Modeling of Infectious Diseases

Jaya, I. Gede Nyoman Mindra; Folmer, Henk; Ruchjana, Budi Nurani; Kristiani, Farah; Andriyana, Yudhie

Published in:
Regional Research Frontiers - Vol. 2

DOI:
10.1007/978-3-319-50590-9_15

Chapter 15
Modeling of Infectious Diseases: A Core Research Topic for the Next Hundred Years

I Gede Nyoman Mindra Jaya, Henk Folmer, Budi Nurani Ruchjana, Farah Kristiani, and Yudhie Andriyana

15.1 Introduction

Incidence of disease is an under-researched topic in regional science. This is unfortunate because it frequently has far-reaching welfare impacts at household, regional, national, and even international levels. For the individual, health problems may range from minor nuisance to death. However, not only the victims but also their family members are affected if they fall ill (e.g., because of an increase in their household tasks or loss of income). Other, mainly financial, implications are related to seeing a doctor or buying medicine. Incidence of disease may also lead to loss of leisure or school days. Another nuisance is restriction of the movement of people to prevent the spread of a disease.

I.G.N.M. Jaya (✉)
Statistics Department, Universitas Padjadjaran, Kabupaten Sumedang, Indonesia
e-mail: mindra@unpad.ac.id

H. Folmer
Faculty of Spatial Science, University of Groningen, Groningen, The Netherlands

B.N. Ruchjana
Mathematics Department, Universitas Padjadjaran, Kabupaten Sumedang, Indonesia

F. Kristiani
Mathematics Department, Parahyangan Catholic University, Kota Bandung, Indonesia

Y. Andriyana
Statistics Department, Universitas Padjadjaran, Kabupaten Sumedang, Indonesia

R. Jackson, P. Schaeffer (eds.), *Regional Research Frontiers - Vol. 2*,
Advances in Spatial Science, DOI 10.1007/978-3-319-50590-9_15

Regional impacts of disease incidence consist in the first place of the impacts on the households that are directly or indirectly affected. However, in addition, there are costs caused by precautionary actions and production losses. In the case of epidemics, such as the Ebola virus disease, a regional system may be paralyzed. Given its welfare impacts and soaring incidence, inter alia, because of climate change, increasing population density, higher mobility, and increasing immunity to several common medicines, the incidence and spread of diseases should become regular research topics in regional science. For recent studies in regional science devoted to the topics, we refer to Ando and Baylis (2013) and Congdon (2013).

Methodological reasons also explain why regional scientists should pay (more) attention to the analysis of the incidence of diseases and its consequences. Although both regional science and epidemiology analyze the spatial distributions of their research topics and apply spatial analytical techniques, interesting methodological differences between them open possibilities for cross-fertilization. Whereas the units of analysis in regional science usually are administrative entities, such as the US states or counties with "large" populations, the spatial units in epidemiology are "small," such as neighborhoods, as required by the effective application of prevention or control measures. Given that the interest in regional science in small region phenomena, such as crime or the development of housing prices at the neighborhood level, is growing, the methods applied in epidemiology may turn out to be applicable in regional science as well. On the other hand, spatial spillover, which is a core issue in regional science for which a large variety of econometric approaches has been developed, has played a less significant role in epidemiology. Considering that infectious diseases tend to spatially spill over, epidemiology may benefit from the spatial spillover models and econometric approaches in regional science.

An important step in the analysis of regional impacts of a disease is the prediction of its incidence. The main objective of this study is to present an overview of the most common statistical methods to predict incidence of *infectious* diseases, to outline their pros and cons and the conditions under which they can be applied. The paper is restricted to infectious diseases. Typical for this type of diseases is that they are transmitted in space (see Sect 15.2). The key concepts in the analysis and prediction of the incidence of an infectious disease are the standardized mortality/morbidity ratio (SMR) and its standard error. In the paper, we discuss three types of approaches that have been used to estimate the key parameters of infectious disease incidence: maximum likelihood (ML), Bayesian methods, and nonparametric methods.

The paper is organized as follows: In Sect. 15.2, we discuss the types of infectious diseases and the basic model used to describe their occurrence. In Sect. 15.3, we discuss the main estimators that have been developed and applied to model the incidence of infectious diseases, i.e., ML, Bayesian smoothing, nonparametric methods, and econometric methods). In Sect. 15.4, we summarize the main findings and present conclusions, including a research agenda.

15.2 Basic Characteristics of Infectious Diseases

Infectious or transmissible diseases are caused by pathogenic microorganisms and transmitted from person to person by direct or indirect contact. Bacteria, viruses, or fungus are examples of the pathogenic agents.

Based on incidence, four types of infectious diseases are usually distinguished. A disease that occurs occasionally in a population is classified as *sporadic*; if it occurs constantly, it is *endemic*; if a large number of victims are infected in a short period, it is *epidemic*; and if it occurs worldwide in a short period, it is *pandemic*.

Infectious diseases have three transmission mechanisms: *contact*, *vehicle*, and *vector transmission*. In the first mechanism, the transmission is by direct person-to-person contact or indirect by contact with nonliving objects (such as contaminated soils) or by mucus droplets in coughing, sneezing, laughing, or talking. In the second mechanism, media, such as air (airborne), food (food-borne), or water (waterborne), are the transmitting agents. Finally, a vector is a mechanism that transports infectious agents from an infected person or animal to susceptible individuals. Vectors consist of two types: biological and mechanical. In the case of a biological vector, the agent reproduces in the vector's body that carries it to the susceptible person. Examples of biological vectors are mosquitoes, ticks, and bugs. A mechanical vector picks up and transports the agent outside of its body. The vector itself is not infected by the agent. An example is a housefly. Vector transmission is the most common transmission mechanism. For more details about transmission and its mechanisms, we refer to, e.g., Chen et al. (2015).

