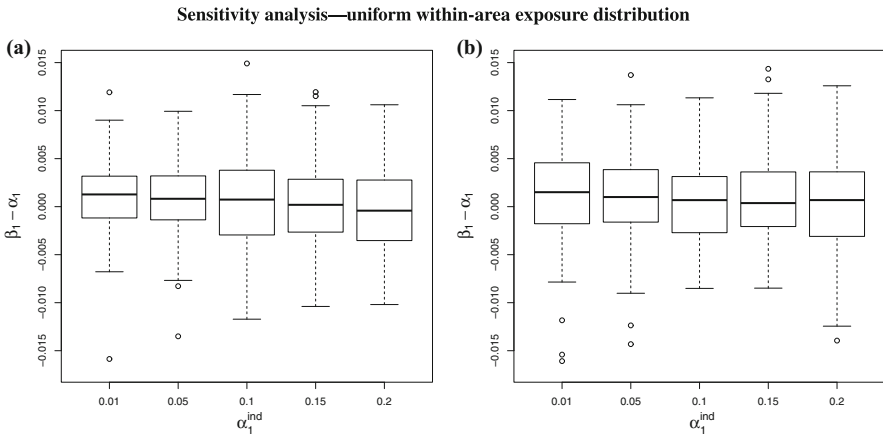
$$\begin{aligned} p_i^k &= \sum_j p_{ij}^k / N_{ik} \\ &\approx E \left( \exp(\alpha_0 + \alpha_1^{ind} x_{ij} + \alpha_1^{con} x_i + u_i + v_i + \gamma_k) | u_i, v_i \right) \\ &= \exp \left( \alpha_0 + \alpha_1^{con} x_i + u_i + v_i + \gamma_k \right) \left( \exp \left( \alpha_1^{ind} (x_i + \delta) \right) \right. \\ &\quad \left. - \exp \left( \alpha_1^{ind} (x_i - \delta) \right) \right) / \left( 2\alpha_1^{ind} \delta \right). \end{aligned}$$

At first glance, this functional form involving  $x_i$  is very different from that from the normal distribution. However, consider when  $\alpha_1^{ind}$  is small and  $x_i$  and  $\delta$  are bounded. Then,

$$\exp \left( \alpha_1^{ind} (x_i + \delta) \right) - \exp \left( \alpha_1^{ind} (x_i - \delta) \right) \approx \exp \left( \alpha_1^{ind} x_i \right) \left( 2\alpha_1^{ind} \delta \right)$$

by Taylor expansion. Thus,

$$p_i^k \approx \exp \left( \alpha_0 + \left( \alpha_1^{ind} + \alpha_1^{con} \right) x_i + u_i + v_i + \gamma_k \right)$$



**Fig. 5** The true value for  $\alpha_1^{ind}$  is on x-axis, and the difference  $\beta_1 - \alpha_1$  is on y-axis. In both cases,  $\delta$  is fixed at 4. **a**  $1/\sigma_u^2 = 111.467$ . **b**  $1/\sigma_u^2 = 25.361$

in this case,  $\beta_1$  corresponds to  $\alpha_1 (= \alpha_1^{ind} + \alpha_1^{con})$ .

