Bayesian spatial scan statistic adjusted for over-dispersion

Deepak Agarwal
Yahoo! Research
February 8th,2007
QMDNS, Virginia.

Yahoo! Proprietary and Confidential

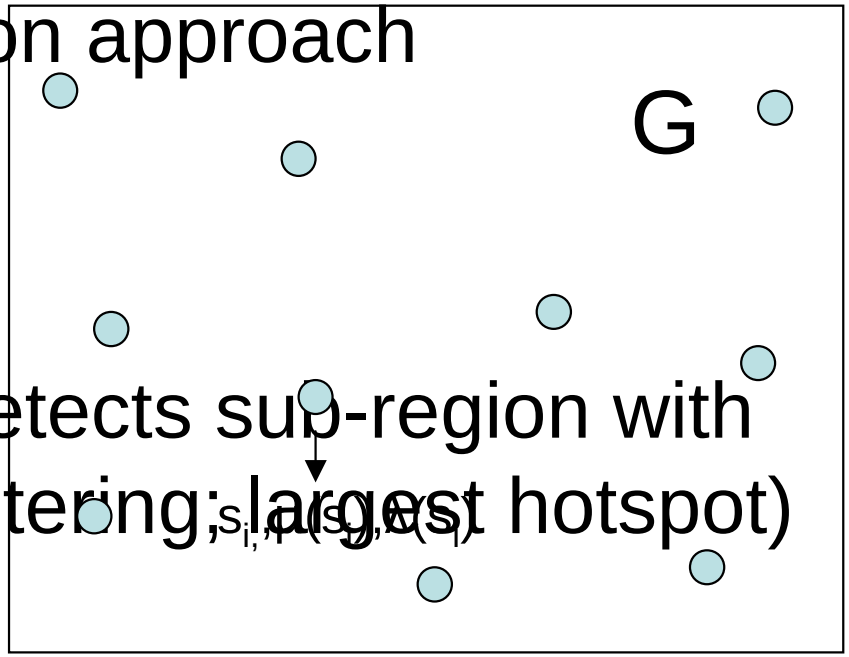
Outline

- Introduction: Spatial Scan statistic
- Bayesian formulation
- Adjusting for over-dispersion
  - Doubly stochastic/Cox process
- Data analysis: SIDs data
- Ongoing work
  - Incorporating spatial correlation

Yahoo! Proprietary and Confidential

Background

- μ(s)=v(s); λ(s)=λ→homogeneous Poisson
- Global tests for spatial homogeneity
  - E.g. Ripley's K-function approach

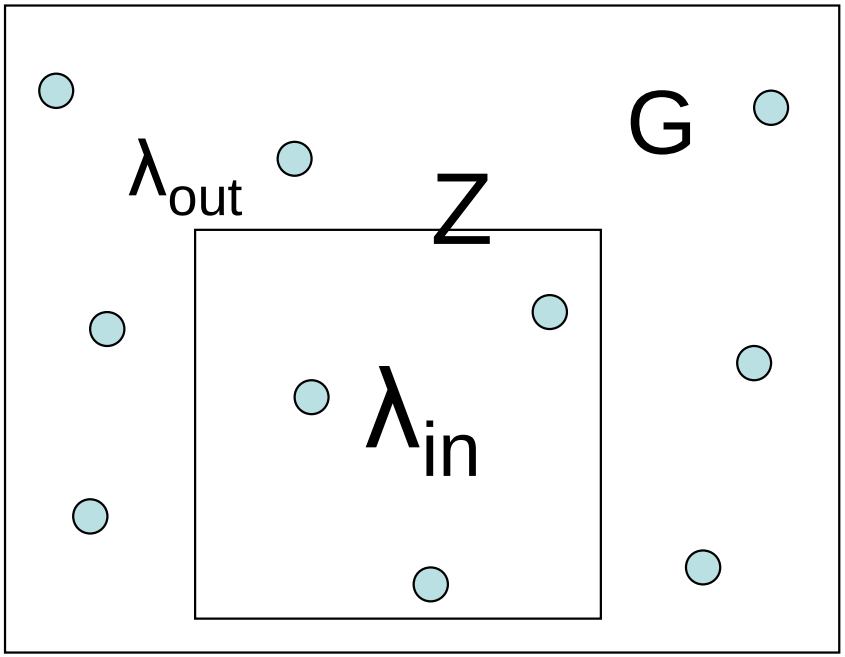


- Scan statistic:
  - Additional nuance (Detects sub-region with maximum spatial clustering; largest hotspot)

