Spatial and Spatio-temporal Epidemiology 8 (2014) 1–10

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Generalizing the spatial relative risk function

W.T.P. Sarojinie Fernando [a], Martin L. Hazelton [a,b,*]

[a] Institute of Fundamental Sciences, Massey University, New Zealand
[b] Infectious Disease Research Centre, Massey University, New Zealand

ARTICLE INFO

Article history:
Received 17 February 2013
Revised 24 December 2013
Accepted 26 December 2013
Available online 7 January 2014

Keywords:
Bandwidth selection
Covariate
Density ratio
Foot and mouth disease
Kernel density estimation
Spatio-temporal

ABSTRACT

The spatial relative risk function is defined as the ratio of densities describing respectively the spatial distribution of cases and controls. It has proven to be an effective tool for visualizing spatial variation in risk in many epidemiological applications over the past 20 years. We discuss the generalization of this function to spatio-temporal case-control data, and also to situations where there are covariates available that may affect the spatial patterns of disease. We examine estimation of the generalized relative risk functions using kernel smoothing, including asymptotic theory and data-driven bandwidth selection. We also consider construction of tolerance contours. Our methods are illustrated on spatio-temporal data describing the 2001 outbreak of foot-and-mouth disease in the United Kingdom, with farm size as a covariate.

1. Introduction

Suppose that an epidemiologist is presented with a dataset comprising the geographical coordinates of cases and controls for some disease of interest. A first step in handling such data would typically be an exploratory analysis, aimed at getting some overall feel for the spatial variation in disease risk. However, this may prove difficult from a simple scatterplot. Consider, for example, the plots in Fig. 1, which display the locations of cases and controls for cancer of the larynx in Chorley–Ribble region of Lancashire, England, for the period 1974–1983. (See Diggle (1990) for details.) As one would expect, the pattern of cases reflects the spatial distribution of the population at risk (as described by the controls), but it is difficult to say much more from the raw data plot. In particular, assessment of spatial variation of risk (including preliminary identification of possible disease hot spots) is problematic. The epidemiologist would be hard pressed to say whether there are tangible differences in the relative risk of disease between the major population centers, for example.

Arguments of this type motivated Bithell (1990) to propose the spatial relative risk function, which provides a smooth 'map' describing the geographical variation in the relative frequency of cases to controls. It is defined formally as follows. For any location $\mathbf{x}$ in some designated geographical region $\mathcal{R} \subseteq \mathbb{R}^2$ of interest, the spatial relative risk at $\mathbf{x}$ is given by

$$r(\mathbf{x}) = \frac{f(\mathbf{x})}{g(\mathbf{x})}, \tag{1}$$

where f is the bivariate probability density of the geographical coordinates of cases of the disease over $\mathcal{R}$, and g is the density of controls over the same region. In practice $r(\mathbf{x})$ must be estimated from case-control data (Bithell, 1990; Bithell, 1991). Plots of estimates of the relative risk function, or more commonly the log-relative risk function $\rho(\mathbf{x}) = \log\{r(\mathbf{x})\}$, have proven to be a highly effective tool for exploratory data analysis, aiding visualization of spatial patterns of risk. See for example Sabel et al. (2000), Prince et al. (2001), Berke (2005), Wheeler (2007), Sanson et al.

* Corresponding author at: Institute of Fundamental Sciences, Massey University, Private Bag 11222, Palmerston North, New Zealand. Tel.: +64 6 3569099x84642; fax: +64 6 3505682.
E-mail address: m.hazelton@massey.ac.nz (M.L. Hazelton).

1877-5845/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.sste.2013.12.002

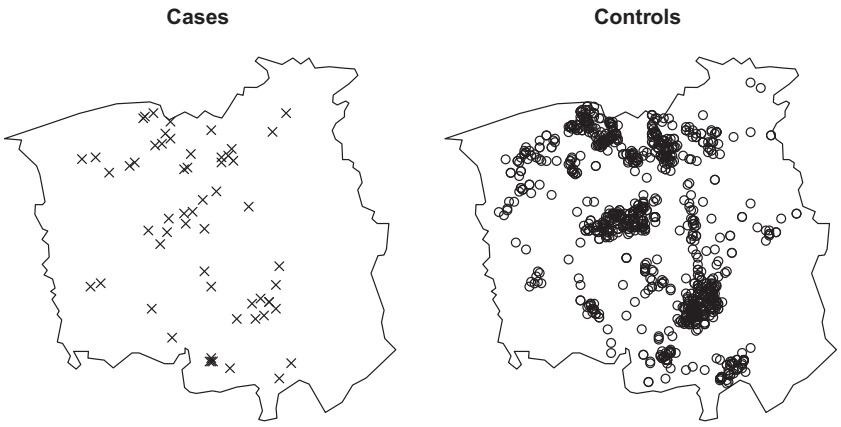

Fig. 1. Locations of cases of cancer of the larynx (left panel) and controls (right panel) in south Lancashire, England.

