Chapter 7

Regional Count Data

Confidentiality restrictions often lead many official agencies to release data as summary counts for a particular set of **enumeration districts** that partition the study area.

How can we detect clusters and/or clustering in disease incidence data available as counts of cases from a set of geographic regions?

Note:

- **Regions** will refer to the enumeration districts.
- **Area** will refer to the study area or a particular collection of regions.

7.1 What Do We Have and What Do We Want?

There are complications to applying point pattern approaches to regional count data.

1. We cannot observe spatial patterns at a scale smaller than the scale of the available data, so we cannot explore the distribution of counts within a census tract without additional information or assumptions.

2. Aggregated data cause problems related to the **ecological fallacy** and the **modifiable areal unit problem (MAUP)**.
    a. The ecological fallacy occurs when associations between outcomes and potential risk factors observed in groups of people are extrapolated to individuals.
        i. E.g., researchers found that breast cancer mortality rates were significantly higher in countries with high fat consumption (in comparison with low consumption). However, for countries with higher fat consumption and higher breast cancer rates, one cannot be sure that the women who died from breast cancer had high fat intake.
        ii. Simpson's paradox is a type of ecological fallacy.

b. The modified areal unit problem occurs when association changes based on the way the data are grouped.
    i. MAUP is particularly problematic when the data contain small clusters associated with increased local risk.
    ii. If a cluster is spread across two regions, the statistical evidence for a cluster can be diluted across these regions.

3. Regional counts must balance the **small-number problem** with the **spatial scale** of the data.
    a. For rare diseases, if the regions are too small, then any occurrence of the disease might be suspicious.
    b. If the regions are too large, then we lose some local information.
    c. The balance between the two is a trade-off between geographic resolution (where we want many small areas) and that statistical stability of estimates associated with these areas (where we want stable local estimates).

Clustering at smaller scales doesn't imply clustering at larger scales, or vice versa.

7.1.1 Data Structure

Regional count data must include:

- A set of **counts observed** for each region.
- Enough information to determine the **counts expected** for each region.

To do a spatial analysis, we need a map of the region borders or the region centroids.

Statistical inference is based on comparing the observed counts to **counts expected** for each region under some null hypothesis.

Assume the observed counts, $Y_1, Y_2, \ldots, Y_N$, represent a realization of independent counts arising from a heterogeneous Poisson process.

- $Y_j \sim \text{Poisson}(E_j)$, with Y_i and Y_j independent when $i \neq j$.

Assume we observe fixed (nonrandom) population counts for each region, denoted $n_1, n_2, \ldots, n_N$.

The expected counts for each region, $E_1, E_2, \ldots, E_N$, are typically estimated using the population counts under a hypothesis of no clusters/clustering.

7.1.2 Null Hypotheses

Because of the properties of Poisson point processes, we assume event counts from non-overlapping regions follow independent Poisson distributions where the underlying intensity function defines the expected value (and variance).

Notes:

- The Poisson distribution assigns a non-zero probability to observing more cases than at-risk population.
- A Binomial distribution may also be used to model the counts.
- If we condition on the total number of cases in the study area, the counts are a realization from a multinomial distribution.

The **total number of cases** is denoted

$$Y_+ = \sum_{i=1}^{N} Y_i.$$

The **total population** is denoted

$$n_+ = \sum_{i=1}^{N} n_i.$$

Note: conditioning on the total number of cases can subtly change the question addressed.

- E.g., “Is there clustering among leukemia cases in upstate New York?” to “Is there clustering among 592 cases distributed at random to people in upstate New York?”

The constant risk hypothesis (CRH) is typically assessed by comparing the counts observed to their corresponding counts expected based on the global incidence rate (global referring to the study area, not the literal globe).

A common estimate of individual disease risk under the constant risk hypothesis (CRH) is $r = Y_+/n_+$.

Under the CRH with constant disease risk r, $E_i = rn_i$ for $i = 1,2, \dots, N$.

- Adjustments based on additional grouping such as age can be made with simple modifications.

When using Monte Carlo simulation to test a CRH, we typically we can either simulate the observed counts as:

- $Y_1 \sim \text{Poisson}(E_1), Y_2 \sim \text{Poisson}(E_2)$, … and all independent of each other.
- $(Y_1, Y_2, \ldots, Y_N) \sim \text{Multinomial}\left(\frac{n_1}{n_+}, \frac{n_2}{n_+}, \ldots, \frac{n_N}{n_+}\right)$

7.1.3 Alternative Hypotheses

Under a heterogeneous Poisson process, our observed regional counts should appear to be (for $i = 1,2, \dots, N$):

1. Independent.
2. Poisson distributed.
3. Have expectations (and variances) E_i.

These three components represent areas for potential deviation from a heterogeneous Poisson process, and these (individually and collectively) represent alternative hypotheses for many of the statistical tests of disease clustering and clusters proposed for count data.

Examples of alternative hypotheses we can consider:

- The counts are correlated.
- The counts are over- or under-dispersed compared to a Poisson distribution.
- The expected counts do not match the observed counts.

We focus primarily on tests of the first and third bullet.

Clarification regarding correlation as a measure of clustering or as an approach to detect clusters

We assume our heterogeneous spatial Poisson point process generates independent events.

- With an infectious disease, we typically attribute clusters of cases to the direct relationship between cases.
- With a non-infectious disease, we typically attribute clusters of cases to putative (presumed) environmental causes.

Does a correlation-based alternative imply an infectious nature to the disease? Not necessarily.

- Suppose we expect the same number of cases in every region of the study area (we have constant risk and the same population size for each region in the study area).
- We observe three contiguous regions that each contain twice the number of cases expected.
- This deviation could be caused by a **trend** in the underlying means of the Poisson counts or by **correlation** between the counts.
- It is impossible to distinguish between heterogeneously distributed independent events and homogeneously distributed dependent events from a single realization.
- Practically, a method assuming independence identifies the deviating regions as evidence of a trend, while a method assuming no trend identifies the deviation as correlation.

- Both methods may “notice” the deviation, but each summarizes the pattern in a different way.
- The best approach depends on the context.

Assessing deviation from a heterogeneous Poisson process in only one category (correlation, Poisson, or expectation) may be inadequate for detecting clusters or clustering.

- A purely goodness-of-fit approach compares observed to expected counts without regard to location.
- A correlation-based method assesses the similarity of neighbors to each other, but not necessarily from their mean.

7.2 Categorization of Methods

There are numerous statistical strategies for detecting clusters or clustering. Methods based on:

- Scanning local rates.
- Global indices of correlation.
- Local indices of correlation.
- Goodness-of-fit tests.
- Combining goodness-of-fit and indices of spatial correlation.

