Geovisual Analytics for the Disease Clusters Detection: U.S. Cervical Cancer Mortality 2000-2004

Jin Chen[1], Robert E. Roth[1], Adam Naito[1], Eugene J. Lengerich[2]
Alan M. MacEachren[1]

[1]GeoVISTA Center, Geography Dept, Penn State University
[2]Department of Health and Evaluation Sciences, Penn State College of Medicine

Outlines

I. Background
   – Visual analytics
   – Mapping U.S. Cervical cancer mortality
II. Spatial cluster detection
III. Scale effects on spatial cluster detection
IV. Geovisual analytics to enhance spatial cluster detection at multiple scales
   – Reliability visualization
   – Value-by-alpha visualization
V. Conclusion

What is Geovisual analytics

- **Visual analytics**
  - "the science of analytical reasoning facilitated by interactive visual interfaces." (Thomas 2005).
  - Proposed by National Visual Analytic Center (NVAC), National Institute of Health (NIH), and NSF
  - Draws upon methods from multiple disciplines: visualization, statistics, data mining, cognition science...
  - Offers new perspectives, approaches for addressing complex questions.
- **Geovisual analytics**
  - focus on solving problems in geographic context

This research

- This research combines visual, statistical, and computational methods to indentify spatial disease clusters at multiple scales.
- Study on U.S. cervical cancer mortality 2000-2004.

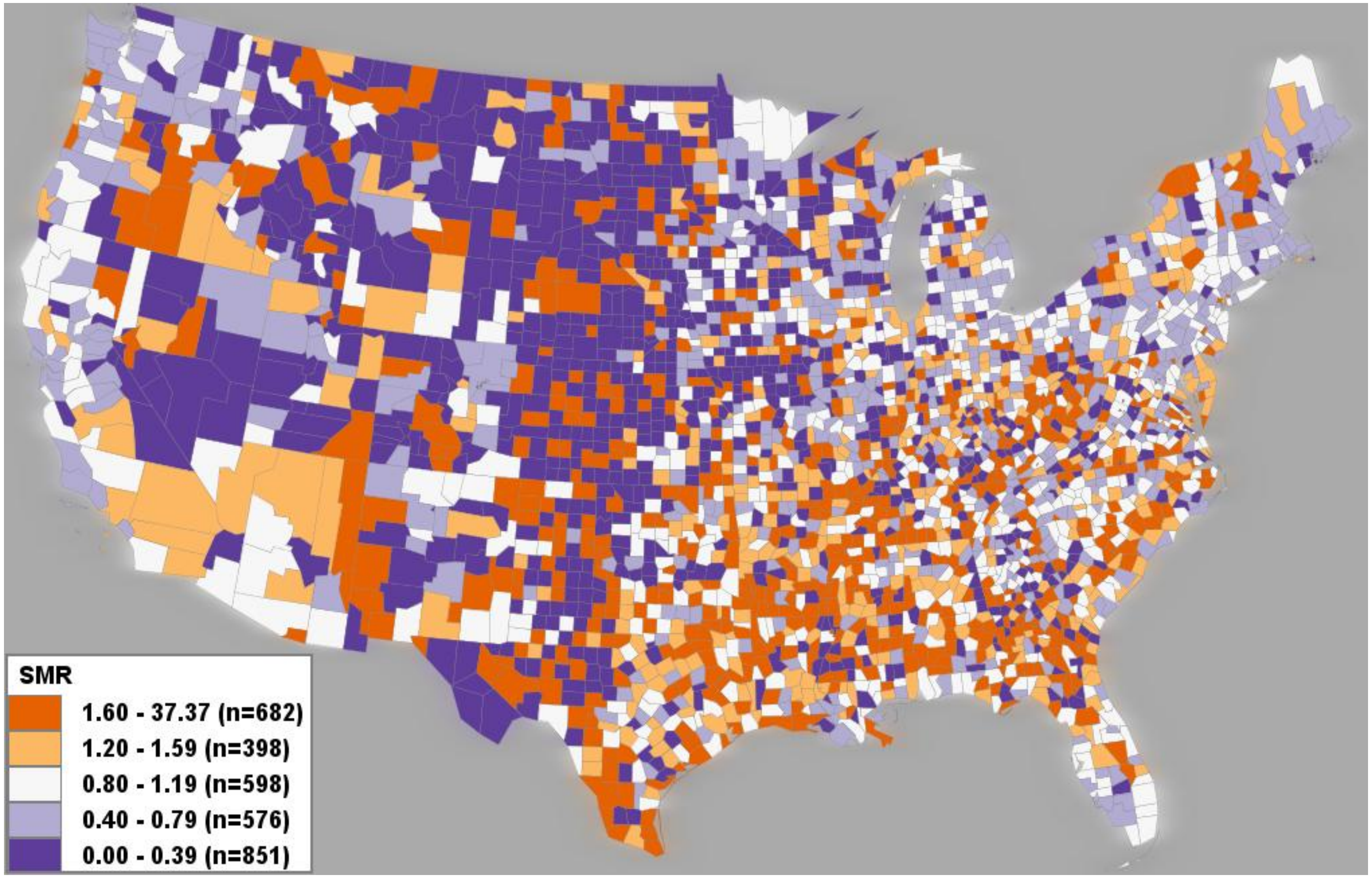
U.S. Cervical Cancer Mortality Ratio

All races, 2000-2004

Standardized Mortality Ratio (SMR)

