Enhance Analysis of SaTScan Results with Geovisual Analytics

Jin Chen[1], Alan M. MacEachren [1], Eugene Lengerich [2]

[1]GeoVISTA Center, Geography Dept, Penn State University

[2]Department of Health and Evaluation Sciences, Penn State College of Medicine

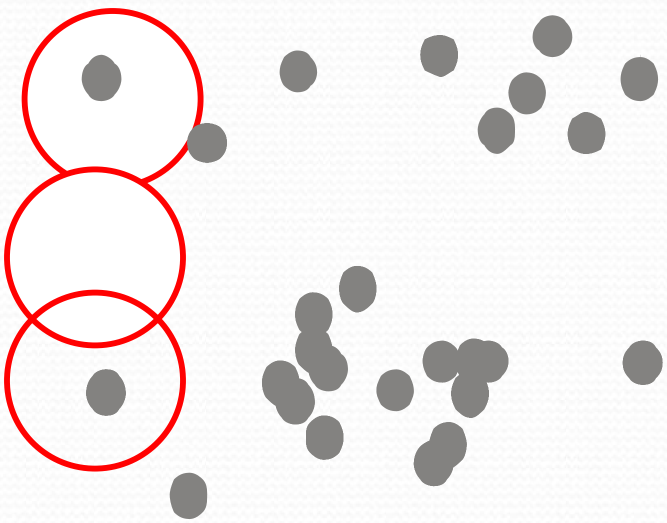
Outline

- Introduction of Kulldorff's spatial scan statistic (SaTScan)
- Limitations with SaTScan
- Geovisual analytics enhances analysis of SaTScan results

GeoVISTA Center

Scan Statistics

- Introduced by Naus(1965). GAM by Openshaw et.al (1988). Rushton and Lolonis (1996).
- Detect a local excess or deficiency of events (e.g. death rate due to a disease).
- Employ a moving "window", collect cases least consistent with null hypothesis (e.g. constant risk of a disease). The cases are most likely clusters.



Kulldorff's spatial scan static

- Introduced by Kulldorff (1997), known as SaTScan
- Similar to GAM, with advantages
  - deterministic - it reports the location and size of clusters
  - inferential - it evaluate statistical significance of the clusters.
- A result reported by SaTScan

Location IDs included.: 12087, 12086, 12021, ... (i.e. FIPS code)
Coordinates / radius..: (33.957701 N, 91.732284 W) / 377.70 km
Population............: 24785853
Number of cases.......: 1047
Expected cases........: 678.20
Observed / expected...: 1.544
...
P-value...............: 0.001

- The method (and the software) is widely used in epidemiology, crime analysis, etc.

Limitations of SaTScan

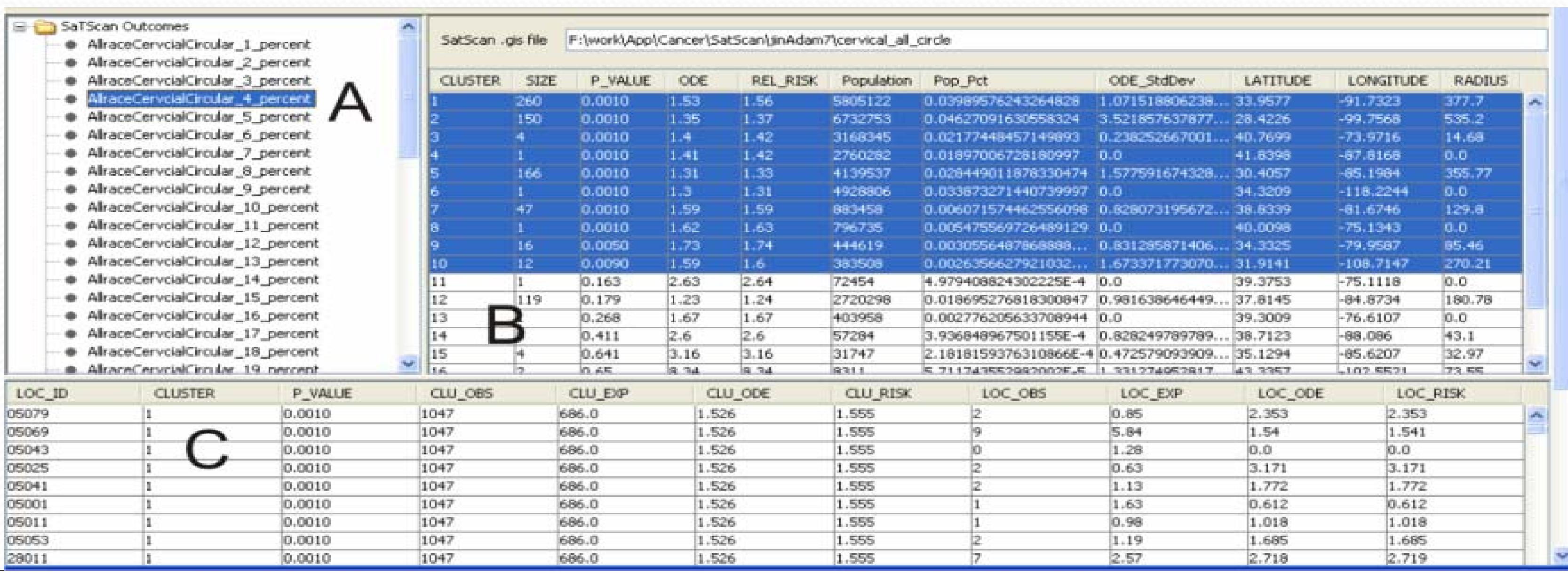
- Lack of cartographic output for displaying clusters

- Produces less usable clusters when inappropriately choosing parameters (i.e., heterogeneous contents)

- Sensitive to user controlled parameter choices
  - e.g., the maximum circle size, defined as the percentage of total population at risk.
  - If a circle covers a region that has N% total population, then the circle size is N%.

