Spatial Data Analysis

Murray Clayton

University of Wisconsin – Madison

What's so special about spatial data?

e.g. Recall z, t tests
$X_1, X_2, \ldots, X_n$ a random sample from $N(\mu, \sigma^2)$. If σ^2 known, then we can test some hypothesis $H_0 : \mu = \mu_0$ using

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

If σ^2 is unknown, use $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ where $s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$

Assumptions:

- data are normal (or close to it)
- data are independent

An Example

Suppose we want to look at soil-water infiltration and test $H_0 : \mu = 20$.
Suppose we know $\sigma^2 = 5$ and will take $n = 10$. We sample along a transect, with 3 m spacing:

---x---x---x---x---x---x---x---x---x---

We observe $\bar{X} = 21.5$. We use:

$$Z = \frac{21.5 - 20}{\sqrt{5}/\sqrt{10}} = 2.12$$

p-value = 0.034.

The Data May Be Dependent

---x---x---x---x---x---x---x---x---x---x---

Let X_i = observation at ith location.

Suppose corr$(X_i, X_j) = .4^{|i-j|}$

e.g. corr$(X_1, X_2) = .4^{|1-2|} = .4^1 = .4$

corr$(X_4, X_5) = .4^{|4-5|} = .4$

corr$(X_2, X_4) = .4^{|2-4|} = .16$

corr$(X_5, X_8) = .4^{|5-8|} = .064$

In that case, we should be using:

$$Z = \frac{21.5 - 20}{\sqrt{2.111}\sqrt{5}/\sqrt{10}} = 1.46$$

p-value = 0.14

Danger!

Ignoring spatial correlation could lead to a very different conclusion: we could reject H_0 when we should accept H_0.

Spatial correlation shouldn't be ignored.

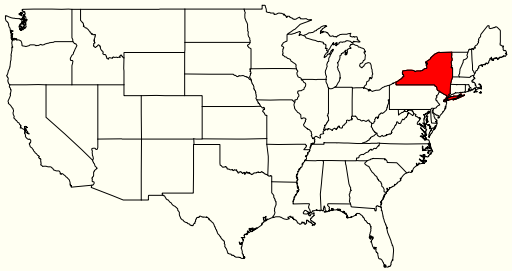
Spatial Clustering

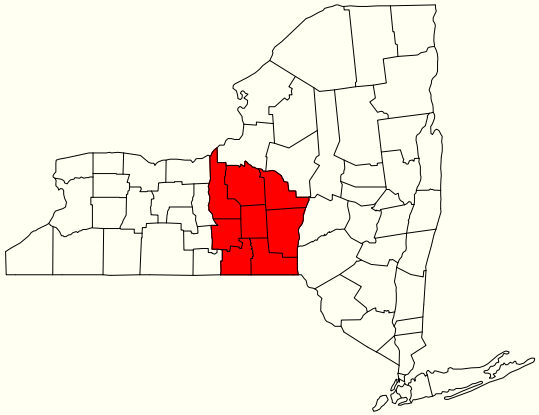
- Detecting and describing spatial clustering of disease

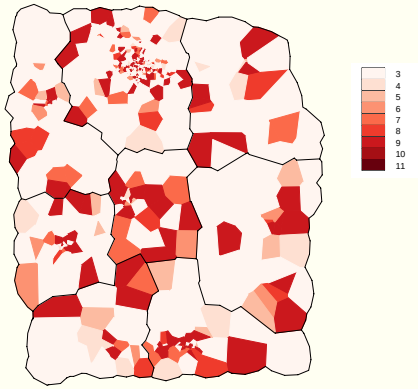
(Ron Gangnon, Ping Yan, Haoda Fu, Junhee Han)

