



CENTER FOR
STATISTICAL ECOLOGY AND
ENVIRONMENTAL STATISTICS

	Stone-Age	Space-Age	Syndrome
	Stone-age data	Space-age data	
Stone-age analysis	+	+	

Spatial Scan Statistic for Geographical and Network Hotspot Detection

C. Taillie and G. P. Patil

Center for Statistical Ecology and Environmental Statistics
Penn State University

Joint Statistical Meetings

Toronto, Canada

August 11, 2004



Examples of Hotspot Analysis

Spatial

- Disease surveillance
- Biodiversity: species-rich and species-poor areas
- Geographical poverty analysis

Network

- Water resource impairment at watershed scales
- Water distribution systems, subway systems, and road transport systems
- Social Networks/Terror Networks

Issues in Hotspot Analysis

- *Estimation:*

Identification of areas having unusually high (or low) response

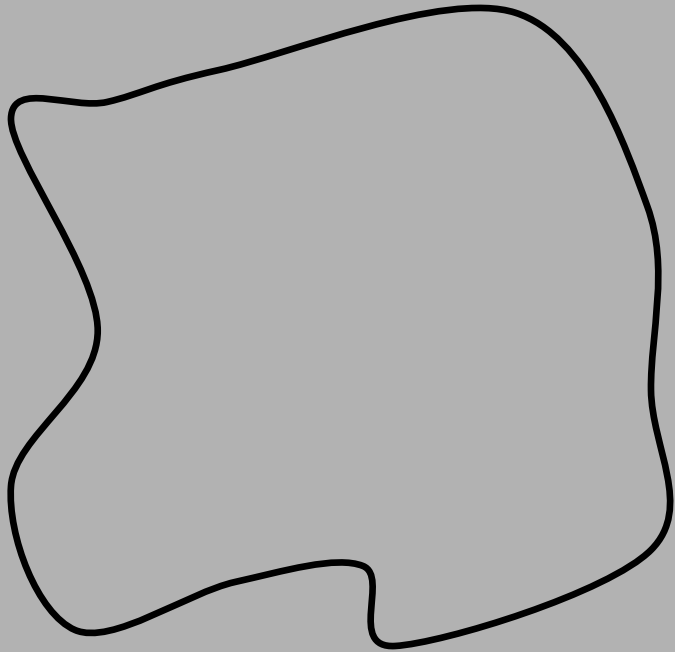
- *Testing:*

Can the elevated response be attributed to chance variation (*false positive*) or is it statistically significant?

- *Explanation:*

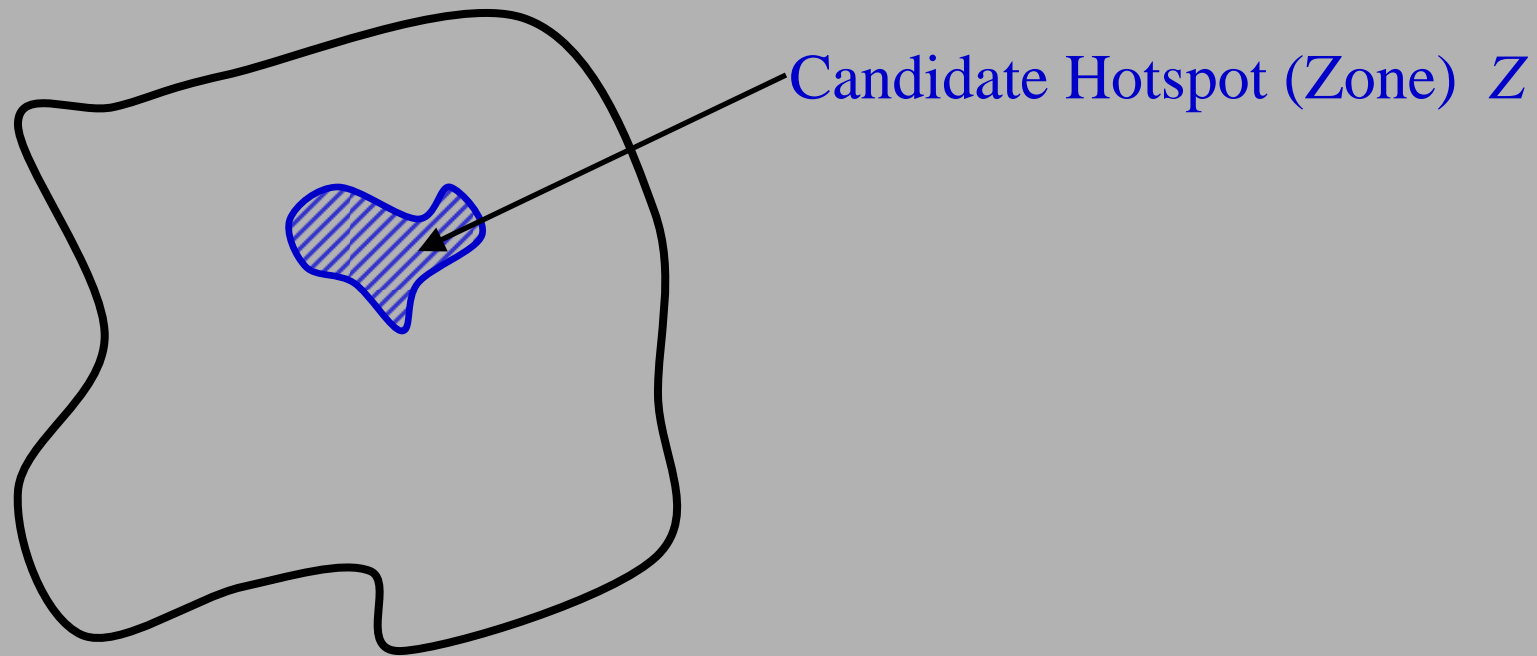
Assess explanatory factors that may account for the elevated response

Spatial Scan Statistic Model



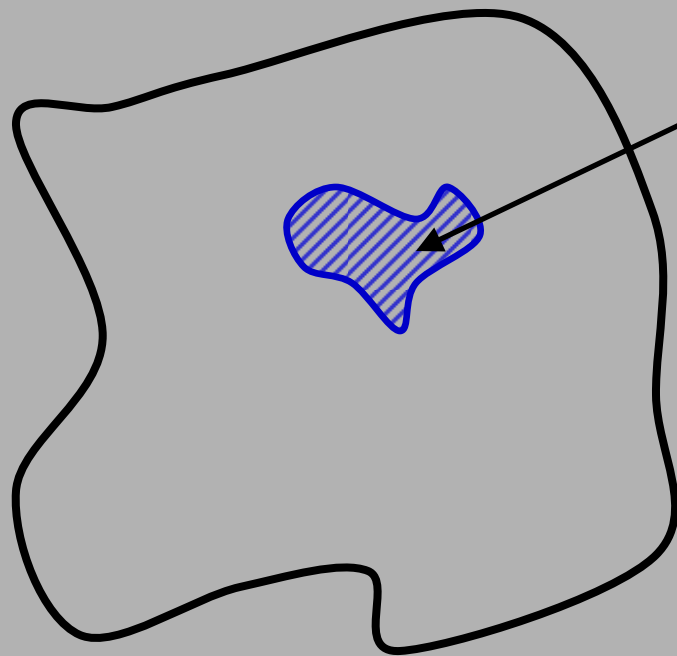
Study Area
spatially distributed
response

Spatial Scan Statistic Model



Study Area
spatially distributed
response

Spatial Scan Statistic Model



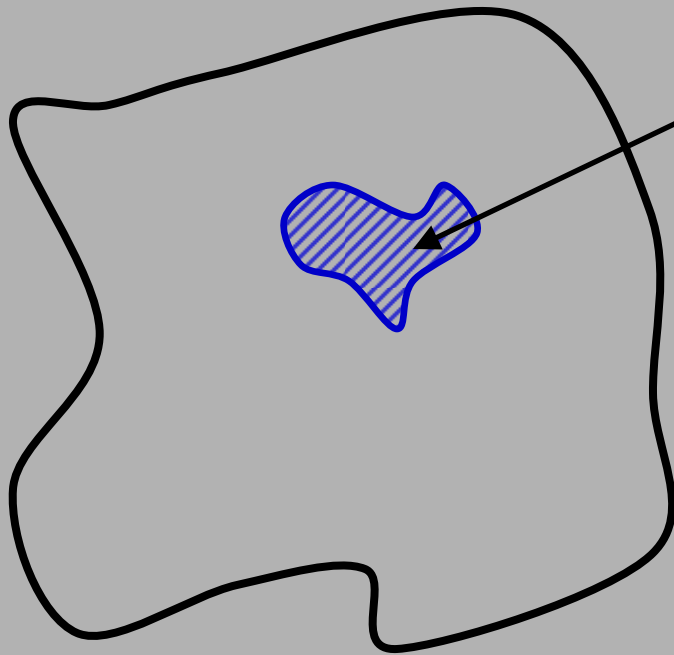
Study Area
spatially distributed
response

