# Upper level set scan statistic for detecting arbitrarily shaped hotspots

G. P. Patil & C. Taillie
Journal of Environmental and Ecological Statistics, Vol. 11, Issue 2, 2004 pp 183-197

Lee Panter

# Introduction & Background

- Three central problems in geographical surveillance for a spatially distributed response variable
  - Identify areas with exceptionally high (low) response
  - Determine whether the exceptional response cases can be attributed to chance variation (false alarm) or are statistically significant
  - Assessment of explanatory factors related to the response anomalies

# Introduction & Background (continued)

- Comparative Questioning
  - Can the Upper Level Set (ULS) statistic be used to answer any of the above questions more: accurately, precisely, or efficiently?
  - Can the ULS statistic be applied to the same (or more varied) applications as the Spatial Scan Statistic?
  - In what scenarios (if any) does the ULS statistic exhibit a higher power than the Spatial Scan Statistic? Lower?

# Definitions and Terminology

- R: A geographically connected region
- T: A set of "cells" forming a partition/tessellation of R
- N: Cardinality of T (also written: $|T|$ )
- $n_1, \ldots, n_N$: cell "sizes"
- $Y_1, \ldots, Y_N$: Responses of interest over cells modeled with independent random variable with realizations $y_1, \ldots, y_N$
- $G_i = \frac{Y_i}{n_i}$: cell "rates" with corresponding realizations $g_i = \frac{y_i}{n_i}$
- $\mathbf{Z}$: A non-empty, connected, zone in R. (A connected union of cells)
- NOTE: Notation is adopted from: PULSE, progressive upper level set scan statistic for geospatial hotspot detection. (G. P. Patil, S. W. Joshi, and R. E. Koli) to avoid long-term confusion.

# Level Sets

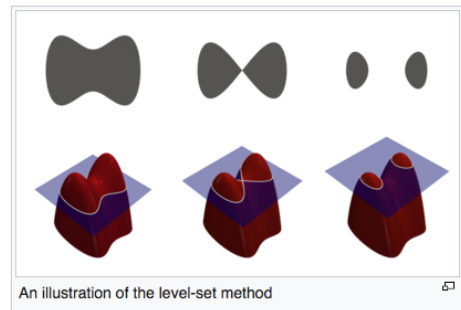
- Definition: Level Set

  A level set of a function $f : \mathbb{R}^n \to \mathbb{R}$ is a set of the form:

  $$L_g = \{(x_1, \ldots, x_N) | f((x_1, \ldots, x_N)) = g\}$$

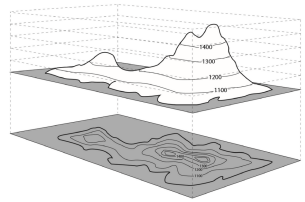
  i.e. the set where the function takes on a given value g.

- When $n = 2$ the level set for each value of g are curves in $\mathbb{R}^2$



An illustration of the level-set method

# Upper Level Sets




- Used to reduce the parameter space in which the search for the cluster MLE takes place (aka parameter space reduction). Note: We first make the ULS parameter space reductions then further limit the reduced parameter space using normal (Population $< 50\%$) constraints.
- We may interpret the cellular "rate" values $g_i$ as a function defined over the vertices of an abstract graph representing the tessellation: $i \mapsto g_i \quad for \quad i \in T$
- Let $G = \{g_i | i \in T\}$, and let $r_1, \ldots, r_m$ be the distinct members of G
- For $j = 1, \ldots, m$ let $T_j = \{i \in T | g_i = r_j\}$
- Then for each value of $r_j$ we have the Upper Level Set: $U_j$ defined as the union of the $T_j$s: $U_j = \bigcup_{k=1}^{j} T_k$

# Upper Level Set Connectedness, and Reduced Parameter Space

- The $U_j$s are defined to be nested sets such that the level sets $T_j$ form the "lower-rate" boundary.
  $U_{j-1} = \bigcup_{k=1}^{j-1}(U_k) \subset \bigcup_{k=1}^{j}(U_k) = U_j$ for $j = 1, \ldots, m$
  with $(U_j) \backslash (U_{j-1}) = T_j$

- We will determine the connected regions of each $U_j$ and define these as $C_j$
  i.e. Let $C_j =$ Set of connected components of $U_j$ for $j = 1, \ldots, m$