Example: New York Leukemia Data

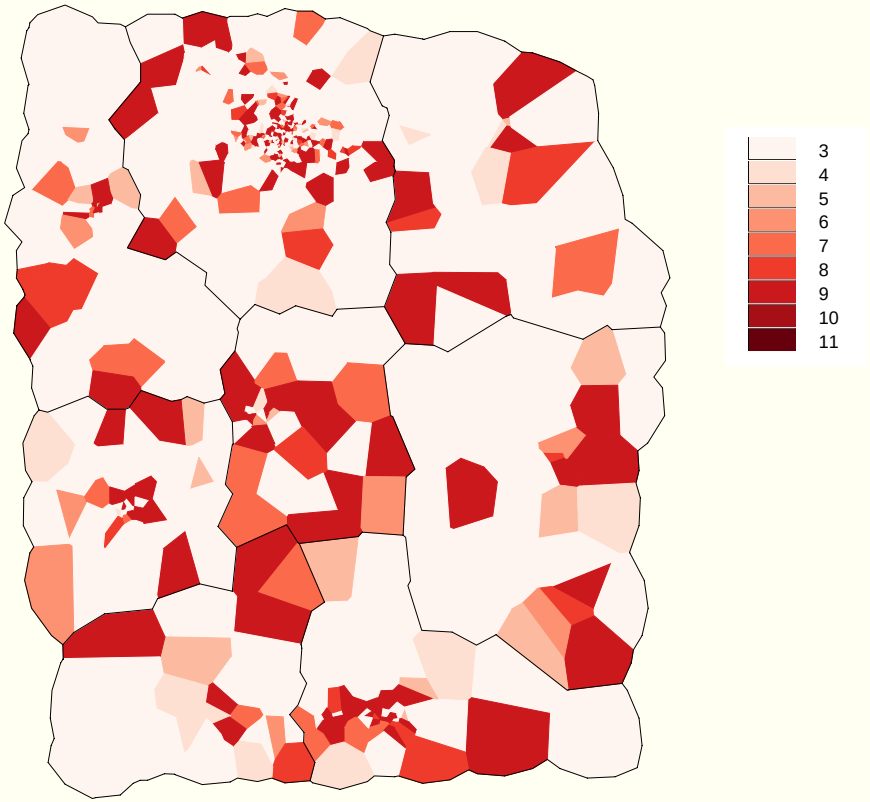
- Eight counties in Upstate New York.
  - Census tracts or census blocks.