Candidate Hotspot (Zone) Z

Given Z , assume:

- response is spatially homogeneous inside Z and outside Z , but with different mean values
- response inside Z and outside Z described by parametric distributions (binomial or Poisson in disease surveillance)

Spatial Scan Statistic Model



Study Area
spatially distributed
response

Candidate Hotspot (Zone) Z

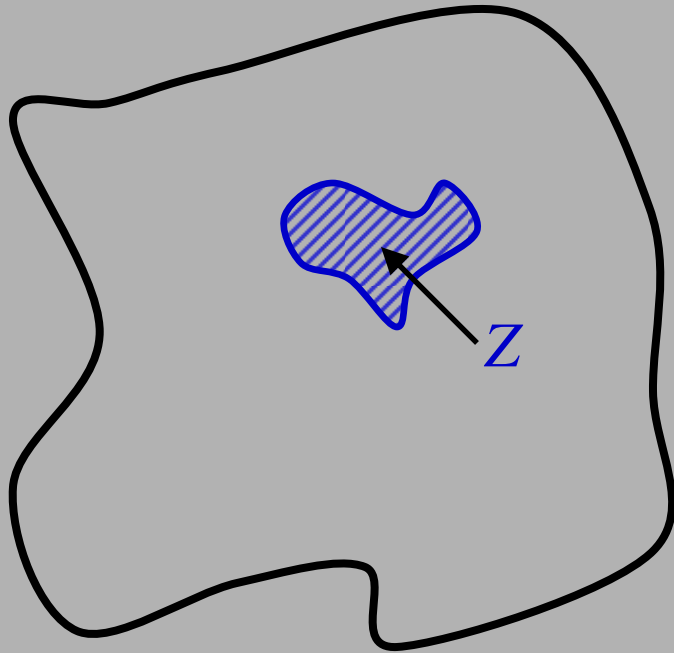
Given Z , assume:

- response is spatially homogeneous inside Z and outside Z , but with different mean values
- response inside Z and outside Z described by parametric distributions (binomial or Poisson in disease surveillance)

Likelihood: $L(Z, p_0, p_1)$

Key Idea: Z is an *unknown* parameter

Likelihood Estimation



Study Area
spatially distributed
response

Maximize $L(Z, p_0, p_1)$

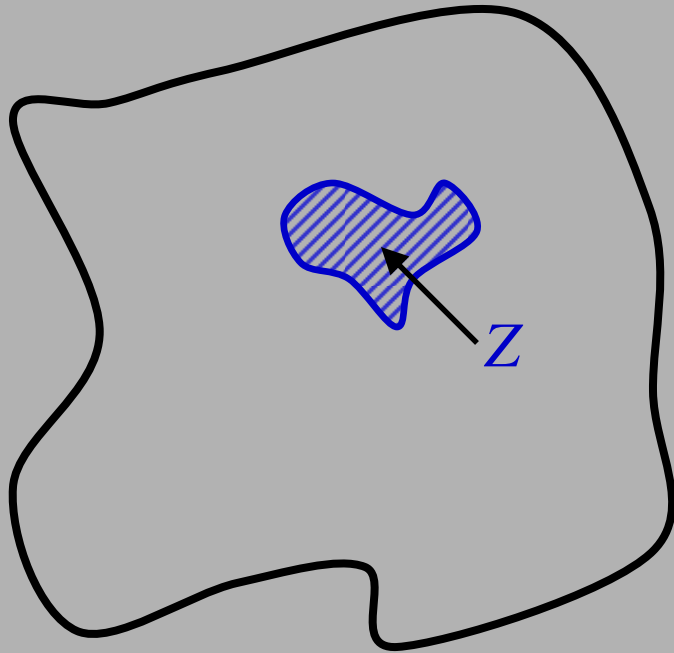
Two stages:

- Fix Z , maximize wrt the conventional parameters.
Gives the *profile likelihood* for Z

$$L(Z) = \max_{p_0, p_1} L(Z, p_0, p_1) = L(Z, \hat{p}_0, \hat{p}_1)$$

- Maximize $L(Z)$ across all candidate zones Z

Likelihood Estimation



Study Area
spatially distributed
response

Maximize $L(Z, p_0, p_1)$

Two stages:

- Fix Z , maximize wrt the conventional parameters.
Gives the *profile likelihood* for Z

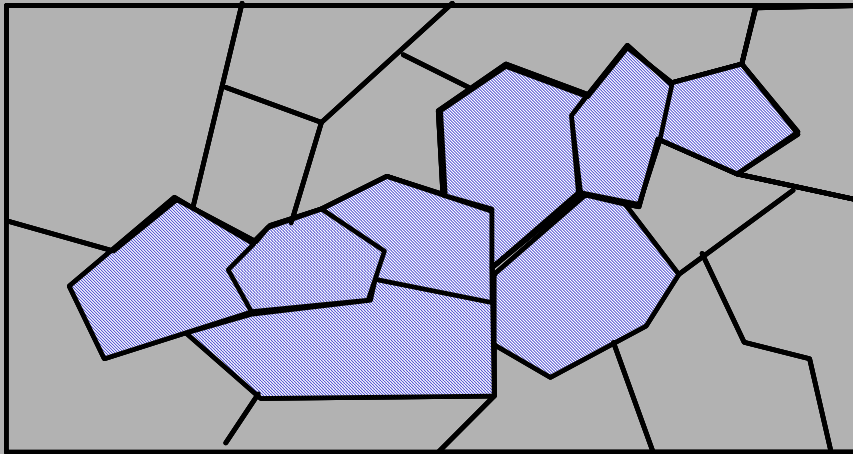
$$L(Z) = \max_{p_0, p_1} L(Z, p_0, p_1) = L(Z, \hat{p}_0, \hat{p}_1)$$

- Maximize $L(Z)$ across all candidate zones Z

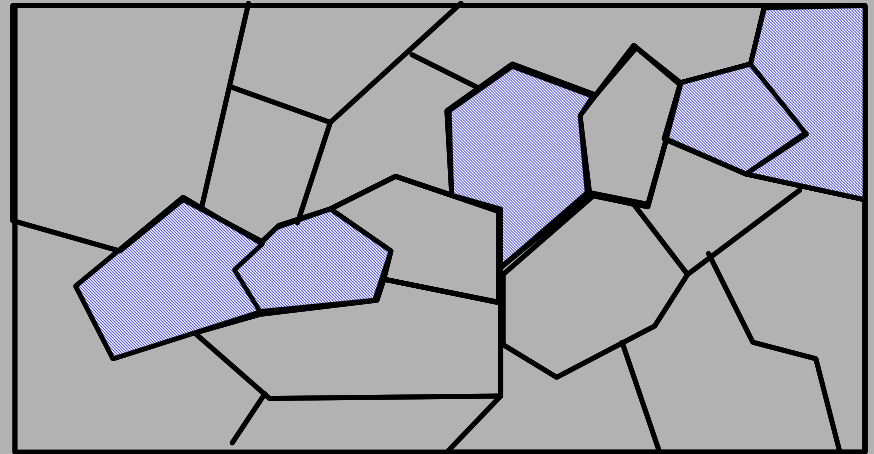
What is a candidate zone ?