15.3 Infectious Disease Modeling

The basic concept in modeling the relative risk of an infectious disease is the SMR. It is used to identify high-risk regions. It is defined as follows: assume y_i and e_i are the observed and expected number of cases in region i, $(i = 1, 2, 3, \ldots, N)$, respectively. The SMR is then defined as follows:

$$SMR_i = \frac{y_i}{e_i}, \tag{15.1}$$

where e_i defined as

$$e_i = N_i \times \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} N_i}, \tag{15.2}$$

and N_i is the size of the population at risk in region i. A larger than one (15.1) SMR means that the region concerned has a larger actual incidence than its expectation; such region is classified as a high-risk region. By contrast, a region with a smaller than one (15.1) SMR is a low-risk region (Tango 2010).

15.3.1 ML

The traditional estimator of relative risk is ML (Shaddick and Zidek 2016). For count data and y_i, a "small" non-negative, discrete number, the Poisson distribution is typically chosen to model infectious disease incidence. With mean and variance $e_i\theta_i$ respectively, where θ_i is the relative risk parameter in region i, the following is obtained:

$$y_i \Big| e_i\theta_i \sim \text{Poisson}\left(e_i\theta_i\right). \tag{15.3}$$

The simplest model assumes no covariate and random term in the model. The ML estimator of θ_i is

$$\widehat{\theta}_i^{ML} = \frac{y_i}{e_i}, \tag{15.4}$$

which is unbiased. The variance is

$$\widehat{V\left(\widehat{\theta}_i^{ML}\right)} = \frac{\widehat{\theta}_i^{ML}}{e_i}. \tag{15.5}$$

For small e_i, (15.4) and (15.5) are "large" which leads to imprecise estimation of relative risk. For example, two similar regions, A and B, have the same population at risk, that is, they have the same expected number of cases, e_i. Suppose that e_i is 0.1 and that in region A one case is found and in B, zero. Hence, $\widehat{\theta}_i^{ML}$ in region A is 10 and in region B, zero. Region A has extreme $\widehat{\theta}_i^{ML}$ compared with region B, while the number of cases differs by 1 only. It follows that the ML-estimated relative risk may be very unstable and lead to wrong conclusions (Pringle 1996). Consequently, more appropriate methods for disease modeling and mapping are required. One class of such methods is smoothing. Smoothing techniques exploit information from neighboring regions to adjust the estimate for a given region. The basic principle is *shrinkage*. That is, ML estimates with small expected rates or high variances will be "shrunk" toward the overall mean, whereas those with small variances will essentially remain unchanged. Smoothing thus decreases the mean squared error (Anselin et al. 2006). Bayesian and nonparametric techniques are two popular smoothing methods used in disease modeling and mapping.

15.3.2 Bayesian Smoothing

Bayesian smoothing methods are statistical approaches to update unknown parameters using information from observations. As a first step, prior information on the parameter of interest is specified in terms of a probability distribution. Next, empirical evidence (data) is obtained and combined with the prior information using Bayes' theorem, which leads to a posterior probability distribution of the parameters. The posterior becomes the basis for statistical inference (Congdon 2010). Specifically, the observed data $y = (y_1, \ldots, y_n)^T$ is assumed to be generated from a probability distribution $f(y_i | \theta_i)$ with unknown parameters $\theta = (\theta_1, \ldots, \theta_n)^T$. The unknown parameters θ, in turn, are assumed to be random variables with prior $f(\theta_i | \boldsymbol{\gamma})$ with unknown hyperparameter $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^T$. The posterior density of θ_i, given the data y_i, the conditional density $f(y_i | \theta_i)$, and the conditional density $f(\theta_i | \boldsymbol{\gamma})$, is

$$f(\theta_i | y_i, \boldsymbol{\gamma}) = \frac{f(y_i | \theta_i) \times f(\theta_i | \boldsymbol{\gamma})}{f(y_i | \boldsymbol{\gamma})}, \tag{15.6}$$

where $f(y_i | \boldsymbol{\gamma})$ is the marginal likelihood of the data given hyperparameter $\boldsymbol{\gamma}$. To ensure that the posterior distribution, $f(\theta_i | y_i)$, is a proper density, the marginal likelihood, $f(y_i | \boldsymbol{\gamma})$, is taken as a normalizing constant, which is found by integrating the likelihood, $f(y_i | \theta_i)$, over the joint prior density:

$$f(y_i | \boldsymbol{\gamma}) = \int f(y_i | \theta_i) \times f(\theta_i | \boldsymbol{\gamma}) \; d\theta_i. \tag{15.7}$$

Based on the above mentioned description, (15.6) can be written as follows:

$$f(\theta_i | y_i, \boldsymbol{\gamma}) \propto f(y_i | \theta_i) \times f(\theta_i | \boldsymbol{\gamma}). \tag{15.8}$$

The estimated posterior density $f(\theta_i | y_i, \widehat{\boldsymbol{\gamma}})$ is used to make inferences about θ_i, where $\widehat{\boldsymbol{\gamma}}$ is an estimate of $\boldsymbol{\gamma}$.

Bayesian approaches are composed of two classes: empirical Bayes (EB) and full Bayes (FB). Each is made up of several types. In the case of EB, parameters $\boldsymbol{\gamma}$ are replaced by point estimates of hyperparameter based on the marginal distribution of y_i. In the case of FB, a prior distribution $f(\gamma_1), \ldots, f(\gamma_k)$, is specified for the hyperparameter $\boldsymbol{\gamma}$ (Hog et al. 2005).

A typical example of each case is presented below.