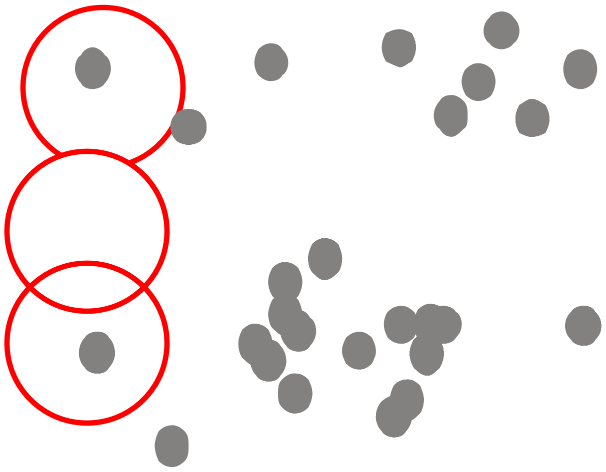
- Measure relative risk: SMR = observed deaths/expected deaths
- In theory, a ratio value of 1.0 means normal risk.

- **High risk**
  - **ratio >1.2**
  - **orange**
- Normal risk
  - ratio = 0.8-1.2
  - white
- **Low risk**
  - **ratio < 0.8**
  - **blue**



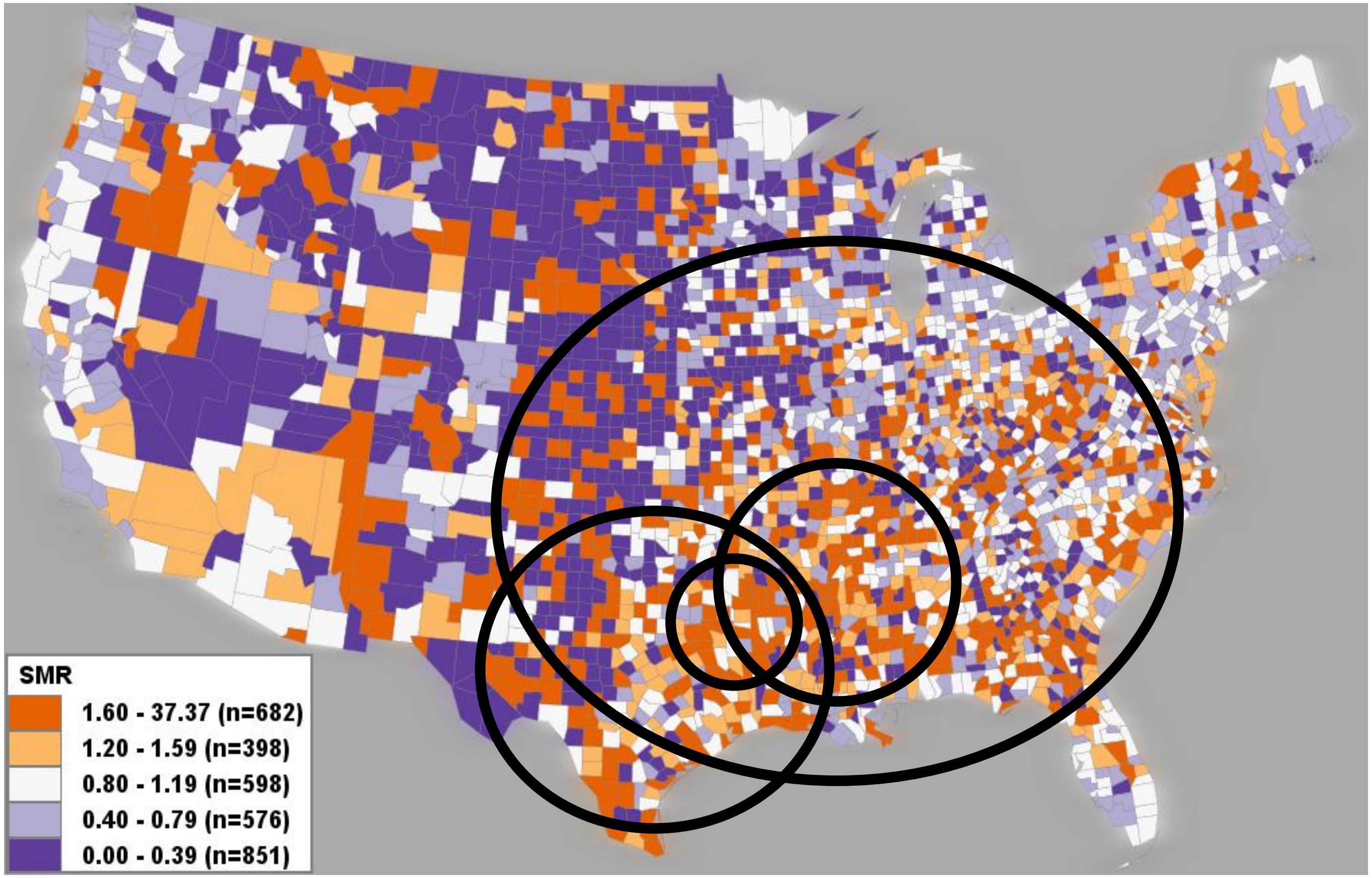
Scan Statistics

- 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.



Scale in spatial cluster identification

- Geographic scale - **size or spatial extent of study area** [McMaster, Goodchild 2004].
- Cluster scale - the size limitation of spatial clusters.
- **Finding appropriate scales is not easy. Where to 'cut'?**



Scale sensitivity

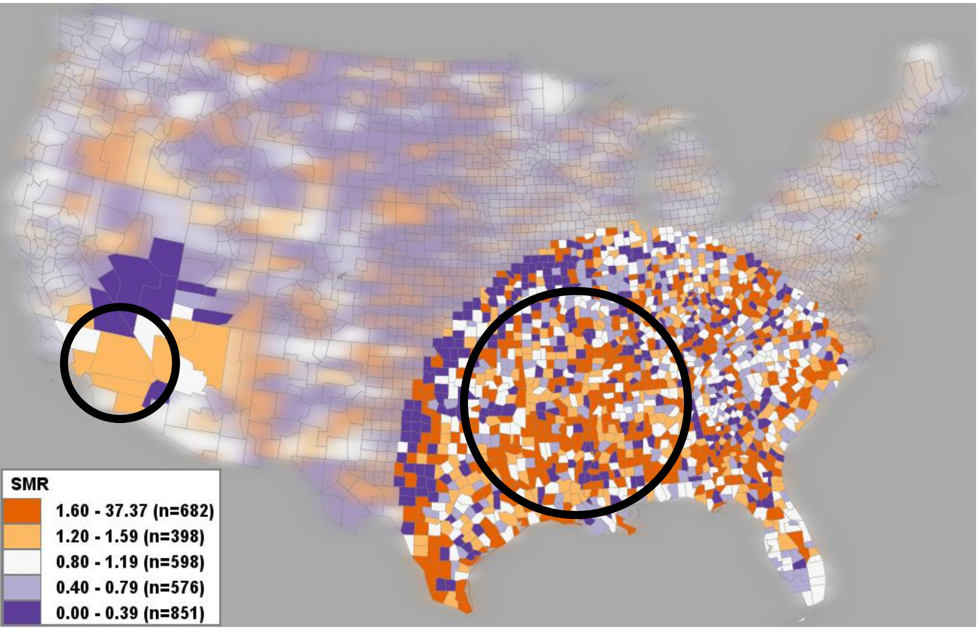
- Spatial cluster detection is sensitive to scale choices
  - at large scales, heterogeneous clusters are often reported.
  - at small scales, clusters are unstable in size and location.
- Scale critically affects spatial cluster identification.
- This research focuses on addressing the scale sensitivity problem.

Scale parameter in scan statistic method

- Explicitly, set the cluster scale as the *maximum circle size* (i.e., diameter) of a cluster.
- Implicitly, set the cluster scale as the ***percentage of population at risk***.
  - E.g., a scale of 50% of population means a cluster can contain at most 50% of total population at risk.
  - This way is often seen in studies of public health, and adopted in this research.