To check the robustness of this approximation, a numerical study is performed. We first generate  $y_i$  from  $Poisson(e_i \exp(\alpha_0 + \alpha_1^{con} x_i + u_i + v_i) (\exp(\alpha_1^{ind} (x_i + \delta)) - \exp(\alpha_1^{ind} (x_i - \delta))) / (2\alpha_1^{ind} \delta))$  which is the aggregated model from the individual level model with (10).  $x_i$  and  $e_i$  are taken from Seoul male tuberculosis data. We use  $\delta = 1, 2$  and 4. To select a reasonable value of  $\delta$ , we need the within-area information on  $x_{ij}$ . However, this information is not available, so we consider  $\delta$  up to 4 where  $2\delta$  covers more than half of the total range of  $x_i$ . We also use  $\alpha_0 = -0.047$ ,  $Var(v_i) = 0.031$  and the reciprocal of the conditional variance of  $u_i$ ,  $1/\sigma_u^2$ , is fixed at 111.467, which is obtained from  $1/Var(E(u_i|y))$  of Seoul male tuberculosis data. We also tried different values of  $1/\sigma_u^2$ , for example,  $1/\sigma_u^2 = 25.361$  in Fig. 5b, the values obtained from R-INLA. The results are similar for different values of  $1/\sigma_u^2$ , therefore for brevity we report only two cases in Fig. 5. To generate  $u_i$  from ICAR, we refer to Rue and Held (2005). For the generated data, we fit the ecological Poisson model with mean  $e_i \exp(\beta_0 + \beta_1 x_i + u_i + v_i)$  and report  $\beta_1 - \alpha_1$  for different values of  $\alpha_1^{ind} = 0.01, 0.05, 0.1, 0.15, 0.2$  and  $\kappa = 0.5$  ( $\alpha_1^{con} = \kappa \alpha_1^{ind}$ ). Each simulation setting is repeated 100 times and the results are summarized in Fig. 5a, b, respectively. Figure 5 shows box-plots where  $\alpha_1^{ind}$  is on the x-axis and the difference  $\beta_1 - \alpha_1$  is on the y-axis when  $\delta = 4$ . It is observed that  $\alpha_1^{ind}$  is sufficiently small, and  $\beta_1$  is very close to  $\alpha_1$  under the misspecification of  $x_{ij}$ . Thus, in Seoul male tuberculosis data, if  $\alpha_1^{ind}$  is believed to be a small positive value, it can be argued that  $\beta_1 \approx 0.05$  means that  $\alpha_1$  is also close to 0.05 even though (10) is true. If  $\alpha_1^{ind}$  is dominant compared to  $\alpha_1^{con}$ , the argument can be stronger, i.e.  $\alpha_1^{ind}$  is close to 0.05 because  $\alpha_1 \approx \alpha_1^{ind}$ . The results for different  $\kappa$  are omitted for brevity because they are similar to the case with  $\kappa = 0.5$ .

### 6.2 Laplace distribution

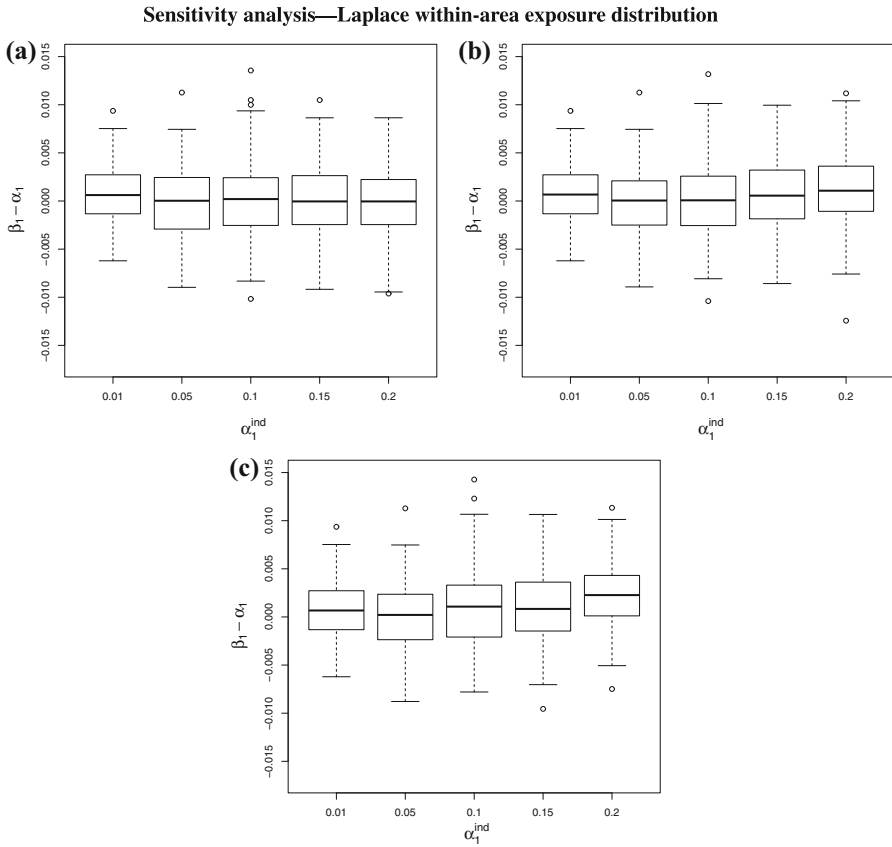
Suppose that

$$x_{ij} \sim \text{Laplace}(x_i, \theta_i) \tag{11}$$

where  $E(x_{ij}) = x_i$  and  $\text{Var}(x_{ij}) = 2\theta_i^2 = s_i^2$ . By using the moment generating function of the Laplace distribution  $E(\exp(\alpha_1^{ind} x_{ij})) = \frac{\exp(\alpha_1^{ind} x_i)}{1 - (\alpha_1^{ind})^2 \theta_i^2}$  for  $|\alpha_1^{ind}| < \frac{1}{\theta_i}$ , we have