15.3.2.1 Empirical Bayes Poisson-Lognormal Model[1]

The empirical Bayes Poisson-lognormal (EBPLN) model was introduced by Clayton and Kaldor (1987). It can be summarized as follows: The prior distribution of the relative risk, θ, is assumed to have a multivariate lognormal distribution. That

[1] Other EB models are the Poisson-Gamma model and the linear empirical Bayes model. See, e.g., Clayton and Kaldor (1987) and Lawson et al. (2000) for details.

is, the log of the relative risk, $\boldsymbol{\zeta} = \log(\theta)\,; \boldsymbol{\zeta} = (\zeta_1, .., \zeta_n)^T$,is assumed to follow a multivariate normal distribution with mean $\boldsymbol{\mu}$and covariance matrix $\boldsymbol{\Sigma}$. Hence, the density function of $\boldsymbol{\zeta}$ is as follows:

$$f(\boldsymbol{\zeta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{1}{2n}} (\theta_1 \ldots \theta_n)^{-1} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(log\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(log\boldsymbol{\theta} - \boldsymbol{\mu})\right\}. \tag{15.9}$$

The EB estimator is obtained from the expectation of the relative risk θ given y, $E(\theta\,|\,y)$. However, the posterior distribution of the Poisson-lognormal is not a closed form, that is, it has no analytical solution for $E(\theta\,|\,y)$. As a way out, Clayton and Kaldor (1987) proposed a quadratic approximation by substituting θ_i for $\exp(\zeta_i)$ to construct the Poisson likelihood $\boldsymbol{\zeta}$ given y. The likelihood thus is

$$L(\boldsymbol{\zeta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\zeta_i) = \prod_{i=1}^{n}\left(\frac{\exp(-e_i\exp(\zeta_i))\,(e_i\exp(\zeta_i))^{y_i}}{y_i!}\right). \tag{15.10}$$

The EB estimator using the quadratic approximation requires the estimate of the vector of parameters $\boldsymbol{\zeta}$. Clayton and Kaldor (1987) proposed ML to estimate $\boldsymbol{\zeta}$. The ML estimator of $\widetilde{\zeta}_i = \log\left(\frac{y_i}{e_i}\right)$.However, this solution does not hold for $y_i = 0$. Therefore, Clayton and Kaldor (1987) suggested to add the constant 0.50 to y_i, that is,

$$\widetilde{\zeta}_i = \log\left(\frac{y_i + 0.5}{e_i}\right). \tag{15.11}$$

Equation (15.11) is an explicit solution of the EB estimate of $\boldsymbol{\zeta}$ based on quadratic approximation. However, the solution is not based on the expectation of the posterior distribution of the Poisson-lognormal model, $f(\boldsymbol{\zeta}\,|\,y\,,\boldsymbol{\mu}\,,\boldsymbol{\Sigma})$. With the quadratic approximation of the likelihood function over the lognormal prior, the posterior distribution of $\boldsymbol{\zeta}$ given the data y is

$$f(\boldsymbol{\zeta}|\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto f(\boldsymbol{y}|\boldsymbol{\zeta})\, f(\boldsymbol{\zeta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{15.12}$$

which follows a multivariate normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ (Leonard 1975; see Clayton and Kaldor 1987, for details). Estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is thus necessary to obtain an explicit solution for $\boldsymbol{\zeta}$ based on $f(\boldsymbol{\zeta}\,|\,y\,,\boldsymbol{\mu}\,,\boldsymbol{\Sigma})$. The EM algorithm can be used for this purpose. In the simplest case, the ζ_i are taken as $i\,.\,i\,.\,d\;N(\mu, \sigma^2)$. Given that the distribution of the ζ_i has two parameters, μ and σ^2, the EBPLN, $\widehat{\zeta}_i^{\text{EBPLN}}$, becomes (Meza 2003):

$$\widehat{\zeta}_i^{EBPLN} = \frac{\widehat{\mu} + \widehat{\sigma}^2\,(y_i + 0.5)\,\widetilde{\zeta}_i - 0.5\widehat{\sigma}^2}{1 + \widehat{\sigma}^2\,(y_i + 0.5)}, \tag{15.13}$$

with

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\zeta}_i^{EBPLN}, \tag{15.14}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\left(\widehat{\sigma}^2\sum_{i=1}^{n}\left[1+\widehat{\sigma}^2\left(y_i+0.5\right)\right]^{-1} + \sum_{i=1}^{n}\left(\widehat{\zeta}_i^{EBPLN} - \widehat{\mu}\right)^2\right). \tag{15.15}$$

The EBPLN estimator of the relative risk is $\widehat{\theta}_i^{\mathrm{EBPLN}} = \exp\left(\widehat{\zeta}_i^{\mathrm{EBPLN}}\right)$.

The EM algorithm to (iteratively) obtain the estimates of μ and σ^2 using Equations (15.13), (15.14), and (15.15) is as follows:

(1) Obtain the initial values of $\left\{\widehat{\zeta}_i,\widehat{\mu},\widehat{\sigma}^2\right\}$:

   (a) $\widehat{\zeta}_i = \log\left(\frac{y_i+0.5}{e_i}\right)$
   (b) $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\zeta}$
   (c) $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\zeta} - \widehat{\mu}\right)^2$

(2) *Expectation* (E) Step: Estimate the relative risk using Equation (15.13).
(3) *Maximization* (M) Step: Update the parameter estimates $\widehat{\mu}$ and $\widehat{\sigma}^2$ using Equations (15.14) and (15.15).
(4) Repeat Steps 2–3 until a predetermined precision is obtained, e.g.,$\left|\widehat{\zeta}_i^{\mathrm{EBLN}(t+1)} - \widehat{\zeta}_i^{\mathrm{EBLN}(t)}\right| \leq 1e - k$, with k a positive integer.

