Chapter 6

Spatial Clusters of Health Events: Point Data for Cases and Controls

Common question related to disease clusters:

- Do cases tend to occur near other cases (perhaps suggesting an infectious agent)?
- Do parts of the study area contain a significant excess of observed cases (perhaps suggesting an environmental risk factor)?
- Where are the most unusual collections of cases (the most likely clusters)?

The question, "Are disease cases clustered?" will be answered via hypothesis testing via the conceptual null hypothesis:

H_0: There are no clusters of cases in the study area

6.1 What do we have? Data types and related issues

Disease cluster case data is usually one of two types:

Case-control point data involve point locations for a set of reported cases and a collection of non-cases (**controls**).

Regional count data provide reported counts of incidents (newly diagnosed) or prevalent (existing) cases residing in regions partitioning the study area (e.g., counties in a state).

For case-control data, the control locations provide background information on spatial patterns of the population at risk.

- The controls are often assumed to be an independent random sample from subjects free of the disease of interest.
- We compare the patterns of the control locations with the patterns of the case locations.

Limited availability of point data for non-diseased persons often necessitates other definitions of controls.

- Sometimes a second disease (or set of diseases) is used as a control.
  - It is assumed the etiology (cause of the disease) is different from the disease under study.
  - E.g., comparing the association of pediatric asthma and traffic flow for pediatric asthma cases among children using California's Medicaid program and controls sampled from the nonrespiratory visits recorded in the same system.
- The choice of controls refines the questions that can be answered by data.

6.2 What do we want? Null and alternative hypotheses

Each test of the null hypothesis that there are no clusters of cases will depend on a set of assumptions, each determining whether observed patterns provide evidence for or against a conclusion of clustering.

CSR is not a satisfactory null hypothesis if the population at risk is distributed heterogeneously across the study area.

- Since this is true of most health applications, standard CSR-based hypothesis tests are inappropriate.

We will use heterogeneous Poisson processes to implement tests.

Consider spatially-varying functions defining the spatial pattern of disease:

- The **disease risk at location *s*, risk(*s*)**, is the probability a person at location *s* contracts the disease of interest (within a specified time interval).
- The **disease rate at location *s*, rate(*s*)**, is the proportion of people at location *s* contracting the disease of interest (within a specified time interval).
- The **relative disease risk at location *s*, relative risk(*s*)**, is the multiplicative increase in disease risk at location *s* compared to the overall disease risk observed.

Risks are unknown and unobserved quantities specific to individuals, often estimated by rates (proportions) observed over groups of people

The **rate ratio** is an estimate of relative risk:

$$\frac{\dfrac{\text{number of incident cases at } s}{\text{number at risk at } s}}{\dfrac{\text{total number of cases}}{\text{total number at risk}}} = \frac{\dfrac{\text{number of incident cases at } s}{\text{total number of cases}}}{\dfrac{\text{number at risk at } s}{\text{total number at risk}}}$$

Once we obtain estimates or summaries of spatially varying risks, rates, or relative risks, we must quantify how a particular estimate varies from what would be expected under the null hypothesis of no clustering of cases.

Two common methods for operationalizing the null hypothesis assuming a heterogeneous Poisson process:

1. The random labeling hypothesis (for point case-control data)
2. The constant risk hypothesis (for regional count data)

The random labeling hypothesis assumes that case and control event locations arise from the same underlying spatial point process.

- The random labeling hypothesis assumes a constant probability of case-control assignment at all locations.
- The random labeling hypothesis is not sensitive to the estimated/assumed background risk, only to the frequency of cases and controls observed.

The constant risk hypothesis (CRH) assumes cases reflect a random sample of the at-risk population where the probability of selection is the same everywhere.

- The CRH assumes a known (or estimated) background risk.
- A test of the CRH is sensitive to how the background risk is estimated (i.e., from an external source or the observed data).

Consider three scenarios for testing the null hypothesis of no clustering among N events, with N_1 case events and N_0 control events.