$$\begin{aligned} p_i^k &= \sum_j p_{ij}^k / N_{ik} \\ &\approx E \left( \exp \left( \alpha_0 + \alpha_1^{ind} x_{ij} + \alpha_1^{con} x_i + u_i + v_i + \gamma_k \right) \mid u_i, v_i \right) \\ &= \exp \left( \alpha_0 + \left( \alpha_1^{ind} + \alpha_1^{con} \right) x_i + u_i + v_i + \gamma_k \right) \end{aligned}$$

because  $\frac{\exp(\alpha_1^{ind} x_i)}{1 - (\alpha_1^{ind})^2 \theta_i^2} \approx \exp(\alpha_1^{ind} x_i)$  when  $\alpha_1^{ind}$  is small. Thus, in this case,  $\beta_1$  corresponds to  $\alpha_1$  again. Like in the previous example, when the effect size of  $\alpha_1^{ind}$  is believed to be small (but nonzero) and the range of its associated covariate is bounded, this approximation leads to results that are quite robust against the distribution misspecification for  $x_{ij}$ . To check the robustness of this approximation, a numerical study is performed. We first generate  $y_i$  from  $\text{Poisson}(e_i \exp(\alpha_0 + \alpha_1^{con} x_i + u_i + v_i) \frac{\exp(\alpha_1^{ind} x_i)}{1 - (\alpha_1^{ind})^2 \theta_i^2})$  which is the aggregated model from the individual level model with (11). Most of the parameters are taken from Seoul male tuberculosis data. For  $s_i^2$ , we considered three models:  $s_i^2 = 1.5$  or  $1.5 + 0.05x_i$  or  $1.5 + 0.1x_i$ . Different values were also studied for the intercept and the slope, which gave similar pattern of results, and therefore details are omitted in this paper. For the generated data, we fit the ecological Poisson model with mean  $e_i \exp(\beta_0 + \beta_1 x_i + u_i + v_i)$  and report  $\beta_1 - \alpha_1$  for different values of  $\alpha_1^{ind} = 0.01, 0.05, 0.1, 0.15, 0.2$  and  $\kappa = 0.5$  ( $\alpha_1^{con} = \kappa \alpha_1^{ind}$ ). Each simulation setting is repeated 100 times and the results are summarized in Fig. 6a–c, respectively. Figure 6 shows box-plots where  $\alpha_1^{ind}$  is on the x-axis and the difference  $\beta_1 - \alpha_1$  is on the y-axis. Like in the previous example, it is observed that  $\alpha_1^{ind}$  is sufficiently small,  $\beta_1$  is very close to  $\alpha_1$  under the various scenarios about the relationship between  $s_i^2$  and  $x_i$ . Thus, in Seoul male tuberculosis data, if  $\alpha_1^{ind}$  is known to be a small nonzero value, it can be argued that  $\beta_1 \approx 0.05$  means that  $\alpha_1$  is also close to 0.05 even though (11) is true. If  $\alpha_1^{ind}$  is dominant compared to  $\alpha_1^{con}$ , the argument can be stronger, i.e.  $\alpha_1^{ind}$  is close to 0.05 because  $\alpha_1 \approx \alpha_1^{ind}$ . The results for different  $\kappa$  are omitted for brevity because they are similar to the case with  $\kappa = 0.5$ . In case of female, we checked similar results of sensitivity analysis and attached the related figures in the Appendix.



**Fig. 6** Data generated from the aggregated model from the individual level model with the Laplace distribution is analyzed with the ecological Poisson model (1). The true value of  $\alpha_1^{ind}$  is on x-axis, and the bias  $\beta_1 - \alpha_1$  is on y-axis. Three scenarios are considered:  $s_i^2 = 1.5$ ,  $s_i^2 = 1.5 + 0.05x_i$ , and  $s_i^2 = 1.5 + 0.1x_i$  are considered for the leftmost, middle and rightmost figures, respectively. **a**  $s_i^2 = 1.5$ . **b**  $s_i^2 = 1.5 + 0.05x_i$ . **c**  $s_i^2 = 1.5 + 0.1x_i$

### 7 Sensitivity analysis for missing binary covariate

It is well known that smoking can affect the occurrence of tuberculosis, so it is desirable to incorporate smoking variable in the individual model. Since we do not have data on the smoking status, in order to account for this, we perform sensitivity analysis with respect to the missing smoking variable.