15.3.2.2 Full Bayesian Poisson-Lognormal Model[2]

Full Bayesian (FB) estimation is more widely used in Bayesian disease modeling than EB because it is more flexible in defining the prior hyperparameter $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^T$, and because it can provide a measure of uncertainty of the estimates of relative risks (Maiti 1998). The quality of the FB estimates depends on the accuracy in determining a hyperprior distribution.

In FB, the posterior parameters can be estimated using Markov chain Monte Carlo (MCMC) simulation, such as the Gibbs sampler and Metropolis-Hastings (M-H) or integrated nested Laplace approximation (INLA). The procedure is as follows: As in the case of EBPLN, FBPLN assumes the log relative risk, ζ_i, to follow a normal distribution, that is, $\zeta_i \sim \mathrm{i.i.d\,Normal}(\mu, \sigma^2)$.

[2] Another FB model is the Poisson Gamma model. See, e.g. Lawson (2006) for an overview.

The basic FBPLN model may be written as follows (Meza 2003):

(i) $y_i \Big| \theta_i \overset{iid}{\sim} \text{Poisson}\,(e_i\theta_i)$

(ii) $\zeta_i = \log\,(\theta_i)\,\Big|\mu, \sigma^2 \overset{iid}{\sim} N\left(\mu, \sigma^2\right)$

(iii) $f\left(\mu, \sigma^2\right) \propto f\,(\mu) f\left(\sigma^2\right)$ with $f\,(\mu) \propto 1; \sigma^{-2} \sim \text{Gamma}\,(a, b)\,; a \geq 0, b > 0$

Commonly, the prior parameters (a, b) are assumed to be known. Obtaining the posterior distribution of $\theta_i|y_i$ involves high-dimensional integrals that are difficult to sample directly from. However, sampling from the full conditional distribution of each parameter is often easy. The Gibbs sampler can be used to estimate the posterior distribute on (Maiti 1998). The full conditional distribution to implement Gibbs sampling can be written as follows:

(i) $f\left(\theta_i|\mu, \sigma^2, y_i\right) \propto \theta_i^{y_i-1} \exp\left[-e_i\theta_i - \frac{1}{2\sigma^2}(\zeta_i - \mu)^2\right]$

(ii) $\left[\mu|\theta_i, \sigma^2, y_i\right] \sim N\left(\frac{1}{n}\sum_i \zeta_i, \frac{\sigma^2}{m}\right)$

(iii) $\left[\sigma^2|\theta_i, \mu, y_i\right] \sim G\left(\frac{n}{2} + a, \frac{1}{2}\sum_i (\zeta_i - \mu)^2 + b\right)$

MCMC samples can be directly generated from (ii) and (iii) using the M-H algorithm. Several software programs can be used to estimate the FBPLN. The WinBUGS software program is generally used.

For computational purposes, ζ_i is decomposed into two components, β_0 and u_i. β_0 is the overall level of the log relative risk, whereas u_i is the residual.

$$\log\,(\theta_i) = \beta_0 + u_i, \tag{15.16}$$

$$u_i \sim i.i.d\; Normal\left(0, \sigma_u^2\right).$$

The parameters β_0 and u_i have a hyperprior distribution as follows:

$$\beta_0 \sim i.i.d\; Normal\left(0, \sigma_{\beta_0}^2\right),$$

$$1/\sigma_u^2 \sim Gamma\,(a, b)\,.$$

Using noninformative prior, the value of $\sigma_{\beta_0}^2$ is usually replaced by a large number, for example, $\sigma_{\beta_0}^2 = 10^5$ and for $a = 0.5$ and $b = 0.0005$ (Tango 2010).

15.4 The Besag, York, and Mollie (BYM) FB Model

ML and the traditional Bayesian approaches do not accommodate spatial trend, covariates, and spatially uncorrelated and spatially correlated heterogeneity. The FBPLN model can be extended to include those components. To consider spatially correlated heterogeneity, Clayton and Kaldor (1987) proposed the conditional autoregressive (CAR) model for the log relative risk. The CAR model is defined as follows:

$$E\left(\zeta_i|\zeta_{j(j\neq i)}\right) = \mu_i + \rho\sum_j w_{ij}\left(\zeta_j - \mu_j\right)$$

$$Var\left(\zeta_i|\zeta_{j(j\neq i)}\right) = \sigma^2, \tag{15.17}$$

where $w_{ij}$is an element of the spatial weights matrix **W.** To simplify computations, μ_i is assumed to be equal to μ.

The "complete" FBLN model to estimate the relative risk was developed by Besag et al. (1991), denoted BYM. Considering its "completeness", it has become a popular model in Bayesian disease modeling and mapping, especially of infectious diseases. The BYM model reads as follows (Lawson et al. 2000):

$$\log\left(\theta_i\right) = t_i + u_i + v_i, \tag{15.18}$$

where t_i denotes the spatial trend and covariates, u_i denotes the spatially uncorrelated heterogeneity, and v_i denotes the spatially correlated heterogeneity (Lawson et al. 2000). A typical spatial trend regression model reads as follows:

$$t_i = \sum_{h=1}^{H}\left(a_h x_i^h + b_h y_i^h\right) + \sum_{k=1}^{K} c_k z_k, \tag{15.19}$$

where $\{(x_i, y_i)\}$ are the centroids of the i-th region, H is the degree of the trend (e.g., $h = 1$: linear trend; $h = 2$:quadratic trend), K is the number of covariates, and z is the vector of covariates.

