Case-Cohort Spatial Estimator

Suzanne Dufault, Nick Jewell

March 2021

Basic idea: use cylinders of radius r and height k around each point in SPP to determine "exposure" status. Use to compute case-cohort relative risk.

Contents

1 Motivation

The sampling mechanism of a test-negative design is often considered a variant of case-control sampling. In particular, there are many shared attributes with case-cohort sampling.

To measure spatial-temporal relative risk, we can consider our n_D VCDs to be a sample of cases. The m test-negatives are a sample of the population at risk. Just as in case-cohort sampling, the exposure status of the VCDs and the test-negatives can then be determined and the results can be displayed as in Table 1.

	VCD (case)	Test-negative (control)
E	$n_D P(E\|D)$	$mP(E)$
$\bar{E}$	$n_D P(\bar{E}\|D)$	$mP(\bar{E})$
	n_D	m

Table 1: Probabilities of cell entries in a 2×2 contingency table for data generated by case-cohort sampling. Source: *Statistics for Epidemiology*

where

$$RR = \frac{P(E|D)P(\bar{E})}{P(E)P(\bar{E}|D)} = \frac{P(D|E)}{P(D|\bar{E})} \tag{1}$$

1.1 Disease Definition

Defining disease for this analysis is a somewhat straightforward task.

At a general level, any VCD will be considered a case and any test-negative will be considered a control.

To incorporate more information on the likely transmission of DENV, we can perform the analysis by serotype of DENV. Specifically, for each of the four serotypes of DENV, we can perform the analysis using *only* the serotype of interest as the cases against the test-negatives as controls.

- **Should the other serotypes be kept in the "cohort"?**

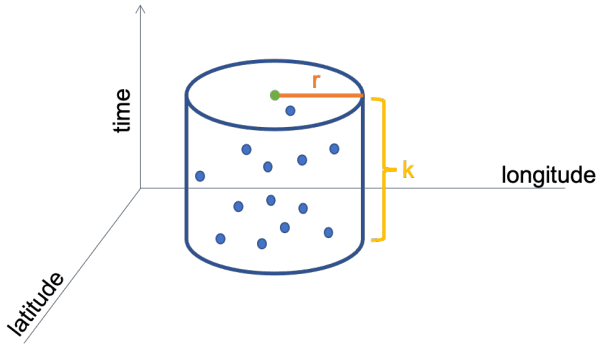
1.2 Exposure Definition

We will define exposure for an individual based on whether a VCD (or serotype-specific VCD) has occurred within close proximity of the individual on 1) space, and 2) time. An individual i that is enrolled at time t_i is **exposed** if a VCD has occurred within the time period $[t_i - k, t_i]$ **AND** has household geocoordinates within r meters of individual i's household. If individual i meets both of these conditions, they are considered exposed. If individual i does not meet both of these conditions, they are considered unexposed.

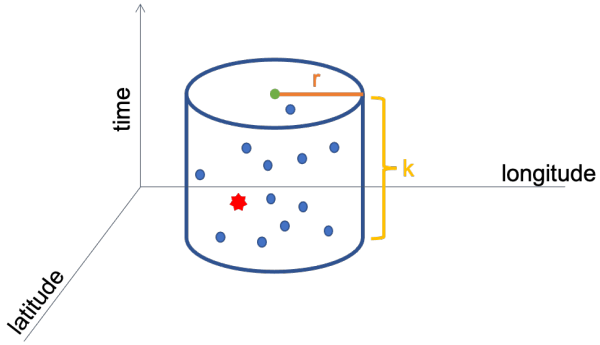
Therefore, exposure is determined in terms of k days and r meters from a (serotype-specific) VCD.

1.2.1 Illustration

For an individual i' (green), we find all individuals with illness onset within k days prior to the illness onset of individual i' and with households within r meters of individual i' (blue).



Individual i''s disease status (VCD/DENV-serotype or test-negative) is already known. We need to determine the exposure status of individual i' in order to complete the contingency table (Table 1).