Visualize SaTScan results

- SaTScan reports clusters in text format. The clusters need to be visualized on a map.
- Visual Inquiry Toolkit (VIT)
  - Visualize SaTScan results without special skills on GIS.
  - Load multiple results
  - Interact with the results



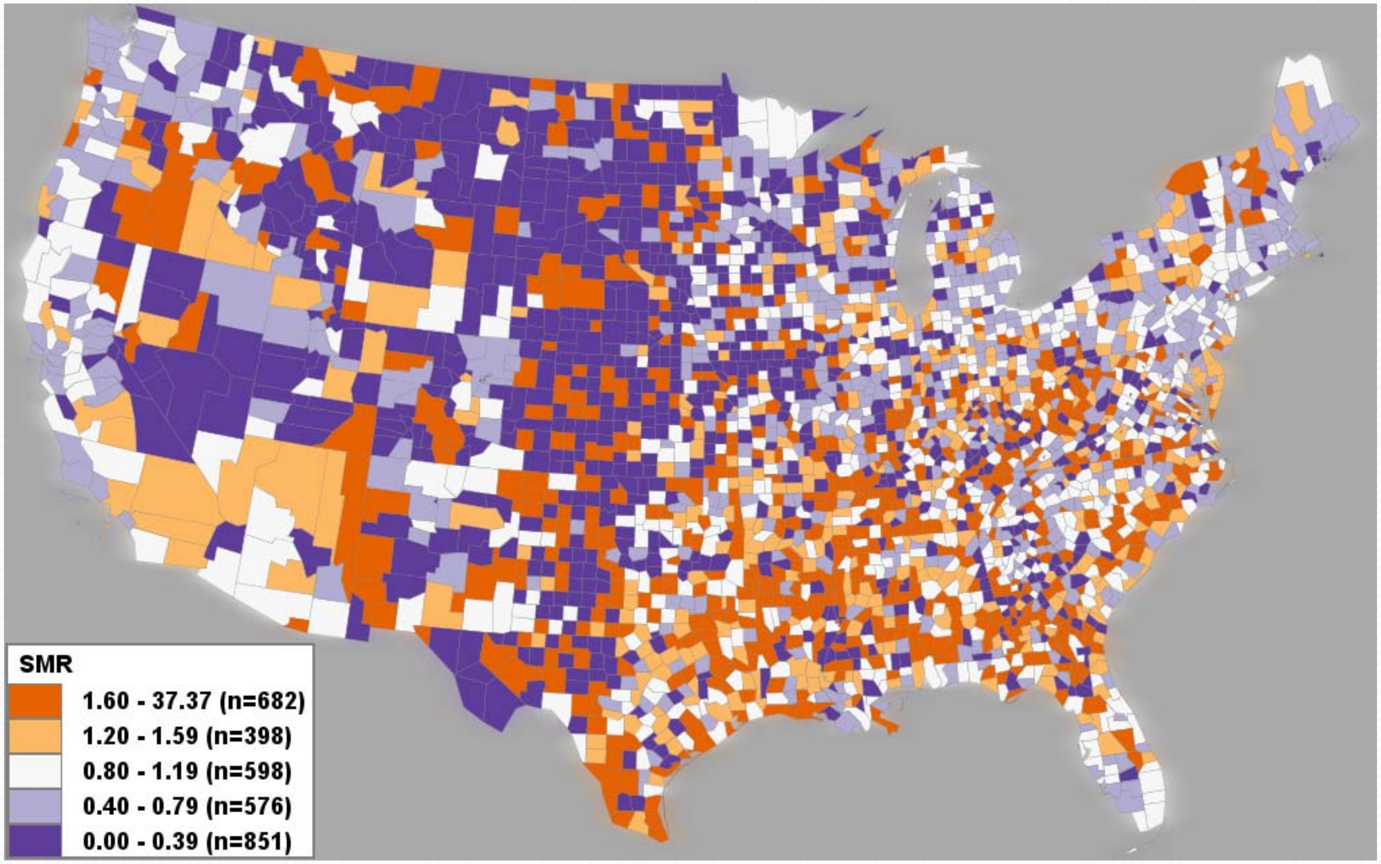
CLUSTER	SIZE	P_VALUE	ODE	REL_RISK	Population	Pop_Pct	ODE_StdDev	LATITUDE	LONGITUDE	RADIUS
1	260	0.0010	1.53	1.56	5805122	0.03989576243264828	1.071518806238...	33.9577	-91.7323	377.7
2	150	0.0010	1.35	1.37	6732753	0.04627091630558324	3.521857637877...	28.4226	-99.7568	535.2
3	4	0.0010	1.4	1.42	3168345	0.02177448457149893	0.238252667001...	40.7699	-73.9716	14.68
4	1	0.0010	1.41	1.42	2760282	0.01897006728180997	0.0	41.8398	-87.8168	0.0
5	166	0.0010	1.31	1.33	4139537	0.028449011878330474	1.577591674328...	30.4057	-85.1984	355.77
6	1	0.0010	1.3	1.31	4928806	0.033873271440739997	0.0	34.3209	-118.2244	0.0
7	47	0.0010	1.59	1.59	883458	0.006071574462556098	0.828073195672...	38.8339	-81.6746	129.8
8	1	0.0010	1.62	1.63	796735	0.005475569726489129	0.0	40.0098	-75.1343	0.0
9	16	0.0050	1.73	1.74	444619	0.00305564878688888...	0.831285871406...	34.3325	-79.9587	85.46
10	12	0.0090	1.59	1.6	383508	0.0026356627921032...	1.673371773070...	31.9141	-108.7147	270.21
11	1	0.163	2.63	2.64	72454	4.979408824302225E-4	0.0	39.3753	-75.1118	0.0
12	119	0.179	1.23	1.24	2720298	0.018695276818300847	0.981638646449...	37.8145	-84.8734	180.78
13		0.268	1.67	1.67	403958	0.002776205633708944	0.0	39.3009	-76.6107	0.0
14		0.411	2.6	2.6	57284	3.936848967501155E-4	0.828249789789...	38.7123	-88.086	43.1
15	4	0.641	3.16	3.16	31747	2.1818159376310866E-4	0.472579093909...	35.1294	-85.6207	32.97