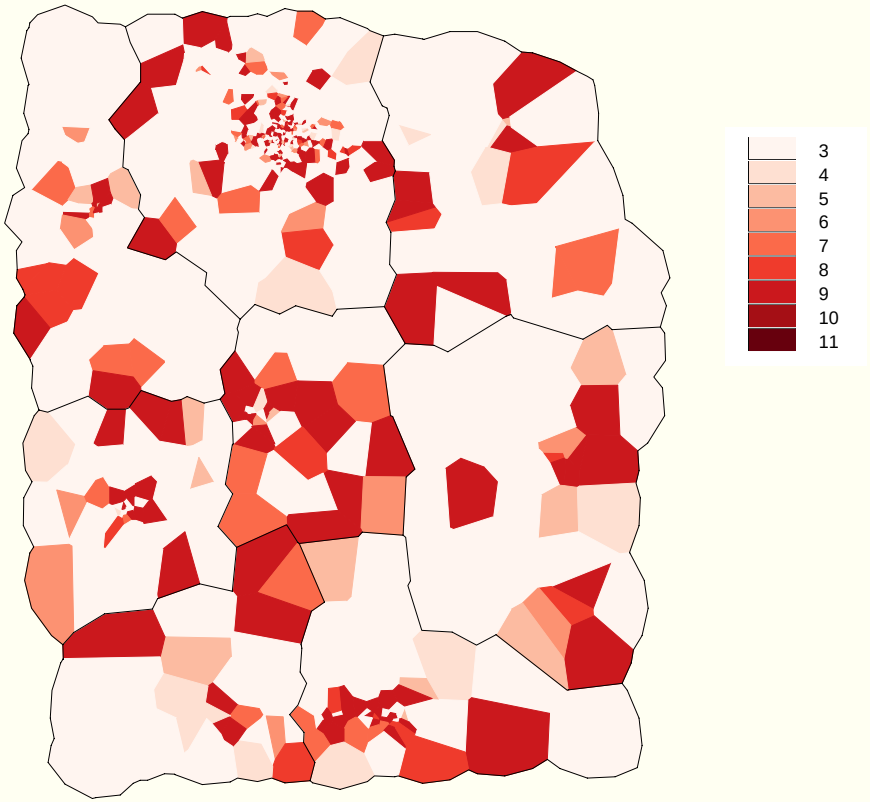


New York Leukemia Data



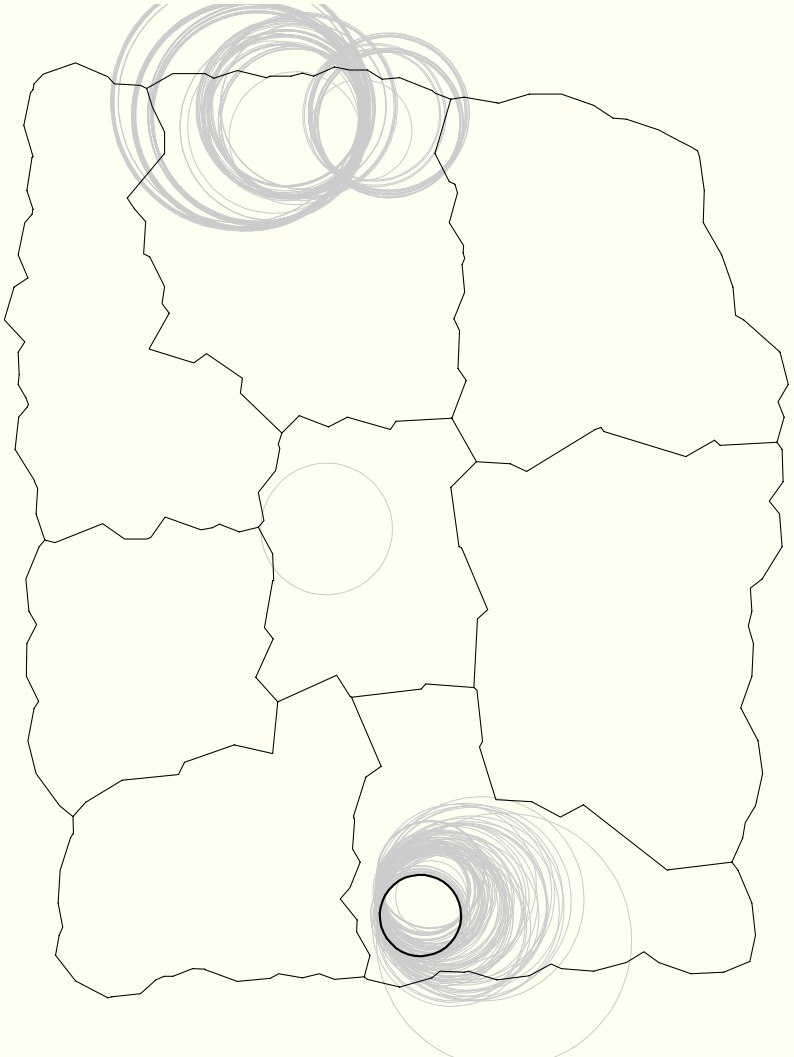
- Is there evidence of clustering?
- If so, where are the clusters located?

A Scan Statistic (Kulldorf)



- Set of possible clusters: e.g. all circles of diameter $\leq$ 20 km; with circle centers on a dense grid within study region.