Individual i' is considered exposed if any of the individuals within close spatial and temporal proximity are VCDs/DENV-serotypes of interest (red star). Otherwise, they are considered unexposed.

1.2.2 Concerns

As with many spatial methods, we are left with attempting to discern an appropriate radius. However, if we simply let $r \to \infty$, then $P(\bar{E}|D) \to 0, P(\bar{E}) \to 0, P(E) \to 1$

2 Notation

(Modeled after Diggle's general notation)

Let $(x_i, t_i, d_i) : i = 1, \ldots, n$ be orderly spatio-temporal point process data where (x_i, t_i, d_i) identifies the household geolocation, time of illness onset, and serostatus for each individual enrolled in the AWED study. Note: $t \in [0, T]$, $x \in$ spatial area A, and $d = \{0, 1\}$ where $d = 1$ denotes a VCD and $d = 0$ denotes a test-negative. This can be extended for serotype-specific DENV. Therefore $(x_i, t_i, d_i) \in A \times [0, T] \times \{0, 1\}$.

Individual i' is determined to be exposed if there exists i for any $i \neq i'$, that meets all of the following conditions:

- $x_i' - x_i \leq r$
- $t_i' - t < k$
- $d_i = 1$

The temporal and spatial windows can be adjusted to examine how the relative risk changes at varying scales.

- **Edge effects?**

3 Preliminary Results

3.1 Overall

The first batch of results uses VCDs and test-negatives from the entire study area and performs two analyses:

1. All dengue versus test-negatives (Fig. 1)
2. Each dengue serotype versus *only* the test-negatives (Fig. 2)

3.2 By Intervention Arm

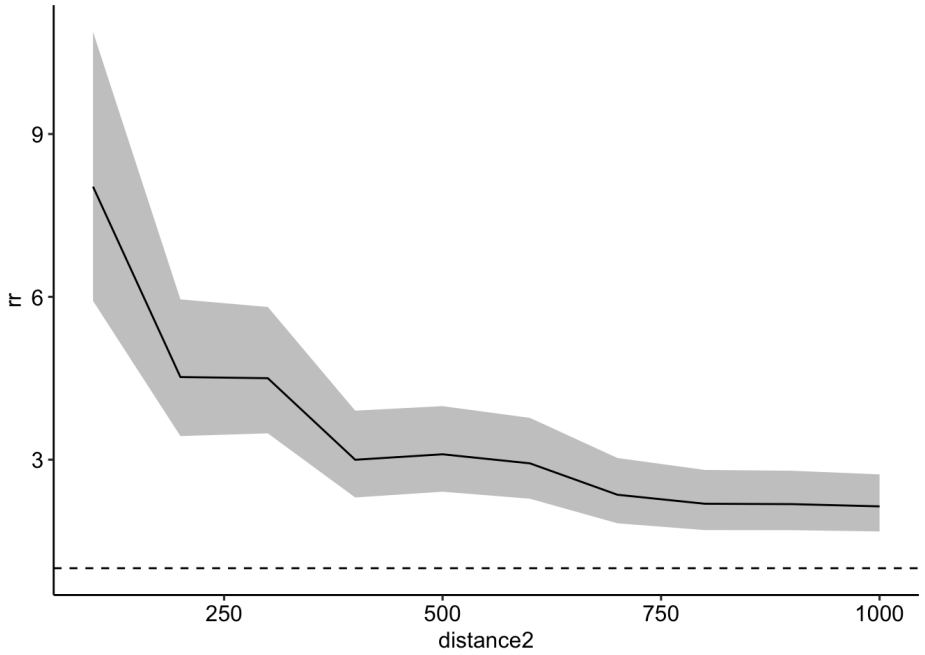


Figure 1: **PRELIMINARY.** Estimating the "relative risk" when $k = 30$ days and r allowed to range in consecutive 100 m bands. This considers the entire study area and uses all dengue versus the test-negatives.