LOC_ID	CLUSTER	P_VALUE	CLU_OBS	CLU_EXP	CLU_ODE	CLU_RISK	LOC_OBS	LOC_EXP	LOC_ODE	LOC_RISK
05079	1	0.0010	1047	686.0	1.526	1.555	2	0.85	2.353	2.353
05069	1	0.0010	1047	686.0	1.526	1.555	9	5.84	1.54	1.541
05043	1	0.0010	1047	686.0	1.526	1.555	0	1.28	0.0	0.0
05025	1	0.0010	1047	686.0	1.526	1.555	2	0.63	3.171	3.171
05041	1	0.0010	1047	686.0	1.526	1.555	2	1.13	1.772	1.772
05001	1	0.0010	1047	686.0	1.526	1.555	1	1.63	0.612	0.612
05011	1	0.0010	1047	686.0	1.526	1.555	1	0.98	1.018	1.018
05053	1	0.0010	1047	686.0	1.526	1.555	2	1.19	1.685	1.685
28011	1	0.0010	1047	686.0	1.526	1.555	7	2.57	2.718	2.719

Relative Risk for US Cervical Cancer Mortality

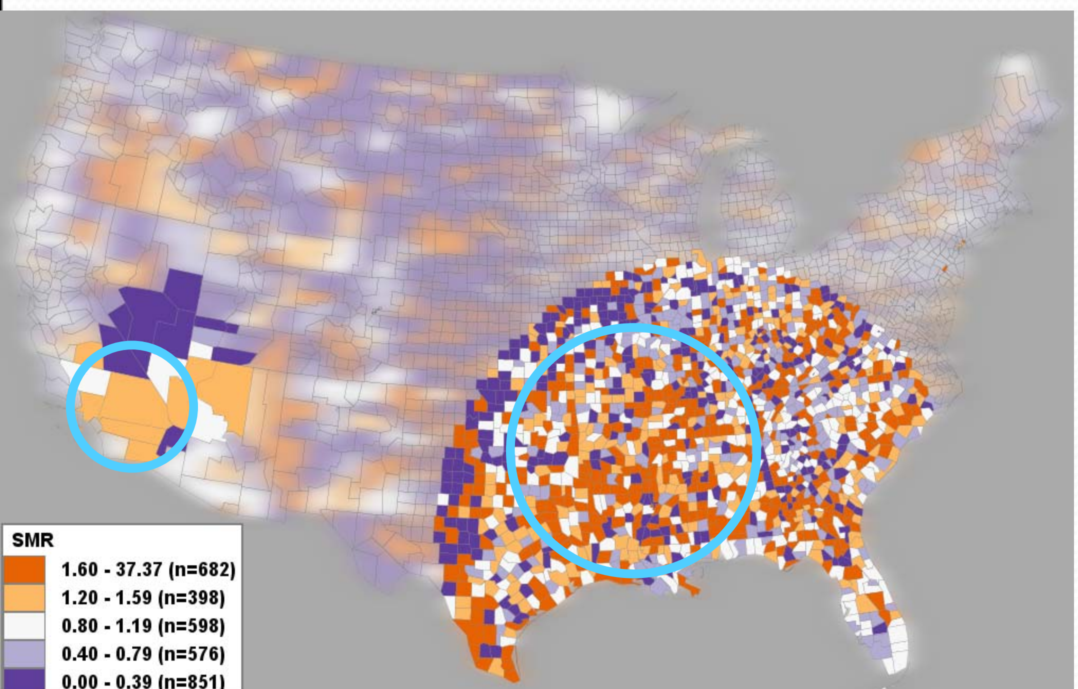
- Measured by Standardized Mortality Ratio (SMR)
  - Equals to observed/expected deaths
  - In theory, a value of 1 means normal risk.

- **Orange (high risk) ratio >1.2**
- White (normal) ratio= 0.8-1.2
- **Blue (low risk) ratio < 0.8**