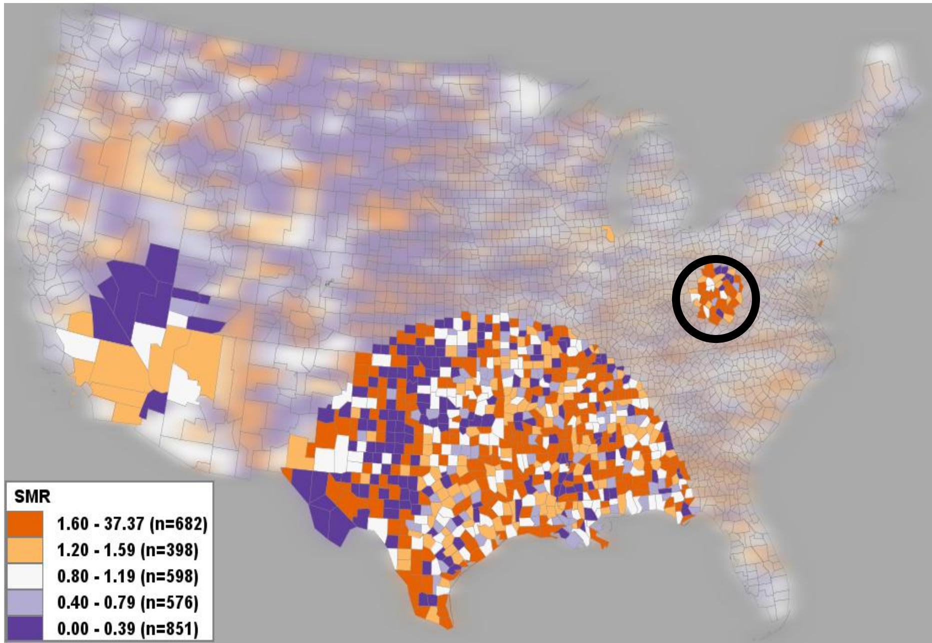
Heterogeneous clusters reported at large scales

- Heterogeneous clusters
  - A high-risk cluster containing considerable number of low risk locations.
- Clusters reported by SaTScan



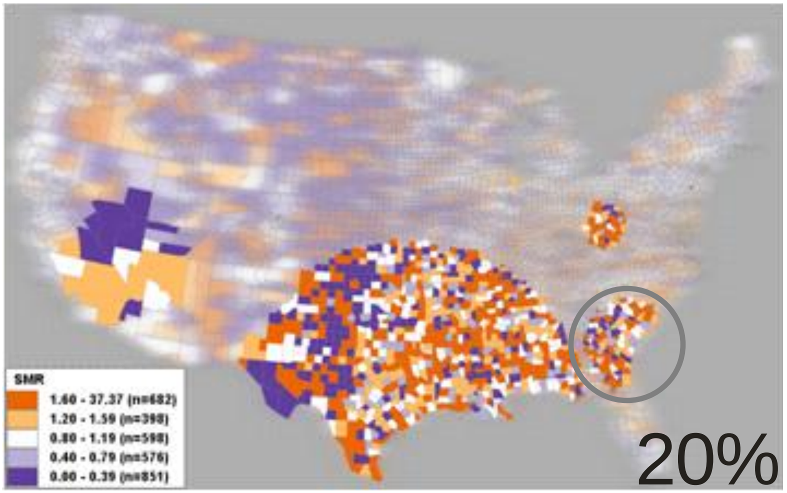
- two clusters are reported at scale of 50%
- they are heterogeneous, less informative
- homogeneous regions in the black circles are more interesting.

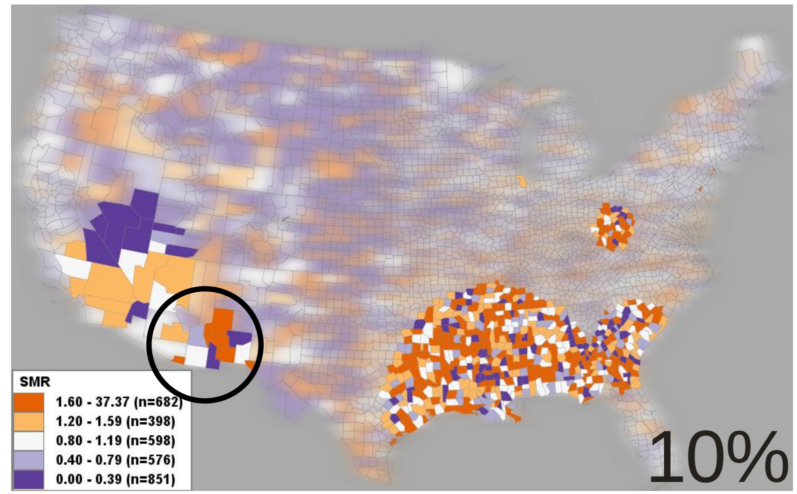
Reduce the **scale to 40%** of population
- three clusters are reported
- one is much more homogeneous