Let  $z_{ij}$  denote the smoking status of  $j$ th person in  $i$ th dong. Then,

$$Y_{ij}|x_{ij}, x_i, z_{ij}, u_i, v_i \sim Ber(p_{ij})$$

where

$$\begin{aligned} p_{ij} &= E(Y_{ij}|x_{ij}, x_i, z_{ij}, u_i, v_i) \\ &= \exp\left(\alpha_0 + \alpha_1^{ind} x_{ij} + \alpha_1^{con} x_i + \alpha_2 z_{ij} + u_i + v_i + \gamma_{k_{ij}}\right). \end{aligned}$$



Consider the individuals in  $k$ th age group only.

$$\begin{aligned}
 p_{ij}^k &= \exp\left(\alpha_0 + \alpha_1^{ind} x_{ij} + \alpha_1^{con} x_i + \alpha_2 z_{ij} + u_i + v_i + \gamma_k\right) \\
 &= \left( (1 - z_{ij}) \exp\left(\alpha_0 + \alpha_1^{ind} x_{ij} + \alpha_1^{con} x_i\right) + z_{ij} \exp\left(\alpha_0 + \alpha_2 + \alpha_1^{ind} x_{ij} + \alpha_1^{con} x_i\right) \right) \exp(u_i + v_i + \gamma_k)
 \end{aligned}$$

Then,

$$\begin{aligned}
 p_i^k &= \sum_j p_{ij}^k / N_{ik} \\
 &= \sum_{j:z_{ij}=0} p_{ij}^k / N_{ik} + \sum_{j:z_{ij}=1} p_{ij}^k / N_{ik} \\
 &\approx \left( (N_{ik0} / N_{ik}) \exp\left(\alpha_0 + \alpha_1 x_i + \frac{1}{2} s_i^2 (\alpha_1^{ind})^2\right) + (N_{ik1} / N_{ik}) \exp(\alpha_0 + \alpha_2 + \alpha_1 x_i + \alpha_1^{ind} \psi + \frac{1}{2} s_i^2 (\alpha_1^{ind})^2) \right) \times \exp(u_i + v_i + \gamma_k) \\
 &= \left( (N_{ik0} / N_{ik}) + (N_{ik1} / N_{ik}) \exp(\alpha_2 + \alpha_1^{ind} \psi) \right) \exp\left(\alpha_0 + \alpha_1 x_i + \frac{1}{2} s_i^2 (\alpha_1^{ind})^2 + u_i + v_i + \gamma_k\right) \tag{12}
 \end{aligned}$$

where  $N_{ik}$  is the population size of  $k$ th age category in  $i$ th dong.  $N_{ik0}$  is the population size with  $z_{ij} = 0$ .  $N_{ik1}$  is the population where  $z_{ij} = 1$ . By definition,  $N_{ik0} + N_{ik1} = N_{ik}$ . The expectations are taken with respect to  $x_{ij}$  given  $z_{ij} = 0$  and  $x_{ij}$  given  $z_{ij} = 1$ , respectively. Here, we assume that the exposure distribution conditioned on  $z_{ij}$  follows  $N(x_i + \psi z_{ij}, s_i^2)$ , which implies that  $\psi$  is the difference of average deprivation index between smoker and non-smoker groups. Kim et al. (2017) gives a hint about a reasonable value for  $\psi$  in Seoul. Their supplemental Fig. 1 provides a linear regression fitting result for the association between male smoking prevalence and deprivation index in Metropolitan at the district level. Approximately 1.2 unit of deprivation index increases as one unit of male smoking prevalence increases. If this value is not far from that of dong level, because males are dominant in the smoker group, we use 1.2 as a proxy value for  $\psi$  as shown in the Appendix. We also consider values larger than 1.2 in the sensitivity analysis.