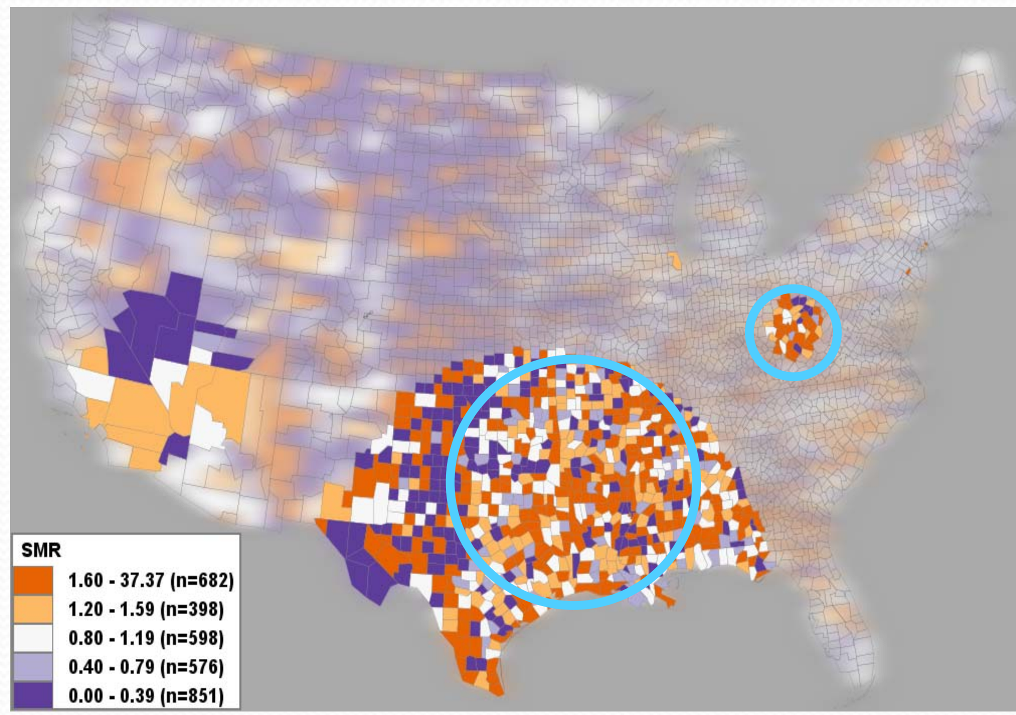


Heterogeneous clusters

- Large, in undesirable scales, less informative
- Mix of high and low risk enumeration units (e.g., counties)
- Core clusters: smaller, more homogeneous high risk