(2011) and Zhang et al. (2012) for applications in both human and veterinary epidemiology.

Estimation of $r(\mathbf{x})$ requires only information on the geographical coordinates of sets of disease cases and controls. However, it is frequently the case that such individual level disease data are indexed by time of occurrence (or at least detection), and accompanied by covariates. When this is the case there is the potential to examine the way in which the spatial pattern of disease varies through time, or changes with other variables. A simple methodology is to compute $r(\mathbf{x})$ from subsets of case-control data stratified by time or covariate of interest. For example, Benschop et al. (2008) investigated the spatio–temporal pattern of Salmonella infection in Danish finisher pig herds by looking at a time-sequence of estimates of $r(\mathbf{x})$ each computed from (large) yearly datasets. However, this type of approach involves unnecessary aggregation of the data, and fails to take advantage of potential stabilization of estimates that can be obtained by smoothing over time or covariate space.

With these comments in mind, we seek to extend the spatial relative risk function to produce tools to help epidemiologists visualize the manner in which geographical patterns of risk change through time, or vary with covariates. We define the *generalized spatial relative risk function* by

$$r(\mathbf{x},\mathbf{z}) = \frac{f(\mathbf{x},\mathbf{z})}{g(\mathbf{x},\mathbf{z})}, \tag{2}$$

where $f(\mathbf{x},\mathbf{z})$ and $g(\mathbf{x},\mathbf{z})$ denote respectively for cases and controls the joint distribution of event coordinates $\mathbf{x} = (x_1, x_2)^{\mathsf{T}}$ and a vector of covariate values $\mathbf{z} = (z_1, \ldots, z_p)^{\mathsf{T}} \in \mathcal{Z}$. We write $\rho(\mathbf{x},\mathbf{z}) = \log\{r(\mathbf{x},\mathbf{z})\}$ for the log-generalized spatial relative risk function. In an analogous manner (indeed, in essence as a special case) we define the *spatio-temporal relative risk function* by

$$r(\mathbf{x},t) = \frac{f(\mathbf{x},t)}{g(\mathbf{x},t)}, \tag{3}$$

where $t \in \mathcal{T}$ is the time of occurrence of the event at location $\mathbf{x}$. As before, $\rho(\mathbf{x},t) = \log\{r(\mathbf{x},t)\}$. In many situations we would expect the control density to remain unchanged through time, in which case (3) becomes

$$r(\mathbf{x},t) = \frac{|\mathcal{T}| f(\mathbf{x},t)}{g(\mathbf{x})}, \tag{4}$$

where $|\mathcal{T}|$ denotes the length of the time period under consideration. We note that a particular implementation of (3) was recently examined by Zhang et al., 2011, although these authors did not study (4).

An alternative generalization of (1) is provided by the *conditional spatial relative risk function*, defined by

$$r(\mathbf{x}|\mathbf{z}) = \frac{f(\mathbf{x}|\mathbf{z})}{g(\mathbf{x}|\mathbf{z})}, \tag{5}$$

where $f(\mathbf{x}|\mathbf{z})$ and $g(\mathbf{x}|\mathbf{z})$ denote, respectively, the conditional density of event location given covariate vector $\mathbf{z}$, for cases and controls. The conditional spatio-temporal relative risk function is defined analogously:

$$r(\mathbf{x}|t) = \frac{f(\mathbf{x}|t)}{g(\mathbf{x}|t)}. \tag{6}$$

If the control density does not change through time then (6) becomes

$$r(\mathbf{x}|t) = \frac{f(\mathbf{x}|t)}{g(\mathbf{x})}.$$

The functions $r(\mathbf{x}|\mathbf{z})$ and $r(\mathbf{x}|t)$ are respectively simply rescaled versions of the generalized spatial and spatio-temporal relative risk functions. For the former function,

$$r(\mathbf{x}|\mathbf{z}) = \frac{f(\mathbf{x},\mathbf{z})g(\mathbf{z})}{g(\mathbf{x},\mathbf{z})f(\mathbf{z})} = r(\mathbf{x},\mathbf{z})\frac{g(\mathbf{z})}{f(\mathbf{z})},$$

where $f(\mathbf{z})$ and $g(\mathbf{z})$ are the marginal densities of the covariate vector for cases and controls respectively. We can think of the difference between $r(\mathbf{x},\mathbf{z})$ and $r(\mathbf{x}|\mathbf{z})$ in terms of normalization. Specifically, $r(\mathbf{x},\mathbf{z})$ is normalized by

$$\int_{\mathcal{R}}\int_{\mathcal{Z}} r(\mathbf{x},\mathbf{z})g(\mathbf{x},\mathbf{z})\,d\mathbf{x}d\mathbf{z} = 1,$$

so that the mean risk is one when averaging over space and covariates. On the other hand,