- Reduced Parameter Space: $\Omega_{ULS} = \bigcup_{k=1}^{m} C_k$

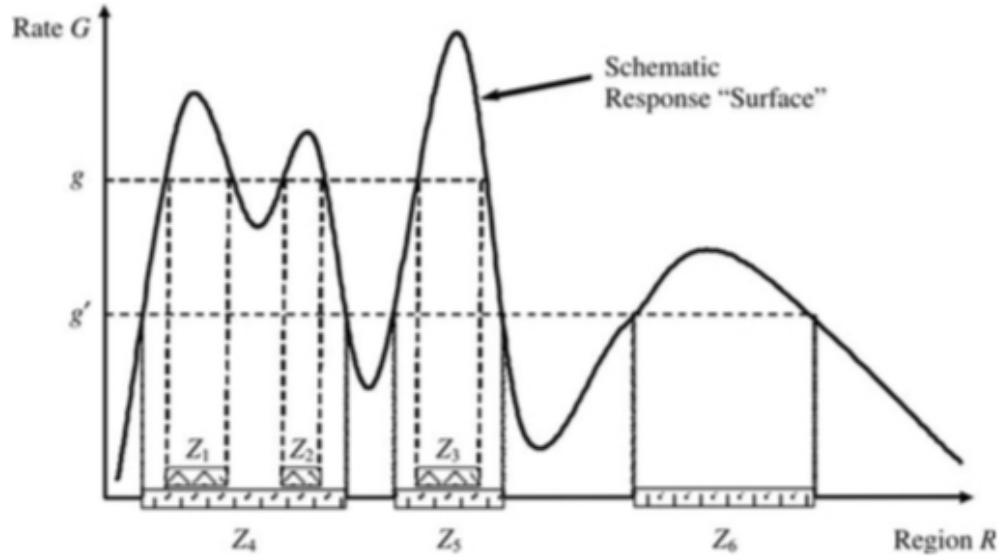
# ULS Tree

- We can define a tree structure associated with $\Omega_{ULS}$ using the following "algorithm":
  - Each level of $r_j$ for $j = 1, \ldots, m$ constitutes a vertical level of the tree.
  - Each level is split into its constituent connected components, which form the unionized sets $C_j$ (call each of these sets $Z_p$ for $p = 1, \ldots, m_j$). Each of these $Z_p$ represents a node in the tree

- Note: Due to the definition of $r_j$s as the unique elements of the N-element set $G = \{g_i | i \in T\}$ we know that: $m \leq N$

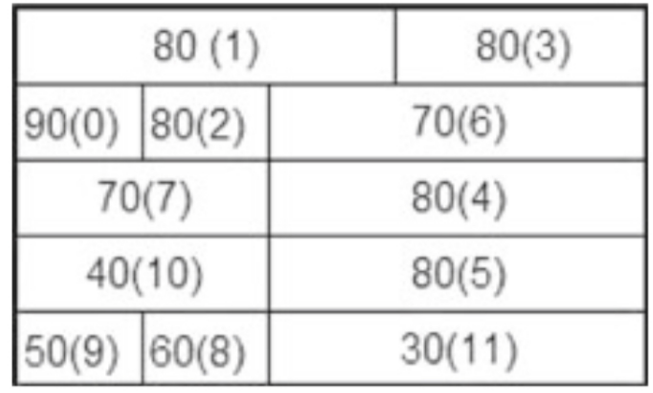
# ULS Tree Properties

- The definition of ULS Sets implies monotone set expansion as the value of j is allowed to increase $j : 1 \rightarrow m$
- There are three ways which the ULS Tree can expand:
  - As $j \rightarrow (j+1)$ the zones $Z_a$ and $Z_b$ merge into a single, connected zone, $Z_c$
  - If $Z_a \subseteq Z_b$, then $j \rightarrow (j+1)$ implies $Z_a \subseteq Z_b$
  - As $j \rightarrow (j+1)$, $U_j = \emptyset$ and $U_{j+1} \neq \emptyset$ with $U_{j+1}$ not connected to any $C_{j+1}$ implies $Z_{j+1} = U_{j+1}$



# ULS Example

- Picture demonstrates connectedness, and relative rates. (Each cell has the same population)
- In order to find the reduced parameter space we need to find:
  - The ordered relative frequency list
  - Spatial adjacency matrix re-ordered with respect to relative frequency