7.3 Scanning Local Rates

The **geographical analysis machine (GAM)** of Openshaw et al. (1988) provides an exploratory tool for both case-control point data and regional data.

Basic process:

- Construct circles of various distances.
- Count the number of cases and the number of at-risk people within the circle.
- Calculate the local incidence proportion (rate).
- Display those circles with local incidence proportions exceeding some user-specified threshold.

Question: Should all or a fraction of the cases in a region intersecting the edge of a circle be included within the circle?

Answer: Depends on the method.

7.3.1 Goals

The goal for methods based on scanning local rates is to identify areas with unusually high (or low) local incidence proportions (rates).

These methods are generally better for detecting clusters of unusual cases than assessing a general pattern of clustering.

7.3.2 Assumptions

The GAM is a useful exploratory tool but not appropriate for statistical inference.

Inferential methods for local rates assume $Y_1, Y_2, \dots, Y_n$ are independent Poisson counts with expectation rn_i for region i, where r is the risk of disease (and constant in all regions).

Local incidence proportions calculated for overlapping circles will be correlated since they contain many of the same counts.

Monte Carlo simulation can be used to account for this correlation across data sets.

- Rates are correlated within a data set, but independent between them.

One complication is population heterogeneity since local incidence proportions have a different number of people at risk, yielding different expectations and variances for each local incidence proportion.

7.3.3 Method: Overlapping Local Rates

Not implemented in R, so I don't think this is very popular. See the book for details.

7.3.4 Method: Turnbull et al.'s CEPP

To make comparison of local incidence rates more comparable, Turnbull et al. (1990) introduced a **cluster evaluation permutation procedure (CEPP)**.

- The method constructs windows originating from each region's centroid with a fixed number of persons at risk, n^*.
- The test statistic is the maximum number of cases across all windows.
- The significance of the test is assessed by simulating data under the CRH and computing the Monte Carlo p-value.
- Since each circle has the same population size, Turnbull et al. (1990), answer the question, "Is the largest observed count greater than what we would anticipate under the constant risk hypothesis?"

Key details:

- To get n^* persons-at-risk in your window, you may need to add fractions of persons-at-risk from neighboring regions until the number of at-risk individuals in the window is n^*.
- How best to add fractions of the population is an open question.
- The simplest approach is to add fractions of persons from the originating region first, from the second nearest region if you need more population, then from the third nearest region, etc.
    - You include the population of an entire region before you start including population fractions from additional regions.
- Fractions of cases are added to the window in the same way.

- Monte Carlo testing based on the constant risk hypothesis could then be used to assess the pointwise significance of each circle.

H_0: There is no window with n^* persons-at-risk that has significantly more cases than what is expected under the constant risk hypothesis.

H_a: There is at least one window with n^* persons-at-risk that has significantly more cases than what is expected under the constant risk hypothesis.

Let w_i denote window i, $i = 1,2, \ldots, N$, i.e., the window that originates from region i.

The population size of w_i is n^*, where

$$n^* = \sum_{j=1}^{N} p_{i,j} n_j \,,$$

and $p_{i,j}$ indicates the proportion of n_j added to the persons-at-risk in w_i.

The simplest approach for determining $p_{i,j}$ is to add fractions of persons from the originating region first, from the second nearest region if you need more population, then from the third nearest region, etc.

e.g., Consider w_7, the window that originates from region 7. Let's say the next nearest neighbor (in terms of inter-centroid distance) is region 9, then 12.

Let's say that $n_7 = 10{,}000, n_9 = 20{,}000, n_{12} = 5{,}000$.

If $n^* = 5{,}000$, then we only need half of the population from region 7 to have n^* cases in w_7. Thus, $p_{7,7} = 0.5$, while $p_{7,j} = 0$ for $j \in \{1,2, \dots ,6,8, \dots , N\}$.

If $n^* = 32{,}000$, then $p_{7,7} = 1, p_{7,9} = 1$, and $p_{7,12} = 0.4$, while all other $p_{7,j}$ are zero.

The test statistic for w_i is

$$T_{w_i} = \sum_{j=1}^{n} p_{i,j} Y_j \,,$$

where the $p_{i,j}$ are identical to the ones described above.

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

$$T_{cepp} \quad = \max_{i=1,\ldots,N} T_{w_i}$$

Simulate N_{sim} data sets under the CRH and compute T_{cepp} for each simulated data set. Denote these statistics $T^{(1)}, \dots, T^{(N_{sim})}$.

The Monte Carlo p-value for the test is

$$\frac{1 + \#\{T^{(j)} \geq T_{cepp}, j = 1,2, \dots, N_{sim}\}}{N_{sim} + 1}.$$

The window that produces largest test statistic is the most likely cluster (MLC).

To identify potential secondary clusters, we compute p-values of significance for each window, w_i, via the formula

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

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).

The clusters are well-defined except that the last region added to the cluster is probably not the complete region.

Data Break: New York Leukemia Data

Consider data related to cases of leukemia (all types) diagnosed in people residing in an eight-county region of upstate New York for the years 1978-1982. Data have been aggregated to the census tract level for all eight counties, leading to 592 cases in 281 regions.

We will use the CEPP method to identify the most likely cluster when considering population radii of 1000, 5000, 10000, and 40000 persons-at-risk.

7.3.5 Method: Besag and Newell Approach

Besag and Newell (1991) address the variability in local incidence proportions within distance-based circles by limiting attention to circles containing a constant number of cases (the local incidence proportion numerator) rather than a constant population radius (the local incidence proportion denominator considered in the CEPP).

- The term **case circles** is used to distinguish circles defined by a case radius from those defined by a distance or population radius.

The method constructs windows originating from each region's centroid until c^* cases are in the window.

To get c^* cases in your window, you may need to add cases from neighboring regions until the number of cases in the window is at least c^*.

- You add all the cases (and population as relevant) from the next nearest region, not fractions like with the CEPP method.
- Let n_{i,c^*} denote the total population included in window w_i.

Let $r = y_+/n_+$ denote the overall disease incidence proportion for the entire study area.

Under the CRH, the expected number of cases in window w_i is $E_i = rn_{i,c}$, i.e., the estimated risk of disease times the population in the window.

The p-value for w_i is probability of observing c^* or more cases among n_{i,c^*} persons at risk for a Poisson random variable,

$$\sum_{j=c^*}^{\infty} \frac{\exp\left(-rn_{i,c^*}\right)\left(rn_{i,c^*}\right)^j}{j!} = \sum_{j=c^*}^{\infty} \frac{\exp(-E_i)(E_i)^j}{j!}$$

Note: You generally want your case circle to cover many regions.

The method provides a p-value associated with each window and hence is primarily a collection of tests to detect individual clusters.

One can also modify this method to make it a focused test.