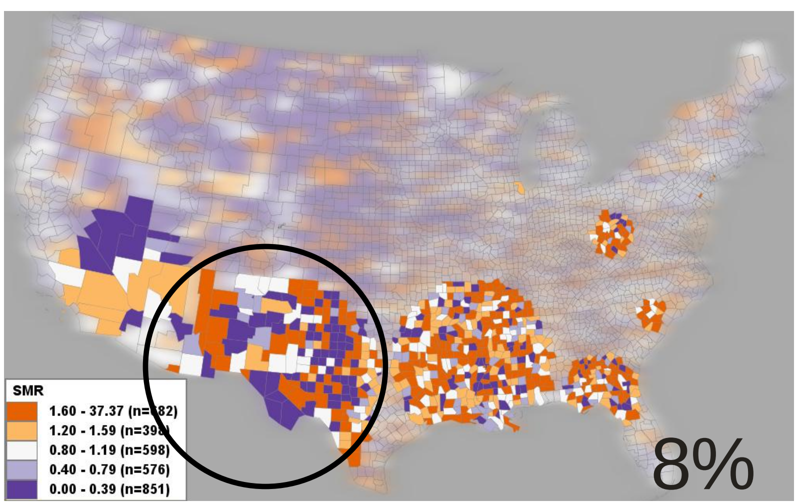


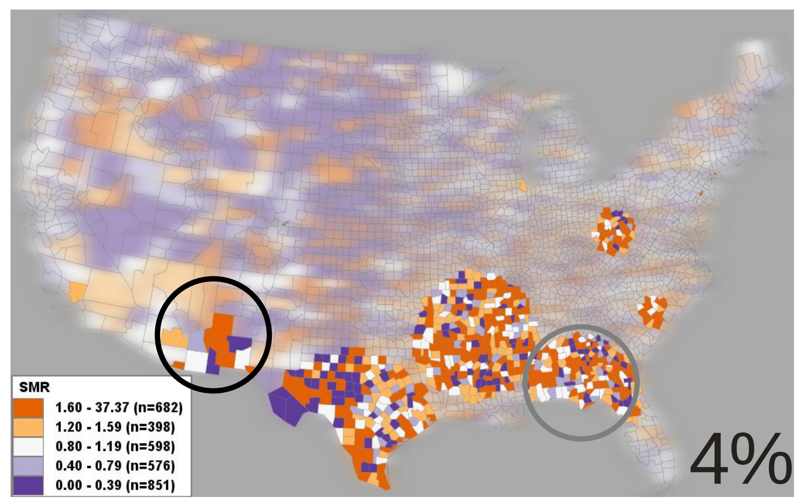
Unstable clusters at small scales

- With smaller scales, some clusters are unstable in size and location.









Multi-scale analysis to confound scale sensitivity

- Instead of searching for a single “optimized” scale, we proposed to run multiple scans at systematically-selected scales.
- Try to find agreement (i.e., high risk locations) among the results produced at different scales.
- High risk locations reported by more results are more reliable.

Reliability Visualization

- Reliability (as high risk)

