Spatial analysis of epidemics: Disease patterns

Everything is related to everything else, but near things are more related than distant things

-- Tobler (1970)
(Tobler's 'Law of Geography')

- A different perspective
- Consider the: arrangement of diseased individuals with each other and with their physical surroundings -- pattern or dispersion
- Generally studied separately from disease gradients
  - For practical purposes, gradients are typically studied when there is one (original) inoculum source
  - Patterns are typically studied where there are "many" or an undefined number of inoculum sources
- The approach (spatial pattern analysis) is very statistical (stochastic) in nature
  - In contrast, for gradients (and disease progress curves), the basic approaches were mostly based on deterministic models (either for \( Y \), or \( dY/ds \), or \( dy/ds \), or \( \partial y/\partial s \)), and statistical versions of models were used primarily for parameter estimation

\[ \ln(y(t)) = \ln(y(0)) - b_2 s + \epsilon \]
• Although the process of disease spread certainly exists for the more complicated (and common) scenarios, it is not easily studied directly.
• Spatial pattern analysis borrows much from ecology, especially statistical ecology, as well as from geostatistics, geography, discrete distribution theory, and other fields.
  ○ Sometimes, in the past, plant pathologists borrowed the wrong methods for pattern analysis (by not appreciating how disease intensity data may differ from the typical data obtained in ecology).
• In recent decades, plant pathologists have become quite sophisticated in spatial pattern analysis, and have developed some specialized methods that properly account for the statistical properties of intensity data.

• Concepts to be considered:
  a. Aggregated, clumped, clustered
  b. Random
  c. Uniform, regular
• Concepts explained later--for now, aggregation indicates spread from plant-to-plant.
• Two important issues:
  o Sampling unit
  o Data (variable) type

Data types -- Recall: Plant disease intensity
• Incidence
  o Disease status of individual plant units -- binary variable
  o Number (or proportion) diseased -- count variable with natural denominator (discrete)
• Count
  o Number of lesions (or other units of infection) -- count variable without a natural denominator (discrete)
• Severity
  o Area (relative or absolute) of plant tissue affected by disease -- continuous variable
  o Ordinal rating of 'degree of infection' -- ordered categorical variable (discrete)
• NOTE: Several types of spatial pattern analysis are dependent on the type of data (type of random variable)!
  o One must be careful.
**Sampling Unit (SU):**
- The entity to be observed or measured (i.e., sampling units are the entities to be sampled from a population for observation/measurement)
  - Leaf, branch (shoot), plant, group of plants, field
  - Each SU can consist of one or multiple individuals (single leaves, leaves on a tree)

**Sample:**
- A selection from a larger population (= a collection of sampling units)
  - We use $N$ for number of sampling units
  - Chosen in various ways (randomly, systematically, etc.)
- **Cluster sampling:**
  - If each sampling unit (e.g., plant) consists of $n$ individuals (e.g., leaves), and observations are made on all $n$ individuals ($n > 1$)
    - i.e., there is more than one individual observed per sampling unit
    - Note: there is a total of $n \cdot N$ individuals in the full sample
  - Disease may or may not be clustered with cluster sampling!

**Census:**
- Observation of all individuals (and obviously all sampling units) in the population
  - All plants in a field (if the plants in the field are the only ones of interest)
  - We still use sampling unit term with a census
    - Think of the field as representative of a super-population

---

**Data collection: sparse sampling versus intensive mapping**

- **Sparse sampling:**
  - Only the observation or measurement (e.g., disease status of plant, number of diseased plants) is recorded for a generally restricted number of sampling units
  - Spatial location of the sampling unit is not recorded or known
  - Spatial analyses for sparse sampling require discrete data! And, statistical methods appropriate for discrete data.
    - (binary, counts [with or without a natural denominator])
    - Indicate heterogeneity (variability) of disease intensity
    - Just one aspect of spatial pattern (so-called small scale pattern)

- **Intensive mapping:**
  - Both the observation (or measurement) and spatial location are recorded for the sampling units
    - (sampling with 'spatially-referenced' sampling units)
    - Distances between sampling units can be determined and utilized
    - Often, a large number of sampling units
  - Allows mapping of observations (for the results of each sampling unit, not necessarily for each observation within the sampling units)
  - Analyses for both discrete and continuous data
  - Indicate pattern over many scales (ultimately) -- explained later.