Example from: Environ Ecol Stat (2010) 17:149-182-PULSE

# ULS Example (continued)

Ordered Relative Frequency List

| Cell Number (i) | Relative Rate |
|---|---|
| 0 | 90 |
| 1 | 80 |
| 2 | 80 |
| 3 | 80 |
| 4 | 80 |
| 5 | 80 |
| 6 | 70 |
| 7 | 70 |
| 8 | 60 |
| 9 | 50 |
| 10 | 40 |
| 11 | 30 |

Reordered Spatial Adjacency Matrix

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | | | | | | | | | | | |
| 1 | 1 | 1 | | | | | | | | | | |
| 2 | 1 | 1 | 1 | | | | | | | | | |
| 3 | 0 | 1 | 0 | 1 | | | | | | | | |
| 4 | 0 | 0 | 0 | 0 | 1 | | | | | | | |
| 5 | 0 | 0 | 0 | 0 | 1 | 1 | | | | | | |
| 6 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | | | | | |
| 7 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | | | | |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | | | |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | | |
| 10 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | |
| 11 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |

$T = \{i = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$

We note that the distinct values of G are the values

$r_i = \{90, 80, 70, 60, 50, 40, 30\}$ which implies that $j = 1, \ldots, 7$

# ULS Example (continued)

We now calculate the sets $T_j$ for $j = 1, \ldots, 7$

$$T_j = \{i \in T | g_i = r_r\}$$

$$\begin{aligned} T_1 &= \{i \in T | g_i = r_1\} \\ &= \{i \in T | g_i = 90\} \\ &= \{i \in T | i = 0\} \end{aligned}$$

$$\begin{aligned} T_2 &= \{i \in T | g_i = r_2\} \\ &= \{i \in T | g_i = 80\} \\ &= \{i \in T | i = 1, 2, 3, 4, 5\} \end{aligned}$$

$$\begin{aligned} T_3 &= \{i \in T | g_i = r_3\} \\ &= \{i \in T | g_i = 70\} \\ &= \{i \in T | i = 6, 7\} \end{aligned}$$

$$\vdots$$

| j | $T_j$ |
|---|---|
| 1 | 0 |
| 2 | 1 2 3 4 5 |
| 3 | 6 7 |
| 4 | 8 |
| 5 | 9 |
| 6 | 10 |
| 7 | 11 |

# ULS Example (continued)

We now calculate the sets $U_j$ for $j = 1, \ldots, 7$

$$U_j = \bigcup_{k=1}^{j} T_k$$

$$U_1 = \bigcup_{k=1}^{1} T_k = T_1 = \{i \in T | i = 0\}$$

$$U_2 = \bigcup_{k=1}^{2} T_k = T_1 \bigcup T_2 = \{i \in T | i = 0, 1, 2, 3, 4, 5\}$$

$$U_3 = \bigcup_{k=1}^{3} T_k = T_1 \bigcup T_2 \bigcup T_3$$

$$= \{i \in T | i = 0, 1, 2, 3, 4, 5, 6, 7\}$$

$$\vdots$$

| j | $U_j$ |
|---|---|
| 1 | 0 |
| 2 | 0 1 2 3 4 5 |
| 3 | 0 1 2 3 4 5 6 7 |
| 4 | 0 1 2 3 4 5 6 7 8 |
| 5 | 0 1 2 3 4 5 6 7 8 9 |
| 6 | 0 1 2 3 4 5 6 7 8 9 10 |
| 7 | 0 1 2 3 4 5 6 7 8 9 10 11 |

# ULS Example (continued)

We now determine the connectedness of each $U_j$ to write the sets $C_j$ as unions of subsets of each $U_j$