In the case of count data, over-dispersion frequently occurs, that is, the variance observed is greater than the mean. Over-dispersion has two types: spatially uncorrelated and spatially correlated heterogeneity (Lawson 2006). Spatially uncorrelated heterogeneity occurs because of observations with small or zero cases, differences in the number of subpopulation, and omitted environmental or ecological factors, such as pollution, rainfall, humidity, temperature, and radiation. Spatially uncorrelated heterogeneity is accommodated by defining a non-informative prior[3] for u_i, usually

[3] A noninformative prior is used to denote lack of information about the parameter of interest (Lawson 2013).

the normal distribution (Lawson et al. 2003):

$$u_i \sim i.i.d \text{ Normal}\left(0, \sigma_u^2\right). \tag{15.20}$$

Spatially correlated heterogeneity, v_i, occurs because of spatial clustering or spatial autocorrelation (Lawson 2006). It can be considered using information relating to adjacent regions, based on the assumption that adjacent regions with similar spatial characteristics have similar relative risks.

A conditional autoregressive (CAR) prior is usually used to capture spatially correlated heterogeneity. Besag et al. (1991) proposed the following CAR prior:

$$v_i \Big| v_{j \neq i} \sim \text{Normal}\left(\frac{\sum_j w_{ij} v_j}{\sum_j w_{ij}}, \frac{\sigma_v^2}{\sum_j w_{ij}}\right), \tag{15.21}$$

where w_{ij} denotes spatial dependence between regions i and j.

A limitation of the Besag prior is that it is only appropriate for strong spatial autocorrelation. If weak spatial autocorrelation exists, the CAR prior produces random effects that are overly smooth (Lee 2013). To overcome this limitation, spatially uncorrelated heterogeneity u_i should be used. To accommodate varying strengths of spatial autocorrelation, Leroux et al. (1999) and Stern and Cressie (1999) proposed alternative CAR priors. The Leroux et al. (1999) CAR prior reads as follows:

$$v_i \Big| v_{j \neq i} \sim N\left(\frac{\rho \sum_j w_{ij} v_j}{\rho \sum_j w_{ij} + 1 - \rho}, \frac{\sigma_v^2}{\rho \sum_j w_{ij} + 1 - \rho}\right), \tag{15.22}$$

The Stern and Cressie (1999) CAR prior is as follows:

$$v_i \Big| v_{j \neq i} \sim N\left(\frac{\rho \sum_j w_{ij} v_j}{\rho \sum_j w_{ij}}, \frac{\sigma_v^2}{\rho \sum_j w_{ij}}\right). \tag{15.23}$$

In both cases, ρ is the spatial autocorrelation parameter. Using the Leroux or Stern and Cressie prior renders spatially uncorrelated heterogeneity $u_i$redundant.

The FBPLN model, including spatial effects, may be written as follows (Rao 2003):

(i) $y_i | \theta_i \sim \text{Poisson}\left(e_i \theta_i\right)$

(ii) $\zeta_i \Big| \zeta_{j(j \neq i)}, \rho, \sigma^2 \sim N\left(\mu + \rho \sum_{il} w_{il}\left(\zeta_l - \mu\right), \sigma^2\right)$

(iii) $f\left(\mu, \sigma^2, \rho\right) \propto f(\mu) f\left(\sigma^2\right) f(\rho)$ with
$f(\mu) \propto 1; \sigma^{-2} \sim \text{Gamma}(a, b); a \geq 0, b > 0, \rho \sim U\left(0, \rho_0\right)$

where ρ_0 denotes the maximum value of ρ in the CAR model and $\boldsymbol{W} = (w_{il})$ is the "adjacency" matrix. Maiti (1998) proposed Gibbs sampling combined with the M-H algorithm to estimate the model.

The BYM model can be summarized as follows:

$$\eta_i = \beta_0 + \boldsymbol{X}_i^T \boldsymbol{\beta} + u_i + v_i, \tag{15.24}$$

where $\eta_i = \log(\theta_i)$, $\beta_0$is the overall relative risk, $\boldsymbol{X}_i^T = (X_{i1}, .., X_{iK})$ is a vector covariates, $\boldsymbol{\beta} = (\beta_1, .., \beta_K)^T$ is a vector regression coefficients, and u_i and $v_i$denote are spatially uncorrelated and spatially correlated heterogeneity, respectively. The following hyperparameter distributions of β_0, u_i and v_i are usually applied:

$$\beta_0, \beta_1, .., \beta_k \sim i.i.d\text{Normal}\left(0, \sigma_\beta^2\right),$$

$$1/\sigma_u^2 \sim \text{Gamma}\,(a, b)\,,$$

$$1/\sigma_v^2 \sim \text{Gamma}\,(a, b)\,.$$

As a non-informative prior, large values for σ_β^2 are usually taken, for example, $\sigma_\beta^2 = 10^5$ and for $a = 0.5$ and $b = 0.0005$ (Tango 2010).

The above-mentioned model only accounts for the spatial pattern of diseases but does not incorporate temporal variation. A model that includes temporal variation is a spatio-temporal model. Spatio-temporal modeling has been widely applied to analyze the spatial distribution of disease incidence and its trend, notably to detect hotspots (Lawson 2014). The most common approach is based on the assumption that a log-linear relationship exists between the relative risk and the calendar time within regions, that is, that the time trend varies from region to region (Lawson 2014). Thus

$$\mathrm{y_{it}}\Big|\mathrm{e_{it}}\theta_{\mathrm{it}} \sim \text{Poisson}\,(\mathrm{e_{it}}\theta_{\mathrm{it}})\,,$$

$$\eta_{\mathrm{it}} = \beta_0 + \mathrm{X}_{\mathrm{it}}^{\mathrm{T}} \boldsymbol{\beta} + \mathrm{u_i} + \mathrm{v_i} + \omega_{\mathrm{t}} + \psi_{\mathrm{t}} + \phi_{\mathrm{it}}, \tag{15.25}$$

where $\eta_{it} = \log(\theta_{it}) u_i$ and v_i denote spatially uncorrelated and spatially correlated heterogeneity, respectively; ω_j and ψ_t denote temporally uncorrelated and temporally-correlated heterogeneity, and ϕ_{ij} is a spatio-temporal interaction effect. This model varies based on the structure of the space-time structure. Model (15.25) is commonly estimated using Bayesian techniques.