H_0: The most compact window (in terms of population size) with at least c^* cases is not significantly more compact than what is expected under the constant risk hypothesis.

H_a: The most compact window (in terms of population size) with at least c^* cases is significantly more compact than what is expected under the constant risk hypothesis.

Data Break: New York Leukemia Data (continued)

We will use the Besag-Newell method to identify the most likely cluster of cases of certain sizes.

We consider case radii of 6, 12, 17, and 23 cases.

7.3.6 Method: Spatial Scan Statistics

The spatial scan test does not require choosing a fixed cluster radius.

Implementation of the spatial scan test for regional count data mirrors that for case-control point data, replacing controls with the population sizes of each region.

The primary goal of the spatial scan test is to identify the cluster (among those considered) least compatible with the null hypothesis of no clusters/clustering.

In a regional count setting, Kulldorf (1997) considered radii distances ranging from the smallest observed distance between a pair of regions (e.g., intercentroid distance) to a user-defined upperbound (e.g., half the width of the study area, half of the population contained in the circle).

A region contributes all its cases and population within the circle if its centroid is contained within the circle.

At each possible circle distance (e.g., at each observed intercentroid difference), we calculate that likelihood ratio statistic testing the constant risk hypothesis versus the specific alternative that risk within regions having their centroid within the circle differs from the risk in the rest of the study area.

With regional count data, we base the likelihood ratio statistic on the Poisson distribution.

Under the CRH, the expected counts consist of age-standardized or of regional population sizes multiplied by an estimator of the overall risk.

Let w_i denote window i, with $i = 1,2, \ldots, N_{win}$.

The test statistic for w_i is

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

where Y_{in} is the number of cases in the window, E_{in} is the expected number of cases in the window, Y_{out} is the number of cases outside the window, E_{out} is the expected number of cases outside the window.

- Typically, $E_{in} = rn_{in}$, where n_{in} is the total population in the window.

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

$$T_{scan} \quad = \max_{i=1,\ldots,N_{win}} T_{w_i}$$

Simulate N_{sim} data sets under the CRH and compute T_{scan} for each simulated data set. Denote these statistics $T^{(1)}, \ldots, T^{(N_{sim})}$.

The Monte Carlo p-value for the test is

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

Notes:

- The statistics are correlated between circles *within* each simulation
- The maximum values are independent *between* simulations
- The p-value for the most likely cluster is the probability of observing a more extreme maximal statistic *anywhere* in the study area (rather than the significance of observing the maximum at a particular location).

How to interpret located clusters: The cluster has a statistically significant test statistic larger than we would expect the most likely cluster to have under the constant risk hypothesis.

H_0: The most likely cluster of cases (in terms of the local rate of cases in the cluster compared to outside the cluster) is consistent with what is expected under the constant risk hypothesis.

H_a: The most likely cluster of cases (in terms of the local rate of cases in the cluster compared to outside the cluster) is more extreme than what is expected under the constant risk hypothesis.

To identify potential secondary clusters, we compute p-values of significance for each window, w_i, via the formula

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

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).

Data Break: New York Leukemia Data (cont.)

We utilize the spatial scan method to identify the census tracts most likely to be a cluster of cases.

7.3.7 Method: Other Spatial Scan Statistics

The original method proposed by Kulldorff (1997) is often called the circular scan method because it tends to identify circular clusters.

There have been many extensions of the spatial scan method.

Kulldorff et al. (2006) proposed the elliptical scan method. This method scans elliptical windows originating from each regions centroids.

Tango and Takahashi (2005) proposed the flexible scan method. This method scans all connected subsets of the K nearest neighbors of an originating region that includes the originating region.

Tango and Takahashi (2012) proposed the restricted flexible scan method. This method updates the flexible scan method to only scan regions that remain after filtering out regions with low estimated risk.

Assuncao et al. (2006) proposed the dynamic minimum spanning tree (DMST) scan method. This method chooses the windows through a greedy search algorithm that adds the connected region that results in the largest possible test statistic.

Costa et al. (2012) improved the DMST method by placing restrictions on the search. The early-stopping DMST method stopped the algorithm when the test statistic goes down. The double connection scan method only considers regions that are connected to at least two of the regions in the current window. The maximum linkage scanning method only considers regions that have the maximum number of connection to the current window.

Abolhassani and Prates (2021) and French et al. (2022) summarize many different variations of scan statistics.

French et al. (2022) performed an extensive simulation study comparing the performance of many scan methods using 126 publicly-available data sets. The methods were compared with respect to power (the probability is detecting ANY cluster), sensitivity (the proportion of the true cluster's population contained in the detected cluster), and PPV (the proportion of the detected cluster's population that overlaps the true cluster).

The graphics below are taken from French et al. (2022)