$$C_1 = U_1 = \{i \in T | i = 0\}$$

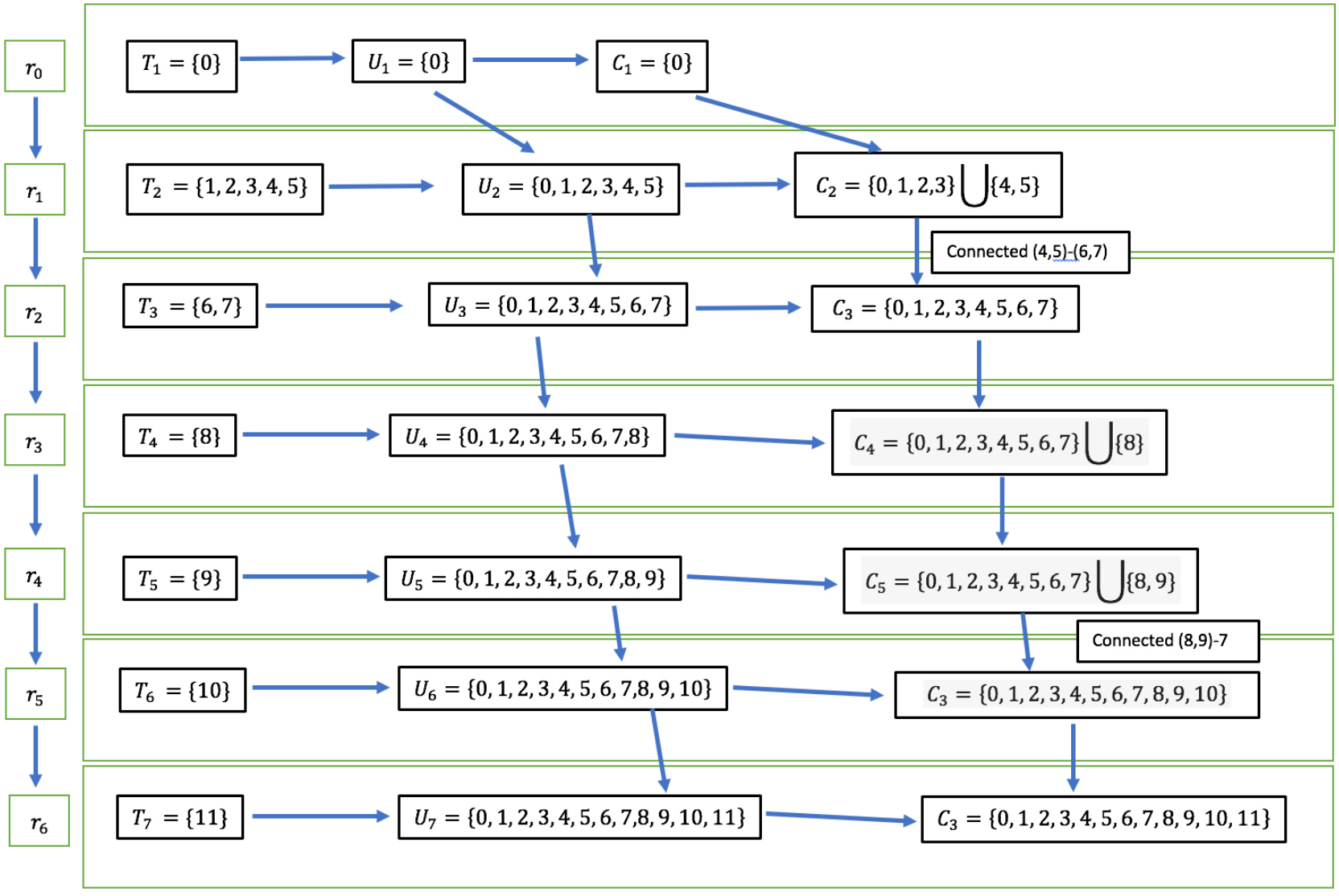
$$C_2 = \{i \in T | i = 0, 1, 2, 3\} \bigcup \{i \in T | i = 4, 5\}$$