15.4.1 Nonparametric Estimation

The most popular nonparametric smoothing technique is the Nadaraya-Watson kernel smoother. It is defined as the weighted average of the ML estimates of the other regions (Lawson et al. 2000):

$$\theta_i^{NP} = \sum_{j \neq i}^{n} \omega_j \theta_j^{ML}, \tag{15.26}$$

with ω_j weights for values of neighboring regions defined as follows:

$$\omega_j = \frac{K\left(\left(\theta_i^{ML} - \theta_j^{ML}\right)/h\right)}{\sum_i^n K\left(\left(\theta_i^{ML} - \theta_j^{ML}\right)/h\right)}, \tag{15.27}$$

where $K(.)$ is a zero mean, radially symmetric probability density function, usually the standard Gaussian distribution:

$$K(z) = (2\pi)^{-1/2}\exp\left(-\frac{z^2}{2}\right), \tag{15.28}$$

with h the bandwidth based on the minimum value of the least squares cross-validation criteria (Simonoff 1999):

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_i^{NP} - \overline{\hat{\theta}_{(-i)}^{ML}}\right)^2, \tag{15.29}$$

Where $\overline{\hat{\theta}_{(-i)}^{ML}}$ denotes the average relative risk estimate using ML without the i^{th} observation.

For an application to relative risk estimation, see Kesall and Diggle (1998). The nonparametric model can be extended to include time variation and spatial dependence as follows:

$$\log\left(\lambda_{it}|y_{it}\right) = \log\left(n_{it}\right) + \log(m) + S_0(t) + \alpha_i + S_i(t), \tag{15.30}$$

where λ_{it} is a mean of Poisson distribution; n_{it} is the population count for the region i in year t; m is the overall mean of the relative risk; $S_0(t)$ is the fixed global of the relative risk trend; α_i is the random spatial effect, which may be spatially correlated; and $S_i(t)$ is the random temporal effect for the region i (MacNab and Dean 2002).

15.4.2 Spatial Econometric Models

The models discussed in the previous sections (explicitly) do not consider spatial dependence even though spatial spillovers are typical for infectious diseases. Particularly, the response variable in one region usually depends on the values of the response variable in neighboring regions (Lawson 2014; Chen et al. 2015), as in the case of dengue fever. Similarly, the status of covariates (e.g., vegetation or water quality) in one region may affect the response variable not only in that region but also in neighboring regions. Finally, spatial dependence may occur among the error terms.

One of the reasons that spatial econometric models have received little attention in epidemiology is that these models have been developed for continuous data rather than count data, especially with respect to the dependent variable. Following Lambert et al. (2010) and Bivand et al. (2014), we specify the spatially lagged (SL) mixed Poisson regression model of relative risk for count data with spatially lagged dependent variable, spatially uncorrelated (u_i) and spatially correlated (v_i) heterogeneity as components of the error term (ε_i), as follows:

$$\boldsymbol{\eta} = \rho_{Lag} \boldsymbol{W} \boldsymbol{\eta} + \beta_0 \mathbf{1}_n + \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{15.31}$$

where $\boldsymbol{\eta} = (\eta_{\mathbf{1}}, .., \eta_n)^T$ with $\eta_i = \log(\theta_i)$, β_0 is the overall relative risk, $\mathbf{1}_n$ is a unit vector of length n, $\boldsymbol{X}$ is a matrix of covariates of size (nxK), $\boldsymbol{\beta} = (\beta_1, .., \beta_k)^T$ is a vector of coefficients, and $\boldsymbol{W}$ is a symmetric adjacency matrix with zero diagonal elements, ρ_{Lag} is the spatial lag parameter that measures infectious disease spillover among regions.

A more general model with wider applicability is the spatial Durbin-Poisson (SD-Poisson) model that allows for spatial spillovers of the covariates in addition to a spatially lagged dependent variable. The SD-Poisson model reads as follows:

$$\boldsymbol{\eta} = \rho_{Lag} \boldsymbol{W} \boldsymbol{\eta} + \beta_0 \mathbf{1}_n + \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{W} \boldsymbol{X} \boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{15.32}$$

where $\boldsymbol{\delta} = (\delta_1, \ \ldots \ , \delta_K)^T$ denotes a vector of coefficients for the spatially lagged covariates $\boldsymbol{WX}$ (Bivand et al. 2014).

From models (15.31) and (15.32), the direct and indirect (spillover) effects can be calculated. To estimate the SL-Poisson model, Lambert et al. (2010) proposed two-step limited information maximum likelihood, and Bivand et al. (2014) developed a Bayesian estimator using INLA.

15.5 Summary and Research Recommendations

Incidences of infectious diseases have been soaring. According to the World Health Organization (2005), climate change, extreme weather, and environmental factors, such as lack of access to clean water and poor sanitation facilities, have contributed to the outbreaks. Socioeconomic conditions, including income, employment, education, and health behavior, are also important factors that influence the transmission of infectious diseases. Increasing urbanization, higher population density, higher mobility, and increasing resistance to several common medicines accelerate the transmission from one location to another because of more contacts between infected and susceptible people (Fong 2013).