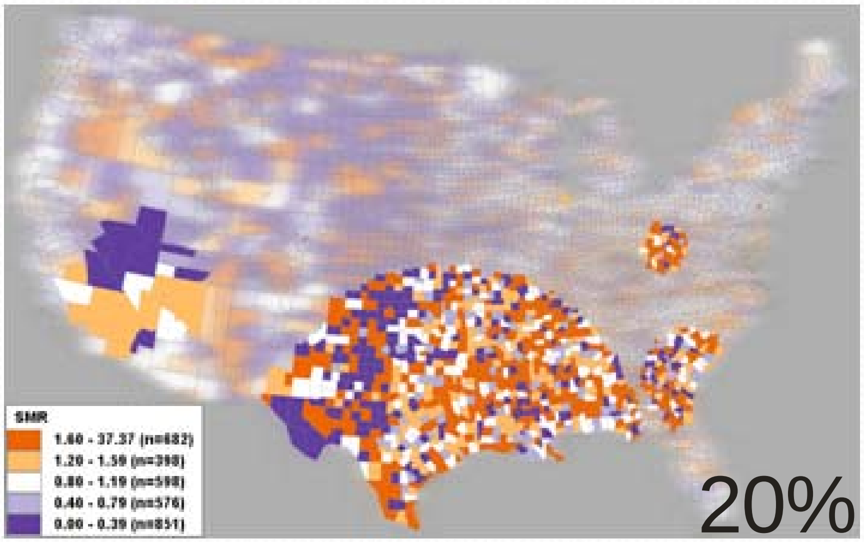
Maximum circle size =50% population

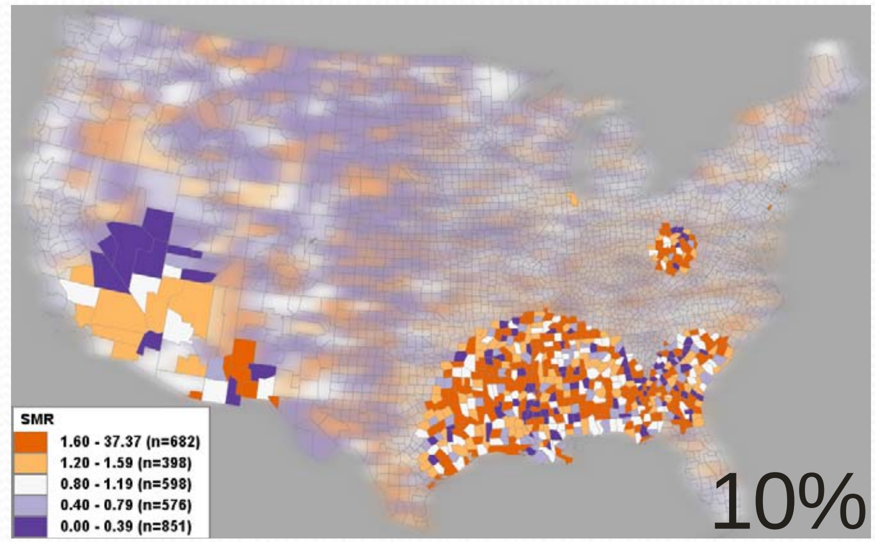


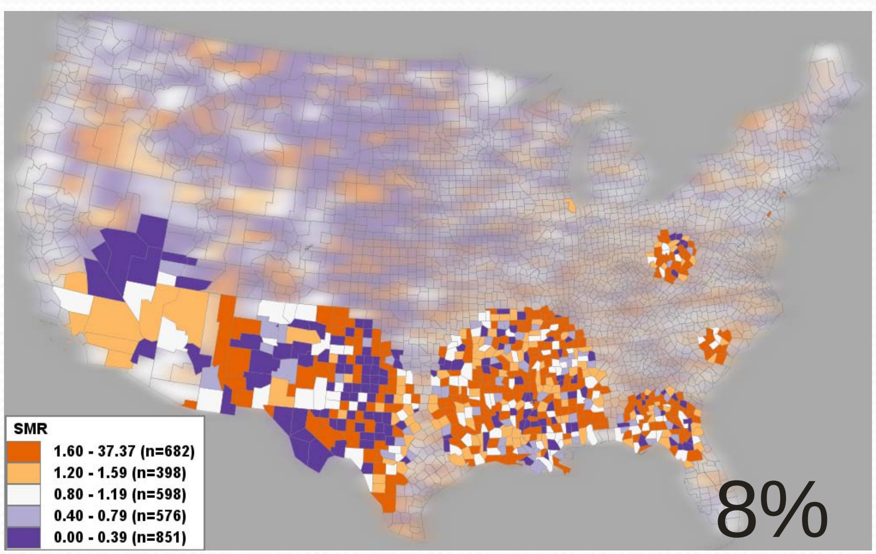
Maximum circle size =40%

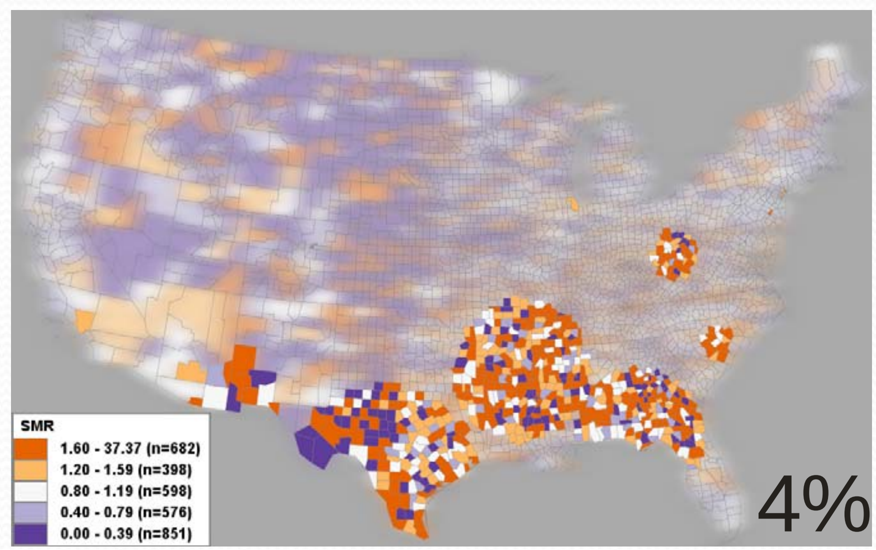
How about reduce the circle size

- Reducing circle size produces small, homogeneous clusters
- SaTScan is sensitive to circle size. Which clusters are more reliable?









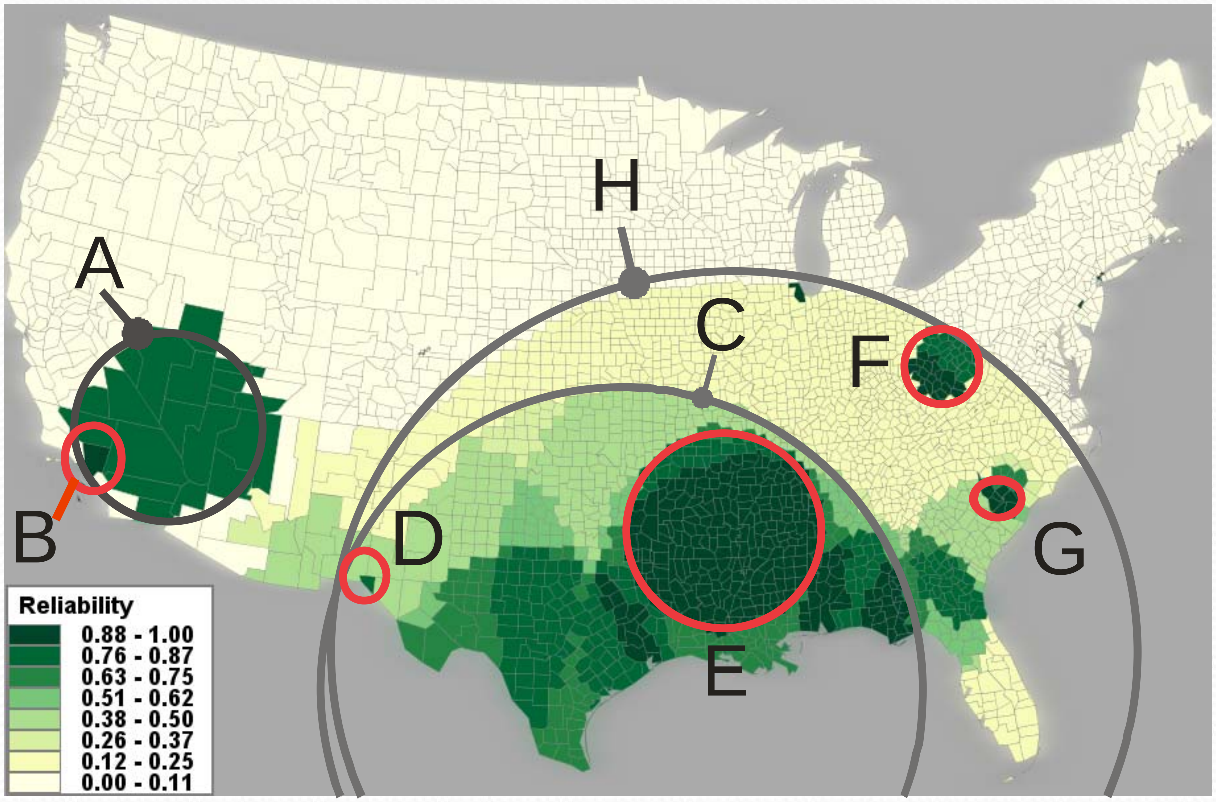
Reliability Visualization

- Estimation of reliability

R = C / S

  - R - reliability score for a unit (e.g., a county)
  - S - total number of scans
  - C - count that a unit is identified as high risk