S1: The random labeling hypothesis addresses the question, "Are the N_1 case locations observed consistent with a random assignment of N_1 events among the N event locations?"

S2: The constant risk hypothesis (with a fixed number of cases) addresses the same question as S1.

S3: The constant risk hypothesis (with a random number of cases) addresses the question, "Are the case locations observed consistent with each of the N locations observed having probability N_1/N of being a case?"

The distinction between S2 and S3 is subtle. It is something like:

S2: Is there evidence of clustering of 592 leukemia cases among 1,057,673 persons at risk?

S3: Is there evidence of clustering of leukemia cases among 1,057,673 persons at risk where each person has a probability of 592/1,057,673 of contracting leukemia in the time interval under study?

We must distinguish between the notions of **clusters** and **clustering**.

- A **cluster** is a collection of cases inconsistent with our null hypothesis of no clusters.
- **Clustering** occurs when overall, the cases tend to cluster together.

A cluster represents an anomaly of cases in the data.

- Typically utilizes multiple tests (multiple p-values) to determine which collection of cases represents the most significant cluster.

Clustering represents a pattern among all or most cases.

- Typically relies on a single assessment of statistical significance of the pattern for the entire area (e.g., a single p-value).

A further distinction in our tests is to distinguish between **general tests and focused tests**.

- A general test tests for clusters/clustering anywhere in the study area.
- A focused test tests for clusters/clustering around predefined foci.

6.3 Categorization of Methods

We will discuss two broad classes of methods for testing whether there are clusters/is clustering in our cases:

1. Methods using first- and second-order summaries of the point process (intensity and K functions).
2. Methods based on scanning local rate estimates.

6.4 Comparing Point Process Summaries

6.4.1 Goals

A comparison of first-order properties for cases and controls is most useful for identifying clusters of cases.

- Are there peaks or valleys of case incidence above and beyond the controls?

A comparison of second-order properties for cases and controls is most useful for identifying clustering of cases.

- The K function summarizes clustering tendency across all events rather than identifying particular collections of events as clusters.

6.4.2 Assumptions and Typical Output

We assume that the case and control locations each represent a realization of two (distinct) heterogeneous Poisson point processes over the same study area.

- Each set of case and control event locations occur independently of other events in the same process, with spatial variation in incidence summarized by the intensity function of the appropriate process.

We will build our inference off the random labeling hypothesis, so we will answer the question, "Is the labeling of event locations observed as cases and controls consistent with a random assignment of labels to this set of locations observed?"

We assume the events within a realization are independent of each other (so they are inappropriate to model infectious outcomes).

We (typically) assume our processes are isotropic, and when estimating the K function, that our processes are stationary.

Because our inference will rely on the random labeling hypothesis, our inference is conditional on the observed set of case and control locations.

6.4.3 Method: Ratio of Kernel Intensity Estimates

Consider observing N_1 case event locations and N_0 control event locations ($N = N_0 + N_1$) distributed throughout the study area.

Define $\lambda_0(s)$ and $\lambda_1(s)$ to be the intensities associated with the controls and cases, respectively.

To assess evidence of spatial clusters, we will compare $\lambda_0(s)$ and $\lambda_1(s)$.

The density function of the control locations is

$$g(s) = \frac{\lambda_0(s)}{\int_D \lambda_0(u)du}.$$

The density function of the case locations is

$$f(s) = \frac{\lambda_1(s)}{\int_D \lambda_1(u)du}.$$

These are just scaled versions of the intensity functions.

We can make inference using

$$r(s) = \log\left\{\frac{f(s)}{g(s)}\right\} = \log\left\{\frac{\lambda_1(s)}{\lambda_0(s)}\right\} - \log\left\{\frac{\int_D \lambda_1(u)du}{\int_D \lambda_0(u)du}\right\}.$$

This is the logarithm of the **relative risk** of observing a case rather than a control at location s.