Candidate Zones Z

- Tessellated study area
- Want zones to be connected



Allowable zone



Not a zone

- Ω = collection of all (allowable) zones
= collection of all connected unions of cells
- Maximize $L(Z)$ for Z in Ω

Maximize $L(Z)$ for Z in Ω

- Ω is a **finite** set

Maximize $L(Z)$ for Z in Ω

- Ω is a **finite** set --- but usually too big for exhaustive search

Maximize $L(Z)$ for Z in Ω

- Ω is a **finite** set --- usually too big for exhaustive search

Possible strategies:

- **Search space reduction**

Replace Ω by a smaller set Ω_0 and do an exhaustive search across Ω_0

Maximize $L(Z)$ for Z in Ω

- Ω is a **finite** set --- usually too big for exhaustive search

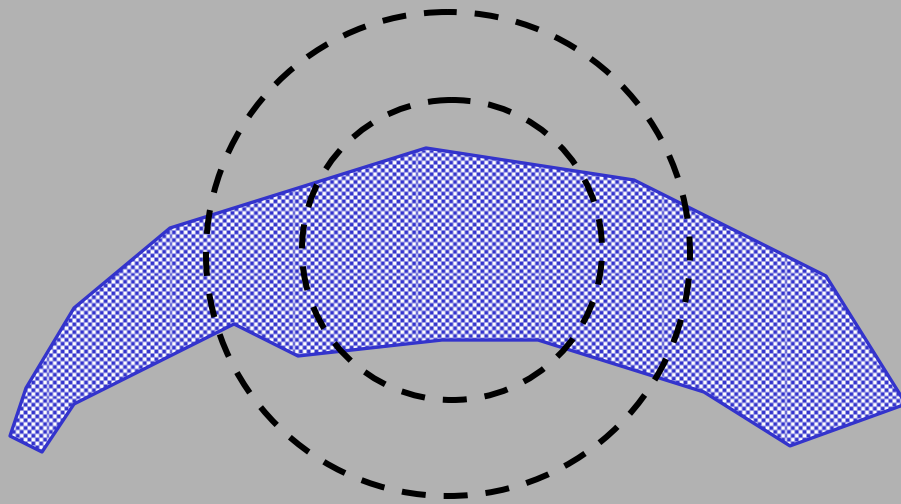
Possible strategies:

- **Search space reduction**

Replace Ω by a smaller set Ω_0 and do an exhaustive search across Ω_0

➤ **Circles (Martin Kulldorff)**

Poor Hotspot Delineation by Circular Zones

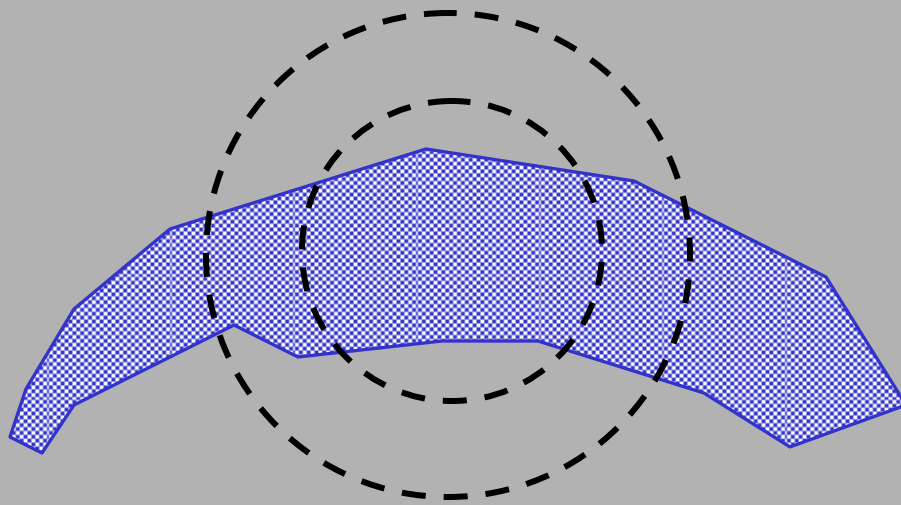


Hotspot




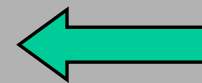
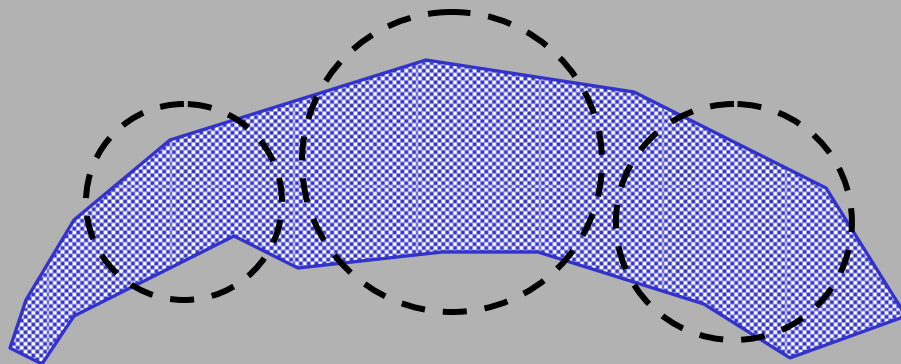
Circular zone
approximations

Poor Hotspot Delineation by Circular Zones



 Hotspot

 Circular zone approximations



Circular zones may represent single hotspot as multiple hotspots

Maximize $L(Z)$ for Z in Ω

- Ω is a **finite** set --- usually too big for exhaustive search

Possible strategies:

- **Search space reduction**

Replace Ω by a smaller set Ω_0 and do an exhaustive search across Ω_0

- Circles (Martin Kulldorff)
- Ellipses

Maximize $L(Z)$ for Z in Ω

- Ω is a **finite** set --- usually too big for exhaustive search

Possible strategies:

- **Search space reduction**

Replace Ω by a smaller set Ω_0 and do an exhaustive search across Ω_0

- Circles (Martin Kulldorff)
- Ellipses
- Upper level sets

Maximize $L(Z)$ for Z in Ω

- Ω is a **finite** set --- usually too big for exhaustive search

Possible strategies:

- **Search space reduction**

Replace Ω by a smaller set Ω_0 and do an exhaustive search across Ω_0

- Circles (Martin Kulldorff)

- Ellipses

- Upper level sets

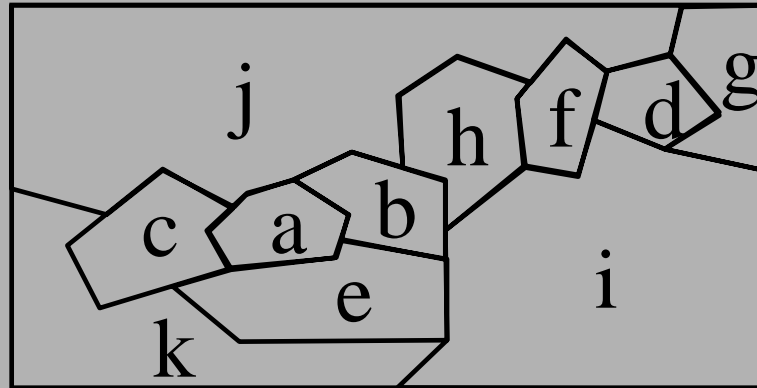
- **Stochastic optimization**

- Simulated annealing (Luis Duczmal)