Spatial scan statistic

- Geometric shape → sub regions
  - Rectangle, Square, Circle, Ellipsoid,…
- Scan G using all possible sub regions

- LLRT: inside-outside
- Max discrepancy region
- Rand test → p-value
- Inhomogeneity
  - Incorporate in μ(s)



Yahoo! Proprietary and Confidential

Model: Poisson case

N points $\{x_i\}$: inhomogeneous Poisson process

$\mu(x)$: Baseline measure; $\lambda(x)$: Intensity function

$$L(\boldsymbol{\lambda};\{x_i\}) \propto (\prod_i \lambda(x_i)\mu(x_i))\exp(-\int_G \lambda(x)d\mu(x))$$

Lattice Data : K regions; n_i = number of points in region i

x_i : Centroid of region i

$\mu(x_i)$: Adjusts inhomogeneity ; e.g., Population at risk

$$L(\boldsymbol{\lambda};\{x_i\}) \propto (\prod_i (\lambda(x_i)\mu(x_i))^{n_i})\exp(-\sum_{i=1}^{K} \lambda(x_i)\mu(x_i))$$

Scan statistic: Max log(LRT)

For region Z

$\lambda(x) = \lambda_{in} \quad x \in Z;\ \lambda(x) = \lambda_{in} \quad x \in G - Z$

$Z_0 : \lambda_{\text{in}} = \lambda_{out}$ vs $Z : \lambda_{\text{in}} > \lambda_{out}$

log(LRT) :

$$D_Z = (n_Z \log(n_Z / \mu_Z) + n_{G-Z} \log(n_{G-Z} / \mu_{G-Z}) - n_G \log(n_G / \mu_G)) 1(\frac{n_Z}{\mu_Z} > \frac{n_{G-Z}}{\mu_{G-Z}})$$

Statistical discrepancy

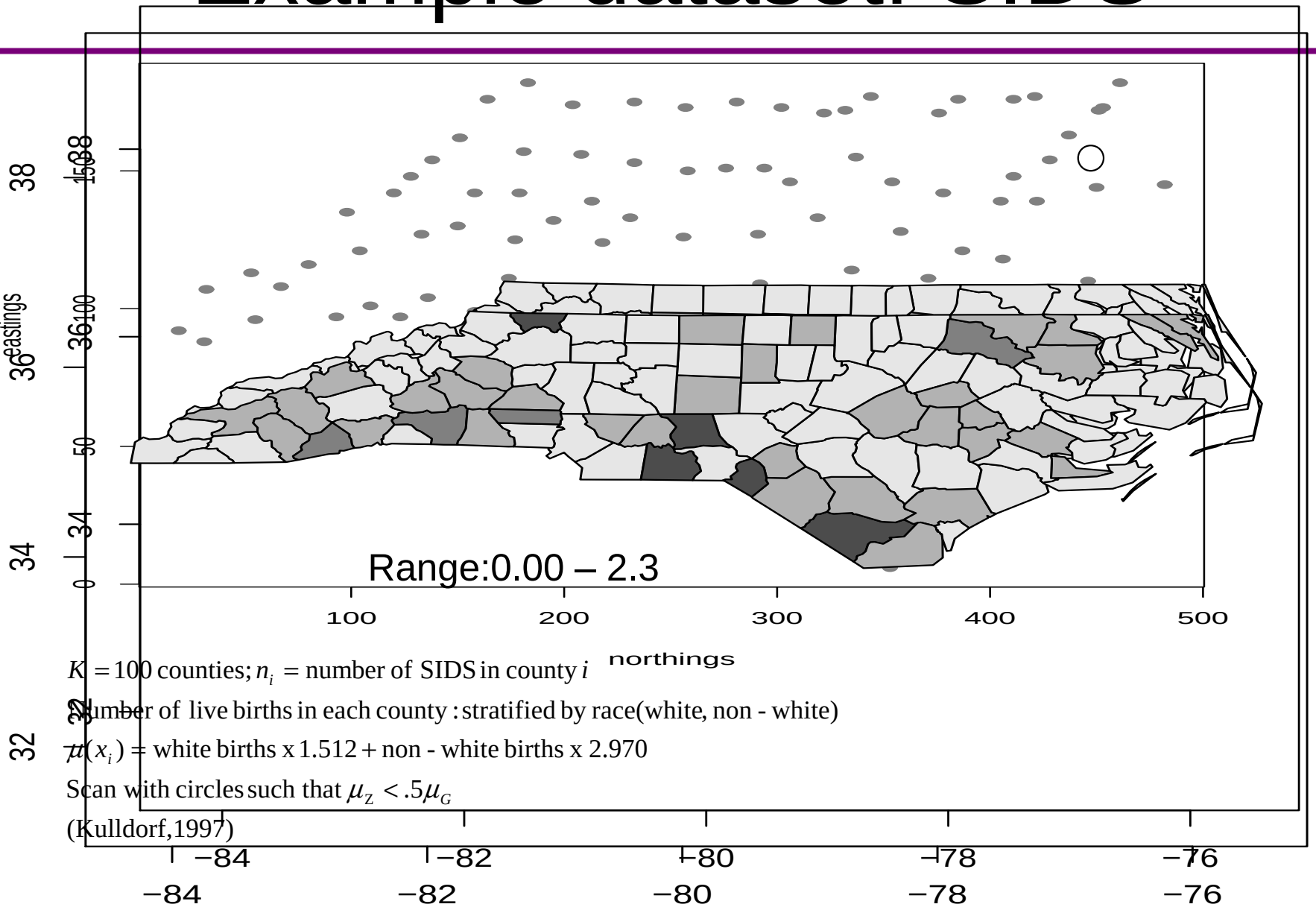
KL : $(n_Z / N, n_{G-Z} / N)$ and $(\mu_Z / \mu_G, \mu_{G-Z} / \mu_G)$ (Agarwal et al, 2006)