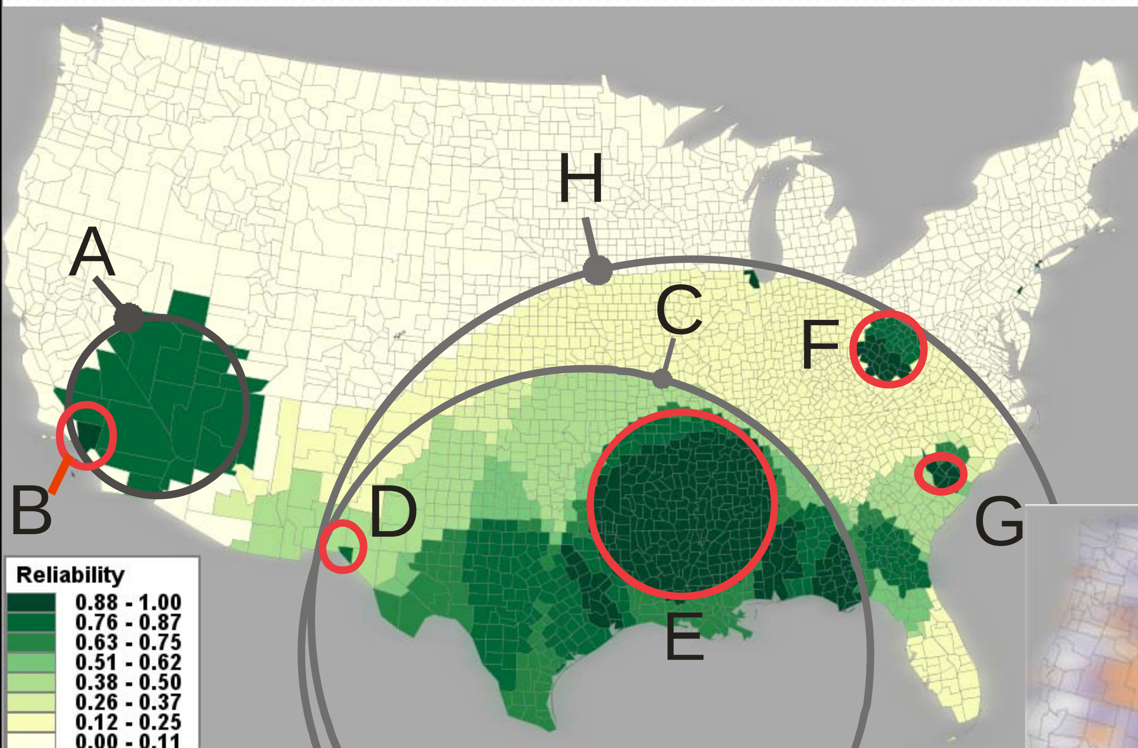
- Core clusters
  - More reliable across scales
  - **Dark, green color**
  - E.g., B, D, E, F, G.



Based on 8 scans: 4%, 6%, 8%, 10%, 20%, 30%, 40%, 50%

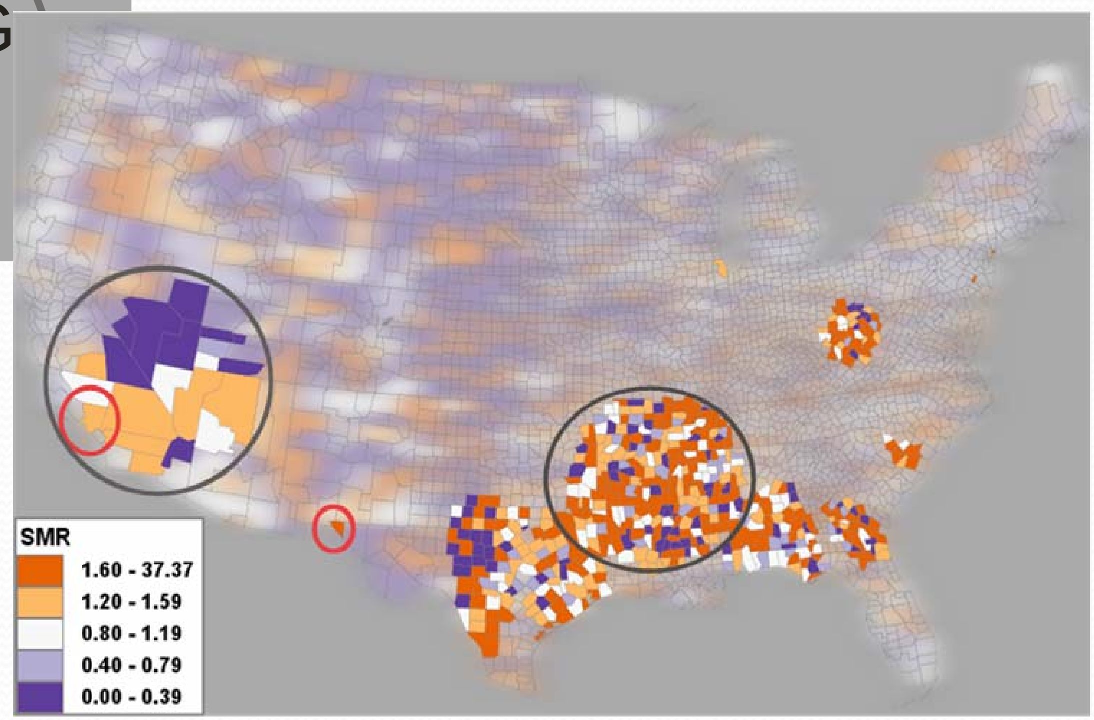
- Reliability - agreement among multiple results. (R-score)
- Validity - if a cluster is a true high risk region. (p-value)