Suppose that  $N_{ik0} = (1 - m_i) N_{ik}$  and  $N_{ik1} = m_i N_{ik}$  where  $m_i$  is the smoking rate in  $i$ th dong. This assumes that the smoking rate is approximately constant across the age groups. Then,

$$\begin{aligned}
 \mu_i &= \sum_k N_{ik} p_i^k \\
 &= \left( \sum_k N_{ik} \exp(\gamma_k) \right) \left( (1 - m_i) + m_i \exp(\alpha_2 + \alpha_1^{ind} \psi) \right) \exp(\alpha_0 + \alpha_1 x_i)
 \end{aligned}$$

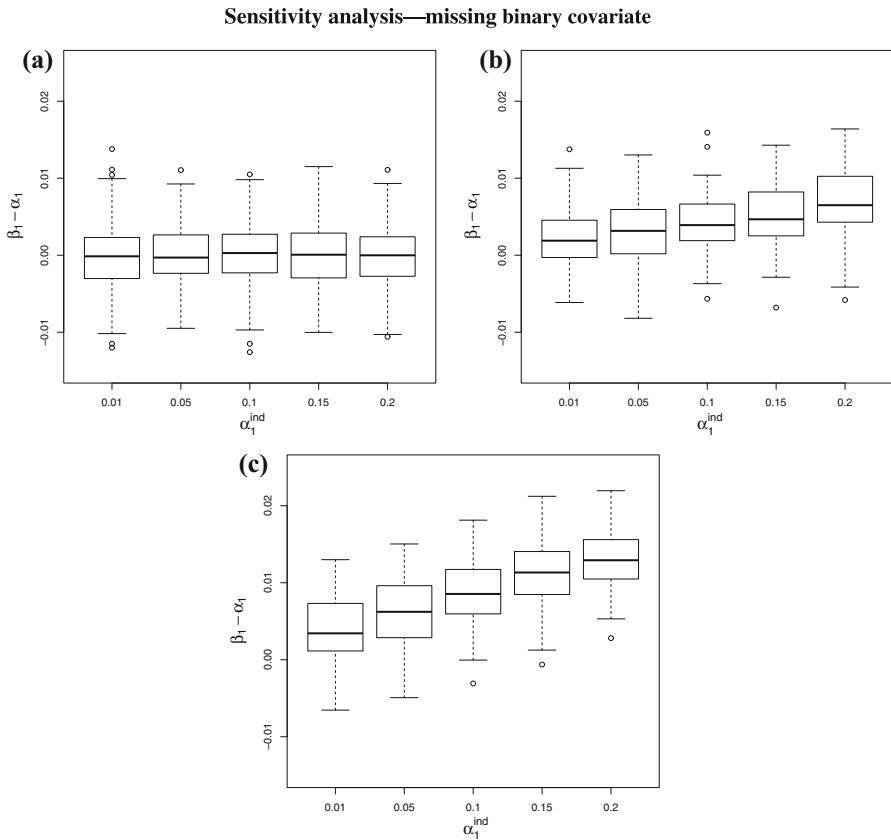
$$\begin{aligned}
& + \frac{1}{2} s_i^2 \left( \alpha_1^{ind} \right)^2 + u_i + v_i \Big) \\
= & e_i \left( (1 - m_i) + m_i \exp \left( \alpha_2 + \alpha_1^{ind} \psi \right) \right) \exp \left( \alpha_0 + \alpha_1 x_i \right. \\
& \left. + \frac{1}{2} s_i^2 \left( \alpha_1^{ind} \right)^2 + u_i + v_i \right) \quad (13)
\end{aligned}$$

$$= e_i \exp \left( \alpha_{0i}^* + \alpha_1 x_i + \frac{1}{2} s_i^2 \left( \alpha_1^{ind} \right)^2 + u_i + v_i \right) \quad (14)$$

where  $\exp(\alpha_{0i}^*) = ((1 - m_i) + m_i \exp(\alpha_2 + \alpha_1^{ind} \psi)) \exp(\alpha_0)$ . Equations (13) and (14) are based on the large sample approximation (12) in each dong.

If  $m_i$  are constant,  $((1 - m_i) + m_i \exp(\alpha_2 + \alpha_1^{ind} \psi))$  will be absorbed in the intercept, so our sensitivity analysis in Sect. 5 can be applied here. When  $m_i$  is an area-specific quantity, their effect will be absorbed into the area-specific random effect  $v_i$ . Thus, it can be argued that the use of area-specific random effect can make our model robust to the missing covariate. In this case, one concern is that the normal assumption for  $v_i$  is robust to the model where  $((1 - m_i) + m_i \exp(\alpha_2 + \alpha_1^{ind} \psi))$  can deviate from the normal distribution.