When one has intensive mapping, one can still use methods appropriate for sparse sampling (for discrete data)
**Data collection:** sparse sampling versus intensive mapping for spatial pattern analysis:

- **Sparse sampling** (discrete variables only--several analytical methods)
- **Intensive mapping** (discrete or continuous variables--several analytical methods, including those for sparse sampling [if discrete])

- The sampling units for both situations can be either single individuals or clusters (cluster sampling)

- Methods for sparse sampling can be applied to intensive maps **[for discrete data]**, but the reverse is NOT true (methods that require location of the SUs cannot be applied when locations are not known!)

---

The intensely-mapped sampling units may or may not be contiguous (touching)

- Refer to sampling units as **quadrats**
- Spatially referenced sampling units
- **Contiguous quadrats**

If there is more than 1 individual observed in each sampling unit: **cluster sampling**

\[
\begin{align*}
N &= 9 \\
N &= 36
\end{align*}
\]
Spatial pattern analysis

- Large field of study
- We will focus on just a few situations...
  - **Discrete data: counts with natural denominator**
    - That is, disease incidence (either as the count itself, or proportion) is determined for each sampling unit
    - Sampling unit consists of \( n \) individuals
      - That is, cluster sampling is used (with \( N \) "clusters" or sampling units)
    - Either sparse sampling or intensive mapping
    - Probably the most common situation

- Book (chapter 9) gives details on:
  - Binary data (incidence) -- that is, not cluster sampling
  - Count data (without natural denominator)
    - Example: Number of lesions, ..., number of spores, ...
      - Most common situation in ecology, entomology, and other fields

Why study patterns?

- Infer the nature of pathogen dispersal or disease spread (spread from plant to plant? Spatial range of inoculum dispersal?)
  - Here, observed pattern itself is of interest (various statistics can indicate the range of dispersal)
    - Pattern is the "response variable"

- Properly estimate mean disease intensity and the variability of disease intensity—**SAMPLING**
  - Determine the required number of sampling units to:
    - Achieve a certain precision, or to
    - Obtain a decision tool with low probability of false positives and false negatives
  - Here, pattern is of interest only because pattern affects standard errors of parameter estimates (sampling distributions of the parameter estimates)

- Properly test for treatment effects (use spatial information)*

- Determine how pattern affects disease development
  - Here, the pattern is the 'treatment' (\( dy/dt \) is the response)

- Determine how disease development affects pattern
  - Here, pattern is the response variable
Analysis of sparsely-sampled incidence data

- Cluster samples (so that there are \( n \) individuals observed in each sampling unit)
- Total of \( N \) sampling units
- \( Y \) is number of diseased individuals (say, for the entire sample)
- \( Y_i \) is number of diseased individuals in the \( i \)-th sampling unit
- \( y_i \) is the proportion of diseased individuals in the \( i \)-th sampling unit (\( y_i = Y_i / n \))

\[
\bar{Y} = \frac{\sum Y_i}{N} = \frac{\sum Y_i}{n \cdot N} = \frac{Y}{n}
\]

Mean proportion is an estimate of probability of a plant (or plant unit [e.g., leaf]) being diseased (\( p \))

\[
p = \bar{Y}
\]

Requires discrete data (incidence here) to interpret in terms of spatial pattern. Can apply methods to sparse sampling or intensive mapping.

\[
\begin{align*}
Y &= 7 + 10 + 5 + 5 + 6 + 3 \\
\bar{Y} &= \frac{Y}{n} = \frac{36}{6} = 6 \\
p &= \frac{\bar{Y}}{n} = \frac{6}{6} = 1
\end{align*}
\]
Random pattern (for discrete data):