- Genetic algorithms

Spatial Scan Statistic Setup

- Tessellation of a geographic region



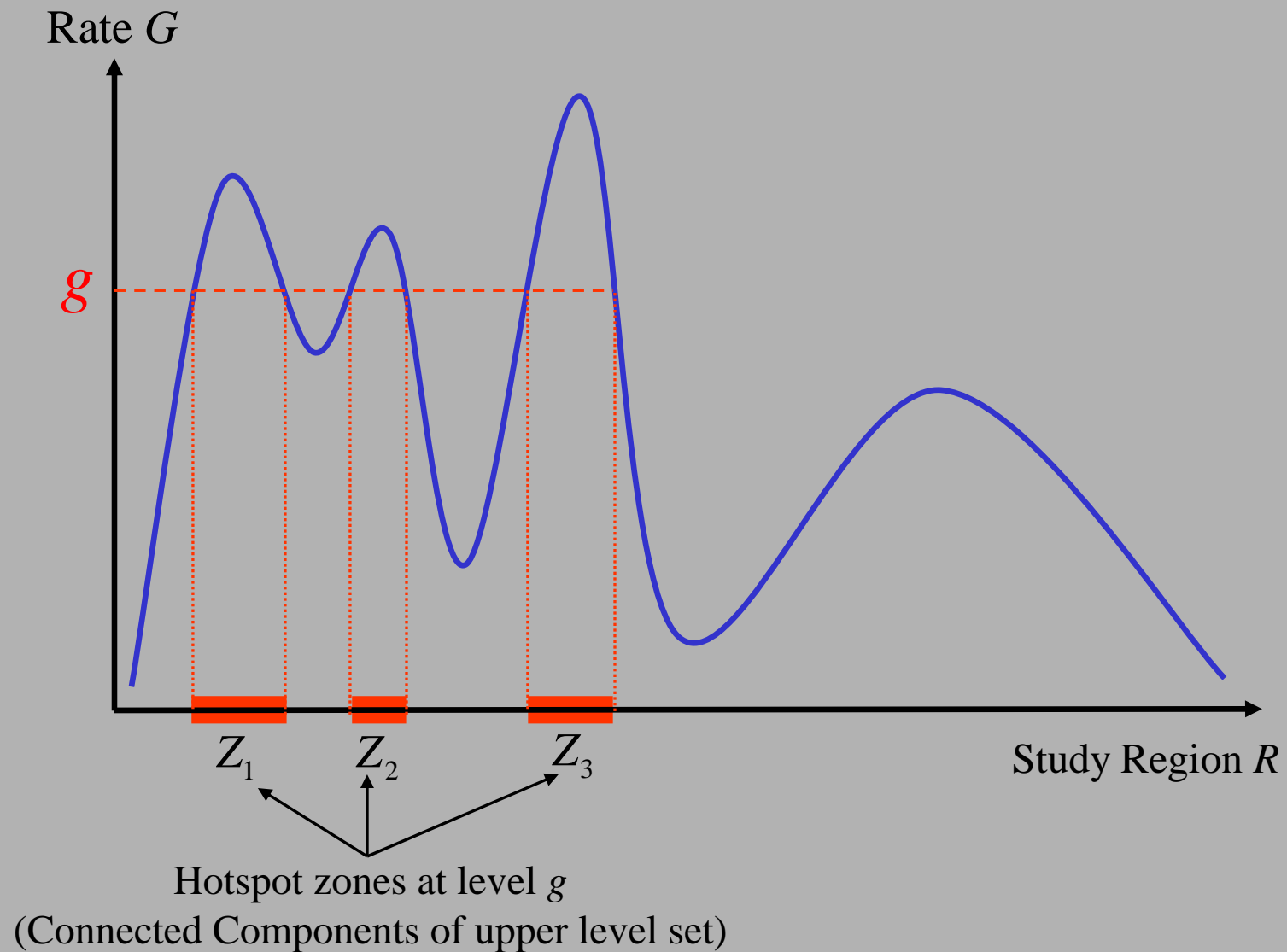
a, b, c, ...
are cell labels

- Region R , Tessellation $T = \{a\}$ of R
- Cell a , Response Y_a , Cell “Size” A_a
- Two distributional settings:
 - Y_a is **Binomial** (N_a, p_a), $A_a = N_a$, $p_a = \text{cell rate}$
 - Y_a is **Poisson** ($\lambda_a A_a$), $\lambda_a = \text{cell rate}$
- Cell sizes A_a are known and fixed
- Cell responses Y_a , $a \in A$, are independent

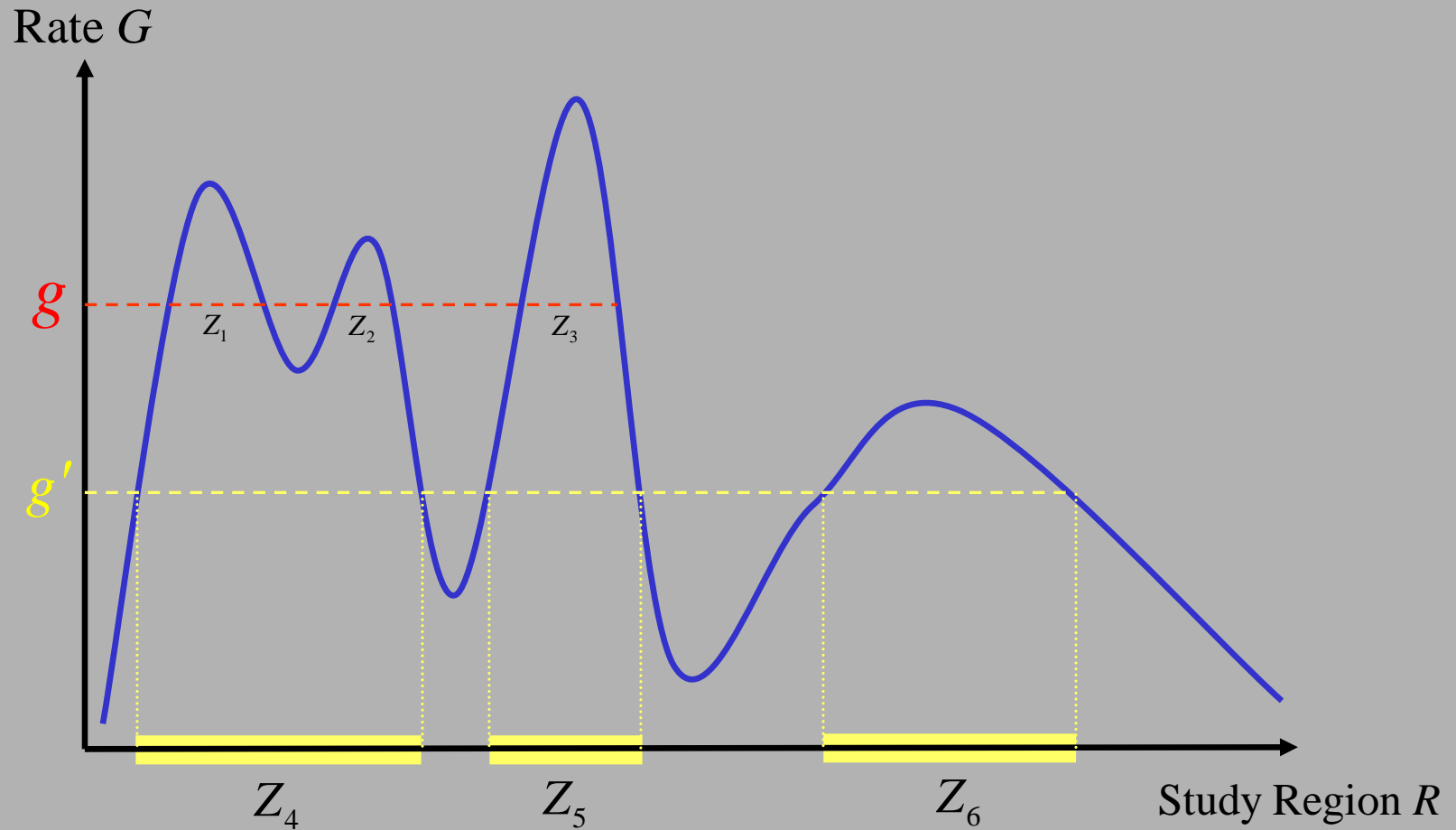
Upper Level Set (ULS)

- **Data-adaptive** approach to reduced parameter space Ω_0
- Zones in Ω_0 are **connected components** of **upper level sets** of the empirical rate function $G_a = Y_a / A_a$
- Upper level set (ULS) at level g consists of all cells a where
$$G_a \geq g$$
- Upper level sets may be disconnected. Connected components are the candidate zones in Ω_0
- These connected components form a rooted tree under set inclusion.
 - Root node = entire region R
 - Leaf nodes = local maxima of empirical rates
 - Junction nodes occur when two zones coalesce

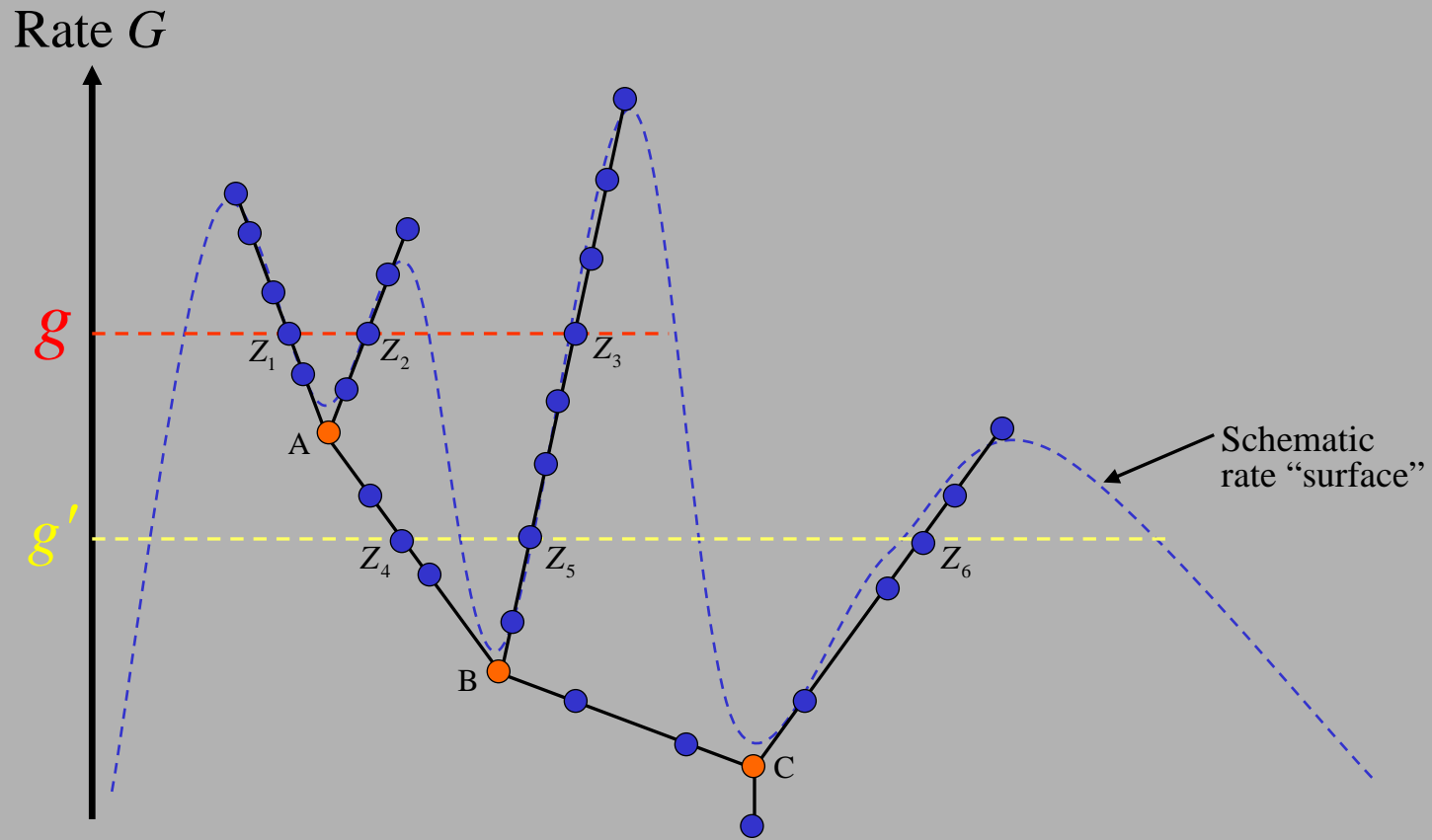
Upper Level Sets (ULS) of Response Surface