If the ratio of $\lambda_1(s)$ and $\lambda_0(s)$ is constant for all s, then $r(s) = 0$.

- If $r(s) > 0$, then there is evidence of a cluster of cases at s.
- The locations with the largest values of $r(s)$ have the highest evidence of a cluster of cases.

We estimate $r(s)$ by

$$\tilde{r}_b(s) = \log\{\widetilde{f}_b(s)/\tilde{g}_b(s)\},$$

where $\tilde{f}_b(s)$ and $\tilde{g}_b(s)$ are estimates of $f(s)$ and $g(s)$, respectively, using bandwidth b.

- In practice, we estimate r on a grid of locations, $s_1, \ldots, s_m$, throughout the study area.
- We will drop the dependency of $\tilde{r}_b(s)$ on b for simplicity (i.e., just call it $\tilde{r}(s)$), even though b is important.

Notes:

- The choice of bandwidth is more important than the choice of the kernel function.
- The bandwidth must be large enough that you don't get 0's in the numerator or denominator.
  - A numerator of zero means there is no risk of getting the disease. We assume this is impossible.
  - A denominator of zero means "dividing by zero", which isn't mathematically possible.
- Make sure that any spikes in $\tilde{r}(s)$ are real, and not just because the denominator is too small. Double check for odd behavior.
- The Gaussian kernel will TEND to have fewer problems since it has infinite support (never returns a zero) but results still need to be double-checked.

- A good bandwidth for estimating $f(s)$ and $g(s)$ may not be a good choice for estimating $r(s)$.
- It is a good idea to use a common bandwidth when $r(s) \approx 0$.
- There are many "automatic" methods for choosing a "good" bandwidth (see the book for details).

A contour plot of $\tilde{r}(s)$ provides a map of areas where cases are more or less likely than controls (with b denoting the selected bandwidth).

We can use the estimated log relative risk to identify local clusters of cases or controls.

- Simulate N_{sim} "null" data sets under the random labeling hypothesis
- Create a grid of locations $\mathcal{S} = \{s_1, \ldots, s_m\}$.
- For simulated null data set i, estimate $r(s_k)$ for each grid point. Denote this $\tilde{r}^{(i)}(s_k)$.
- At grid point s_k, a one-sided Monte Carlo p-value to identify locations where the cases are more clustered than the controls at location s is

$$\frac{1 + \#\left\{\tilde{r}^{(i)}(s_k) \geq \tilde{r}(s_k), i = 1,2, \ldots, N_{sim}\right\}}{N_{sim} + 1}.$$

  - A significant p-value indicates evidence of a local cluster of cases relative to controls at location s_k.

Alternatively, one can construct Monte Carlo envelopes to identify cluster locations at each grid point:

- e.g., compute the $\alpha/2$ and $1-\alpha/2$ quantiles of
$$\left\{\tilde{r}^{(1)}(s_k), \tilde{r}^{(2)}(s_k), \ldots, \tilde{r}^{(N_{sim})}(s_k), \tilde{r}(s_k)\right\}.$$
- If $\tilde{r}(s_k)$ is larger than the $1-\alpha/2$ quantile, then the cases are clustered relative to the controls at locations s_k [beyond what is expected under the random labeling hypothesis] at a significance level of α.
- If $\tilde{r}(s_k)$ is smaller than the $\alpha/2$ quantile, then the controls are clustered relative to the cases at locations s_k [beyond what is expected under the random labeling hypothesis] at a significance level of α.
- You can construct one-sided limits as well.

The locations where $\tilde{r}(s)$ exceeds the upper envelope are locations where there is a cluster of cases relative to controls.

- i.e., the relative proportion of cases to controls is greater than what we expect under the random labeling hypothesis.

The locations where $\tilde{r}(s)$ falls below the lower envelope are locations where there is a cluster of controls relative to cases.

- i.e., the relative proportion of controls to cases is greater than what we expect under the random labeling hypothesis.

Kelsall and Diggle (1995) suggest a global test of whether the clustering patterns are the same for the cases/controls.