Max discrepancy region : $Z^* = \arg\max_Z D_Z$

Significance testing :

p - value based on Monte - Carlo simulations (expensive)

Example dataset: SIDS



$K = 100$ counties; n_i = number of SIDS in county i

Number of live births in each county : stratified by race(white, non - white)

$\pi(x_i)$ = white births x 1.512 + non - white births x 2.970

Scan with circles such that $\mu_Z < .5\mu_G$

(Kulldorf,1997)

Bayesian approach

$$\text{Model space}: \mathrm{M} = \{\mathrm{Z}_0\} \cup \{\mathrm{Z}_\mathrm{i} : i = 1,.., N_{reg}\}$$

$$\pi^{\mathrm{M}} : \text{Prior distribution on M}$$

$$f(D \mid Z_i) = \text{Marginal likelihood under } Z_i$$

$$\text{Posterior}: f(Z_i \mid D) \propto f(D \mid Z_i)\pi^{\mathrm{M}}(Z_i)$$

$$f(D \mid Z_i) = \int f(D \mid Z_i, \theta_{Z_i})\pi(\theta_{Z_i})d\theta_{Z_i}$$

$$\text{Detect "hotspots" using a threshold on } f(Z_i \mid D)$$

$$\text{Threshold}: \text{Expert decides; decision theoretic approach}$$

Yahoo! Proprietary and Confidential

Poisson case

$$Z_i : \theta_{Z_i} = (\theta_{in} = \log(\lambda_{in}), \theta_{out} = \log(\lambda_{out}))$$

$$\pi(\theta_{in}, \theta_{out}) \sim N((\mu_0, \mu_0), (\delta, \delta), \rho = 0) 1(\theta_{in} \geq \theta_{out})$$

$$\delta = 100 * \mathrm{Var}(n_i / \mu_i) * m^2 ; \text{ where } m = mean(n_i / \mu_i);$$

(proper but vague)

$$\pi^{\mathrm{M}}(Z \mid P) = P1(Z = Z_0) + (1 - P)\sum_{i=1}^{N_{reg}} 1(Z = Z_i) / N_{reg}$$