$$C_3 = \{i \in T | i = 0, 1, 2, 3, 4, 5, 6, 7\}$$

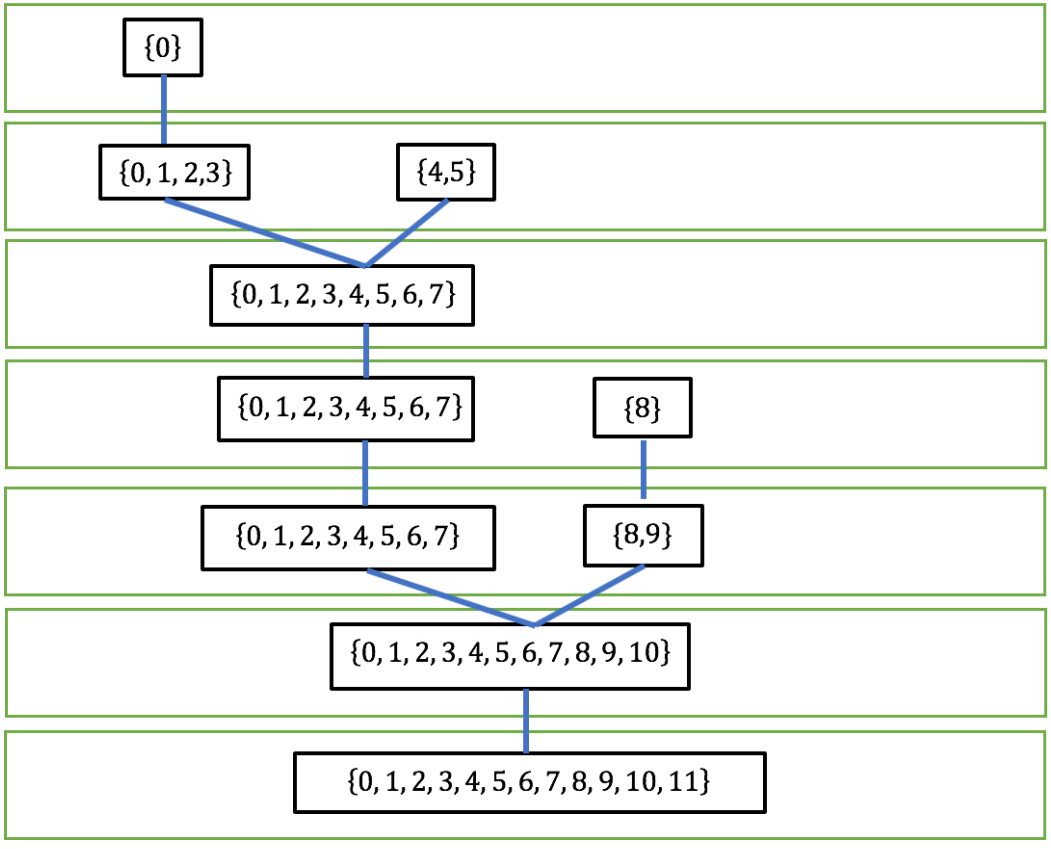
$$\vdots$$

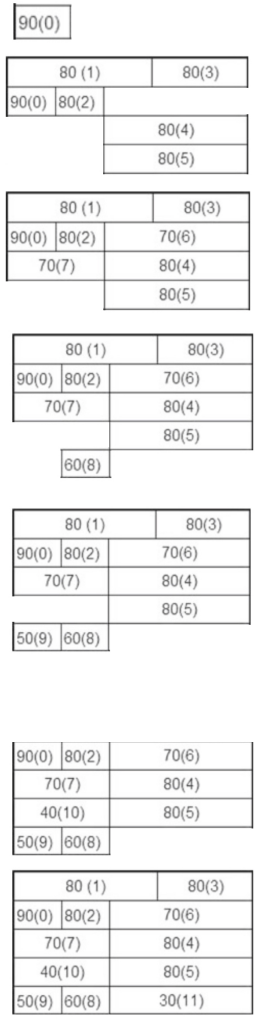
| j | $C_j$ |
|---|---|
| 1 | $\{0\}$ |
| 2 | $\{0,1,2,3\} \bigcup \{4,5\}$ |
| 3 | $\{0,1,2,3,4,5,6,7\}$ |
| 4 | $\{0,1,2,3,4,5,6,7\} \bigcup \{8\}$ |
| 5 | $\{0,1,2,3,4,5,6,7\} \bigcup \{8,9\}$ |
| 6 | $\{0,1,2,3,4,5,6,7,8,9,10\}$ |
| 7 | $\{0,1,2,3,4,5,6,7,8,9,10,11\}$ |

# ULS Example (continued)



# Tree Structure



# Parameter Space Reduction-$\Omega_{ULS}$

- "The difficult part of hotspot estimation lies in maximizing $L(p_0, p_1|\mathbf{Z})$ [the unrestricted likelihood function under $H_1$] as $\mathbf{Z}$ varies over the collection $\Omega$ "
- Parameter Space Reduction: Replace the full parameter space by a subspace $\Omega_0 \subset \Omega$ of a more manageable size. The profile likelihood $L(p_0, p_1|\mathbf{Z})$ is then maximized by exhaustive search across $\Omega_0$ (provided that $\Omega_0$ contains the MLE for the full $\Omega$, or at least an approximation to it).
- By defining $\Omega_0 = \Omega_{ULS}$ we can substantially reduce the search area.