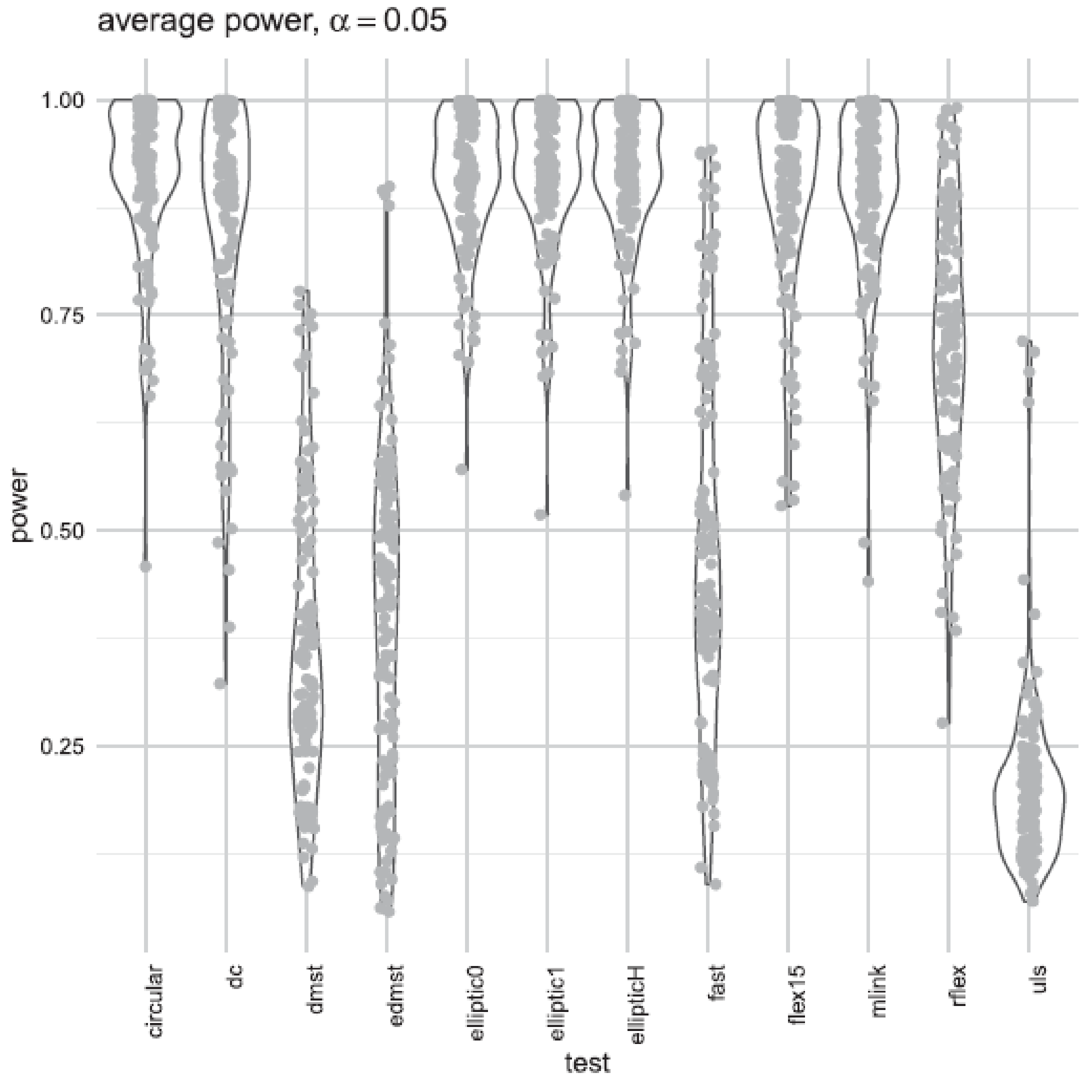

Figure 5. Average power of each method across all 126 cluster models.

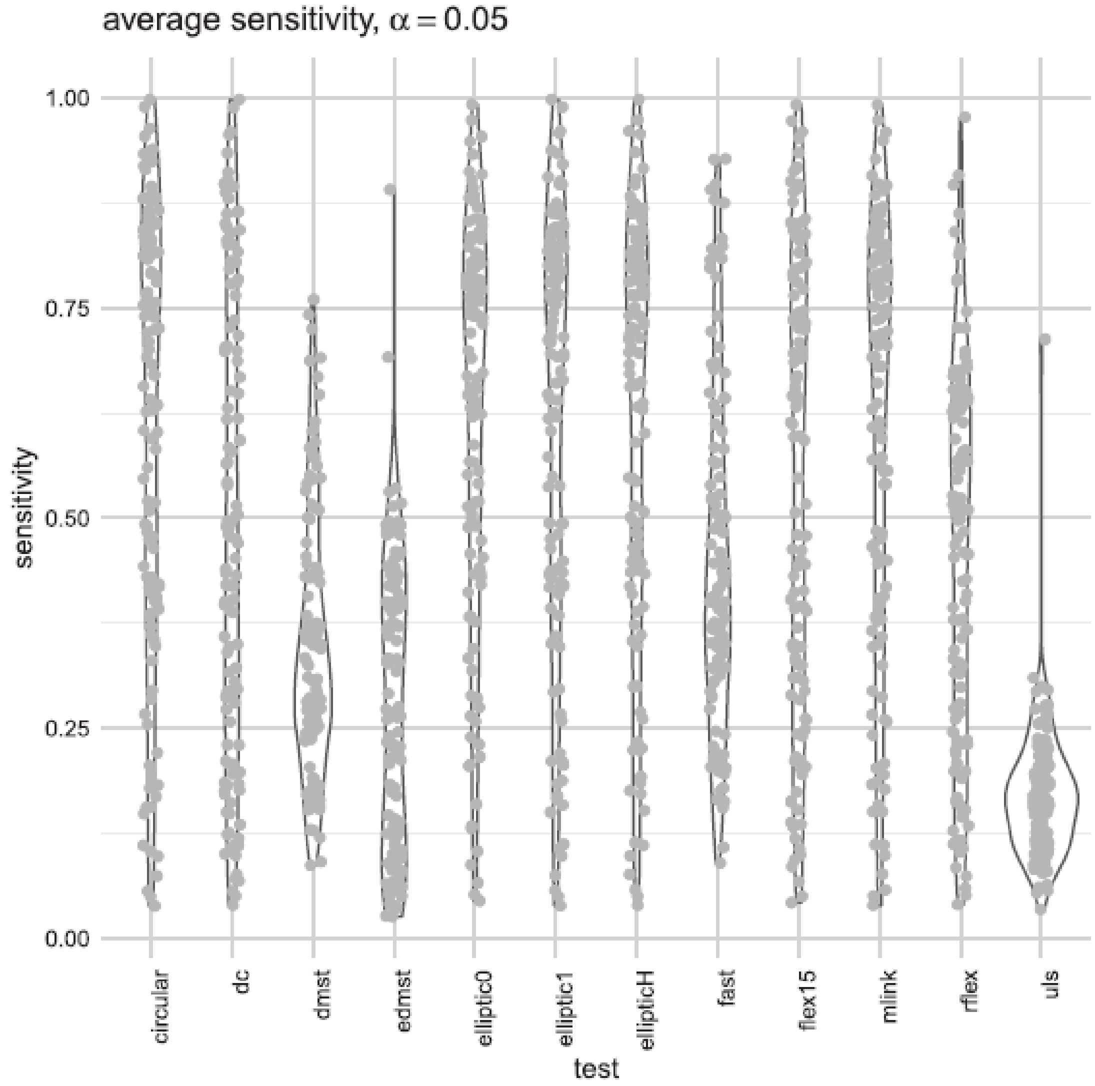

Figure 6. Average sensitivity of each method across all 126 cluster models.