A global test of

$$H_0: r(s) = 0 \text{ for all } s \text{ in } D,$$

can be used to assesses whether the intensities of cases and controls are always in the same relative proportion.

The alternative hypothesis is

$$H_a: r(s) \neq 0 \text{ for at least one } s \text{ in } D.$$

Kelsall and Diggle (1995) suggest using the statistic

$$T = \int_D \{\tilde{r}(u)\}^2 \, du \,,$$

which averages the squared $\tilde{r}(s)$ surface across the study area.

Failing to reject the null hypothesis is the same as failing to reject the null hypothesis that the case and control densities are equal for all locations in the study area.

- Alternatively, that the log relative risk of cases relative to controls equals zero for all locations in the study area.

Concluding the alternative means that the spatial densities of the cases and controls differ for at least one location in the study area.

- Alternatively, that the log relative risk of cases relative to controls differs from zero for at least one location in the study area.

A Monte Carlo test can implemented by simulating N_{sim} data sets under the random labeling hypothesis, computing the associated test statistics, $T^{(1)}, \ldots, T^{(N_{sim})}$, and computing the Monte Carlo p-value

$$\frac{1 + \#\left\{T^{(i)} \geq T, i = 1, \ldots, N_{sim}\right\}}{N_{sim} + 1}.$$

6.4.4 Method: Difference Between K Functions

A way of comparing the second-order properties of the case and control event locations is by looking at the difference in the K functions of the processes, i.e., considering

$$KD(h) = K_{case}(h) - K_{control}(h).$$

We estimate $KD(h)$ by $\widehat{KD}(h)$, replacing the K functions with their (edge-corrected) estimates.

Under the random labeling hypothesis, the expected value of $KD(h)$ is zero for any distance h.

- Positive values of $KD(h)$ suggest spatial clustering of cases over and above any clustering observed in the controls.
- Distances for which $KD(h)$ exceeds zero prove insight into the spatial scale of any clustering observed.
- The locations of the clustering cannot be determined from $KD(h)$.

We can easily construct Monte Carlo pointwise interval estimates of $KD(h)$ under the random labeling hypothesis.

- Generate realizations of the "null" data using random labeling.
- Calculate $\widehat{KD}(h)$ for each realization.
- Calculate envelopes based on percentiles of $\widehat{KD}(h)$ from the simulated data (or the minimum and maximum values).

If $\widehat{KD}(h)$ is above the envelopes for certain ranges of h, there is evidence of clustering of cases relative to controls at that spatial scale.

If $\widehat{KD}(h)$ is below the envelopes for certain ranges of h, there is evidence of clustering of controls relative to cases at that spatial scale.

Suppose we wish to construct a global test for whether the relative clustering behavior for cases is greater than the clustering behavior of controls over a range of distances.

More formally, our null hypothesis is

$H_0: KD(h) = 0$ for all $h \in [0, h^*]$.
$H_a: KD(h) > 0$ for at least one $h \in [0, h^*]$.

Diggle and Chetwynd (1991) suggest the test statistic

$$KD_+ = \sum_{k=1}^{m} \widehat{KD}(h_k) / \sqrt{\text{Var}[\widehat{KD}(h_k)]},$$

where the h_k are the distances for which we estimate $KD(h)$.

Next, calculate KD_+ for each of the N_{sim} random labelings.

- Denote these $KD_+^{(1)}, \ldots, KD_+^{(N_{sim})}$.

The Monte Carlo p-value for this test of clustering of cases relative to controls is the proportion of simulated and observed KD_+ at least as large as the observed KD_+, i.e.,

$$\frac{1 + \#\left\{KD_+^{(i)} \geq KD_+, i = 1,2, \ldots N_{sim}\right\}}{N_{sim} + 1}.$$

If we fail to reject H_0, then there is no evidence of clustering of cases above and beyond the controls at spatial scales between 0 and h^*.