- The probability of a plant (or plant unit [e.g., leaf]) being diseased is independent of the disease status of other plants
  - That is, \( p \) is constant (across SUs)
- The disease status of a plant is unrelated to disease status of neighboring plants (correlation of disease status of a given plant with neighboring plants is zero)
- Knowing the disease status of a plant provides no information on the disease status of other plants
- If \( p \) is constant, diseased individuals per sampling unit (\( Y \)) has a binomial distribution
- Distribution, or probability distribution:
  - (For discrete random variables): a mathematical formula that gives the probability of each value of the variable
- Distributions are evaluated by comparing the observed frequency (\( O \)) of diseased individuals to predicted (expected; \( E \)) frequency

Phomopsis leaf blight of strawberry (Turechek & Madden, 1999, and other papers). Transect through field. \( N=59, n=15; \ p=0.226 \ (=200/[15 \times 59]) \Rightarrow \chi^2 = \sum \frac{(O-E)^2}{E} = 3.39\)

Spatial pattern analysis

Methods for discrete data (count with natural denominator) and cluster sampling (\( n \) individuals in each SU)

Requires only sparse sampling, but can be used with intensive mapping (just not using the location information)
Phomopsis leaf blight of strawberry (Turechek & Madden, 1999, and other papers). - Transect through field. $N=59$, $n=15$; $p = 0.226$ ($=200/15\times59$)

$$\hat{p} = \frac{X}{n} = \frac{3}{15} = 0.226$$

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = 0.0325$$

Sample Variance

Diseased leaflets per sampling unit

Distribution

Observed $d(0)$

Map

Epidem18_13_ Page 15
Binomial distribution:
- Estimated mean $y$: $\hat{y} \approx 2.56$
- Estimated variance of proportion: $\hat{p}(1-\hat{p})/n = s_{bin}^2 = \sqrt{\hat{p}(1-\hat{p})/n}$
- Do not confuse $s_{bin}^2$ with the observed variance: $s^2 = \frac{\sum(y_i^2)}{n} - \frac{\left(\sum y_i\right)^2}{n} = 0.032$

What happens if the data are truly binomial?
- Among other things, the variance is determined completely by $p$ and $n$.

Some evidence of nonrandomness
- Predictions too high in middle, and too low at low and high values of $Y$

Nonrandom pattern (for discrete data):
- The probability of a plant (or plant unit [e.g., leaf]) being diseased is not independent of the disease status of other plants
  - That is, the probability is NOT constant across SUs, but is a random variable
- The disease status of a plant is related to disease status of neighboring (other) plants (correlation of disease status of a given plant with neighboring plants is nonzero)
  - If a given plant is diseased, there is a tendency for neighboring plants to be diseased.
- Knowing the disease status of a plant provides some information on the disease status of other plants
  - Call this aggregated, clumped, clustered, ...
- One approach to put into practice: specify a statistical distribution for $p$ (since it is now a random variable)
  - If $p$ has a beta distribution, then $Y$ (number diseased per SU) has a so-called beta-binomial distribution
    - A messy formula, but this does not matter much (after all, the normal distribution has a messy formula)
**Beta-binomial** distribution (two parameters):
- $p$ (mean probability of a plant being diseased)
- $\theta$ (heterogeneity or aggregation parameter)
  - $\theta=0$ (reduce to binomial)
  - $\theta>0$ (aggregated, clustered)
  - Can go to infinity, but even 0.2 is large

Fitting discrete distributions to data:
- **Maximum likelihood** is generally best, but this may require specialized computer programs (an iterative method for the beta-binomial)
- Simpler methods work well for many purposes
  - So-called **moment** method ("method of moments")
  - Sometimes, the estimates of parameters are very close for different methods
  - Moment methods can be done using nothing more than the estimates of the mean and variance of a sample!

For the Phomopsis example:

- Observed variance increases relative to the binomial variance, $\theta$ increases.