$$f(D \mid Z) : \text{Tierney - Kadane approximation}$$

$$f(P) = (\alpha + 1)P^{\alpha} \text{ (Scott and Berger, 2006)}$$

Choose α such that median = .9

Yahoo! Proprietary and Confidential

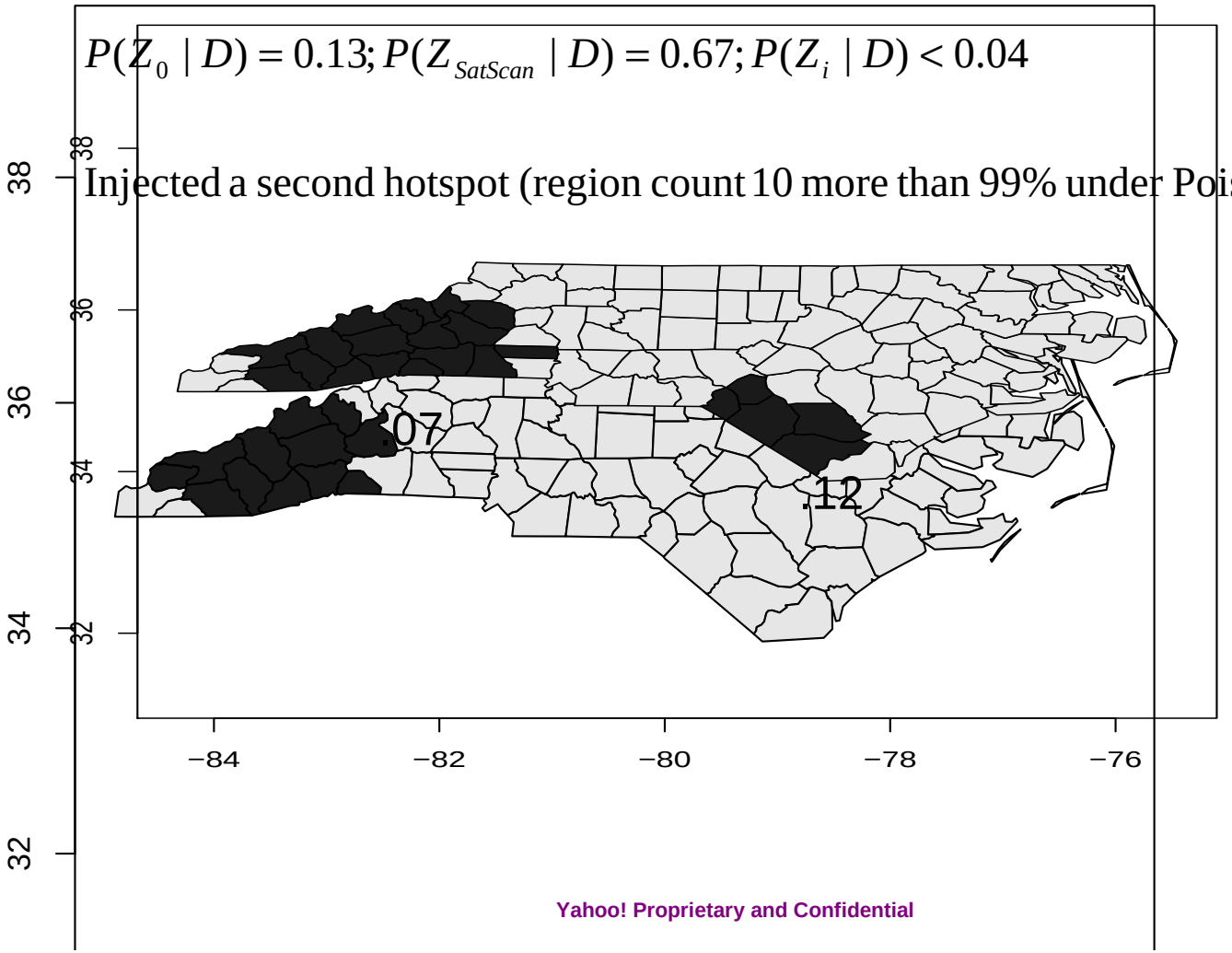
Analysis for Poisson