# Final Reductions

- We have significantly reduced $\Omega$, but we still need to make sure that any "candidate cluster regions" we are checking satisfy the assumption that: collectively, the constituent regions do not constitute a majority of the underlying population.
- We check that no single connected subset of any $C_j$ has a population greater than $0.5 * N = \frac{\sum_{i=1}^{N} N_i}{2}$ in the progression $j : 1 \rightarrow m$
- Note $N = \sum_{i=1}^{N} N_i = 810$ so no single candidate region can have a population greater than 410
  - $C_0 \rightarrow \{0\} \rightarrow N_0 = 90 < 410$
  - $C_1 \rightarrow \{0, 1, 2, 3\} \bigcup \{4, 5\}$
    - $\{0, 1, 2, 3\} \rightarrow N_0 + N_1 + N_2 + N_3 = 330$ We consider all increasing subsets
    - $\{4, 5\} \rightarrow N_4 + N_5 = 160$ We consider all increasing subsets

# Final Reduction Results

$$\Omega_{ULS}^{red} : \{0\}, \{0,1\}, \{0,1,2\}, \{0,1,2,3\}$$
$$\{4\}, \{5\}, \{4,5\}$$
$$\{8\}, \{8,9\}$$

# Hypothesis Testing-Binomial (Review)

- In the Binomial setting, it is assumed that:
  - $n_i \in \mathbb{N}$ (cell populations) and $Y_i \sim Binomial(n_i, p_i)$ (cell responses) where $p_i$ is an unknown parameter attached to cell $i$ with $0 < p_i < 1$
- Under the null hypothesis, we assume constant risk across all the $i = 1, \ldots, N$ cells, which means that:
  - $p_i$ is the same for all cells in R.
- If $n_+ = \sum_{i=1}^{N} N_i$ is the total population in R, and $Y_+ = \sum_{i=1}^{N} Y_i$ is the total number of cases observed, then under the null hypothesis, $Y_i \sim Binomial\left(n_i, \frac{Y_+}{n_+}\right)$

# Hypothesis Testing-Poisson (Review)

- In the Poisson setting, it is assumed that:
  - $Y_i \sim Poisson(\lambda_i)$ where the $Y_i$s are independent and $\lambda_i > 0 \ \ \forall \ i = 1, \ldots, N$ is an unknown parameter attached to cell i
- We use the expected cell counts - $E_i$ - to estimate $\lambda_i$:

$$\hat{r} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} n_i} = \frac{Y_+}{n_+}$$

$$E_i = \hat{r} Y_i$$

- Then under the null hypothesis we model each $Y_i \sim Poisson(E_i)$

# Test Statistic Calculation

- ULS has successfully reduced the parameter space $\Omega \longrightarrow \Omega_{ULS}$ which we have managed to even further refine using standard parameter space constraints (Population Upper Bound for Cluster).
- We now wish to calculate a test-statistic for each of the sets in our final reduction results.
- The highest valued observed test statistic will be used for comparison in Monte Carlo simulation to compute a p-value.
- We consider a region $\mathbf{Z_0} \in \Omega_{ULS}^{red}$ to be a candidate hot spot region if we cannot reject:
  - $\tilde{H}_0 : \exists\ p_1 \geq p_0$ *such that* :

$$p_a = \begin{cases} p_1 & \forall\ i\ \in Z_0 \\ p_0 & \forall\ i\ \notin Z_0 \end{cases}$$

# Test Statistic Calculation (continued)

- The previous Hypothesis test is equivalent to:
  - $\tilde{H}_0 : \ \mathbf{Z} = \mathbf{Z_0}$ *for* $\mathbf{Z} \in \Omega_{ULS}$
  - $\tilde{H}_1 : \ \mathbf{Z} \neq \mathbf{Z_0}$ *for* $\mathbf{Z} \in \Omega_{ULS}$
- We consider the hotspot estimate to be the set consisting of all zones $\mathbf{Z_0}$ for which $\tilde{H}_0$ could not be rejected
- The Likelihood Ratio Test statistic for the above hypothesis test framework can be computed for both the Binomial and Poisson models by estimating the required parameters through conditioning.