New York Leukemia Data – Scan Statistic

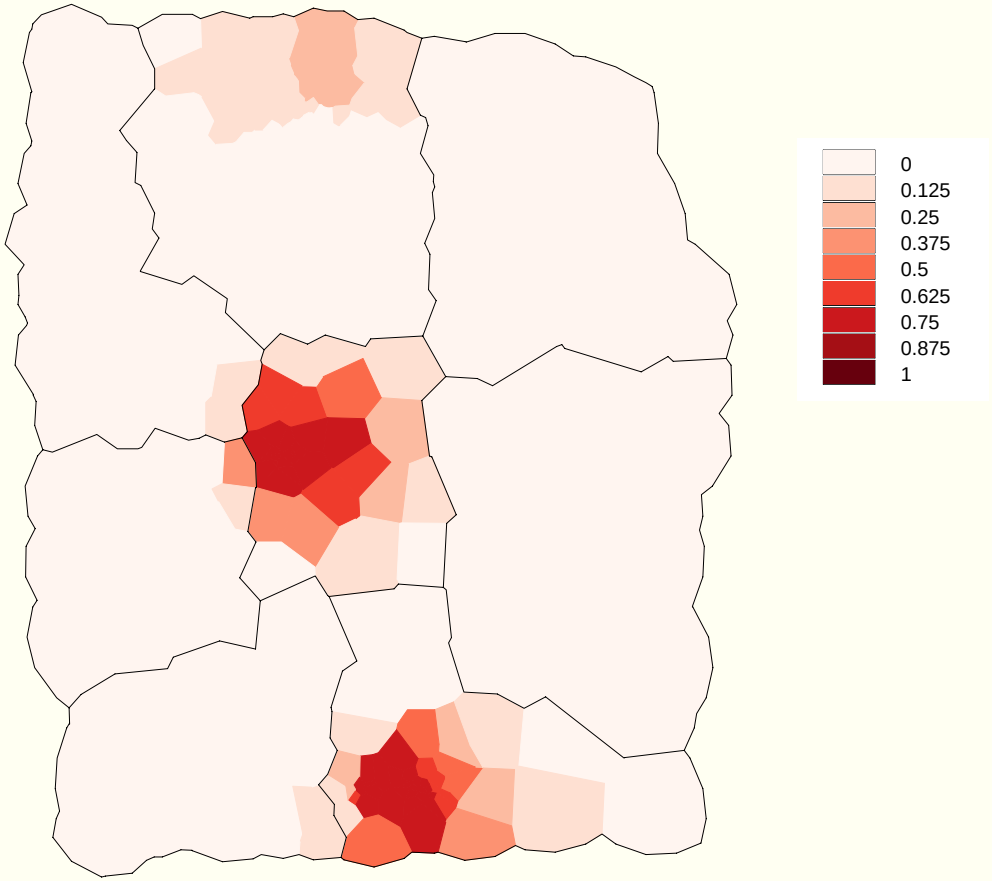


Problem: The scan statistic doesn't seem to use all of the evidence available

Hierarchical Model

- k possibly overlapping clusters.
- $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$ denote sets of cells belonging to each cluster.
- $$\log(\rho_i) = \mu + \sum_{j=1}^{k} \theta_j \mathbf{1}_{\{i \in \mathbf{c}_j\}} + \epsilon_i$$
- Assume prior, and use RJMCMC to simulate posterior. (A sort of scan statistic.)

Posterior Prob of Cluster Membership



RJMCMC (Variable k)