Infectious diseases often have serious direct and indirect effects at the individual, household, and regional levels ranging from increased morbidity and mortality to the paralysis of an entire region or even a country. Early identification of an endemic is an important first step to prevent its transmission and to reduce its effects. Implementation of such early warning systems (EWSs), including roadmaps to prevent or restrict the spread of an infectious disease, is still in its infancy in most (developing) countries (Lowe et al. 2011). Therefore, the development and implementation of EWSs based on information about when and where outbreaks will occur and what factors influence transmission is a high-priority research topic. A related research topic is how to use EWS information in taking appropriate and efficient actions to manage transmission and to prevent epidemics. The development and implementation of an EWS requires intensive interaction between natural and social regional scientists.

An important component of an EWS is the identification of high-risk regions and spatial clustering. For this purpose, predictive models are required (Chen et al. 2015). In this paper, an overview of the most common approaches in disease incidence modeling has been presented. Four types of approaches have been discussed, namely, ML, Bayesian smoothing, nonparametric smoothing, and spatial econometric methods. An important conclusion that emerges from the overview presented in Sect. 15.3 is that the first three types of models do not adequately account for the basic characteristic of infectious diseases, i.e., spatial spillover. Admittedly, several of the approaches that have been commonly applied in infectious disease modeling account for similarities among spatial units, notably climate and environmental conditions, which significantly affect habitat suitability and distribution of vectors. However, this is not the same as accounting for spatial spillover. Spatial spillover means that the sheer presence of an infectious disease in one region, at present or in the past, increases the likelihood of occurrence in neighboring regions. Another type of spatial dependence relates to the covariates in that covariates in one region affect the response variable not only in that region but also in neighboring regions.

A major research topic for the immediate future is the development of models that can explain and predict the spatio-temporal distribution of infectious diseases. For that purpose, epidemiological and spatio-temporal econometric models could be combined. The basic structure of such a model that links the log of the relative risk to its predictors is as follows:

$$\eta_{it} = \beta_0 + \rho_1 \sum_{j=1}^{n} w_{ij}\eta_{jt} + \rho_2 \sum_{j=1}^{n} w_{ij}\eta_{jt-1} + \rho_3\eta_{it-1} + \sum_{k=1}^{K} \beta_{1k}X_{kit} + \sum_{k=1}^{K} \beta_{2k}X_{kit-1}$$
$$+ \sum_{k=1}^{K} \beta_{3k} \sum_{j=1}^{n} w_{ij}X_{kjt} + \sum_{k=1}^{K} \beta_{4k} \sum_{j=1}^{n} w_{ij}X_{kjt-1} + u_i + v_i + \omega_t + \psi_t + \phi_{it}, \tag{15.33}$$

where $\eta_{it} = \log(\theta_{it})$; ρ_1 and ρ_2 denote the spatial lag coefficients of the log relative risk without and with time lag, respectively; ρ_3 denotes a temporal lag coefficient of the log relative risk; β_{1k} and β_{2k} denote the regression coefficients with and without temporal lag of the k_{th} covariates, respectively; β_{3k} and β_{4k} denote the spatial lag coefficients of the covariates with and without temporal lag, respectively; u_i and v_i denote spatially uncorrelated and spatially correlated heterogeneity, respectively; ω_j and ψ_t denote temporally uncorrelated and temporally correlated heterogeneity and ϕ_{it} is a spatio-temporal interaction effect. Correlated heterogeneity is variability that occurs because of spatial or temporal dependence; uncorrelated heterogeneity is variability that occurs because of random spatial or temporal variation (Lawson 2006; Bernardinelli et al. 1995).

Model (15.33) is a complex model with a discrete (Poisson distributed) dependent variable, involves many covariates, and is influenced by location and time heterogeneity. Spatial panel econometrics comes to mind to estimate model (15.33). However, spatial panel econometrics has been developed for continuous response variables, while epidemiological data are commonly measured in count format. Therefore, models such as (15.33) cannot be estimated by conventional approaches. The development of appropriate estimators of such models is an important topic for further research. We expect that Bayesian statistics will be increasingly used in epidemiology and regional science models of count data (see also Congdon 2013). For complex models, such as the spatio-temporal varying coefficient model, the calculation of the likelihood function, along with the problem of identifiability of the parameters, is very difficult. The Bayesian method can solve this problem (Martinez and Achcar 2014).

We also expect the random effect generalized linear mixed model and Bayesian inference with INLA to become popular in infectious disease modeling. INLA is a relatively new approach to Bayesian statistical inference for latent Gaussian Markov random fields. The main advantage of the INLA approach over MCMC is that it can compute significantly faster (Rue et al. 2007).

References

Ando AW, Baylis K (2013) Spatial environmental and natural resource economics. In: Fischer MM, Nijkamp P (eds) Handbook of regional science. Springer, New York, pp 1029–1048

Anselin L, Lozano N, Koschinsky J (2006) Rate transformations and smoothing. University of Illinois, Urbana

Bernardinelli L et al (1995) Bayesian analysis of space-time variation in disease risk. Stat Med 14:2433–2443

Besag J, York J, Mollié A (1991) Bayesian image restoration with two applications in spatial statistics. Ann Inst Stat Math 43:1–59

Bivand RS, Gómez-Rubio V, Rue H (2014) Approximate bayesian inference for spatial econometrics models. Spatial Statistics 9:146–165

Chen D, Moulin B, Wu J (2015) Sntroduction to analyzing and modeling spatial and temporal dynamics of infectious diseases. In: Chen D, Moulin B, Wu J (eds) Analyzing and modeling spatial and temporal dynamics of infectious diseases. Wiley, Hoboken, NJ, pp 3–17

Clayton D, Kaldor J (1987) Empirical bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43(3):671–681

Congdon P (2010) Bayesian hierarchical method. Tylor & Francis Group, New York

Congdon P (2013) Bayesian spatial statistical modeling. In: Fischer MM, Nijkamp P (eds) Handbook of regional science. Springer, New York, pp 1419–1434