Note: as the observed variance increases, relative to the binomial variance, $\theta$ increases.
Spatial Pattern Analysis — review

- *An* approach for describing and understanding plant diseases in space
  - Often done when spread from a well-defined inoculum source cannot be studied
  - The methodology is *statistical*
    - Thus, the *form of the random variable* is of paramount importance -- in choosing the appropriate analysis and in interpreting results
    - Most of the spatial analysis has been done with *disease incidence* -- *discrete random variable with natural denominator* -- or with *counts* (*not considered in lecture*)
- The form of the sampling unit, and the type of sampling, are very important in spatial pattern analysis
  - We consider *'cluster sampling'*; either with *'sparse sampling'* or *'intensive mapping'*
    - This is a typical situation with pattern analysis
- With a random pattern, the probability of being diseased is a *constant*, and the *binomial distribution* is appropriate for number diseased per sampling unit
- With aggregated pattern, the probability of being diseased is a *random variable*, and the *beta-binomial distribution* is appropriate for number diseased per sampling unit
Although it is useful to directly fit distributions to data, and determine their goodness of fit, such an approach is not necessary. In particular, one can utilize properties of the beta-binomial distribution to test for aggregation, and quantify the degree of aggregation.

For instance, it is very important to consider the estimated variance of the discrete data, $s_y^2$ or $\hat{s}^2$. It can be shown that the variance of a variable with the beta-binomial distribution is:

$$s_y^2 = \hat{s}^2 \frac{(1+n\theta)(1+\theta)}{n(1+2\theta)}$$

Thus, the beta-binomial variance equals the binomial variance (for a random pattern) times a scaling factor that goes up with increasing aggregation.

- At $\theta=0$, beta-binomial variance equals binomial variance.
  - In fact, here the two distributions are the same.
- At $\theta>0$, beta-binomial variance is larger than the binomial variance.
  - Increasing variability (heterogeneity) means greater aggregation.
- Since the scaling factor can take on any value (in principle), the beta-binomial variance can always be made to equal the actual variance of a sample, $s_y^2 \approx \hat{s}^2$.
  - In fact, rearrangement of above formula is used to get the moment estimate of $\theta$. 

\[ \theta \uparrow \hat{s}_y^2 \uparrow \]
A very useful statistic for characterizing aggregation is the ratio of the so-called observed variance (which is not based on any assumptions about the distribution) and the estimated variance for a variable with a binomial distribution (i.e., for a random situation)

\[ D = \frac{s^2}{\hat{\pi}(1-\hat{\pi})/n} = \frac{s^2}{\hat{E}(\hat{\pi})/n} = \frac{s^2}{s^2_{\text{bin}}} = \frac{s^2_{\text{bin}}}{s^2_{\text{bin}}} = \frac{1 + n\theta}{1 + \theta} \]

- \( D \) is known as the **index of dispersion**

**Some alternative expressions**

- **Very simple evaluation of aggregation:**
  - Get \( D \) (from observed variance and binomial variance)
    - \( D=1 \): random
    - \( D>1 \): aggregated
    - \( D<1 \): regular (uniform)
      - Minimum \( D \) is 0
  - Test of aggregation:
    - \((N-1)D\) has chi-square distribution with \( N-1 \) df if random (i.e., if binomial)
    - Null hypothesis: random (binomial)
    - If \((N-1)D > \text{critical chi-square}\), then conclude aggregated

- \( H_0: \text{random} \)
  - \( H_a: \text{aggregated} \)

\[ \overline{\text{equivalent}} \]

- \( H_0: D=1 \)
- \( H_a: D>1 \)

\[ H_o'; \theta=0 \]
\[ H_a'; \theta>0 \]
\[ Y = \hat{\theta} = 0.226 \]

\[ \hat{\theta} = 0.146 \]

\[ \chi^2 = \frac{0.0325}{0.0117} = 2.78 \]

\[ (Y-1) = (59-1) \times 2.78 = 161 \]

\[ \chi^2_{58} = 76.8 \]

\[ \alpha = 0.05 \]

161 >> 76.8, thus, reject Ho in favor of Ha

\[ df = N - 1 = 58 \]