Changing Connectivity of ULS as Level Drops

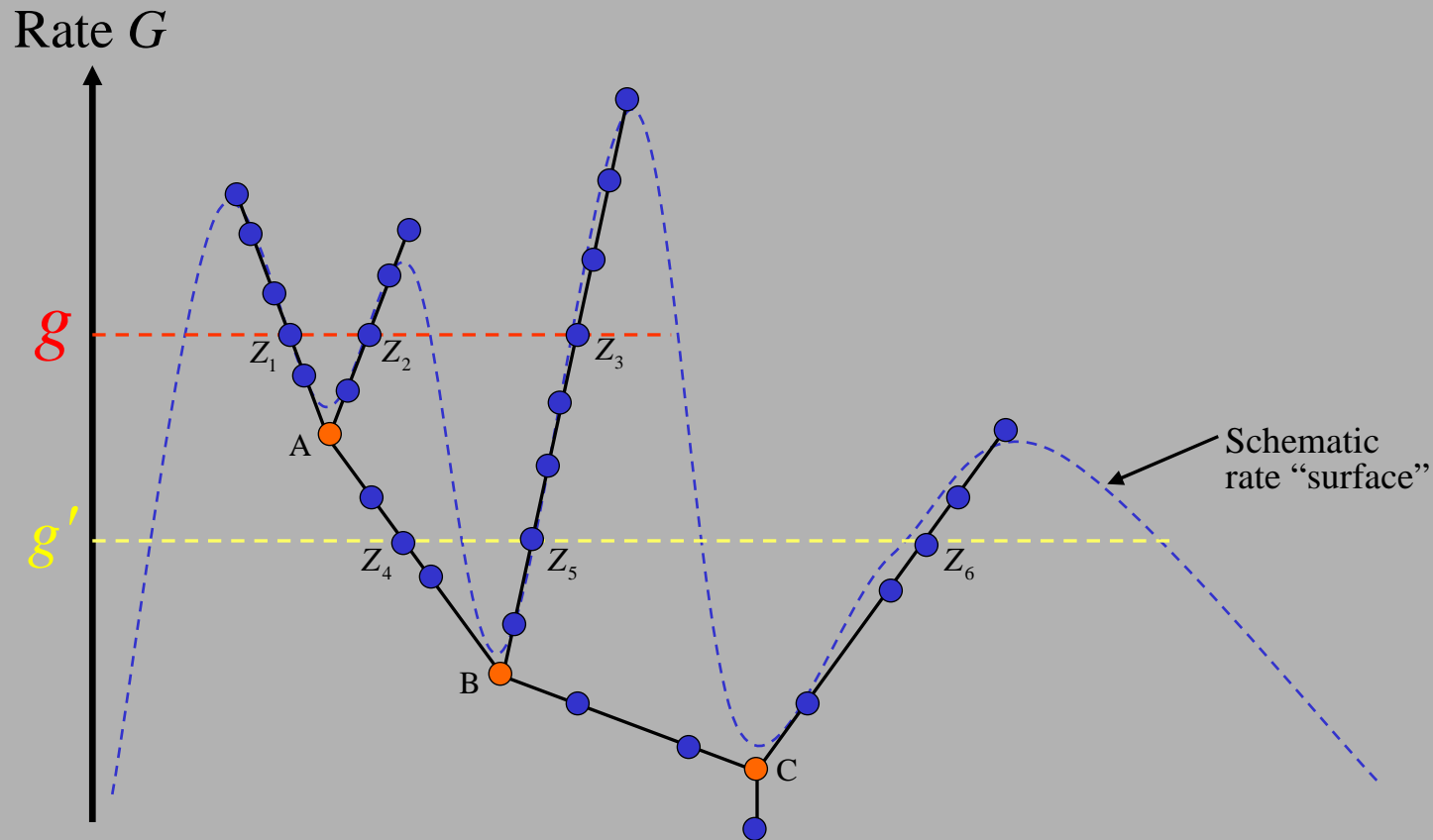


ULS Zonal Tree



A, B, C are junction nodes where multiple zones coalesce into a single zone

ULS Zonal Tree

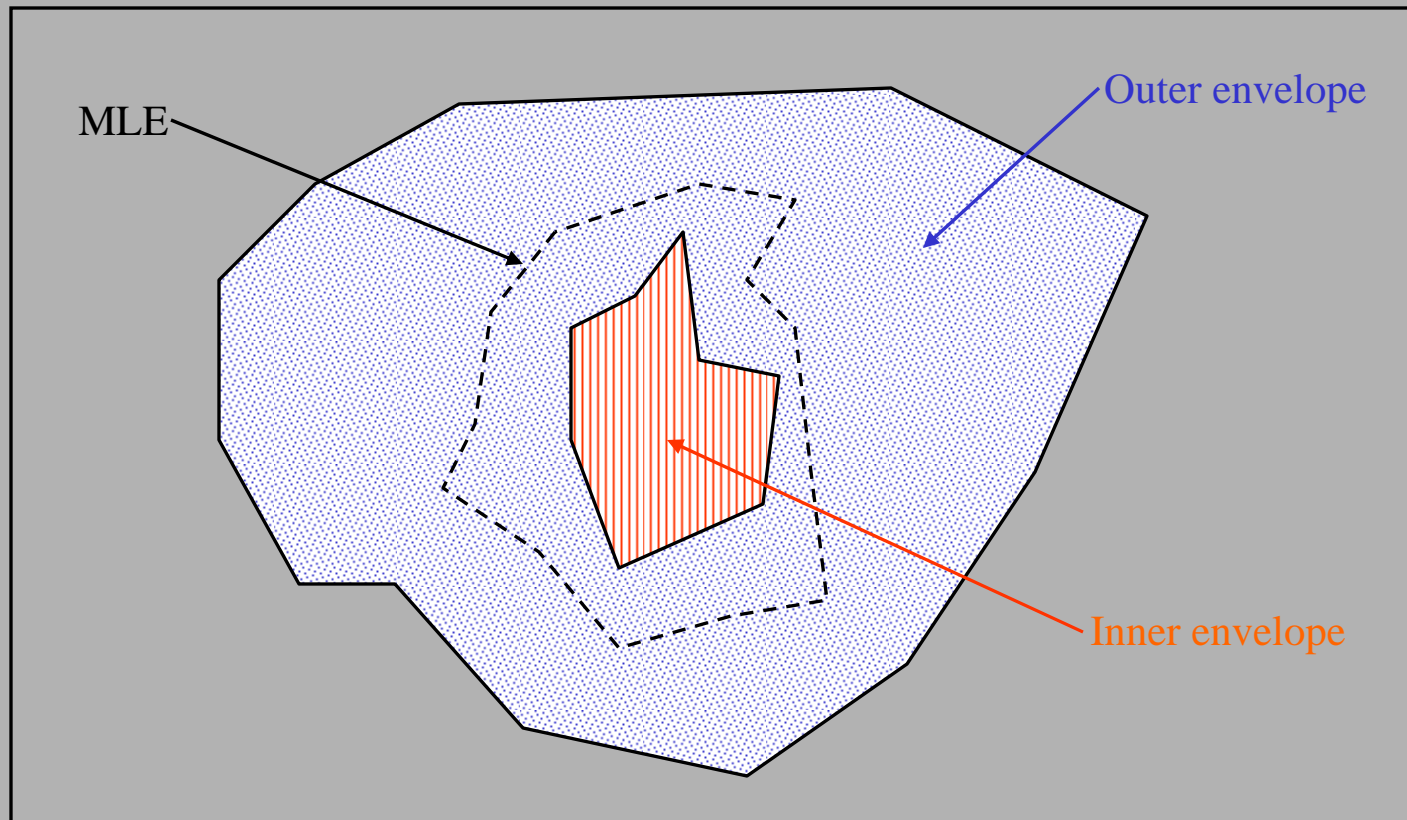


- Maximize profile likelihood $L(Z)$ across the tree
- Typical behavior of $L(Z)$
- Number of nodes does not exceed number of cells in tessellation