Reliability Map Extract Core Clusters



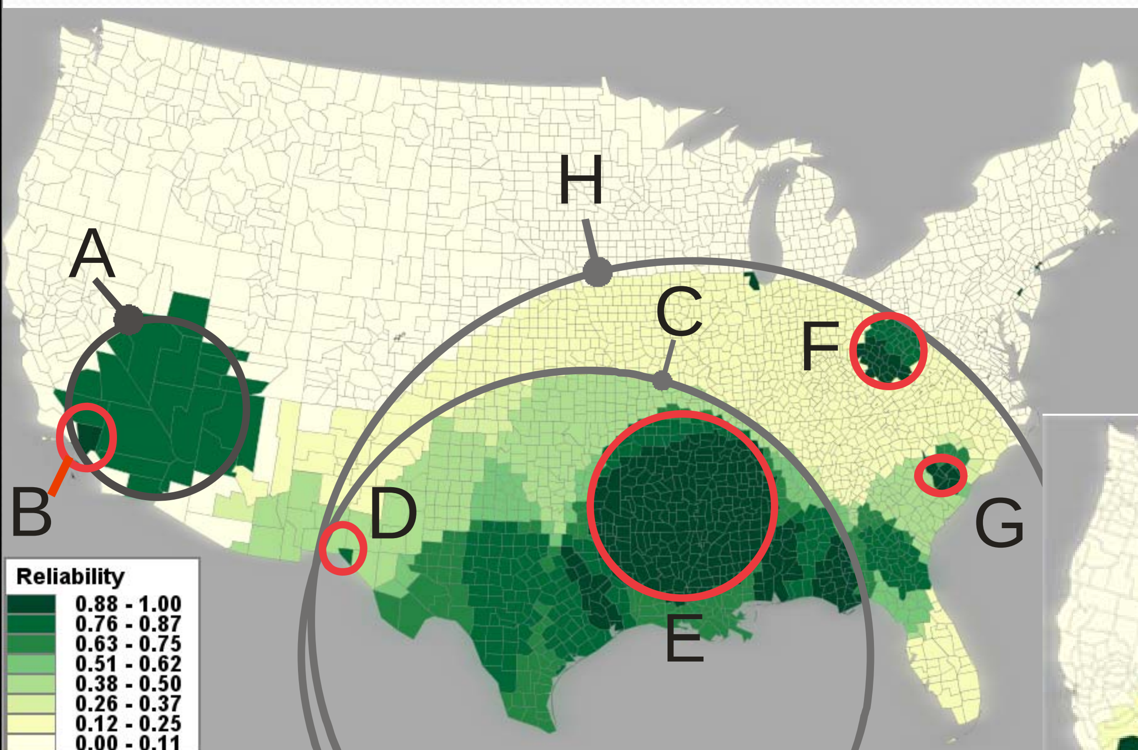
High reliable core clusters are
- stable across multiple runs
- smaller in size
- more homogeneous high risk

4%, 6%, 8%, 10%, 20%, 30%, 40%, 50%



SMR map displays only clusters with high reliability scores

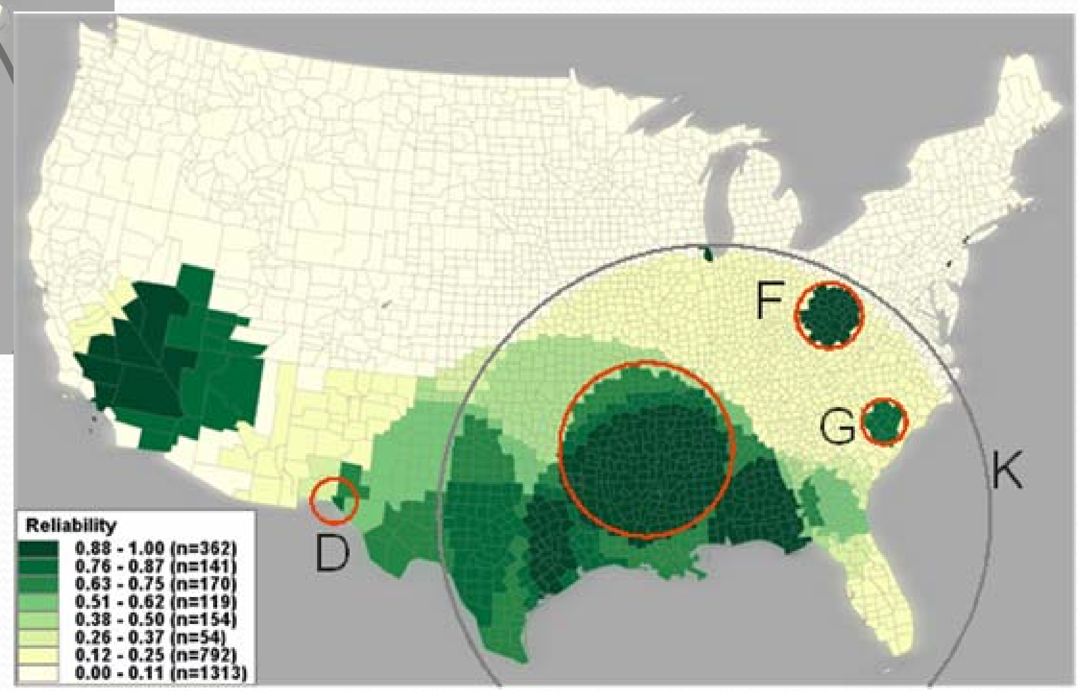
Reliability Map Alleviate Sensitivity



4%, 6%, 8%, 10%, 20%, 30%, 40%, 50%

Both results report same or similar core clusters (e.g., D, E, F, G)