To check the robustness of this misspecification, a numerical study is performed. Each simulation setting are repeated 100 times. We first generate  $y_i$  from Poisson  $(e_i ((1 - m_i) + m_i \exp(\alpha_2 + \alpha_1^{ind} \psi)) \exp(\alpha_0 + \alpha_1 x_i + u_i + v_i))$  which is the aggregated model from the individual level model with (14) with  $s_i^2 = 0$ . In order to focus on the effect of the misspecification of  $v_i$ , we do not consider within-area exposure variability here.  $m_i$  is generated from  $N(0.458 + c x_i, 0.005^2)$  where  $c = 0, 0.02$  or  $0.04$ . 0.458 is the average smoking rate of Seoul male in 2008 (Korea Centers for Disease Control and Prevention 2008), and  $c > 0$  reflects that smoking rate is higher in disadvantaged social classes. For example,  $c=0.04$  implies that on average 0.04 unit of smoking rate increases as one unit of deprivation index increases. For  $\alpha_2$ , we use 0.01, 0.05 and 0.1. For brevity, we report the result when  $\alpha_2 = 0.1$  only because the results are quite similar for different  $\alpha_2$ . The other parameters are the same as the previous settings. For the generated data, we fit the ecological Poisson model with mean  $e_i \exp(\beta_0 + \beta_1 x_i + u_i + v_i)$  and report  $\beta_1 - \alpha_1$  for different values of  $\alpha_1^{ind} = 0.01, 0.05, 0.1, 0.15, 0.2$ . Figure 7a–c are box-plots where  $\alpha_1^{ind}$  is on the x-axis and the difference  $\beta_1 - \alpha_1$  is on the y-axis. All the simulation results considered here show that  $\beta_1 - \alpha_1$  is close to 0. We also tried 0.417, which was the smoking rate of Seoul male in 2013. Since the simulation results are similar to those of 2008, they are omitted for brevity. From the simulation, we argue that in Seoul male tuberculosis data, the variability due to  $m_i$  does not show a substantial difference between  $\beta_1$  and  $\alpha_1$  even when misspecified normal distribution is used for  $v_i$ . However, it is notable that the bias  $\beta_1 - \alpha_1$  increases with the coefficient  $c$ . Therefore, in the case that the deprivation index affects the smoking rate strongly, i.e.  $c$  is large enough, this sensitivity analysis warns us to be careful in the interpretation of  $\beta_1 \approx \alpha_1$ .