Confidence Sets and Hotspot Rating

- Determine a **confidence set** for the hotspot
- Each member of the confidence set is a zone which is a **statistically plausible** delineation of the hotspot at specified confidence
- Confidence set lets us **rate individual cells** a for hotspot membership
- Rating for cell a is percentage of zones in confidence set that contain a
- Map of cell ratings:
 - Inner envelope = cells with 100% rating
 - Outer envelope = cells with positive rating

Hotspot Rating



Confidence Set Determination

- Confidence set is all null hypotheses that cannot be rejected

- As hypotheses, use $\tilde{H}_0 : \text{hotspot } Z = Z_0$
 $\tilde{H}_1 : \text{hotspot } Z \neq Z_0$

where $Z_0 \in \Omega_0$ is a given zone.

- Confidence set is all $Z_0 \in \Omega_0$ for which \tilde{H}_0 cannot be rejected.

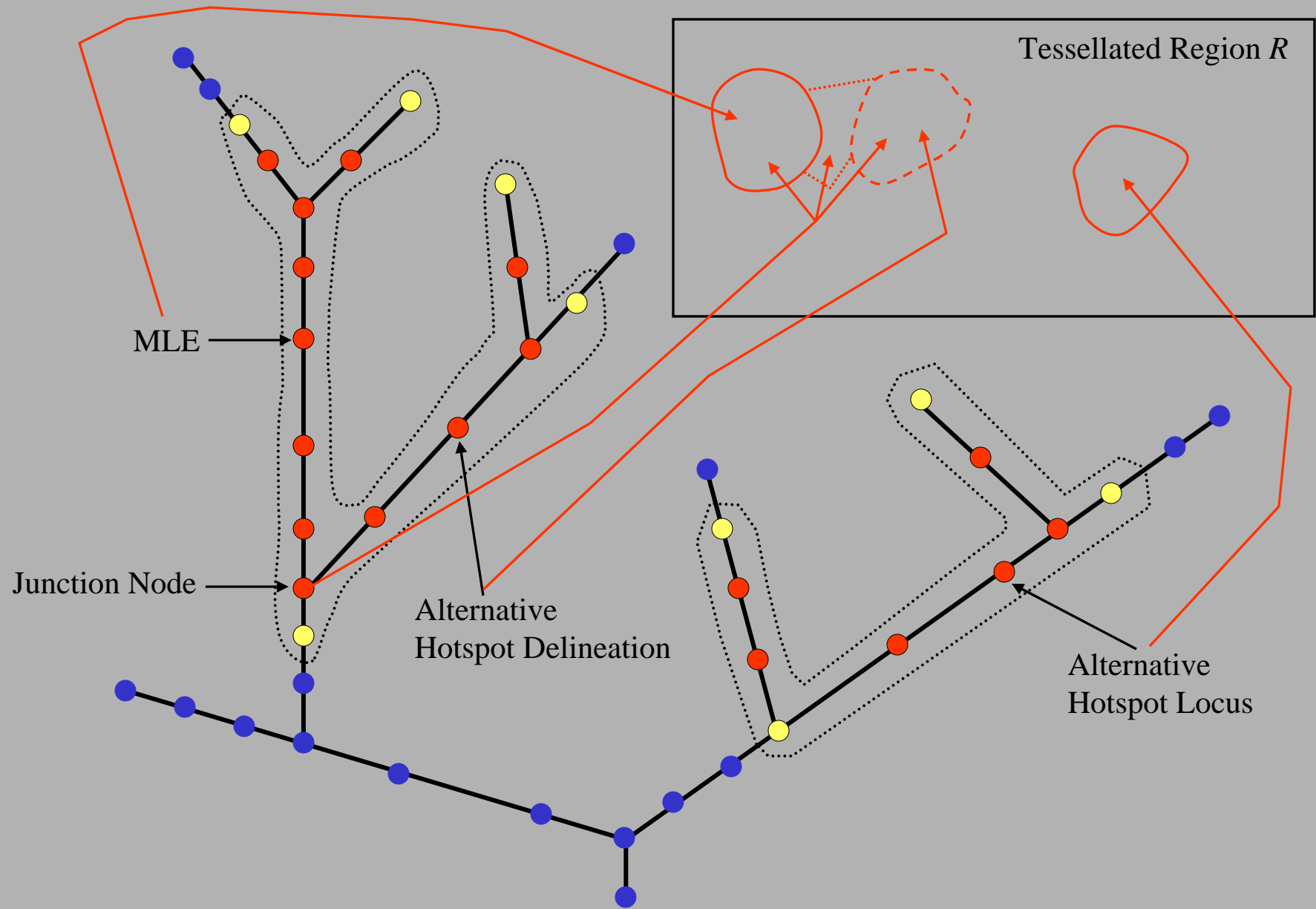
- Likelihood ratio test:

Test statistic: $LR = L(Z_0) / L(\hat{Z})$ where $\hat{Z} = \text{MLE under } \tilde{H}_0 \cup \tilde{H}_1$

Reject H_0 when LR is small

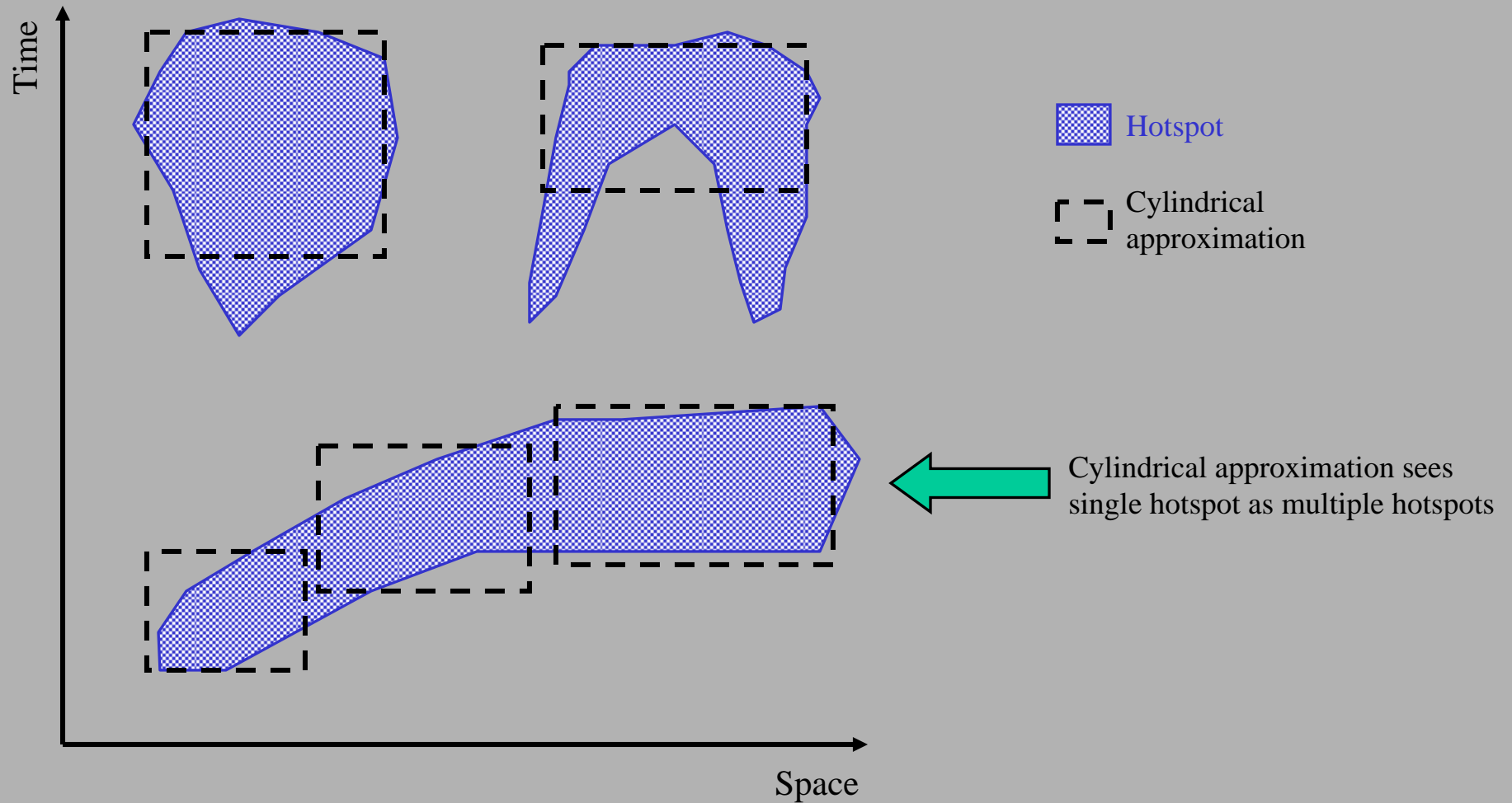
- Null distributions have to be determined by simulation