Program uses MLE (population) not moment estimates

\[ \hat{\theta} = 0.1379 \]

\[ s_e(\hat{\theta}) = 0.0414 \]

\[ H_0: \theta = 0, \quad H_a: \theta > 0 \quad \text{based on } BBD \]

\[ t = \frac{\hat{\theta} - 0}{s_e(\hat{\theta})} \]

\[ N - 1 = 58 \text{ df} \]
Meaning of aggregation based on variances and/or discrete distributions

- The beta-binomial is not the only distribution that can describe aggregated (clustered) disease incidence data, but is the most common
  - It has very useful theoretical properties (not discussed here)
  - A special case ($\theta=0$) is the binomial (random)

- Distributions of this type explicitly characterize heterogeneity of the random variable
  - When $\theta > 0$, the variable is overdispersed (so that $\theta$ is a measure of overdispersion or degree of heterogeneity)

- The beta-binomial (or similar distribution) characterizes the pattern of disease at the spatial scale of the sampling unit or smaller
  - That is, $\theta$ represents aggregation of diseased individuals within sampling units, not across sampling units.
    - SMALL-SCALE PATTERNS (e.g., small patches of disease)
      - If the data were mapped (not required, because all this works for sparse-sampling data collection), one would not necessarily see big patches of high disease and gaps of low (no) disease

- Greater aggregation within sampling units is manifested by greater variability between sampling units!
Meaning of aggregation based on variances and/or discrete distributions (continued)

- Small scale pattern can be made clear by considering the **intra-cluster correlation** ($\rho$)
  - The correlation of disease status of individuals within a sampling unit (an average, of sorts, across all sampling units)
  - Tendency for individuals within a sampling unit to have the same value
- It can be shown that: $\rho = \theta/(1+\theta)$
  - $\rho = 0$: no correlation (no aggregation)
  - $\rho > 0$: aggregation (maximum of 1)

Can also be derived with no consideration of the statistical distribution.
Probability of any two randomly selected individuals from the same sampling unit (SU) having the same disease status (\( P_{\text{same}} \)):

\[
1 - 2p(1-p)(1-p)
\]

**Summary (so far):**
- There are multiple ways of saying *almost* the same thing about a data set in terms of heterogeneity/overdispersion — **small scale pattern**
  - Sometimes just a matter of preference
- A full fitting of a distribution model to data is more informative (more information than just means and variances), but more challenging
  - Goodness of fit cannot always be determined
    - See book (if \( n \) is not fixed)
- **NOTE:** There is much more to pattern than just small-scale aggregation.
  - Example, large patches that extend over multiple sampling units
  - Or, mixture of small and large patches
  - **Topic covered next**.....
- **Caution:** appropriate statistical distributions for unbounded counts, and associated heterogeneity indices (\( D \), etc.), are different from those presented in class
  - Random: Poisson distribution
  - Aggregated: Negative-binomial distribution
Disease patterns (continued)

- Analyses based on intensively mapped data
  - Consider only cluster sampling (n individuals per sampling unit), N sampling units – disease incidence (Y_i / n = y_i)
- Once again, concerned with arrangements of diseased individuals
  - The previous methods provide no information on larger spatial scales (e.g., larger areas)
    - In fact, any arrangement of the N counts (but not the n individuals within the sampling units) gives the same D, ρ, etc.
  - Now our interest is in:
    - Tendency for observations (values) from nearby locations to have similar magnitude compared with locations farther apart
      - A form of aggregation
- For the typical analyses here, $Y$ (or $y$) can be either a discrete (count; not just binary) or continuous random variable
  - Disease incidence, density, or severity
  - Key: spatial referencing of the sampling units
    - Intensive mapping

- Three (general) major methods:
  - Spatial autocorrelation analysis (including Moran’s $I$ and Geary’s $g$)
  - Semivariogram analysis (mirror image of autocorrelation)
    - Geostatistics
  - SADIE (counts only -- with or without natural denominator)
    - Joe Perry et al. (Rothamsted)
    - Turechek, Madden, Xiangming Xu (see book)