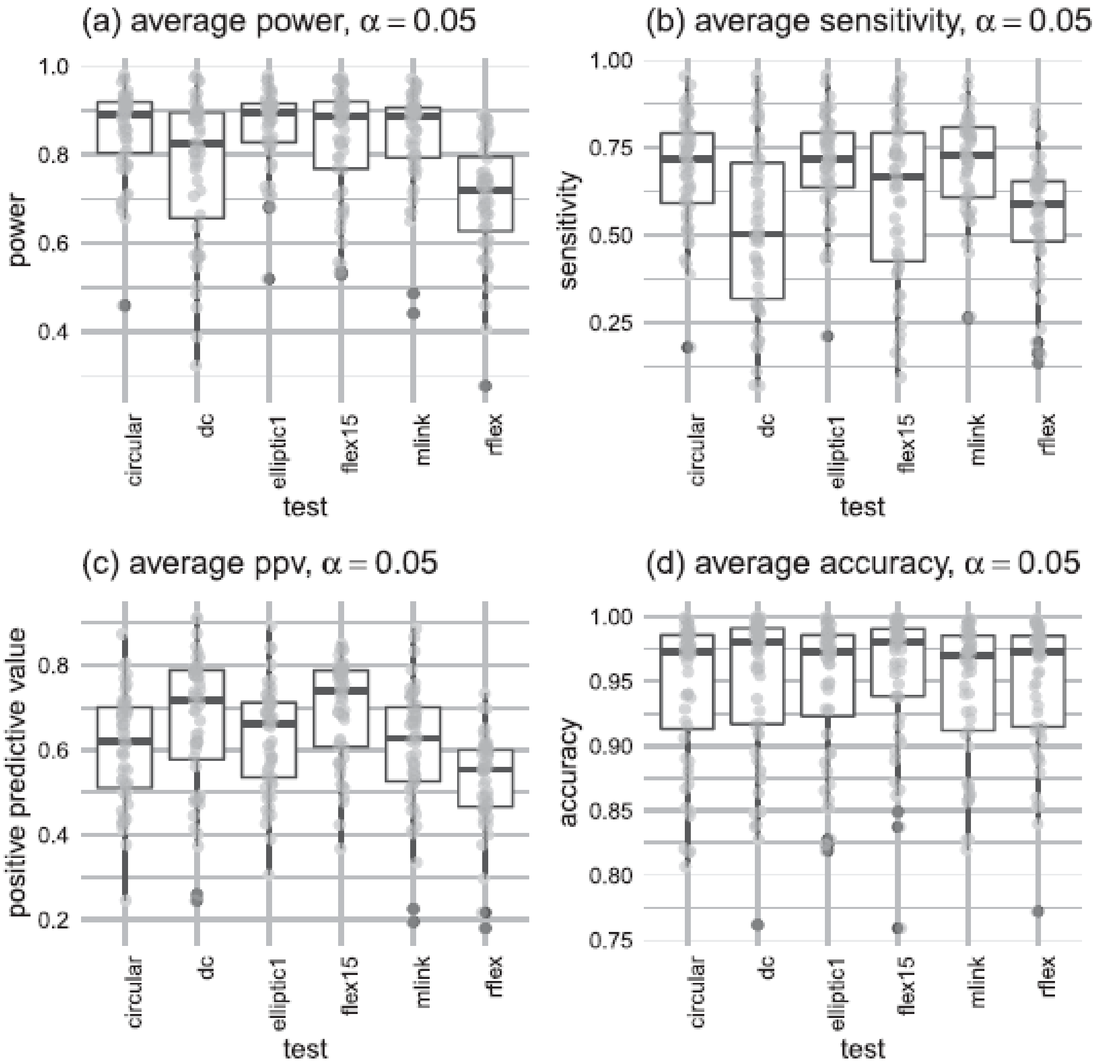

Figure 8. Power-related measures for several methods for the irregularly-shaped cluster models.

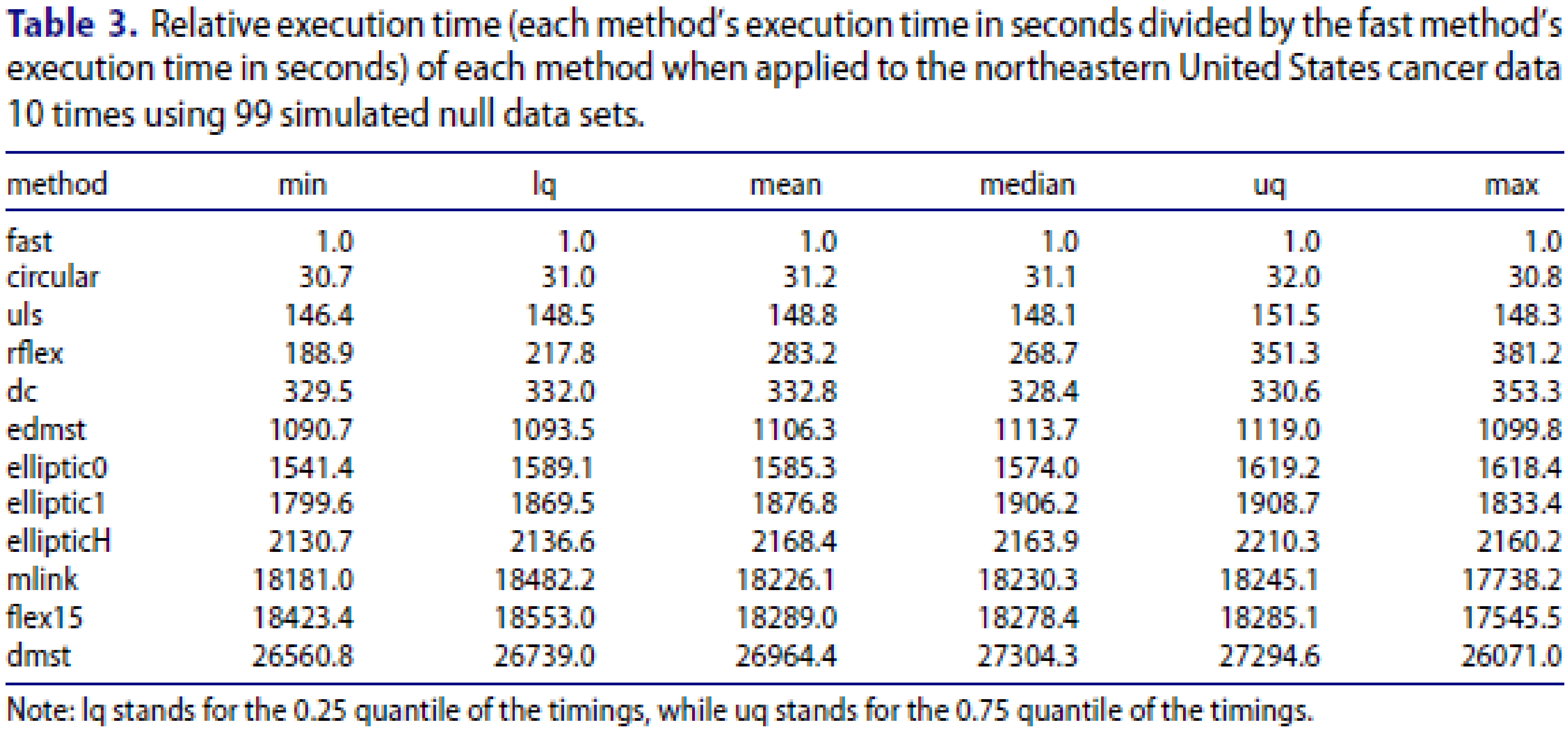

Table 3. Relative execution time (each method's execution time in seconds divided by the fast method's execution time in seconds) of each method when applied to the northeastern United States cancer data 10 times using 99 simulated null data sets.