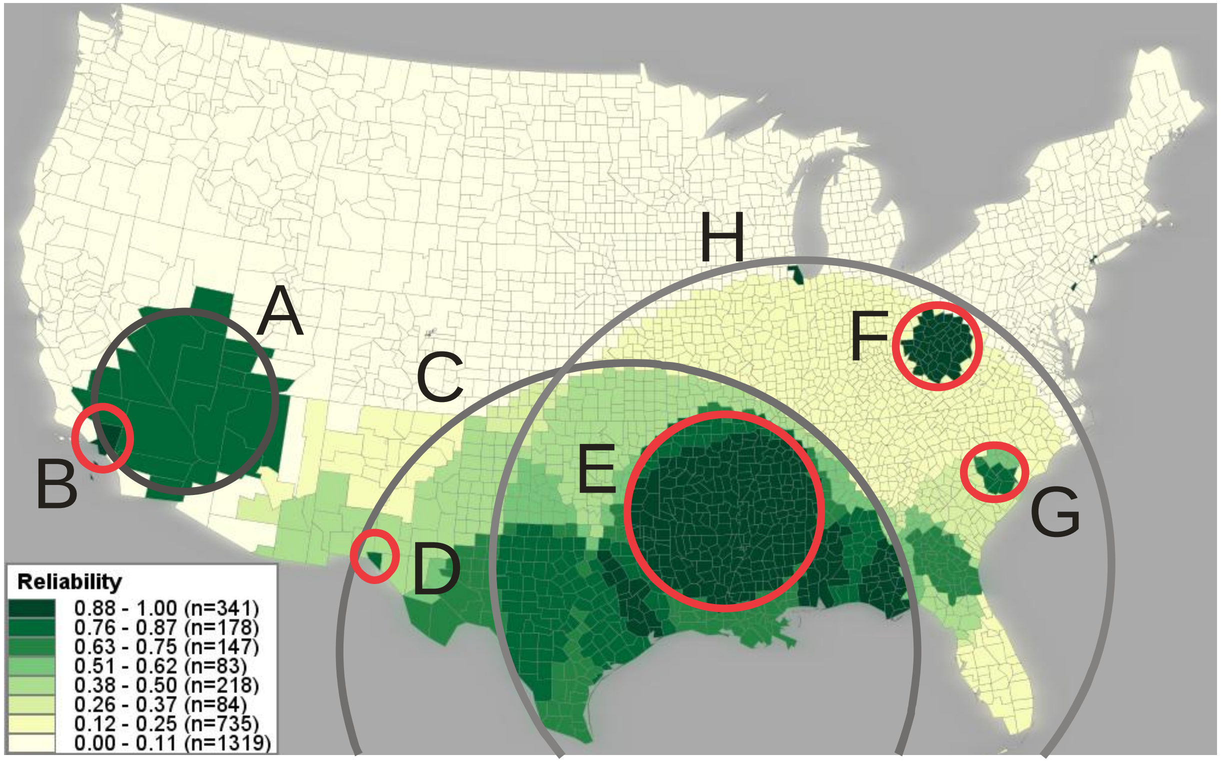
  R = C / S

  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

- Reliable clusters
  - Stable across scales
  - dark-green color
  - e.g., cluster B, D, E, F, G in red circles

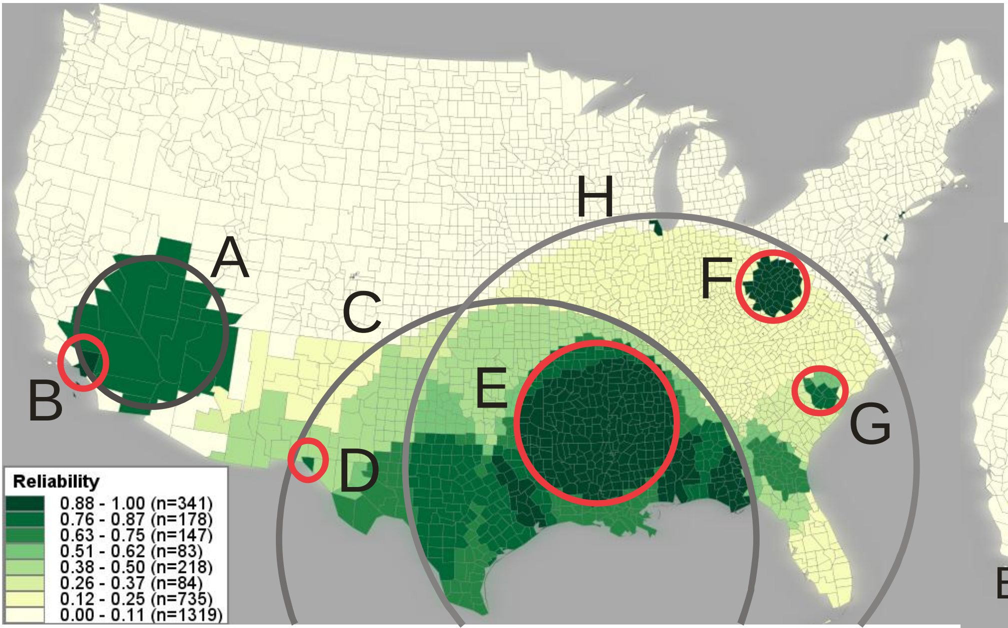


Summary of 8 scans with 8 scale values: 4%, 6%, 8%, 10%, 20%, 30%, 40%, 50%

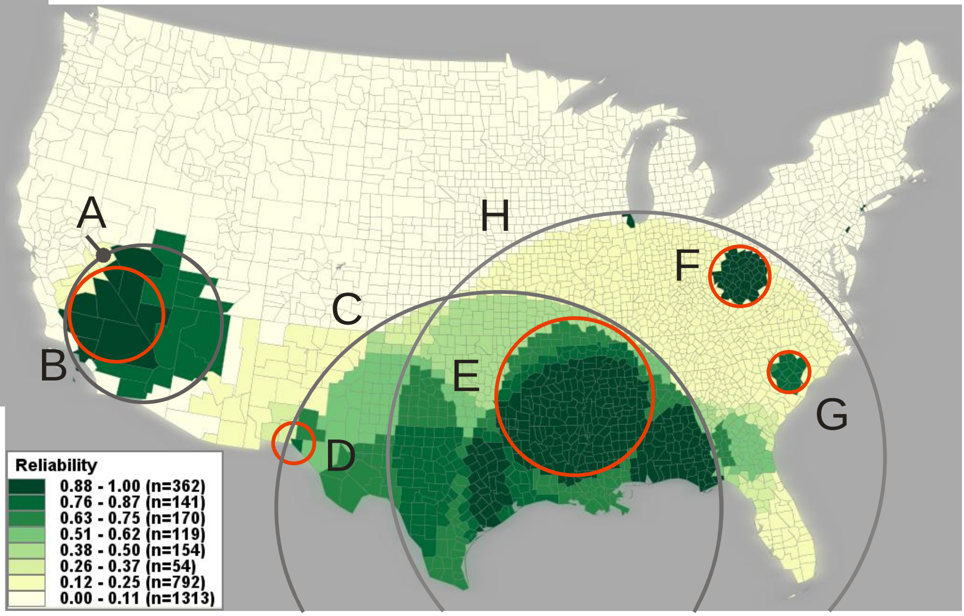
- Note: reliability is different from validity, the later indicates if a cluster is a true high risk region.