Fong I (2013) Emerging infectious diseases of the 21st century, challenges in infectious diseases. Springer, Toronto

Hog RV, McKean JW, Craig AT (2005) Introduction to mathematical statistics. Pearson Prentice Hall, Upper Saddle River, NJ

Kesall JE, Diggle PJ (1998) Spatial variation in risk of disease: a nonparametric binary regression approach. Appl Stat 47(2):559–573

Lambert DM, Brown JP, Florax RJ (2010) A two-step estimator for a spatial lag model of counts: theory, small sample performance and an application. Reg Sci Urban Econ 40(4):241–252

Lawson AB (2006) Statistical methods methods in spatial epidemiology. Wiley, Chichester

Lawson AB (2013) Bayesian disease mapping, hierarchical modeling in spatial epidemiology, 2nd edn. CRC Press/Taylor & Francis Group, Boca Raton, FL

Lawson AB (2014) Hierarchical modeling in spatial WIREs. Comput Stat. doi:10.1002/wics.1315

Lawson AB, Biggeri B et al (2000) Disease mapping models: an empirical evaluation. Stat Med 19:2217–2241

Lawson AB, Browne WJ, Rodeiro CL (2003) Disease mapping with WinBUGS and MLwiN. Wiley, Chichester

Lee D (2013) CARBayes: an R package for bayesian spatial modeling with conditional autoregressive priors. J Stat Softw 55(13):1–24

Leonard T (1975) Bayesian estimation methods for two-way contingency tables. J R Stat Soc ser B 37:23–37

Leroux B, Lei X, Breslow N (1999) Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: Halloran ME, Berry D (eds) Statistical models in epidemiology, the environment, and clinical trials. Springer, New York, pp 135–178

Lowe R, Bailey TC, Stephenson DB et al (2011) Spatiotemporal modeling of climate-sensitive disease risk: towards an early warning system for dengue in Brazil. Comput Geosci 37:371–381

MacNab YC, Dean C (2002) Spatiotemporal modeling of rates for the construction of disease maps. Stat Med 21:347–358

Maiti T (1998) Hierarchical bayes estimation of mortality rates disease mapping. J Stat Plan Inference 69(2):339–348

Martinez EZ, Achcar AJ (2014) Trends in epidemiology in the 21st century: time to adopt Bayesian methods. Cad Saúde Pública 30(4):703–714

Meza JL (2003) Empirical bayes estimation smoothing of relative risks in disease mapping. J Stat Plan Inference 11:43–62

Pringle D (1996) Mapping disease risk estimates based on small number :an assessment of empirical bayes techniques. Econ Soc Rev 27(4):341–363

Rao J (2003) Small area estimation. Wiley, Ottawa

Rue H, Martino S, Chopin N (2007) Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations. Statistics Report No 1. Norwegian University of Science and Technology

Shaddick G, Zidek JV (2016) Spatiotemporal methods in environmental epidemiology. CRC Press/Taylor & Francis Group, New York

Simonoff JS (1999) Smoothing methods in statistics. Springer, New York

Stern H, Cressie N (1999) Inference for extremes in disease mapping. In: Lawson AB, Biggeri A, Bohning D et al (eds) Disease mapping and risk assessment for public health. Wiley, New York, pp 63–84

Tango T (2010) Statistical methods for disease clustering theory and methods. Springer, London

WHO (2005) Using climate to predict infectious disease epidemics. WHO, Geneva

I Gede Nyoman Mindra Jaya is a lecturer, Department of Statistics, Universitas Padjadjaran, Bandung, Indonesia. His research interests include research methodology, spatial and spatiotemporal econometrics, spatial and spatiotemporal disease mapping, and Bayesian modeling. He is a Ph.D. student, in Faculty of Spatial Science, University of Groningen, The Netherlands since 2013.

Henk Folmer is a professor, Department of Economic Geography, Faculty of Spatial Sciences, Groningen University, The Netherlands, and professor and academic dean, College of Economics and Management. Northwest Agriculture and Forestry University, China. His research interests include research methodology, (spatial) econometrics, environmental and resource economics, life satisfaction and subjective wellbeing. He earned the Ph.D. in economics from the University of Groningen in 1984. He holds an honorary doctorate from the University of Gothenburg, Sweden, and received the Outstanding Foreign Expert Award for Economic and Social Development of the Province of Shaanxi, China, in 2014.

Budi Nurani Ruchjana is a professor, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia. Her research interest are stochastic process, time series analysis, geostatistics, spatiotemporal modeling and its applications. She earned the Ph.D. in Mathematics and Natural Sciences from Institut Teknologi Bandung at 2002 with a concentration in applied statistics. She is a Dean of Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran for period 2012–2016. She is also a President of the Indonesian Mathematical Society (IndoMS) period 2014–2016 and a member of Commission Developing Country International Mathematical Union (CDC IMU) for period 2015–2018.

Farah Kristiani is a lecturer, Department of Mathematics, Parahyangan Catholic University, Bandung, Indonesia. Her primary research interests are applied statistics in dengue disease mapping; Bayesian modeling; and actuarial science in life insurance. She is a Ph.D. student, in Mathematics Department from Sultan Idris Education University, Malaysia since 2013.

Yudhie Andriyana is a lecturer at Statistics Department, Universitas Padjadjaran, Indonesia. He is currently assigned as the head of Master Program in Statistics, Faculty of Mathematics and Natual Sciences Universitas Padjadjaran. His research interest is Nonparametric Regression, especially working on quantile objective function in varying-coefficient models. He earned his Ph.D in Statistics from KU Leuven, Belgium, in 2015.