method	min	lq	mean	median	uq	max
fast	1.0	1.0	1.0	1.0	1.0	1.0
circular	30.7	31.0	31.2	31.1	32.0	30.8
uls	146.4	148.5	148.8	148.1	151.5	148.3
rflex	188.9	217.8	283.2	268.7	351.3	381.2
dc	329.5	332.0	332.8	328.4	330.6	353.3
edmst	1090.7	1093.5	1106.3	1113.7	1119.0	1099.8
elliptic0	1541.4	1589.1	1585.3	1574.0	1619.2	1618.4
elliptic1	1799.6	1869.5	1876.8	1906.2	1908.7	1833.4
ellipticH	2130.7	2136.6	2168.4	2163.9	2210.3	2160.2
mlink	18181.0	18482.2	18226.1	18230.3	18245.1	17738.2
flex15	18423.4	18553.0	18289.0	18278.4	18285.1	17545.5
dmst	26560.8	26739.0	26964.4	27304.3	27294.6	26071.0

Note: lq stands for the 0.25 quantile of the timings, while uq stands for the 0.75 quantile of the timings.

7.4 Global Indices of Spatial Autocorrelation

We move from methods for smoothing local incidence proportions (rates) to methods that summarize the extent of observed spatial similarity between nearby regions.

A **global index of spatial autocorrelation** provides a summary over the entire study area of the level of spatial similarity among neighboring observations.

7.4.1 Goals

A global index of spatial autocorrelation summarizes the degree to which similar observations tend to occur near each other.

- Extreme values of the index in one direction suggest positive spatial correlation, while the opposite direction suggests negative spatial correlation.

Global indices of spatial correlation can be used to detect clustering in disease patterns, but not detect individual clusters.

Autocorrelation among disease counts or incidence proportions may reflect real association between cases due to infection, or perceived association based on a spatial aggregation of similar values.

7.4.2 Assumptions and Typical Output

Indices of autocorrelation share a common basis structure:

- The similarity between values at locations i and j is measured.
  - High values suggest greater similarity. Low values dissimilarity.
- A weight related to the proximity of locations i and j.

Let sim_{ij} denote the similarity between data values Y_i and Y_j.

Let w_{ij} denote a weight describing the proximity between locations i and j.

Global indices of spatial autocorrelation build on the basic form

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\mathrm{sim}_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}},$$

the weighted average of the similarity between observations.

When calculating a spatially averaged **incidence proportion**, to stabilize the incidence proportion in region i, we typically assume that the data observed in region i should substantially contribute to this weighted average. Hence, we set $w_{ii} > 0$.

When calculating a spatially averaged **similarity** between nearby observations, we omit the similarity between a region and itself, so we set $w_{ii} = 0$.

Note:

- sim_{ij} is a function of random variables, so it has a distribution.
- w_{ij} are fixed quantities based on the underlying geography of the regions.

Changing sim_{ij} yields new methods, while different w_{ij} simply yields a different result for a specific method.

Spatial Proximity Matrices

The collection of weights w_{ij} is known as a **spatial proximity matrix** (or spatial connectivity or spatial weight matrix).

The w_{ij} quantify the spatial dependence between regions i and j, and collectively, they define a neighborhood structure.

The simplest neighborhood definition is provided by the **binary connectivity matrix**, in which

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ share a boundary} \\ 0 & \text{otherwise.} \end{cases}$$

This matrix is symmetric, and $w_{ii} = 0$.

An alternative is to expand neighborhoods to include regions that are close, but not necessarily adjacent. E.g.,

$$w_{ij} = \begin{cases} 1 & \text{if the centroid of region } j \text{ is one of the } q \\ & \text{nearest to the centroid of region } i \\ 0 & \text{otherwise.} \end{cases}$$

The regions for which $w_{ij} = 1$ are called the q nearest neighbors of region i.

This matrix does not have to be symmetric.

Neighbors can be defined by some parametric function of distance. If d_{ij} is a measure of distance (not necessarily Euclidean) between centroids, we could choose:

$$w_{ij} = \begin{cases} 1 & \text{if } d_{ij} < \delta \\ 0 & \text{otherwise} \end{cases}$$

or

$$w_{ij} = \begin{cases} d_{ij}^{-\alpha} & \alpha > 0 \\ 0 & \text{otherwise} \end{cases}$$

for some power α and threshold distance δ.

These matrices will be symmetric.

A more exotic definition of neighborhood structure is based on the fraction of region i's border that is shared with region j, e.g.,

$$w_{ij} = \begin{cases} l_{ij}/l_i & \text{if regions } i \text{ and } j \text{ share a border} \\ 0 & \text{otherwise,} \end{cases}$$

where l_{ij} is the length of the border between regions i and j, and l_i is the perimeter of region i.

Such a structure may arise as a model of the flow of goods, people, or disease between two regions.

This spatial proximity matrix will not be symmetric, in practice.

Sometimes, we may want to adjust for the total number of neighbors in each region and employ a **row standardized matrix**, where we divide each w_{ij} by the sum of neighbor weight for region i, giving a matrix W_{std}, where

$$w_{std,i,j} = w_{ij} / \sum_{j=1}^{N} w_{ij}.$$

W_{std} does not need to be symmetric.

Null Distributions

Inference for a global index of spatial autocorrelation derives from the null distribution (i.e., the distribution of the index under the null hypothesis).

Observed values of the index falling in the tails of this distribution suggest significant spatial autocorrelation.

The null distribution must be accurately determined in order for valid statistical inference to be made.

Asymptotic results for the distributions of global indices of spatial autocorrelation exist under certain assumptions, but these fail in most public health contexts.

Two common assumptions for deriving the null distribution:

1. **Normality assumption**. Assumes all observations are i.i.d. Gaussian.
2. **Randomization assumption**. Randomly assigns the observes values among the N locations. (Sort of like the random labeling hypothesis).
    a. This implicitly assumes the data are i.i.d. from some distribution.

These assumptions are generally violated because we observe counts (non-Gaussian) and the population sizes differ (so the means vary, meaning the assumption of i.i.d data is naïve).

We will use Monte Carlo hypothesis testing in a similar mechanism to before.

- We compare our observed index to the index we observe for N_{sim} data sets simulated under the constant risk hypothesis.

7.4.3 Method: Moran's I

Moran's I is a widely used measure of global spatial autocorrelation.

- It has connections with likelihood ratio tests and best invariant models for particular models of correlation for normally-distributed random variables.