Extensions

- space-time
- MCCF
- computational issues

Regression Models for Spatial Images

Murray Clayton

April 13, 2010

Gypsy Moth Defoliation Example

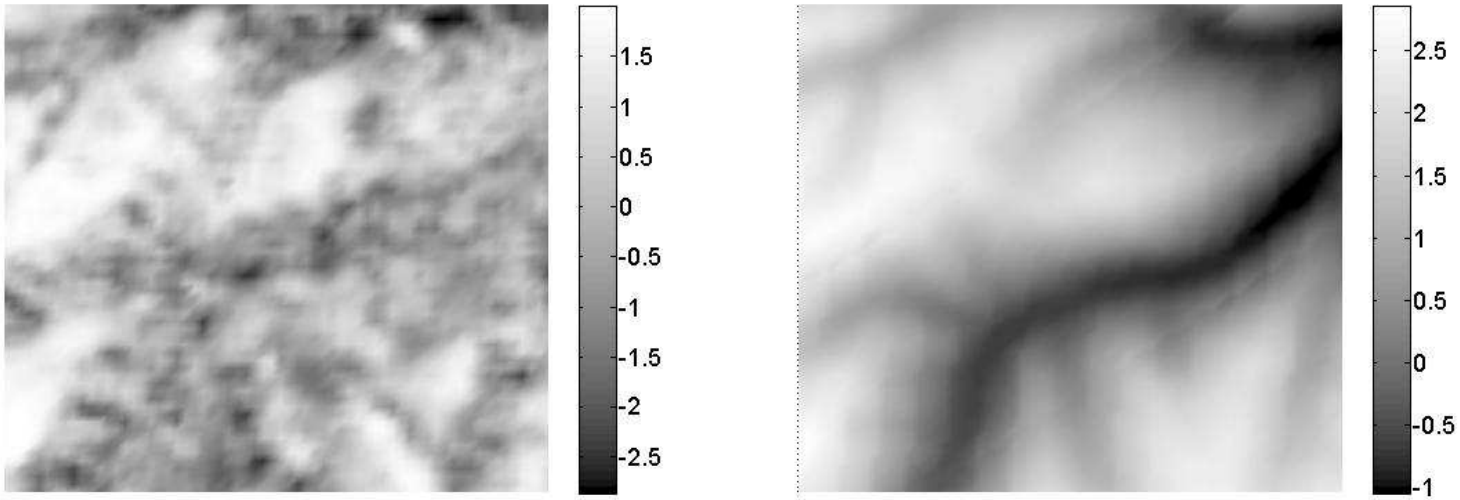


Figure: *Left: satellite image for gypsy moth defoliation rates. Right: elevation. Researchers have observed that gypsy moth defoliation rates increase with increased elevation. We can see that defoliation rates are generally higher on ridges than in the valleys.*

Modeling Possibilities

Each pixel has a location (s, t).
We could fit a model like:

$$y(s,t) = a + b\,x(s,t) + e(s,t)$$

But we probably don't believe it's right. (For two reasons.)

We Doubt a Geostatistical Model, Too

$$y(s,t) = a + b\,x(s,t) + e(s,t)$$

where the $e(s,t)$ are spatially correlated.

How About a Point-wise Model?

$$y(s,t) = a(s,t) + b(s,t)x(s,t) + e(s,t)$$

Functional Concurrent Linear Model (FCLM)

Assume we have n pairs of images, (Y_i, X_i), $i = 1, \ldots, n$, and $n \geq 1$ and consider the model

$$y_i(s,t) = a(s,t) + x_i(s,t)b(s,t) + e_i(s,t)$$

which we rewrite in matrix form:

$$Y_i = A + X_i \circ B + E$$

and expand A and B with a 2-D discrete wavelet expansion

$$Y_i = \sum_{j=1}^{H} v_j \phi_j + X_i \circ \left\{ \sum_{j=1}^{H} w_j \phi_j \right\} + E = \sum_{j=1}^{H} v_j \phi_j + \sum_{j=1}^{H} w_j (X_i \circ \phi_j) + E$$

We can construct a shrinkage procedure based on LASSO:

Consider a standard regression:
$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e.$

LASSO

$$\min \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{1i} + \cdots + b_k x_{ki})]^2 + \lambda (\sum_{j=1}^{k} |b_j|)$$