# Test Statistic Calculation-Poisson Model

- In the case of the Poisson Model, the assumption that $\vec{Y} = (Y_1, \ldots, Y_N)$ and $\vec{\lambda} = (\lambda_1, \ldots, \lambda_N)$ with each $Y_i \sim \mathit{Poisson}(\lambda_i)$, independent, allows us to conclude that:
  - $L\left(\vec{\lambda}|\vec{Y} = \vec{y}\right) = \prod_{i=1}^{N} P\left(Y_i = y_i|\lambda_i\right)$

    $\mathfrak{l}\left(\vec{\lambda}|\vec{Y} = \vec{y}\right) = \sum_{i=1}^{N} \log(P\left(Y_i = y_i|\lambda_i\right))$
- We calculate the likelihood ratio test statistic using:
  - $\Lambda = \frac{\sup_{\vec{\lambda}\in\Theta_0} \; L\left(\vec{\lambda}|\vec{Y}=\vec{y}\right)}{\sup_{\vec{\lambda}\in\Theta} \; L\left(\vec{\lambda}|\vec{Y}=\vec{y}\right)}$
  - Where $\vec{\lambda} \in \Theta_0 \Rightarrow (\lambda_1, \ldots, \lambda_N) = (E_1, \ldots, E_N) = \vec{E}$

    $\Rightarrow \sup_{\vec{\lambda}\in\Theta_0} \; L\left(\vec{\lambda}|\vec{Y} = \vec{y}\right) = L\left(\vec{E}|\vec{Y} = \vec{y}\right)$

    $\Rightarrow \sup_{\vec{\lambda}\in\Theta_0} \; \mathfrak{l}\left(\vec{\lambda}|\vec{Y} = \vec{y}\right) = \mathfrak{l}\left(\vec{E}|\vec{Y} = \vec{y}\right)$

# Test Statistic Calculation-Poisson Model (continued)

- $\hat{\lambda}_{MLE} = \sup_{\vec{\lambda} \in \Theta} \mathfrak{l}\left(\vec{\lambda} \mid \vec{Y} = \vec{y}\right) \Rightarrow 0 = \frac{\partial}{\partial \lambda_i}\left[\mathfrak{l}\left(\vec{\lambda} \mid \vec{Y} = \vec{y}\right)\right]$ $\forall i = 1, \ldots, N$
- $0 = \frac{\partial}{\partial \lambda_i}\left[\mathfrak{l}\left(\vec{\lambda} \mid \vec{Y} = \vec{y}\right)\right]$

  $\Leftrightarrow 0 = \sum_{i=1}^{N} \frac{\partial}{\partial \lambda_i}\left[\log\left(P\left(Y_i = y_i \mid \lambda_i\right)\right)\right]$

  $\Leftrightarrow 0 = \sum_{i=1}^{N} \frac{\partial}{\partial \lambda_i}\left[y_i \log(\lambda_i) - \lambda_i - \log(y_i!)\right]$

  $\Leftrightarrow 0 = \sum_{i=1}^{N} \left(\frac{y_i}{\lambda_i} - 1\right)$

  $\Leftrightarrow 0 = \frac{1}{\lambda_1}\left(y_1 - \lambda_1\right) + \frac{1}{\lambda_2}\left(y_2 - \lambda_2\right) + \cdots + \frac{1}{\lambda_N}\left(y_N - \lambda_N\right)$

  $\left(\lambda_i > 0 \ \forall i = 1, \ldots, N\right)$

  $\Leftrightarrow 0 = \left(y_1 - \lambda_1\right) + \left(y_2 - \lambda_2\right) + \cdots + \left(y_N - \lambda_N\right)$

  $\Leftrightarrow y_i = \lambda_i \ \forall i = 1, \ldots, N$
- $\sup_{\vec{\lambda} \in \Theta} \mathfrak{l}\left(\vec{\lambda} \mid \vec{Y} = \vec{y}\right) = \mathfrak{l}\left(\vec{y} \mid \vec{Y} = \vec{y}\right)$