If we conclude H_a, then there is evidence of clustering of cases above and beyond the controls at some spatial scale between 0 and h^*.

6.5 Scanning Local Rates

We now consider an approach for identifying clusters of cases for case-control point data by scanning local rate estimates (essentially, the proportion of events that are cases in some window).

- We identify a cluster of cases if this proportion is significantly higher than what we expect under the random labeling hypothesis or constant risk hypothesis.

6.5.1 Goals

The primary goal of comparisons of local rates (case/control ratios) is to determine areas where the observed rate appears inconsistent with the rate observed over the rest of the study area.

These approaches primarily detect **clusters** rather than clustering.

6.5.2 Assumptions and Typical Output

Methods using scans of local rates or case/control ratios typically seek to find the most unusual aggregation of cases (i.e., most likely clusters).

These methods build off standard GIS operations like calculating distances from a point and counting case and control events occurring within a specified polygon or circle.

Tests based on local rates or case/control ratios often condition on the set of all locations and operationalize the null hypothesis of no clustering through a random labeling hypothesis or a constant risk hypothesis.

6.5.3 Method: Geographical Analysis Machine

6.5.4 Method: Overlapping Local Case Proportions

Both of these methods essentially look at case/control ratios within circles on a set of gridded locations in the spatial domain.

If this ratio is unusual in some way, then we plot the associated circle.

The idea is simple and not too difficult to implement, but we won't discuss either of these methods in detail.

6.5.5 Method: Spatial Scan Statistics

A scan statistic involves defining a moving "window" and a statistical comparison of a measurement (e.g., a count or rate) within the window to the same sort of measurement outside the window.

Kulldorf (1997) defined a scan statistic built to find the collection of case(s) least consistent with the null hypothesis of of no clustering.

On a grid of locations in the spatial domain (typically, the set of observed event locations), we consider circular windows of variable radii ranging from 0 to some user-defined upper bound (e.g., one-half the width of the study area).

Let $N_{1,in}$ and $N_{in} = N_{0,in} + N_{1,in}$ denote the number of case locations and persons at risk inside a particular window, respectively.

Define $N_{1,out}$ and $N_{out} = N_{0,out} + N_{1,out}$ denote the number of case locations and persons at risk outside a particular window, respectively.

Let w_i denote window i, $i = 1,2, \ldots, N_{win}$, where N_{win} is the total number of windows considered.

For window i, w_i, compute the following statistic measuring how unusual the proportion of cases in the windows is to the proportion of cases outside the window:

$$T_{w_i} = \left(\frac{N_{1,in}}{N_{in}}\right)^{N_{1,in}} \left(\frac{N_{1,out}}{N_{out}}\right)^{N_{1,out}} I\left(\frac{N_{1,in}}{N_{in}} > \frac{N_{1,out}}{N_{out}}\right),$$

where $I(\cdot)$ is the indicator function that returns the value 1 if the argument is true and 0 otherwise.

- In practice, this statistic is computed on the natural log scale for numerical stability.

To account for multiple comparisons, Kulldorff (1997) suggests the overall test statistic (used to identify clustering):

$$T_{scan} = \max_{\text{all windows}} \left(\frac{N_{1,in}}{N_{in}}\right)^{N_{1,in}} \left(\frac{N_{1,out}}{N_{out}}\right)^{N_{1,out}} I\left(\frac{N_{1,in}}{N_{in}} > \frac{N_{1,out}}{N_{out}}\right)$$
$$= \max_{i=1,\ldots,N_{win}} T_{w_i}.$$

- Practically, we only maximize over windows where the observed rate inside the window exceeds that outside the window.
- This determines which windows are "most likely to be a cluster".

We determine whether there is convincing evidence for clustering using a Monte Carlo test.