Gypsy Moth Defoliation Example

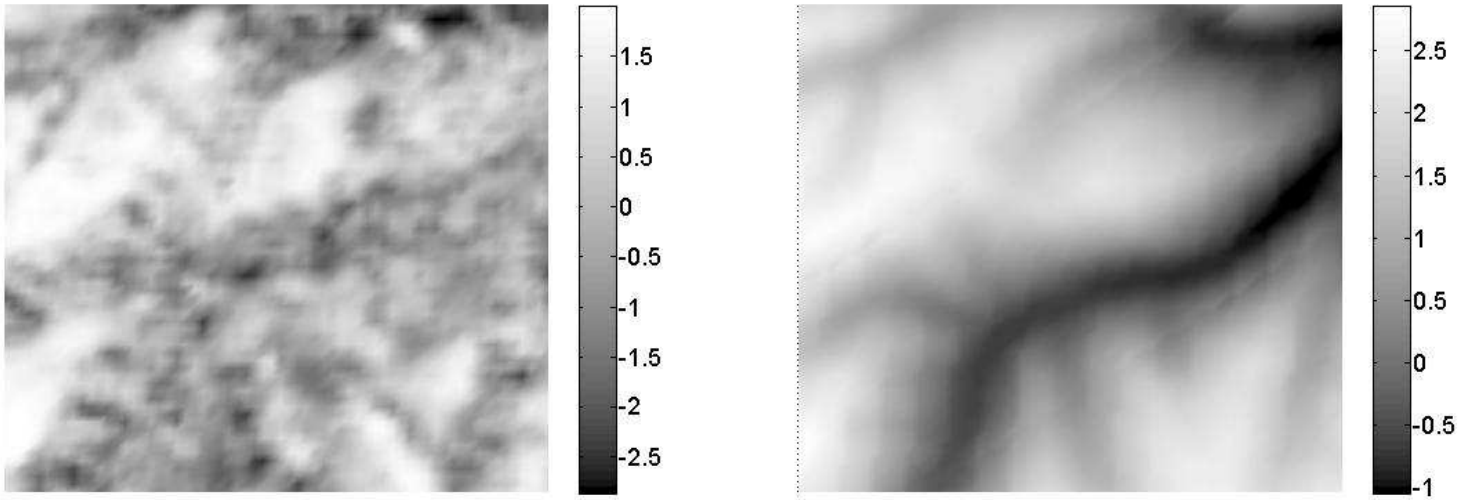
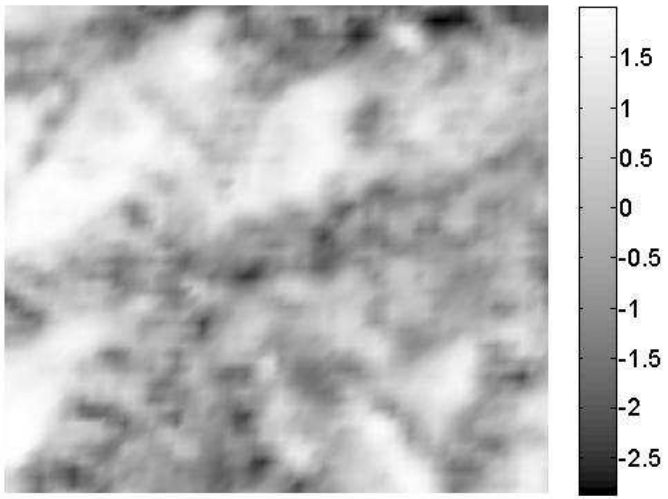
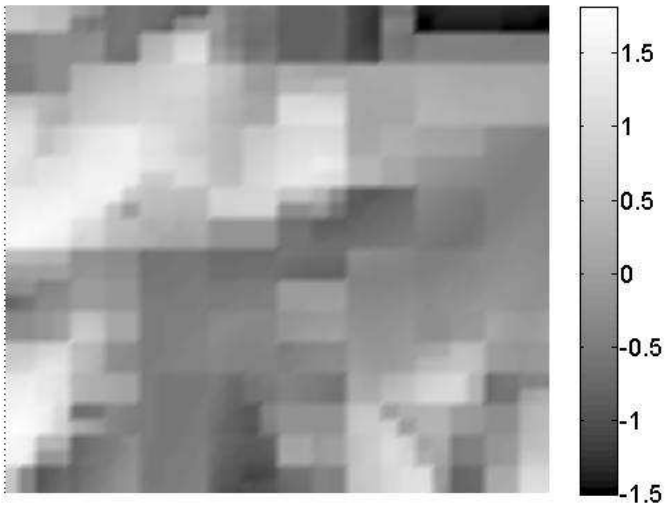


Figure: *Left: satellite image for gypsy moth defoliation rates. Right: elevation. Researchers have observed that gypsy moth defoliation rates increase with increased elevation. We can see that defoliation rates are generally higher on ridges than in the valleys.*