$H_0: \rho = 0$
$H_a: \rho \neq 0$ (or > 0 or < 0),

where ρ is the population correlation between responses $Y_1, \ldots, Y_N$.

The similarity measure used for Moran’s I is

$$\mathrm{sim}_{ij} = (Y_i - \bar{Y})\left(Y_j - \bar{Y}\right)/s^2,$$

with $\bar{Y} = \sum_{i=1}^{N} Y_i/N$ and $s^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$.

Moran’s I is calculated using the formula

$$I = \left(\frac{1}{s^2}\right)\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}}.$$

Some details:

- Large values of I indicate positive correlation, small values indicate negative correlation.
- I is NOT constrained to [-1, 1] like Pearson's correlation, though there are connections between Pearson's correlation coefficient and Moran's I.
- There are many distributional results related to Moran's I, though they don't normally apply in public health contexts.

Caution is warranted when applying Moran's I to public health region count data:

- Differences from the sample mean may not measure similarity when there are drastically different population sizes.
- Variation in the mean may be due to population differences, not autocorrelation.
- Apparent autocorrelation in the data may simply be due to relationships among the population sizes and not to any spatial pattern in disease counts.

One solution: Work with crude incidence proportions (local rates).

- This removes heterogeneities in the value expected under the constant risk hypothesis.
- The variances still depend on the population size of each region, so some heterogeneity remains.
- Various adaptions of Moran's I for incidence proportions have been proposed, with associated variations of the associated null distribution in the presence of heterogeneous population sizes.

Walter (1992) suggests a modified Moran's I statistic based on the CRH:

$$I_{cr} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \frac{(Y_i - rn_i)}{\sqrt{rn_i}} \frac{\left(Y_j - rn_j\right)}{\sqrt{rn_j}}}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}},$$

where r denotes the overall disease incidence proportion and n_i is the population size for region i.

I_{cr} assesses the residual spatial autocorrelation, i.e., autocorrelation among deviations of values observed from what we expect.

- I_{cr} is Moran's I applied to the Pearson residuals from a Poisson GLM with local mean equal to rn_i.
- A generalized version of I_{cr} can be constructed for any generalized linear model, replacing rn_i by the appropriate expectation.

Data Break: New York Leukemia Data (cont)

We apply Moran's I as an index of spatial association to the New York Leukemia data.

We will test for spatial association using various null hypotheses (normality assumption, randomization assumption, CRH) and statistics (standard Moran's I, incidence rates, and I_{cr}).

7.4.4 Method: Geary's C

Geary's C tests

$H_0\colon \rho = 0$ versus $H_a\colon \rho \neq 0$ (or > 0 or < 0),

where ρ is the population correlation between $Y_1, \ldots, Y_N$.

Geary's C measures similarity with

$$\mathrm{sim}_{ij} = \left(Y_i - Y_j\right)^2 \frac{N-1}{2\sum_{i=1}^{N}(Y_i - \bar{Y})^2}.$$

sim_{ij} will be small when the counts are close to one another, and larger the less similar the values are.

Geary's c is defined by

$$c = \frac{N-1}{2\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(Y_i - Y_j\right)^2}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}}.$$

Geary's c ranges from 0 to 2, with 0 indicating perfect positive spatial correlation ($Y_i = Y_j$ for any pair of region with nonzero w_{ij}) and 2 indicating perfect negative spatial autocorrelation.

Geary's c is not a true correlation coefficient but is a spatial analogue of the Durbin-Watson d statistic to assess serial autocorrelation in regression and time series.

A few notes:

1. Geary’s c is small for positive spatial autocorrelation, and larger for negative.
2. Geary’s c can be adjusted for heterogeneous population sizes, like we saw previously for Moran’s I. (The details are important, but we won’t discuss them).
3. Asymptotic arguments for the distribution of the statistic are likely to be inappropriate, so Monte Carlo testing is recommended.

Data Break: New York Leukemia Data (cont.)

We apply Geary's c to the New York leukemia data to assess for global evidence of clustering

7.5 Local Indicators of Spatial Association

There are lot of local indices of spatial association (LISA), but we won’t discuss them.

7.6 Goodness-of-fit Statistics

We now consider goodness-of-fit tests for clustering/cluster detection that are similar to indices of spatial autocorrelation but assess a different aspect of the data.

7.6.1 Goals

The goal of any goodness-of-fit statistic is to summarize deviation between observed data and their expected values under some probabilistic model.

The **Pearson's $\boldsymbol{\chi^2}$ statistic** is defined as

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i},$$

where N is the number of cells, O_i is the values observed for cell i, and E_i is the value expected in cell i.

The statistic is generally used in the context of contingency tables with independent rows and columns (under the null hypothesis).

The expected values are often derived using binomial or multinomial distributions.

We can adapt this idea to our setting by thinking of each spatial region as a “cell” and replacing independence of rows and columns of a contingency table with the constant risk hypothesis.

7.6.2 Assumptions and Typical Output

The default Pearson's χ^2 statistic doesn't account for spatial patterns in the deviations.

- We must incorporate a measure of spatial structure in our setting to correct for this.

Most goodness-of-fit tests assume independence of counts in both the null and alternative hypotheses.

Measures of spatial autocorrelation assume independence of counts in the null hypothesis, but not the alternative.

Under a null hypothesis with independent counts, inference for goodness-of-fit tests typically draws from the distribution of a sum of independent, standardized regional observations.

- Pearson's χ^2 statistic has a chi-square distribution with $N - 1$ degrees of freedom (asymptotically).
- The modifications we make to the statistic may invalidate these asymptotic results, so Monte Carlo testing under the constant risk hypothesis is the safest bet.

Since a goodness-of-fit test results in a single p-value, it summarizes the evidence for clustering across all regions.

- Tests can be adapted to identify local features.

7.6.3 Method: Pearson's χ^2

Pearson's statistic seeks to detect any sort of deviation from the null hypothesis and makes no distinction between nonspatial and spatially structured collections of deviations.

Indices of spatial association may not be useful for assessing clustering because they only compare deviations between pairs of regions and do not assess the magnitude of lack of fit within each region (since typically the spatial weight $w_{ii} = 0$).

7.6.4 Method: Tango's Index

Tango (1995) considers the region proportions

$$\left(\frac{Y_1}{Y_+}, \dots, \frac{Y_N}{Y_+}\right),$$

where $Y_+ = \sum_{i=1}^{N} Y_i$ is the total number of observed cases.

- Note that these values reflect the proportion of cases in each region, not the incidence proportion in each region.

We also obtain the expected proportions under the constant risk hypothesis, the vector

$$\left(\frac{n_1}{n_+}, \dots, \frac{n_N}{n_+}\right),$$

where $n_+ = \sum_{i=1}^{N} n_i$ denotes the total population at risk.