- We simulate N_{sim} data sets under the random labeling hypothesis.
- For each simulated data set, compute T_{scan}. Denote these statistics $T^{(1)}, \ldots, T^{(N_{sim})}$.
- The Monte Carlo p-value for the test is the proportion of observed and simulated test statistics at least as large as the observed test statistic, i.e.,

$$\frac{1 + \#\left\{T^{(i)} \geq T_{scan}, i = 1,2, \ldots, N_{sim}\right\}}{N_{sim} + 1}.$$

What is being tested?

H_0: There are no cluster of cases in the study area.
H_a: There is at least one cluster of cases in the study area.

More specifically:
H_0: There are no windows where the most likely cluster is more unusual than what is expected under the random labeling hypothesis.
H_a: There is at least one window where the most likely cluster is more unusual than what is expected under the random labeling hypothesis.

Since each window is determined by a well-defined set of event locations, the window that produces the largest test statistic is also the most likely cluster (MLC).

If we reject H_0, we can identify the MLC of cases!

The spatial scan statistic thus tests for both clustering and can be used to identify clusters.

In fact, we can determine secondary clusters (other windows that seem like clusters of cases but are not the MLC).

Compute p-values of significance for each window, w_i, via the formula

$$\frac{1 + \#\left\{T^{(j)} \geq T_{w_i}, j = 1,2, \ldots, N_{sim}\right\}}{N_{sim} + 1}.$$

The second MLC has the second largest test statistic among all windows that don't overlap the MLC (and is significant).

The third MLC has the third largest test statistic among all windows that don't overlap the MLC and second MLC (and is significant).

6.6 Nearest-Neighbor Statistics

6.6.1 Goals

These methods examine local patterns of cases in the vicinity of other cases.

Evidence for clustering involves observing more cases among the nearest neighbors of cases than one would expect under the random labeling hypothesis.

This method answers the questions, “Are there more cases than one would expect under random labeling in the q locations nearest each case?”

6.6.2 Assumptions and Typical Output

Nearest-neighbor statistics summarize clustering behavior across the study area.

Nearest-neighbor-based tests are generally operationalized under the random labeling hypothesis.

Output from these tests includes an overall p-value summarizing the significance of the clustering compared to patterns expected under random labeling.

6.6.3 Method of q Nearest Neighbors of Cases

We will utilize a test statistic that represents the number of the q nearest neighbors that are also cases.

Let

$$w_{i,j} = \begin{cases} 1 \text{ if location } j \text{ is among the } q \text{ nearest neighbors of location } i \\ 0 \text{ otherwise.} \end{cases}$$

Let $T_q = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{i,j} \delta_i \delta_j$, where $\delta_i = 1$ if event location i is for a case, and 0 otherwise.

T_q simply accumulates the number of times a case is within the q nearest neighbor of another case.

The null hypothesis of this test is that the number of cases within the q nearest neighbors of other cases isn't larger than what is expected under the random labeling hypothesis.

The alternative hypothesis of this test is that the number of cases within the q nearest neighbors of other cases is greater than what is expected under the random labeling hypothesis.

If $T_q^{(1)}, \ldots, T_q^{(N_{sim})}$ are the test statistics for N_{sim} data sets simulated under the random labeling hypothesis, then the Monte Carlo p-value is

$$\frac{1 + \#\left\{T_q^{(i)} \geq T_q, i = 1,2, \ldots, N_{sim}\right\}}{N_{sim} + 1}.$$

Different values of q may generate different results, possibly indicating the scale (in the sense of the number of nearest neighbors, not geographic distance) of any clustering observed.

- Since the test statistics T_{q_1} and T_{q_2} are correlated for $q_1 < q_2$ (since the q_2 nearest neighbors include the nearest q_1 nearest neighbors), we consider the contrasts $T_{q_2} - T_{q_1}$.
- The contrasts are interpreted as excess cases between the q_1 and q_2 nearest neighbors of cases.
- If the p-value for the difference $T_{q_2} - T_{q_1}$ is large, then there is no significant difference in clustering between the q_2 and q_1 nearest neighbors.
    - i.e., the clustering is caused by the q_1 nearest neighbors, even though we see an affect for the q_2 nearest neighbors.