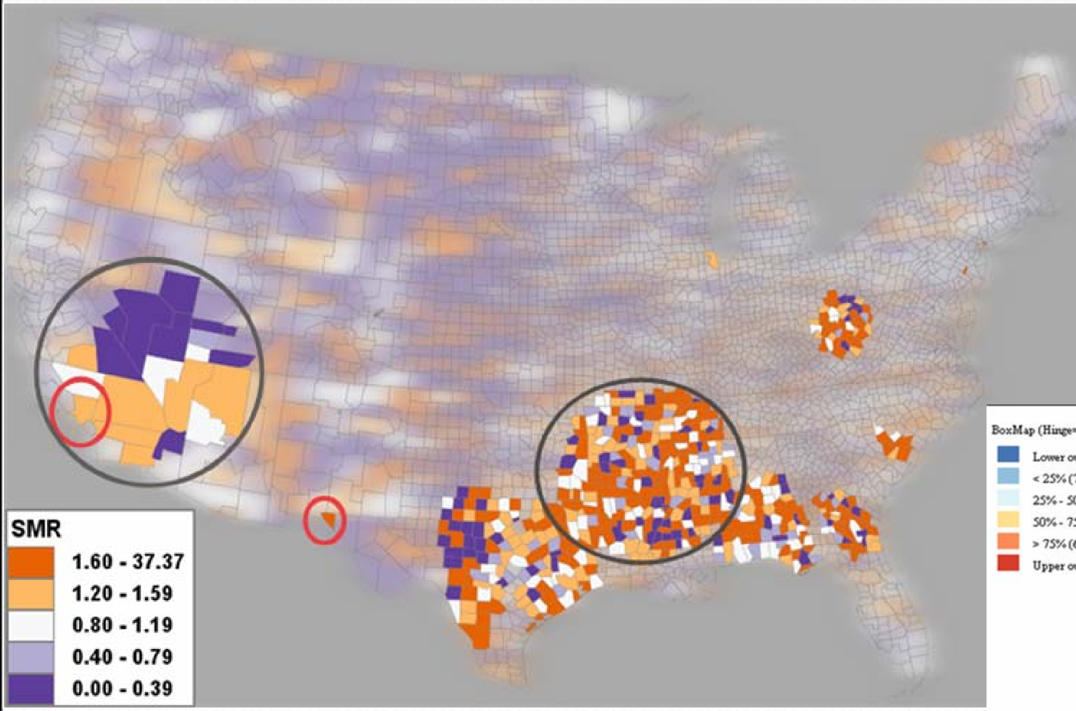
Clusters (H, K) of low R-score vary



5%, 7%, 9%, 11%, 19%, 29%, 39%, 49%

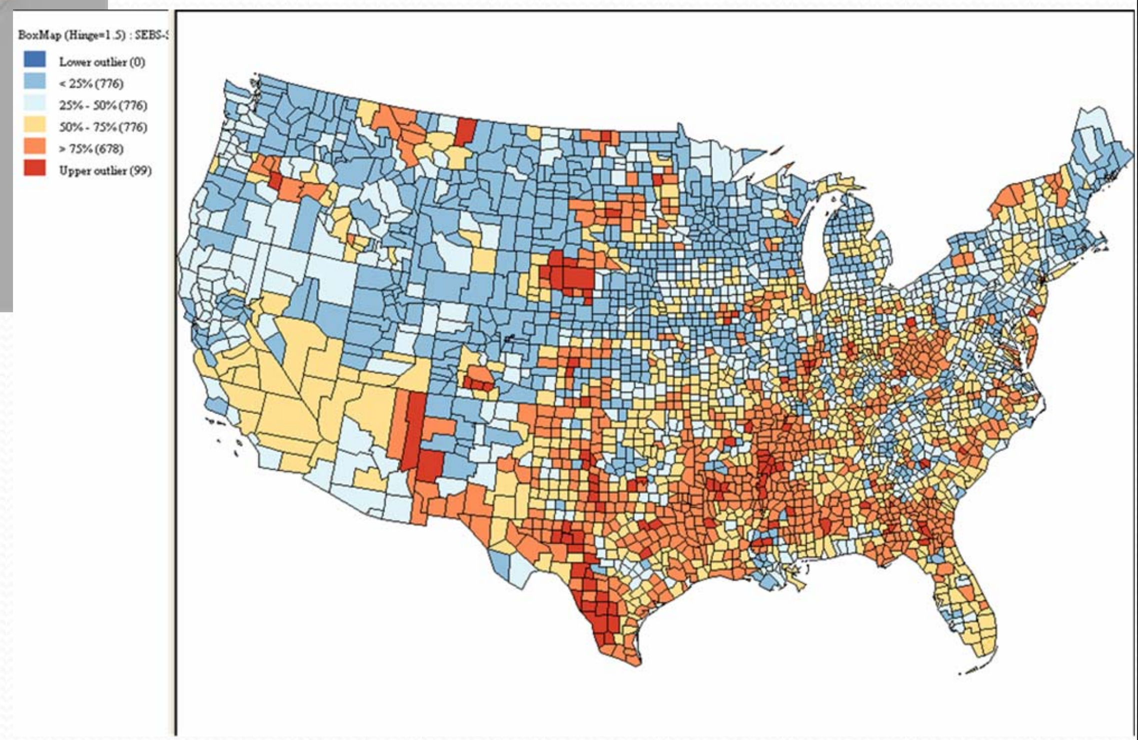
GeoVISTA Center

Compare with other method



- The two methods produce similar results
- A combination of SaTScan and reliability visualization produces more accurate clusters in terms of location and size

Core clusters by reliability visualization



Kafadar’s spatial smoothing method (in GeoDa)

Summary

- No a single optimized maximum circle size.
- Need run multiple scans with different sizes
- Reliability visualization can help
  - identify smaller, more homogeneous clusters
  - alleviate sensitivity of SaTScan results

GeoVISTA Center

Many thanks for your attention.

Questions?