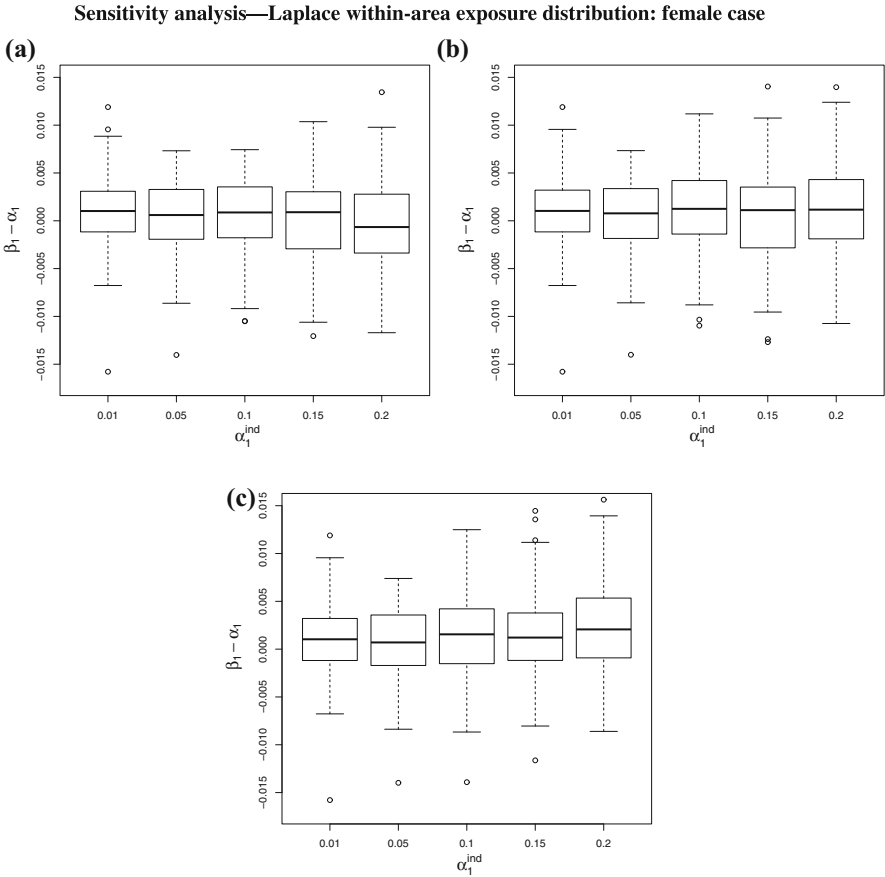


**Fig. 7** Box plots of  $\beta_1 - \alpha_1$  versus  $\alpha_1^{ind}$ . For the three figures,  $\alpha_2$  is fixed at 0.1. **a**  $m_i \sim N(0.458, 0.005^2)$ . **b**  $m_i \sim N(0.458 + 0.02x_i, 0.005^2)$ . **c**  $m_i \sim N(0.458 + 0.04x_i, 0.005^2)$

## 8 Concluding remarks

It is ideal to work with individual level data in order to correctly deal with the ecological bias issue, however, obtaining individual data is not feasible in a lot of ecological studies. In case when individual level data are not available, sensitivity analysis can be an alternative way to provide some justification for assumptions that are made for conveying the findings from the ecological model to the individual level model. In particular, it is useful to compare the fitted ecological model with the derived ecological model from the individual level. However, there are still limitations of sensitivity analysis because it is impossible to fully consider all possible scenarios. Wakefield (2007) also pointed out that there may still be undiscovered factors that can distort the ecological analysis. For example, in Sect. 5, the deprivation index in the individual level may depend on age group.

By analyzing Seoul tuberculosis data, we found that the deprivation index is likely to have a small positive effect on the occurrence risk of tuberculosis at the individual



**Fig. 8** Data generated from the aggregated model from the individual level model with the Laplace distribution is analyzed with the ecological Poisson model (1). The true value of  $\alpha_1^{ind}$  is on x-axis, and the bias  $\beta_1 - \alpha_1$  is on y-axis. Three scenarios are considered:  $s_i^2 = 1.5$ ,  $s_i^2 = 1.5 + 0.05x_i$ , and  $s_i^2 = 1.5 + 0.1x_i$  are considered for the leftmost, middle and rightmost figures, respectively. **a**  $s_i^2 = 1.5$ . **b**  $s_i^2 = 1.5 + 0.05x_i$ . **c**  $s_i^2 = 1.5 + 0.1x_i$

level in Seoul. We considered this in various aspects by performing sensitivity analysis: (1) contextual effect, (2) the functional relationship between mean and variance of the individual level exposure, (3) different distribution assumption for the individual exposure variable and (4) missing binary covariate correlated with the individual exposure variable. Intensive numerical studies support our theoretical analysis. Since the direction of the effect of the deprivation index is consistent across various scenarios, our finding is considered to be robust to some degree. Ultimately, this sensitivity analysis should be corroborated by confirmatory epidemiological studies to investigate association at an individual level (Fig. 8).

**Acknowledgements** This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-

2016R1D1A1B03936100) and the Bio-Synergy Research Project (NRF-2017M3A9C4065964) of the Ministry of Science, ICT and Future Planning through the National Research Foundation.

## Appendix

In Sect. 7, we assume that  $x_{ij}|z_{ij} \sim N(x_i + \psi z_{ij}, s_i^2)$ . Note that

$$E(x_{ij}) = E(E(x_{ij}|z_{ij})) = x_i + \psi E(z_{ij}) = x_i + \psi m_i$$

where  $m_i$  denotes the smoking rate in  $i$ th dong. The difference of mean deprivation indices between two different dongs ( $i$  and  $i'$ ) becomes

$$E(x_{ij}) - E(x_{i'j}) = (x_i - x_{i'}) + \psi(m_i - m_{i'}).$$

Therefore, we have

$$\psi = \frac{E(x_{ij}) - E(x_{i'j}) - (x_i - x_{i'})}{(m_i - m_{i'})}.$$

Take two dongs where  $m_i > m_{i'}$ . If we assume that  $\psi > 0$  and  $x_i$  monotonically increases with  $m_i$ , then