Confidence Region on ULS Tree

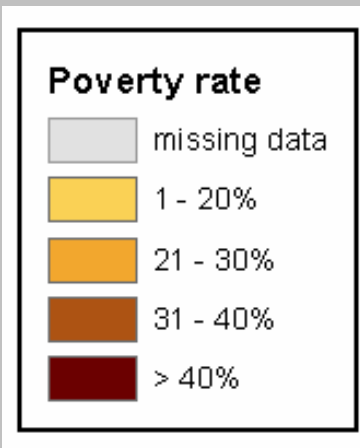
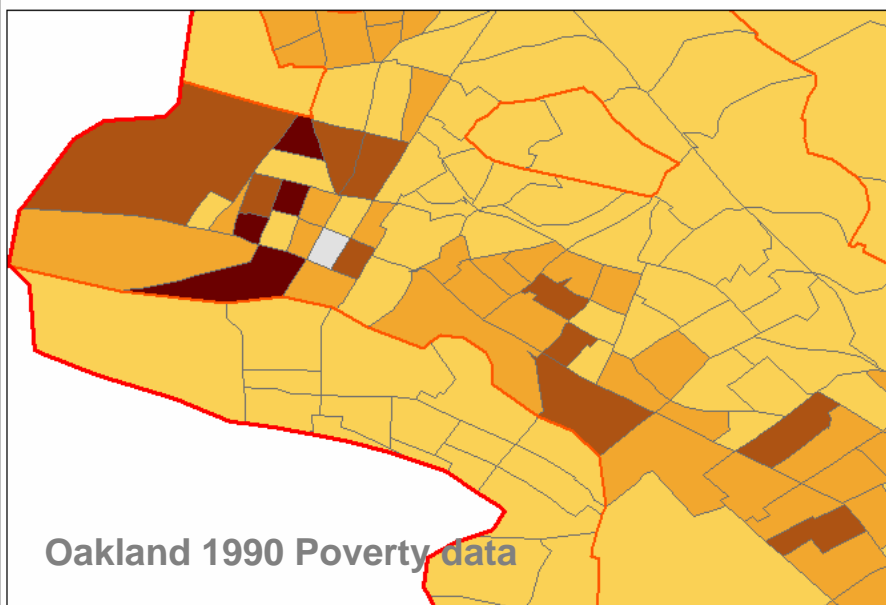
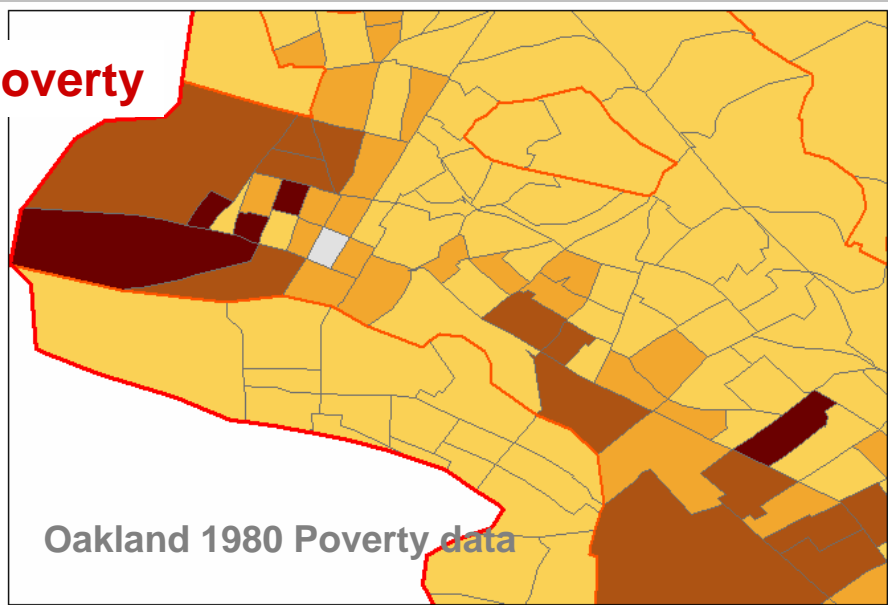
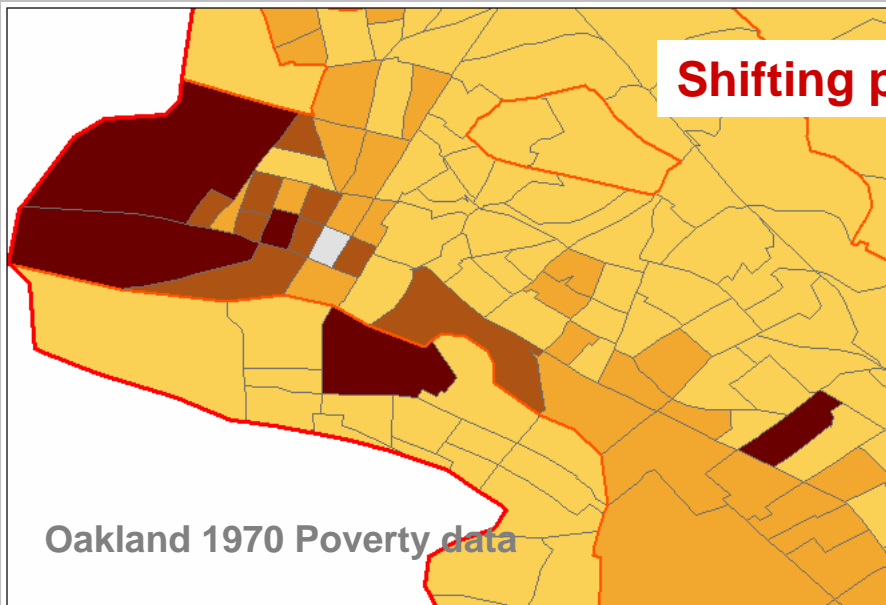


Space-Time Generalizations

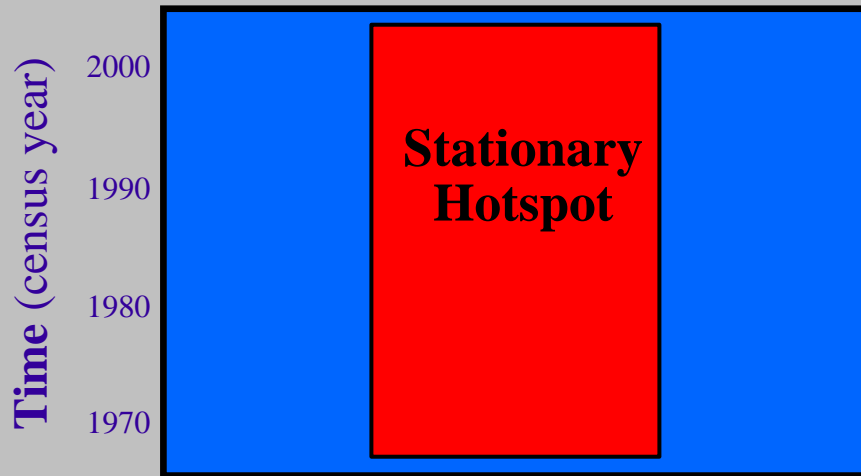
Some Space-Time Hotspots and Their Cylindrical Approximations