Reliability visualization alleviate scale sensitivity

- Compare two reliability maps that were produced at different set of 8 scales.
- Similar reliable, high risk clusters (in red circles) are reported by both maps.
- Therefore, the results produced by a reliability map are **less sensitive** to the scaling choices.



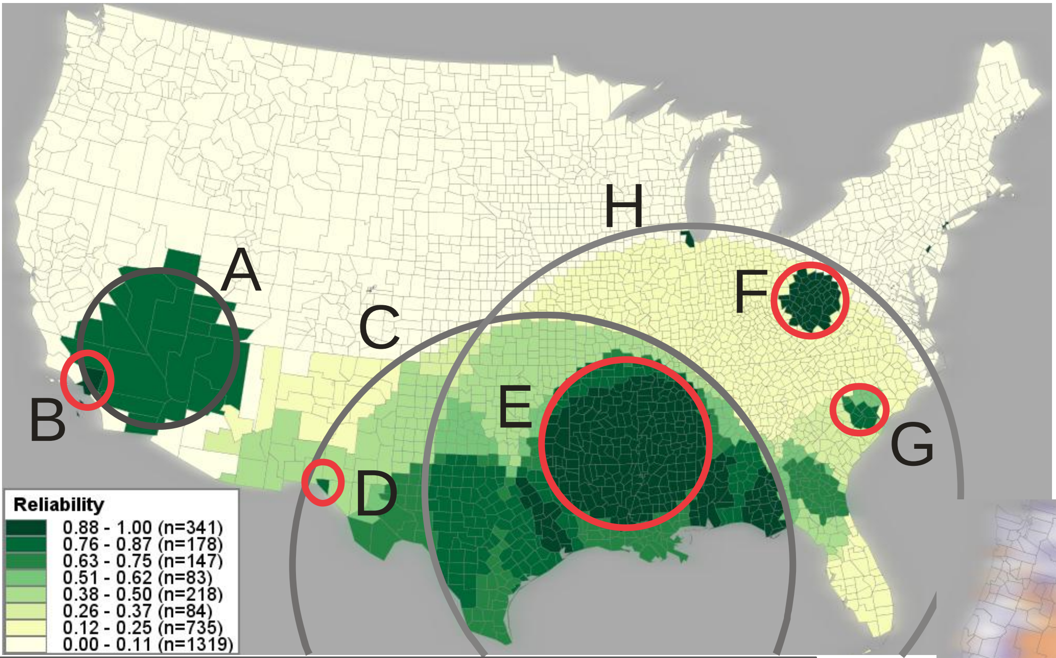
8 scales: 4%, 6%, 8%, 10%, 20%, 30%, 40%, 50%



8 scales: 5%, 7%, 9%, 11%, 19%, 29%, 39%, 49%

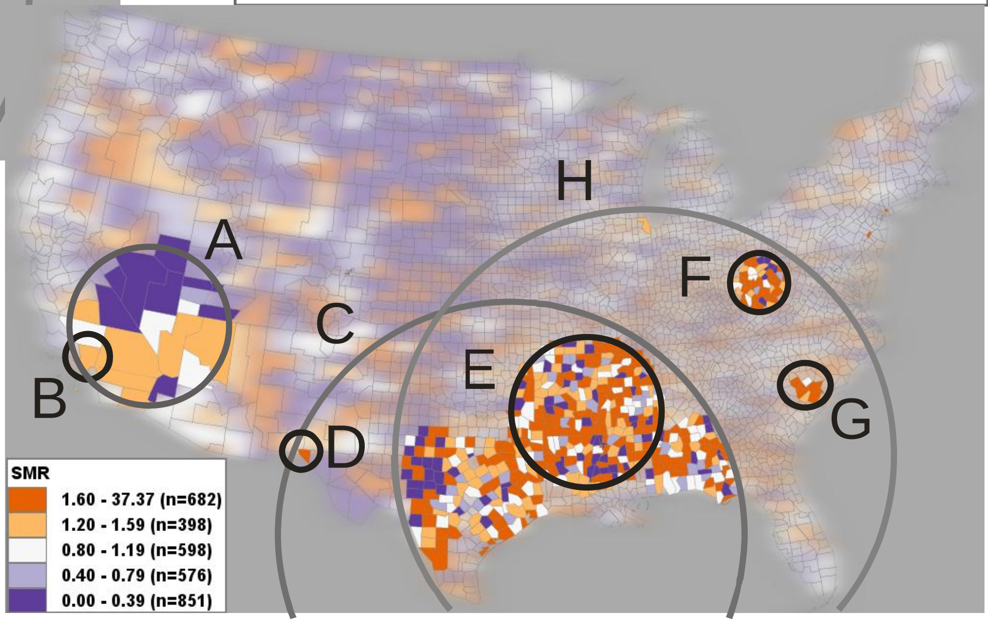
Reliability visualization extracts homogenous clusters