$$\psi \leq \frac{E(x_{ij}) - E(x_{i'j})}{(m_i - m_{i'})}.$$

## References

- Besag J (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J R Stat Soc Ser B* 36:192–236
- Besag J, Higdon D (1999) Bayesian analysis of agricultural field experiments (with discussion). *J R Stat Soc Ser B* 61:691–746
- Besag J, Kooperberg C (1995) On conditional and intrinsic autogressions. *Biometrika* 82:733–746
- Besag J, York J, Mollié A (1991) Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 43:1–20
- Degeun S et al (2010) A small-area ecologic study of myocardial infarction, neighborhood deprivation, and sex—a Bayesian modeling approach. *Epidemiology* 21:459–466
- Diggle PJ, Elliott P (1995) Statistical issues in the analysis of disease risk near point sources using individual or spatially aggregated data. *J Epidemiol Community Health* 49:S20–27
- French C, Kruijshaar M, Jones J, Abubakar I (2009) The influence of socio-economic deprivation on tuberculosis treatment delays in England, 2000–2005. *Epidemiol Infect* 137:591–596
- Greenland S (1992) Divergent biases in ecologic and individual-level studies. *Stat Med* 11:1209–1223
- Greenland S (2001) Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *Int J Epidemiol* 30:1343–1350
- Greenland S (2002) A review of multilevel theory for ecologic analyses. *Stat Med* 21:389–395
- Greenland S, Morgenstern H (1987) Ecological bias, confounding, and effect modification. *Int J Epidemiol* 18:269–274
- Haneuse S, Wakefield J (2004) Ecological inference incorporating spatial dependence. In: King G, Rosen O, Tanner MA (eds) *Ecological inference: new methodological strategies*. Cambridge University Press, New York, pp 266–301
- Jackson C, Best N, Richardson S (2005) Improving ecological inference using individual-level data. *Stat Med* 25:2136–2159

- Jee S, Golub J, Jo J, Park I, Ohrr H, Samet J (2009) Smoking and risk of tuberculosis incidence, mortality, and recurrence in south Korean men and women. *Am J Epidemiol* 170:1478–1485
- Kim J, Yim J (2015) Achievements in and challenges of tuberculosis control in South Korea. *Emerg Infect Dis* 21:1913–1920
- Kim I, Bahk J, Yoon T, Yun S, Khang Y (2017) Income difference in smoking prevalences in 245 districts of South Korea: patterns by area deprivation and urbanity, 2008–2014. *J Prevent Med Public Health* 50:100–126
- Korea Centers for Disease Control and Prevention (2008) The main results. Community Health Survey website. <https://chs.cdc.go.kr/chs/index.do>. Accessed 8 June 2015
- Le Cam L (1960) An approximation theorem for the Poisson binomial distribution. *Pac J Math* 10:1181–1197
- Lopez De Fede A, Stewart J, Harris M, Mayfield-Smith K (2008) Tuberculosis in socio-economically deprived neighborhoods: missed opportunities for prevention. *Int J Tuberc Lung Dis* 12:1425–1430
- McLennan D, Barnes H, Noble M et al (2011) The English indices of deprivation 2010. Department for Communities and Local Government, London
- Richardson S, Stucker I, Hemon D (1987) Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *Int J Epidemiol* 16:111–120
- Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman and Hall/CRC, Boca Raton
- Salway R, Wakefield J (2005) Sources of bias in ecological studies of non-rare events. *Environ Ecol Stat* 12:321–347
- Spiegelhalter D, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). *J R Statist Soc B* 164:583–639
- Steele JM (1994) Le Cam's inequality and Poisson approximations. *Am Math Mon* 101:48–54
- Thomas A, Best N et al (2004) GeoBugs user manual. Medical Research Council Biostatistics Unit. <http://www.mrc-bsu.cam.ac.uk/software/bugs/thebugs-project-geobugs/>. Accessed 8 June 2015
- Townsend P (1987) Deprivation. *J Soc Policy* 16:125–146
- Wakefield J (2003) Sensitivity analysis for ecological regression. *Biometrics* 59:9–17
- Wakefield J (2007) Disease mapping and spatial regression with count data. *Biostatistics* 8:158–183
- Wakefield J, Salway R (2001) A statistical framework for ecological and aggregate studies. *J R Stat Soc A* 164:119–137
- World Health Organization (2014) Global tuberculosis report 2014. WHO website. <http://www.who.int/tb/country/en/>. Accessed 8 June 2015

**Eunjung Song** is a graduate student at Inha University in the Department of Statistics.

**Soeun Kim** is Assistant Professor at University of Texas Health Science Center in the Department of Biostatistics and Data Science.

**Seungsik Hwang** is Associate Professor at Seoul National University Graduate School of Public Health in Department of Public Health Sciences.

**Woojoo Lee** is Associate Professor at Inha University in the Department of Statistics.