# Test Statistic Calculation-Poisson Model (continued)

$$\Lambda = \frac{\sup_{\vec{\lambda}\in\Theta_0} \; L\left(\vec{\lambda}|\vec{Y}=\vec{y}\right)}{\sup_{\vec{\lambda}\in\Theta} \; L\left(\vec{\lambda}|\vec{Y}=\vec{y}\right)} = \frac{\mathfrak{l}\left(\vec{E}|\vec{Y}=\vec{y}\right)}{\mathfrak{l}\left(\vec{y}\,|\vec{Y}=\vec{y}\right)}$$

$$\mathfrak{l}\left(\vec{\lambda}\,|\vec{Y}=\vec{y}\right) = y_i \log(\lambda_i) - \lambda_i - \log(y_i!)$$

$$\Rightarrow \; \mathfrak{l}\left(\vec{E}|\vec{Y}=\vec{y}\right) = y_i \log(E_i) - E_i - \log(y_i!)$$

$$\Rightarrow \; \mathfrak{l}\left(\vec{y}|\vec{Y}=\vec{y}\right) = y_i \log(y_i) - y_i - \log(y_i!)$$

$$\Lambda = \frac{y_i \log(E_i) - E_i - \log(y_i!)}{y_i \log(y_i) - y_i - \log(y_i!)} = \left[\frac{E_i^{y_i} e^{-E_i}}{y_i!}\right]\left[\frac{y_i!}{y_i^{y_i} e^{-y_i}}\right]$$

$$= \left(\frac{E_i}{y_i}\right)^{y_i} * exp\{-E_i - y_i\}$$

# Simulation of Null Hypothesis Data

- We simulate count data under the Null distributions (Poisson or Binomial)
- Under the Null Hypothesis of the example above we have:
  - $\hat{r} = \frac{1}{1200} \sum_{i=1}^{12} Y_i = \frac{810}{1200} = .675$
    (The average rate over the cells)
  - Therefore $E_i = .675 \ \forall \ i = 1, \ldots, 12$
- We simulate 999 random samples: $\vec{Y}^k = (Y_1^k, \ldots, Y_{12}^k)$ where each $Y_i^k \sim Poisson(.675) \ \forall i = 1, \ldots, 12 \ \ \forall k = 1, \ldots, 999$
- $p = P\left(\{T_{sim}\} \bigcup T_{MLC} \geq T_{MLC}\right)$

# Comparison of ULS to Spatial and Flexible Scan Statistics

- The Upper Level Set scan estimate of cluster regions is compared against the Spatial and Flexible scan statistics under the most homogeneous conditions achievable for each test.
- The settings applied uniformly to all comparisons include:
  - 50% population upper bound on cluster size
  - Significance level of $\alpha = 0.10$ to determine presence of cluster in Hypothesis Testing
  - Comparison against 999 Null-distribution simulations for computations of p-values.
- The settings above are repeated for $k = 5, 7, 10$ maximal region size tests for the Flexible Scan Statistic

# Comparison of Region ID Selections by Method & P-values

| | Significant Local Indices | P-Value |
|---|---|---|
| ULS | [1] 77 76 78 12 5 11 31 256 9 17 18 33 53 49 15 27 13 46 37 43 38<br>[22] 47 51 2 44 35 1 40 54 52 14 41 16 7 255 82 28 55 30 176 211 201<br>[43] 89 90 86 92 85 93 259 252 88 153 106 103 102 230 232 237 155 225 171 166 167<br>[64] 170 240 224 144 159 226 228 146 135 210 208 120 139 132 131 150 151 130 138 126 124<br>[85] 117 123 125 119 219 216 111 220 217 113 114 205 206 209 115 143 164 83 278 80 118<br>[106] 21 87 156 234 169 218 204 84 213 258 152 250 223 104 133 99 81 182 6 253 239<br>[127] 4 207 79 | 0.01 |
| SatScan | 52 50 53 38 49 48 15 39 37 1 44 16 47 40 14 2 51 13 43 45 17 55 11 3 12 46 36 35 54 10 | 0.271 |
| | 88 87 92 86 89 91 93 85 | 0.01 |
| Flex | K=5 86 92 88 89 | 0.017 |
| | K=7 86 92 88 89 85 | 0.019 |
| | k=10 88 92 86 89 93 85 90 | 0.004 |
| | 38 44 40 43 53 37 52 46 | 0.063 |

# Comparison of Physical Cluster Appearance by Method