Detect one hotspot, exactly the one by SatScan

99% for region = 112; observed = 139

$P(Z_0 \mid D) = 0.13; P(Z_{SatScan} \mid D) = 0.67; P(Z_i \mid D) < 0.04$

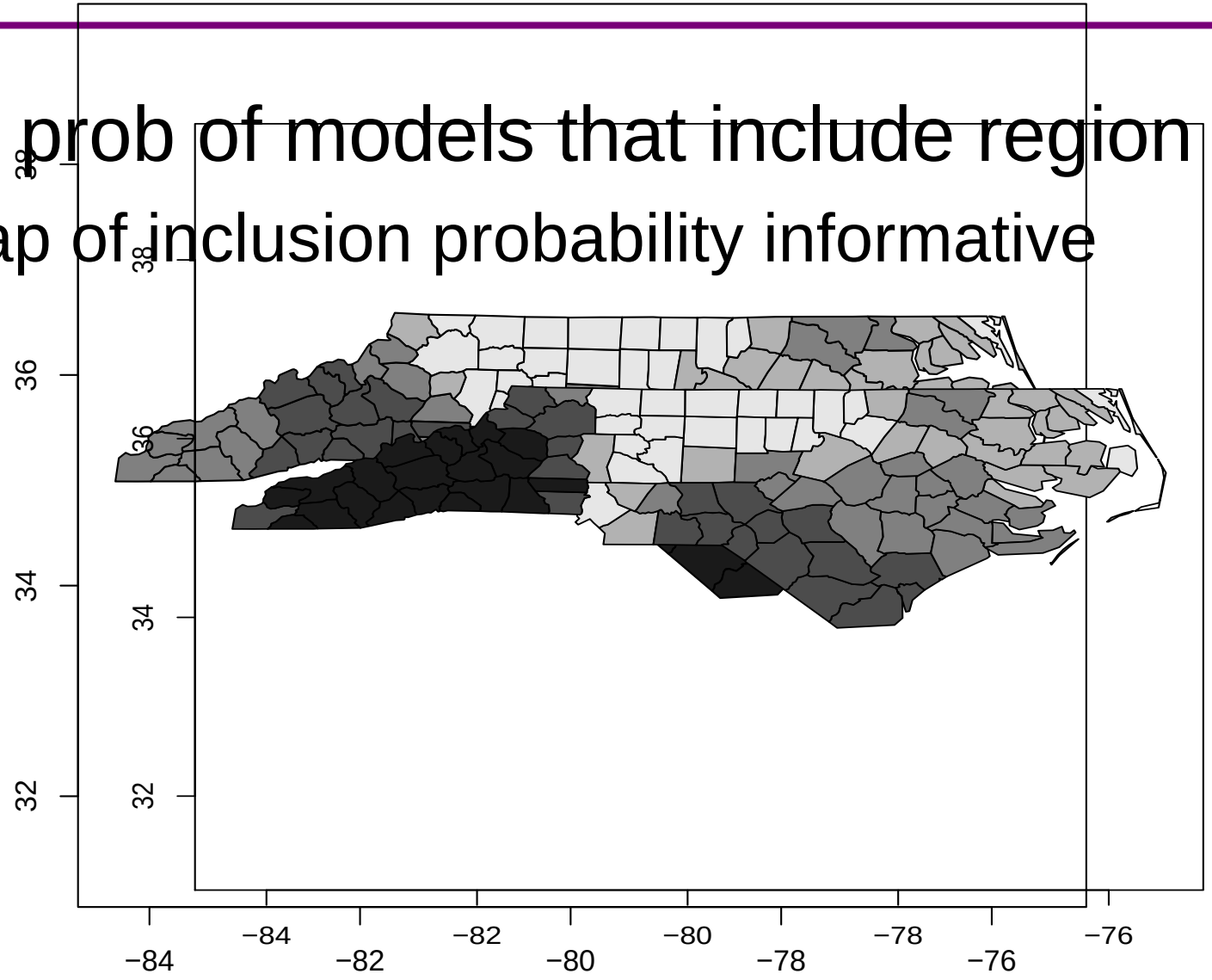
Injected a second hotspot (region count 10 more than 99% under Poisson)