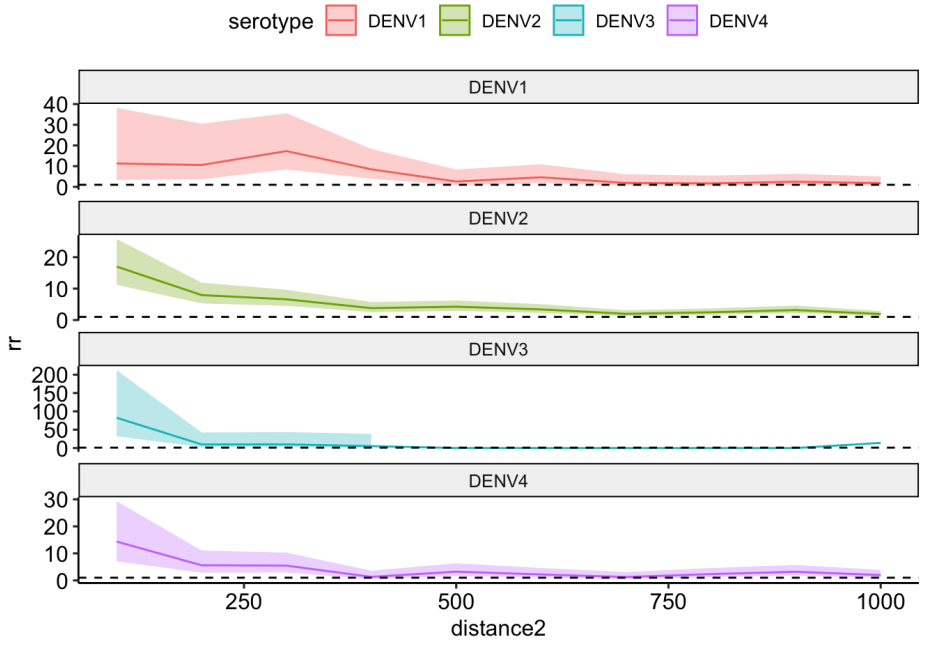


Figure 2: **PRELIMINARY.** Estimating the "relative risk" when $k = 30$ days and r allowed to range in consecutive 100 m bands. This considers the entire study area and uses serotype-specific dengue (excluding individuals with alternative serotypes) and the test-negatives.

A Pseudo-code

To optimize code, note that we only need to know a and b of Table 1, since n_D and m are fixed. Let $N = n_D + m$ Since there are far fewer VCDs than test-negatives, and VCDs are necessary in determining the source of exposure (whereas test-negatives alone provide no information as to the exposure status of another point in the point process):

1. Order participants such that $t_1 < t_2 < \ldots < t_N$ where t_i is the associated time of illness onset of the ith event
2. Build a distance matrix, $\mathbb{A}_{N\times N}$, of pairwise distances. Note: this matrix is symmetric.
3. Build a time matrix, $\mathbb{B}_{N\times N}$, of pairwise differences in illness onset time (scale = days). Note:
4. For given time horizon, k,
5. Identify each VCD and build cylinders forward in time $(t_i + k)$ to determine which unique individuals, e_i, are exposed by each VCD.
6. Identify the disease status (d_i) for each individual in e_i.
7. $a =\mid e_i \mathbb{I}\{d_i = 1\} \mid$, (the length of the set that fits the conditions)
8. $b =\mid e_i \mathbb{I}\{d_i = 0\} \mid$
9. $c = n_D - a$
10. $d = m - b$
11. $RR = ad/bc$
12. $\text{var}(\log RR) = 1/a + 1/b + 1/c + 1/d$
    - given rare disease setting