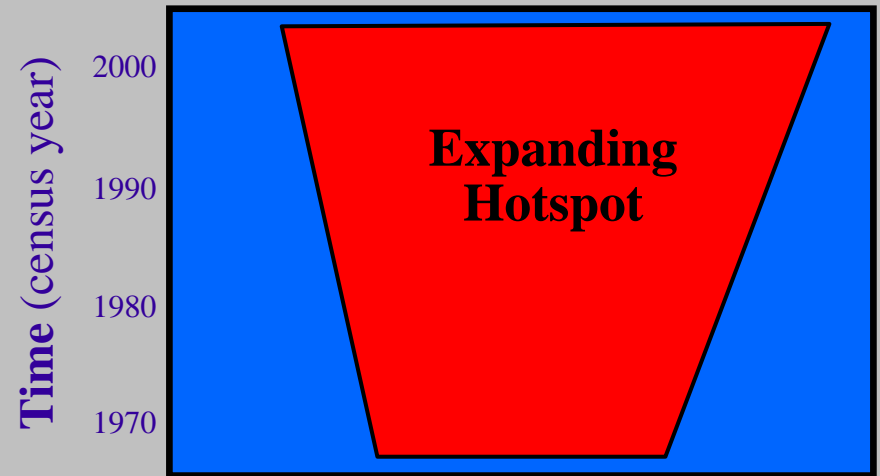
Shifting poverty



Typology of Space-Time Hotspots



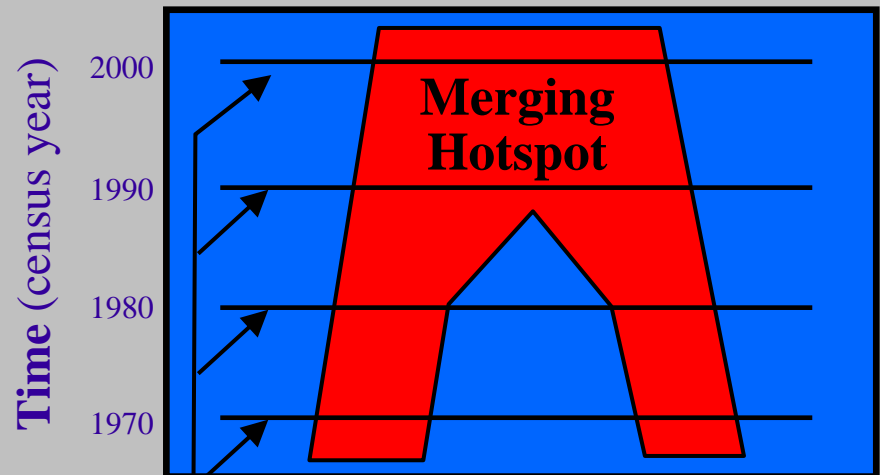
Space (census tract)



Space (census tract)



Space (census tract)

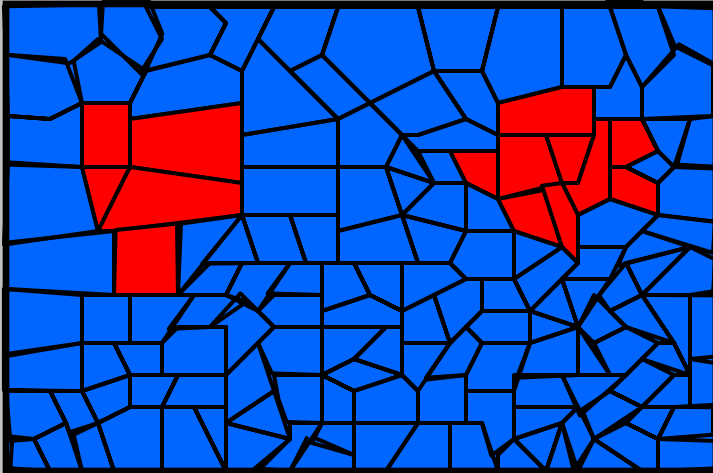


Space (census tract)

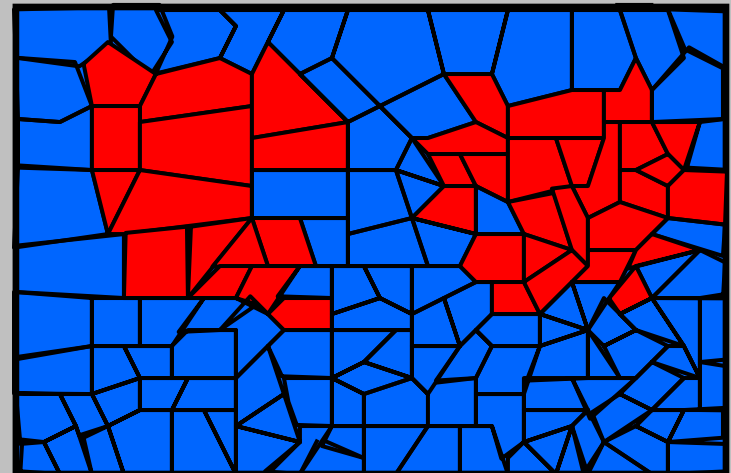
Time slices of
space-time hotspot

Trajectory of a Merging Hotspot

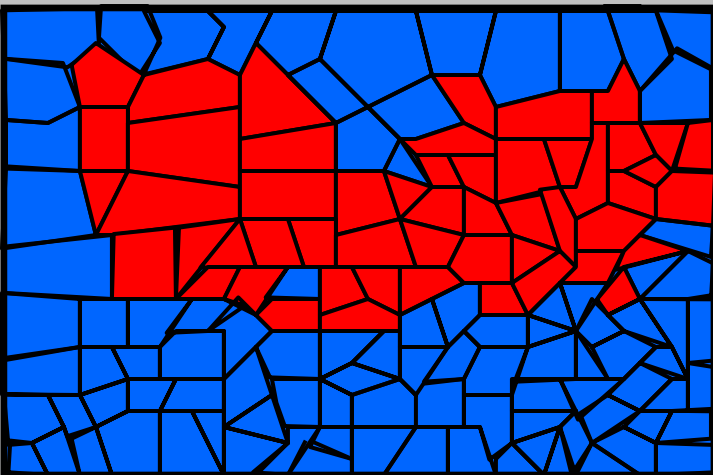
1970



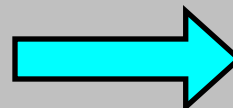
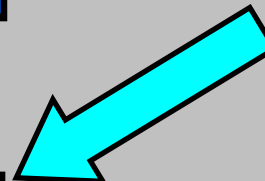
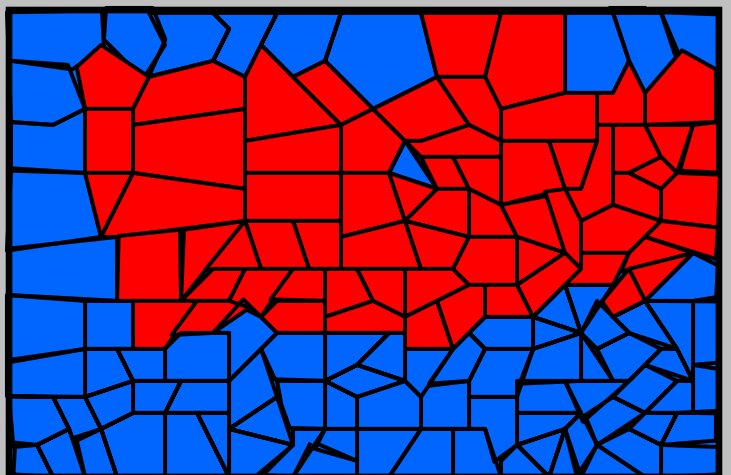
1980



1990



2000



Hotspot Detection for Continuous Responses

- Human Health Context:
 - Blood pressure levels for spatial variation in hypertension
 - Cancer survival (censoring issues)
- Environmental Context:
 - Landscape metrics such as forest cover, fragmentation, etc.
 - Pollutant loadings
 - Animal abundance

Hotspot Model for Continuous Responses

- Simplest distributional model:

$$Y_a \sim \text{Gamma}(k, \beta)$$

- Additivity with respect to the index parameter k suggests that we model k as proportional to size:

$$k_a = A_a / c.$$

- Scale parameter β takes one value inside Z and another outside Z
- Other distribution models (e.g., lognormal) are possible but are computationally more complex and applicable to only a single spatial scale

Circles vs ULS

- Circles capture only compactly shaped clusters
 - ❖ Want to identify clusters of arbitrary shape
- Circles provide point estimate of hotspot
 - ❖ Want to assess estimation uncertainty (hotspot confidence set)
- Circles handle only synoptic (tessellated) data
 - ❖ Want to also handle data on a network

Features of ULS Scan Statistic

- Identifies arbitrary shaped clusters
- Applicable to data on a network
- Confidence set, hotspot rating
- Computationally efficient
- Applicable to continuous response
- Generalizes to space-time scan