Yahoo! Proprietary and Confidential

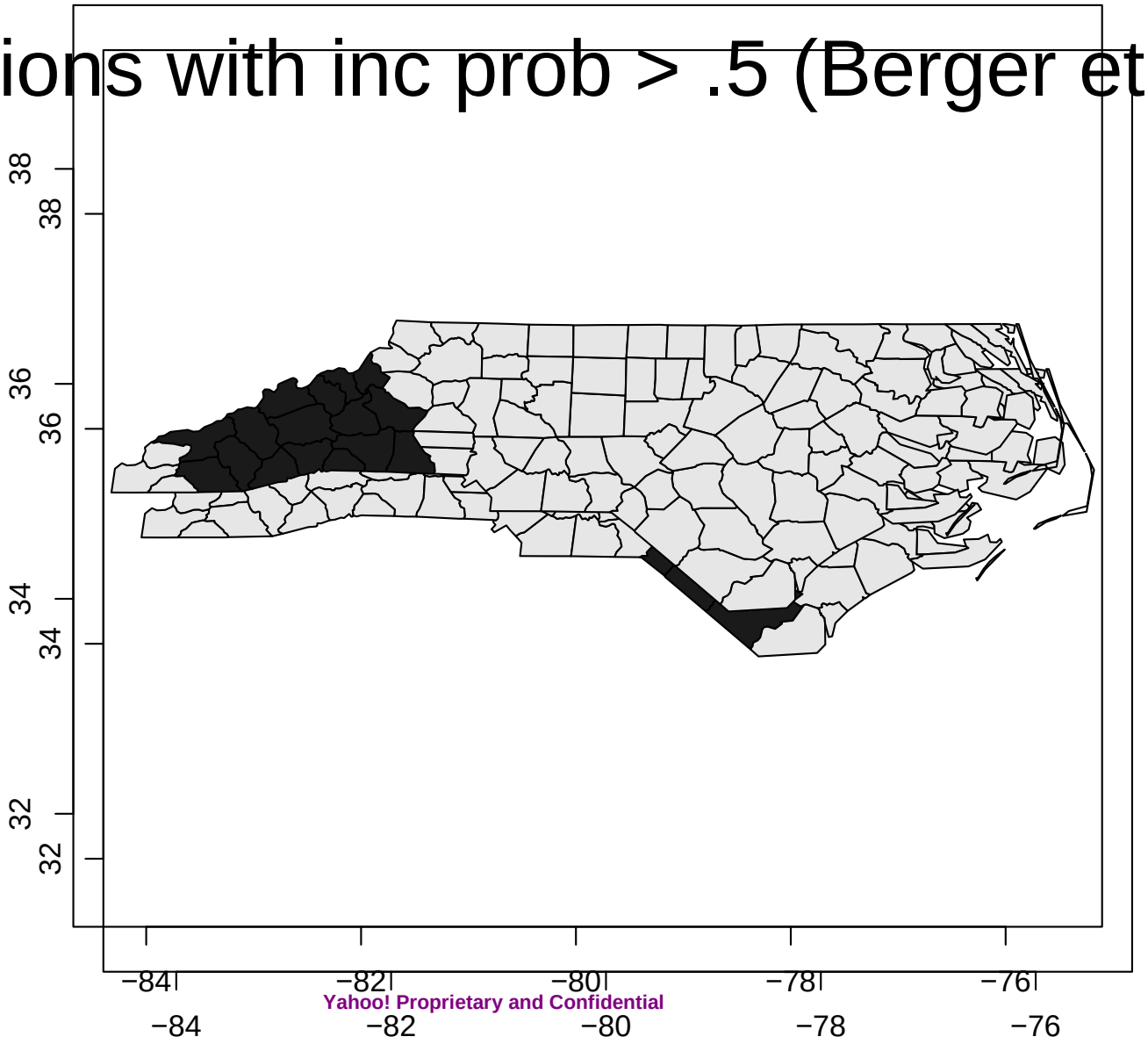
Inclusion probability

- Post prob of models that include region
  - Map of inclusion probability informative



Median probability model

- All regions with inc prob > .5 (Berger et al)



Yahoo! Proprietary and Confidential

Adjusting for over-dispersion

- Over-dispersion in data is routine
  - Adjust to reduce false +ve's
- SIDS data; negative binomial better