Both Y_+ and n_+ are known constants, and we will condition our inference on them.

- Under the constant risk hypothesis, the population proportion provide the expected cell probabilities for a multinomial distribution.

Tango’s index compares the case proportions observed to those expected under the constant risk hypothesis.

H_0: The observed case proportions in the enumeration districts are consistent with what is expected under the CRH.

H_a: The observed case proportions in the enumeration districts are inconsistent with what is expected under the CRH.

The null hypothesis is a form of “no clustering of cases” in our enumeration districts.

The alternative hypothesis is a form of “clustering of cases” in our enumeration districts.

Tango's index is defined as

$$T_{\mathrm{ti}} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}^{*}\left(\frac{Y_i}{Y_+} - \frac{n_i}{n_+}\right)\left(\frac{Y_j}{Y_+} - \frac{n_j}{n_+}\right),$$

where the w_{ij}^{*} denote spatially defined weights providing a measurement of closeness between regions i and j.

- E.g., $w_{ij}^{*} = \exp(-d_{ij}/\kappa)$, where κ is a dependence scaling parameter.

Rogerson (1999) rewrote Tango's index as

$$T_{\mathrm{ti}} = \sum_{i=1}^{N} w_{ii}^{*} \left(\frac{Y_i}{Y_+} - \frac{n_i}{n_+}\right) + \sum_{i=1}^{N} \sum_{j \neq i}^{N} w_{ij}^{*} \left(\frac{Y_i}{Y_+} - \frac{n_i}{n_+}\right)\left(\frac{Y_j}{Y_+} - \frac{n_j}{n_+}\right).$$

Tango's index is clearly the sum of two parts:

- the first measures the goodness-of-fit in each region
- the second measures spatial similarity between two regions.

Tango’s index may be large (indicating clustering) because of lack-of-fit within regions (poor fit), spatial similarity in deviations from expectation (spatial autocorrelation) between regions, or both.

- A large lack-of-fit component suggests the constant risk hypothesis is false (clustering within the region).
- A large spatial association component suggests similarity in incidence proportion among regions close to each other (clustering between regions).
- Both conclusions are evidence of clustering.
- Rogerson (1999) weights Tango’s index in a way to connect it more directly to Moran’s I.
    - There is a lot of interesting discussion in the book about what this means in terms of testing.

There are some asymptotic results for Tango's test of clustering (some of them actually reasonable), but there are better procedures.

1. Apply a Monte Carlo testing approach, where we simulate regional counts conditional on the fixed total Y_+.
   a. Each simulation assigns a total of Y_+ cases to the regions where the probability of each case falling in region i is n_i/n_+ for $i = 1, \dots, N$).
2. Tango also provides a procedure based on the chi-square approximation that is adequate with as few as one case expected per region.
   a. This can be implemented using simple matrix algebra, though we refer to the book for details.

Data Break: New York Leukemia (cont.)

We apply Tango’s method to identify clustering of disease for the New York leukemia data.

7.6.5 Method: Focused Score Tests of Trend

See book for details.

7.7 Statistical Power and Related Considerations

Power is the probability of concluding the alternative hypothesis when the alternative hypothesis is true.

- In our context, detecting clusters/clustering of cases.

We would like to use the most powerful test, holding Type I error constant.

7.7.1 Power Depends on the Alternative Hypothesis

There are many different alternative hypotheses that capture clusters/clustering of cases.

No one method can detect them all (equally well).

There is no uniformly most powerful test for all types of clusters/clustering simultaneously.

- We tend to think of lack-of-fit tests as detecting first-order clustering (deviation from expected counts).
- Spatial autocorrelation indices detect second-order clustering (similarity between nearby counts).
- The distinction is a bit vague, as both can be used to detect the other.

Some tests are most powerful for detecting specific alternative hypotheses but may not be for alternative scenarios.

7.7.2 Power Depends on the Data Structure

It is easier to detect a cluster/clustering of cases if the clusters/clustering occur in regions with high population.

- We get better estimates of relative risk at these locations.

7.7.3 Theoretical Assessment of Power

Theoretical results related to the comparison of power between methods may be done in certain contexts, but results are relatively few.

7.7.4 Monte Carlo Assessment of Power

Due to the challenges of theoretical power comparisons, Monte Carlo simulation can be used to compare power of different methods for specific alternative hypotheses.

This is tedious and unglamorous but can be very informative.

It is most helpful to do this for many different alternative hypotheses to get a clear picture of how the methods perform in different scenarios.

7.7.5 Benchmark Data and Conditional Power Assessments

Kulldorf et al. (2003) propose the creation and use of **benchmark data sets** incorporating a wide variety of types of clusters and clustering made available through the Internet to allow researchers to compare and contrast the performance of new and existing methods on the same sets of data.

This provides valuable insights into the power of existing and new methodologies, though different data sets may produce different results.

Data Break: New York Leukemia (cont.)

Are there clusters of elevated leukemia rates in upstate New York? Possibly.

Our conclusions depend on:

- the validity of the test statistics
- the types of clusters they detect
- the degree to which our data support the assumptions the test statistics require and provide sufficient statistical power to detect the deviations observed

Against:

- Many tests of spatial autocorrelation are not significant. But we think the data violate testing assumptions.
- Assumptions of Poisson counts may be unrealistic. Are the data overdispersed? ($\text{Var}(Y_i) > E(Y_i)$)
- Have we accounted for spatial patterns in important covariates possibly affecting local leukemia rates?

For:

- Most of the tests seem to indicate clustering, and generally in the same areas.

7.8 Additional Topics and Further Reading

7.8.1 Related Research Regarding Indices of Spatial Association

We only covered the most well-known indices of spatial association.

- There has been a lot more work on indices applied to regions with heterogeneous population sizes.
- There has been some interesting work on "exact" methods and more accurate, less time-consuming alternatives to Monte Carlo testing.

7.8.2 Additional Approaches for Detecting Clusters and/or Clustering

Tests of Overdispersion Cluster identification can be performed by comparing the mean and variance of the observed data, or to do a formal test for overdispersion.

Other Modifications to Pearson's χ^2 Adjustments for multiple comparisons, assessing temporal clustering, among others, have been proposed in tests for clusters/clustering.

Stone's Tests This family of tests offers a way of constructing statistical tests of fit for particular alternatives of interest, relative to certain foci.

Other methods include **Weighted Likelihood Ratio Tests, Bayesian Cluster Detection**, and **Classification Methods**

7.8.3 Space-Time Clustering and Disease Surveillance

People are frequently interested in identifying disease clusters in space and time.

Recent increased interest in bioterrorism and emerging infection diseases has led to increased efforts to detect **emerging** disease clusters.

- There is a lot of research on prospective disease surveillance.