B Background Reading

B.1 Use of Cylinders in Spatio-Temporal Literature

B.1.1 For estimating intensity functions and conditional intensity functions

Diggle Chapter from Nick

> Firstly, let C be a cylinder centred on the point (x, t), with radius u and height v. The intensity function of the process, denoted by $\lambda(x, t)$, is the limit of the ratio
>
> $$\lambda(x,t) = \lim_{u\to 0, v\to 0} \frac{E[N(C)]}{\pi u^2 v}$$
>
> as $u \to 0$ and $v \to 0$. Informally, $\lambda(x, t)$ is the expected number of events per unit volume in the immediate neighbourhood of the point (x, t).
>
> Empirical counterparts of the various intensity functions defined above can be calculated by, in essence, replacing expectations by observed counts. For example, an estimate of a a spatio-temporal intensity can be calculated by a direct analogy with the defining equation (1.1), replacing the expectation in the numerator by the observed number of events in a cylinder of radius u and height v, as the centre of its base, (x, t), moves over the region of interest. Second-order functions can be estimated in a similar fashion by considering a moving pair of cylinders. In practice, this approach needs considerable refinement if it is to produce useful results, taking into account

issues such as the choices for u and v, the treatment of edge-effects and, in the non-stationary case, the inherent ambiguity between first-order and second-order effects. A good example of this is the careful discussion in Baddeley, Møller and Waagepetersen (2000, [5]) concerning the joint estimation of first-order and second-order intensities of an intensity-reweighted stationary spatial point process.

On conditional intensities:

Note that the events $(x_i, t_i) : i = 1, 2, \ldots$ of an orderly spatio-temporal point process can be labelled unambiguously in order of increasing event-times t_i. The history of the process at time t is the set $H_t = \{(x_j, t_j) : j < t\}$. In Section 1.2, equation (1.1) we defined the intensity of a spatio-temporal process as the limit of the ratio $E[N(C)]/(\pi u^2 v)$ as u and v both tend to zero, where C is a cylinder of radius u and height v, and $N(C)$ is the number of events in C. The conditional intensity function of the process, $\lambda(x, t|H_t)$, is similarly defined, but replaces the unconditional expectation of $N(C)$ in the numerator of (1.1) by its conditional expectation given H_t. An important theoretical result is that any orderly process is completely specified by its conditional intensity function. Also, defining a model by specifying its current properties conditional on its past is a natural approach when the scientific objective is to find a causal explanation for the behaviour of the process being studied.

B.1.2 Space-time scan statistics

Ansari, et al. 2020

Space–time scan statistics (Kulldorff 1997, 2018) are employed to identify statistically significant hotspots in spatiotemporal data. The cylinder is used to scan space–time to identify candidate hotspots and then hypothesis testing is performed. For each candidate hotspot, the log-likelihood ratio is calculated and the highest likelihood ratio is evaluated using a significance value. The spatiotemporal hotspots have many diverse application domains, ranging from public health to criminology (Shekhar et al. 2015).

Shi, et al. 2019

Space–time scan statistics was extended from space scan statistics to detect clusters with the highest likelihood ratio by moving a cylinder as a scan window to scan ST data [43,44].

...

Space–time scan statistics considers the time dimension and is an extension of space scan statistics in that a three-dimensional cylinder instead of a two-dimensional circle is used. The time interval between events is the height of cylinder.

As with the space scan statistic, the null hypothesis is that the spatiotemporal distribution of events is random. The scan window of the cylinder was changed with different radii and height, looking for the maximum value of log likelihood ratio of all the circles as the cluster region. The formulation was:

$$S = \log\left(\frac{n_z}{u_z}\right)^{n_z}\left(\frac{N-n_z}{N-u_z}\right) I\left(\frac{n_z}{u_z} > \frac{N-n_z}{N-u_z}\right)$$

where S was the log likelihood of cylinder, n_z and u_z were the observed and expected number of points, respectively, N was the total number of observed points, and I was the indicator function. If the left side was larger than right side, I was equal to 1, otherwise equal to 0. Many distribution functions could be used, one of which was the Poisson distribution. To obtain the simulated distribution for significance testing of clusters, Monte Carlo replications of data were used to obtain likelihood ratio statistics S. It was necessary to obtain p values by generating replications such as 999 or even higher to calculate the probability of a random appearance of an observed high-density cluster in a cylindrical window.