Revised Schedule:

- **Tuesday April 26th** –
  - Lecture Kat Grace – Modeling Kenyan Malnutrition
  - Lab Catch up Diego Pedreros
- **Thursday**
  - Lecture Chris Funk - Modeling variograms
  - Lab 5 Chris Funk – Analyzing East Africa rainfall

Lab Policy – New lab centricism

- First 20 minutes spent on new lab
- Priority given to questions on new lab

Lecture today

- Output of Fest function
- Index of dispersion and $X^2$ distribution
- Variance matrix re-explained
- Variograms
Fest Output

A data frame containing up to seven columns:

- **r**
  - the values of the argument r at which the function F(r) has been estimated

- **rs**
  - The border corrected estimator of F(r)

- **km t**
  - The spatial Kaplan-Meier estimator of F(r)

- **hazard**
  - The hazard rate lambda(r) of F(r) by the spatial Kaplan-Meier method

- **cs**
  - The Chiu-Stoyan estimator of F(r)

- **raw**
  - The uncorrected estimate of F(r), i.e. the empirical distribution of the distance from a random point in the window to the nearest point of the data pattern X

- **theo**
  - The theoretical value of F(r) for a stationary Poisson process of the same estimated intensity.
**G function definition**: Proportion of event-to-nearest-event distances $d_{\text{min}}(u_i)$ no greater than given distance cutoff $d$

- $\hat{G}(d) = \frac{\#[d_{\text{min}}(u_i) \leq d]}{n}$

**F function definition**: Proportion of point-to-nearest-event distances $d_{\text{min}}(t_p)$ no greater than given distance cutoff $d$

- $\hat{F}(d) = \frac{\#[d_{\text{min}}(t_p) \leq d]}{m}$

Expected $G$ and $F$ function under CSR for relatively small distances to avoid edge effects:

$$\mathbb{E}\{G(d)\} = \mathbb{E}\{F(d)\} = 1 - e^{-\lambda \pi d^2}$$
\[ \tilde{X} = X - \overline{X} \]

\[ \Sigma = \frac{1}{n} \tilde{X}^T \tilde{X} \]
First & Second Order Effects

**First-order effects:** influence of external or environmental factors on process outcomes; e.g., abundance of plants within a sub-region could depend on soil type.

- Note: first-order effects are typically assumed to influence the magnitude of process outcomes at each location, and hence are associated with the mean of all possible process outcomes at each location.

**Second-order effects:** influence of process outcomes at one location on possible process outcomes at nearby locations; e.g., non-contagious versus contagious diseases.

- Note: second-order effects typically express some measure of “similarity” between possible process outcomes at different locations once the first-order effects have been removed, and are often associated with the covariance or correlation coefficient between different random variables.
Distance-based Metrics

- Under assumptions of stationarity (e.g. \( \mu = 0 \) or \( \lambda = \) some constant), location itself can be discarded, and the relationship between locations (e.g. distance) used in instead.

- Examples include the K-function and semivariograms.
1. construct set of concentric circles (of increasing radius \( d \)) around each event
2. compute # of events in each distance “band”, excluding event at the center
3. cumulative number of events up to radius \( d \) around all events becomes the sample \( K \) function \( K(d) \)

\[
K(d) = \frac{\mathbb{E}\{ \text{# of events within distance } d \text{ of any arbitrary event} \}}{\mathbb{E}\{ \text{# of events within study area} \}} \approx \frac{1}{\lambda} \frac{1}{n} \# \{d_{ij} \leq d, \ i = 1, \ldots, n, \ j = 1, \ldots, n\} = \hat{K}(d)
\]


CSR-Expected K Function

- K(d) & L(d) functions under CSR
  \[ \mathbb{E}\{K(d)\} = \frac{\lambda \pi d^2}{\lambda} = \pi d^2 \]
  
  this can become a very large number (due to \(d^2\)), and consequently small differences between \(K(d)\) and \(\mathbb{E}\{K(d)\}\) cannot be easily resolved.

- use L function instead:
  \[ \hat{L}(d) = \sqrt{\frac{\hat{K}(d)}{\pi}} - d \]

- With \(\mathbb{E}\{L(d)\} = 0\)

- Interpreting the L function
  - \(L(d) > 0\) implies clustering
  - \(L(d) < 0\) implies stratification

- Watch out for edge effects
  - Reality tends to be ‘patchy’
  - Can we use Monte Carlo simulations instead of edge effect corrections?
Sample and Population Semivariogram

**Setting:** Set of $N$ measurements $\{y(s_n), n = 1, \ldots, N\}$ of attribute $Y$ which varies continuously in space; $y(s_n)$ denotes $n$-th measurement obtained at $n$-th sample location with coordinate vector $s_n$.

**Inference objective:**
- Go beyond sample semivariogram to infer a model of spatial association for the population, i.e., the entire $Y$-attribute field; most often, one is talking about parametric semivariogram models which are expressed as functions of Euclidean distance.