- Of course, there are other analyses for sampling units consisting of single individuals
- Important reminder: methods unique to intensely mapped data (spatial referencing) cannot be applied to sparse sampling (i.e., if the spatial locations are not known, one cannot use methods that require the spatial locations!)
  - But methods of sparse sampling can be applied to intensely mapped data, if $Y$ is discrete (although these would not take advantage of the spatial referencing)

**Larger scale spatial patterns:**
- Tendency for observations (values) from nearby locations to have similar magnitude compared with locations farther apart
  - A form of aggregation

autocorrelation

nearest neighbor

Nearest neighbor

Spatial log order

White

Bishop

Queen

Square

Across-row

Within-row

0 5 1 0 3 6 4
2 7 8 2 4 9 4
6 8 5 2 3 5 9 7
2 6 4 1 0 4 1 8
1 3 3 2 2 2 6 5
0 1 0 1 3 2 3
Spatial autocorrelation:
- The degree of association in Y (or y) between neighboring sampling units
- For immediate neighbors (each sampling unit with those next to it), \( r(1) \)
  - Most commonly determined
- For the next most immediate neighbors, \( r(2) \)
  - And so on
- The larger the \( r() \), the more aggregated

\[
r(i) = \frac{\sum (Y_i - \bar{Y}) (Y_{i+k} - \bar{Y})}{\sum (Y_i - \bar{Y})^2}
\]

\( N_i \): number of pairs for immediate neighbors

A lot of specialized software programs

**Example Calculation:**

\[ r(1) = 0.19 \]

\[ SE(r(1)) = \sqrt{\frac{1}{N(1)-3}} \]
Spatial autocorrelation

- Magnitude of $\hat{\rho}(1)$ is evaluated.
  - Large values indicate aggregation, at a spatial scale larger than the size of the sampling unit.
    - Standard error of $\hat{\rho}(1)$: $\approx (1/N)^{1/2}$

- Note: The analyses here are for a different scale than the heterogeneity analyses (beta-binomial, $D$, etc.) presented earlier.

- Thus, the methods can give different results, since they are not characterizing the same thing.

\[
\hat{r}(1) = \frac{\sum (y_i - \bar{y})(y_j - \bar{y})}{\sum (y_i - \bar{y})^2}
\]

$N$: number of units

$J$: nearest neighbors

$J+1$: one more unit

$\text{Denominator:}$

Max. Like. (ML)

\[
\begin{aligned}
\text{estimate of variance of } r(1) \\
\text{SE}(\hat{r}(1))
\end{aligned}
\]
Synthesis

- There are many ways of characterizing spatial patterns of organisms, including diseased individuals.
- Some approaches are dependent on the type of random variable (for discrete data only) — the nature of $Y$.
  - Generally: Appropriate methods for characterizing small-scale patterns (with sparse sampling or intensive mapping).
- Some approaches depend on intensive mapping (spatially referenced sampling units) — the nature of the sampling method.
  - Appropriate methods for characterizing larger-scale patterns.
    - Can be for a wider range of disease variables.
- Reminder: we only discussed methods for sparse sampling and intensive mapping when there is a count ($Y$ out of $n$) in each sampling unit.
- Unless a pattern is truly random, there is a scale to the pattern.
  - That is, there may be small clumps (not even visible in a sense) from a map, but quantifiable using discrete-data analyses.
  - There may be large and very large clumps (patches) dispersed over the area of interest, quantifiable through autocorrelations, etc.
- Results of different classes of analyses may be complementary, not contradictory.
- In general, when intensive maps are available with discrete data, one should always assess small scale and larger scale patterns.

---

Epidem18_13_ Page 46
Clarification: by \textit{discrete} here, I mean counts of individuals out of \(n\) in each sampling unit -- other distributional approaches are used when there is not a natural denominator (but still discrete) -- see textbook.
Reading assignment:
Sections 9.1 - 9.3 in Chapter 9 (pages 235-238)

For background (optional):
Sections 9.4 and 9.9