$$\text{mean}(\mu)/\gamma = 15.03/16.44 = .94$$

Cox/doubly stochastic process

Conditional on an error process $\nu(x)$, inhomogeneous Poisson

$$\lambda(x_i) = \lambda_{in}\nu(x_i) \quad x_i \in Z$$

$$\lambda(x_i) = \lambda_{out}\nu(x_i) \quad x_i \in G - Z$$

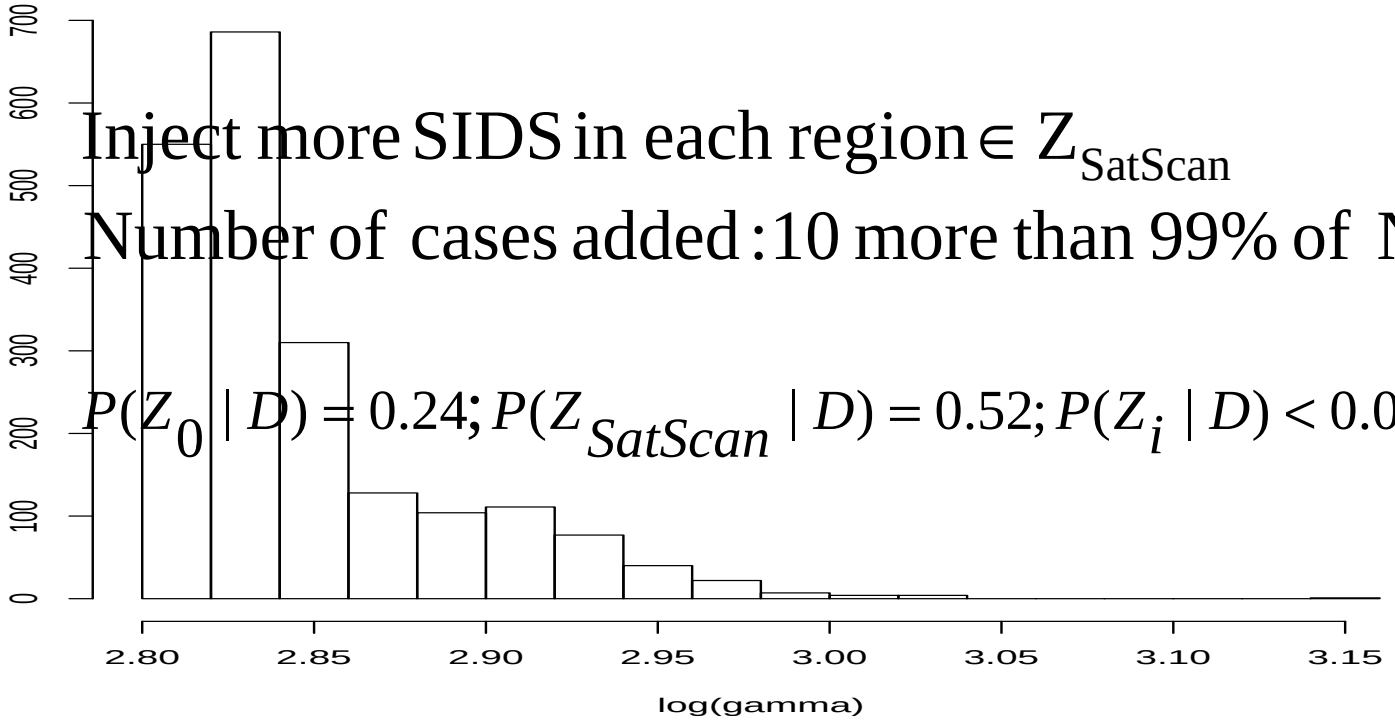
$\nu(x_i)$ iid Gamma(γ_z, γ_z) accounts for overdispersion

Results with negative binomial

$$P(Z_0 \mid D) = 0.84; P(Z_{SatScan} \mid D) = 0.02; P(Z_i \mid D) < 0.01$$

Hotspot does not exist anymore!!

99% for SatScan region under NB = 153 > Obs = 139



Inject more SIDS in each region $\in Z_{\text{SatScan}}$

Number of cases added: 10 more than 99% of NB

$$P(Z_0 \mid D) = 0.24; P(Z_{SatScan} \mid D) = 0.52; P(Z_i \mid D) < 0.05$$

Yahoo! Proprietary and Confidential

Neil et al (NIPS 2005)

$$n_Z \sim NB(mean = m * n_G / \mu_G; \text{var} = 1/\mu_Z);$$

$$n_{G-Z} \sim NB(mean = n_G / \mu_G; \text{var} = 1/\mu_{G-Z})$$

Biased towards regions with larger measure;

we estimate appropriate parameter from data

No global model; hard to incorporate other features like spatial correlation. Trivial in our framework

Fact : The measure for SatScan region 20th percentile

Model	Poisson	NB	N+M
Null	0.13	0.84	0.74
SatScan	0.67	0.02	0.07
penultimate	0.04	0.01	0.04

Ongoing Work

- Incorporating spatial association
  - Put a spatial prior on the error process
- Loss functions to rank discrepant regions
  - Important for multiple hotspots
- Large scale power tests