**Objectives**
- To pinpoint the limitations of a sample semivariogram/covariogram/correlogram.
- To highlight some conditions that can be used to check whether an arbitrary function of distance is as valid semivariogram model.
- To survey some of the most frequently used semivariogram models in practical applications.
Consider a set of $L$ lag distance classes; let $\{h_l, l = 1, \ldots, L\}$ denote the set of average distances between data pairs in each class.

Compute sample semivariance $\gamma(h_l)$, or moment of inertia of $h_l$-specific scatter-plot of lagged $y$-attribute values, for each distance class $h_l$:

$$\hat{\gamma}(h_l) = \frac{1}{2N(h_l)} \sum_{n=1}^{N(h_l)} [y(s_n) - y(s_n + h_l)]^2$$

where $N(h_l)$ denotes # of data pairs whose inter-distances fall in the $l$-th distance class $h_l$; $h_l$ is the separation vector with magnitude $h_l = ||h_l||$.

Plot average distances $\{h_l, l = 1, \ldots, L\}$ versus corresponding sample semivariance values $\{\gamma(h_l), l = 1, \ldots, L\}$; such a plot is called a sample (empirical) semivariogram.
Semi-variogram Example

NDJ 1981-82 average precipitation (in mm)

Sample semivariogram of precipitation

Distance (degrees)
Squared semi-differences cloud

Structure cloud

squared attribute semi-differences vs distance
Empirical Semivariogram

2D sample auto-semivariogram

- Semivariance $\gamma(h)$
- Lag distance $h$
- Semivariogram values: 2D distribution of data points
Conversion between models under 2nd-order stationarity, with $\sigma(0) = \gamma(\infty)$ being the sill of the semivariogram model:

- Semivariogram $\rightarrow$ covariogram: $\sigma(h) = \sigma(0) - \gamma(h)$

- Covariogram $\rightarrow$ correlogram: $\rho(h) = \frac{\sigma(h)}{\sigma(0)}$

- Semivariogram $\rightarrow$ correlogram: $\rho(h) = 1 - \frac{\gamma(h)}{\sigma(0)}$

- Covariogram $\rightarrow$ semivariogram: $\gamma(h) = \sigma(0) - \sigma(h)$
Empirical Correlogram

2D sample auto-correlogram

Correlation $\rho(h)$ vs lag distance $h$ for omnidirectional correlation.

Plot shows the relationship between correlation $\rho(h)$ and lag distance $h$ for omnidirectional correlation.
Conversion between models under 2nd-order stationarity, with $\sigma(0) = \gamma(\infty)$ being the sill of the semivariogram model:

- **Semivariogram $\rightarrow$ covariogram:**
  \[ \sigma(h) = \sigma(0) - \gamma(h) \]

- **Covariogram $\rightarrow$ correlogram:**
  \[ \rho(h) = \frac{\sigma(h)}{\sigma(0)} \]

- **Semivariogram $\rightarrow$ correlogram:**
  \[ \rho(h) = 1 - \frac{\gamma(h)}{\sigma(0)} \]

- **Covariogram $\rightarrow$ semivariogram:**
  \[ \gamma(h) = \sigma(0) - \sigma(h) \]
Recall:

\[
\hat{\sigma}_X = \frac{1}{N} \sum_{n=1}^{N} x_n^2 - \hat{\mu}_X^2
\]

\[
\hat{\sigma}_Y = \frac{1}{N} \sum_{n=1}^{N} y_n^2 - \hat{\mu}_Y^2
\]

\[
\hat{\sigma}_{XY} = \frac{1}{N} \sum_{n=1}^{N} x_n y_n - \hat{\mu}_X \hat{\mu}_Y
\]

Expanding:

\[
\hat{\gamma}_{XY} = \frac{1}{N} \sum_{n=1}^{N} (x_n - y_n)^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n^2 + y_n^2 - 2x_n y_n)
\]

\[
= \frac{1}{N} \sum_{n=1}^{N} x_n^2 + \frac{1}{N} \sum_{n=1}^{N} y_n^2 - 2 \frac{1}{N} \sum_{n=1}^{N} x_n y_n
\]

\[
= \hat{\sigma}_X + \hat{\mu}_X^2 + \hat{\sigma}_Y + \hat{\mu}_Y^2 - 2 \hat{\sigma}_{XY} - 2 \hat{\mu}_X \hat{\mu}_Y
\]

\[
= \hat{\sigma}_X + \hat{\sigma}_Y - 2\hat{\sigma}_{XY} + [\hat{\mu}_X - \hat{\mu}_Y]^2
\]

What’s the difference: To estimate the moment of inertia \(\gamma_{XY}\) you do not need to know the mean values \(\mu_X\) and \(\mu_Y\); these two mean values are required for estimating the covariance \(\sigma_{XY}\).