Model 1: $Y = A + X_1 \circ B_1 + E$



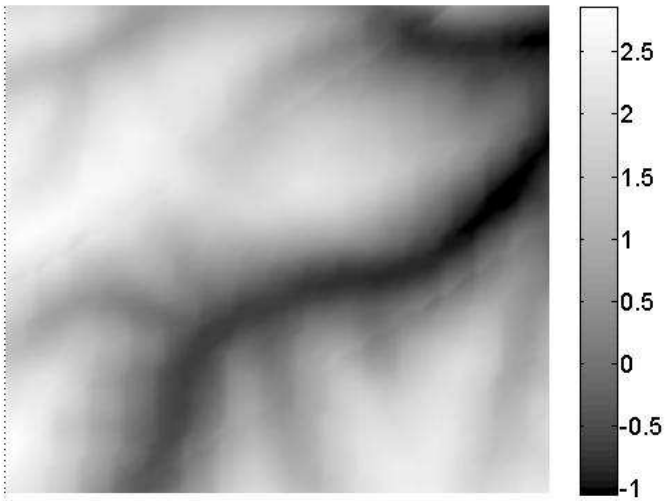
Y: defoliation rate



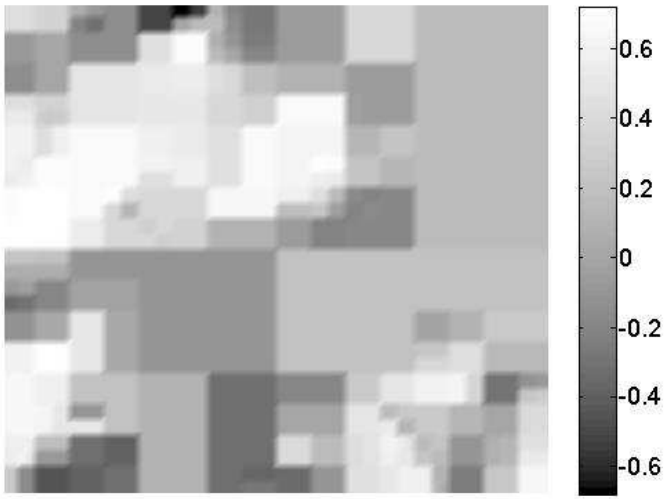
$\widehat{Y}$

Figure: *The model is* $Y = A + X_1 \circ B_1 + E$. *Elevation is the explanatory variable* X_1. *Defoliation rate is the response* Y.

Model 1: $Y = A + X_1 \circ B_1 + E$



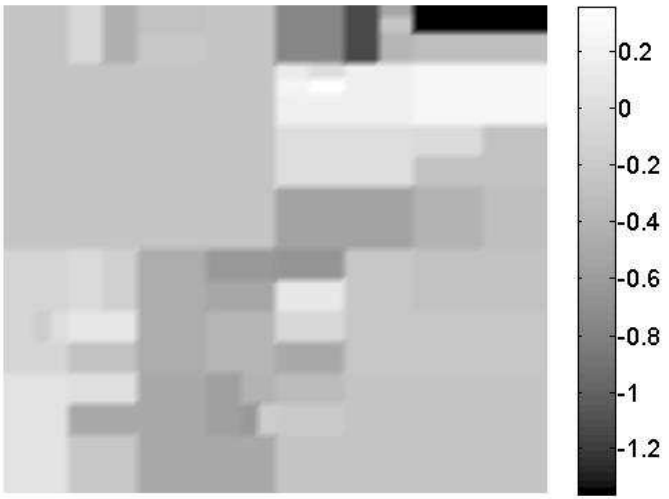
X_1:elevation



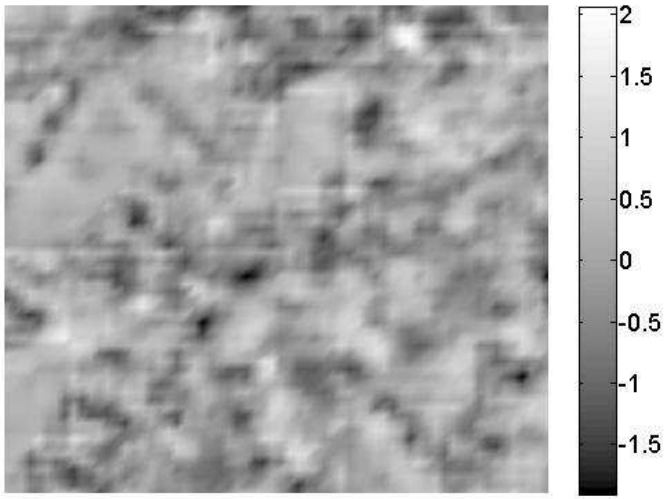
$\widehat{B}_1$

Figure: *The model is* $Y = A + X_1 \circ B_1 + E$. *Elevation is the explanatory variable* X_1. *Defoliation rate is the response* Y.

Model 1: $Y = A + X_1 \circ B_1 + E$



$\widehat{A}$



residual plot

Figure: *The model is* $Y = A + X_1 \circ B_1 + E$. *Elevation is the explanatory variable* X_1. *Defoliation rate is the response* Y.

What Else?

- Can add more covariates.
- Residual plots are important.
- Residual plots are not sufficient (degenerate cases).
- Missing data.
- Inference.