High reliable clusters are more homogeneously in high risk



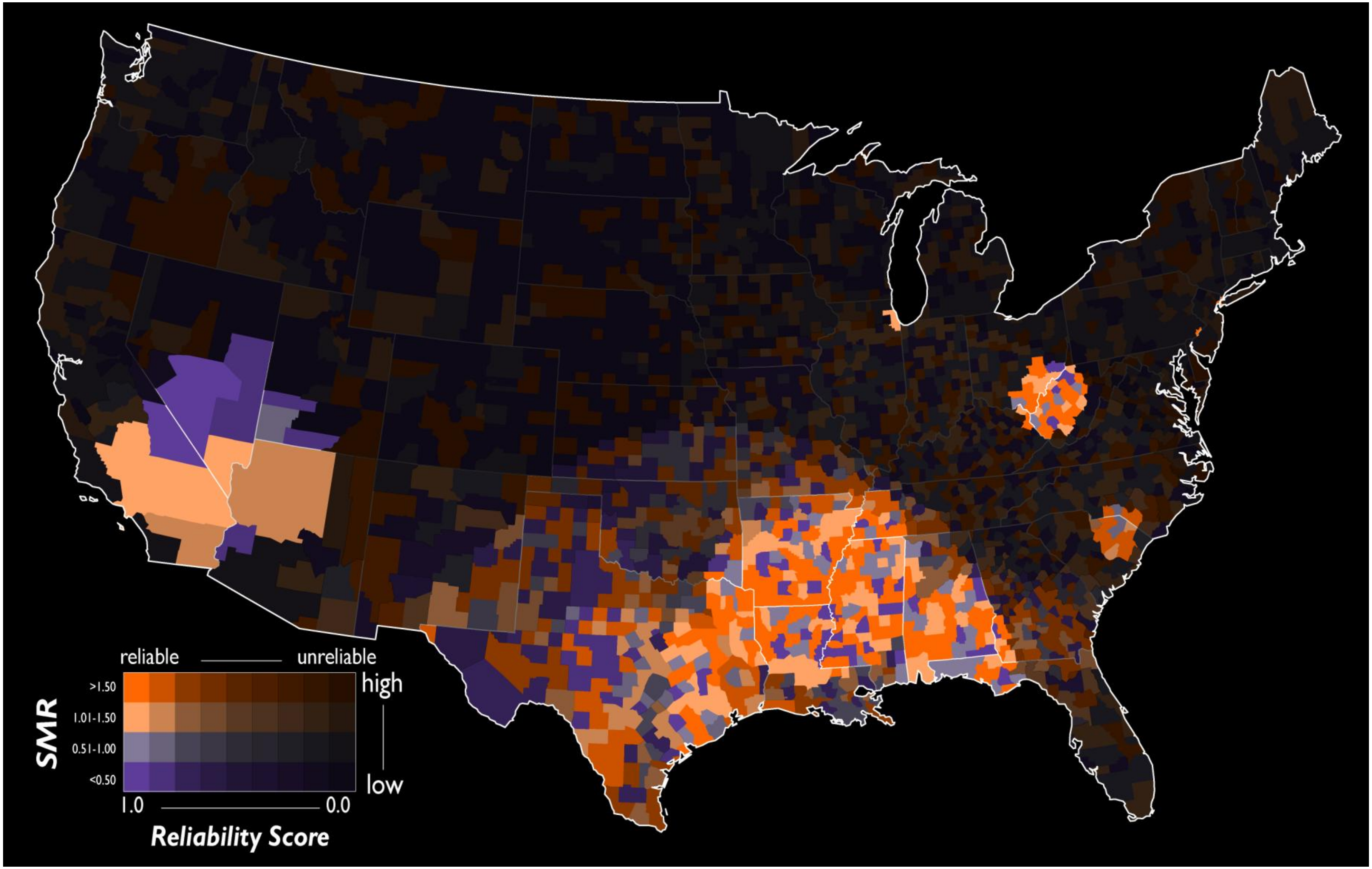
4%, 6%, 8%, 10%, 20%, 30%, 40%, 50%

The SMR map below displays reliable, high risk clusters are in black circles

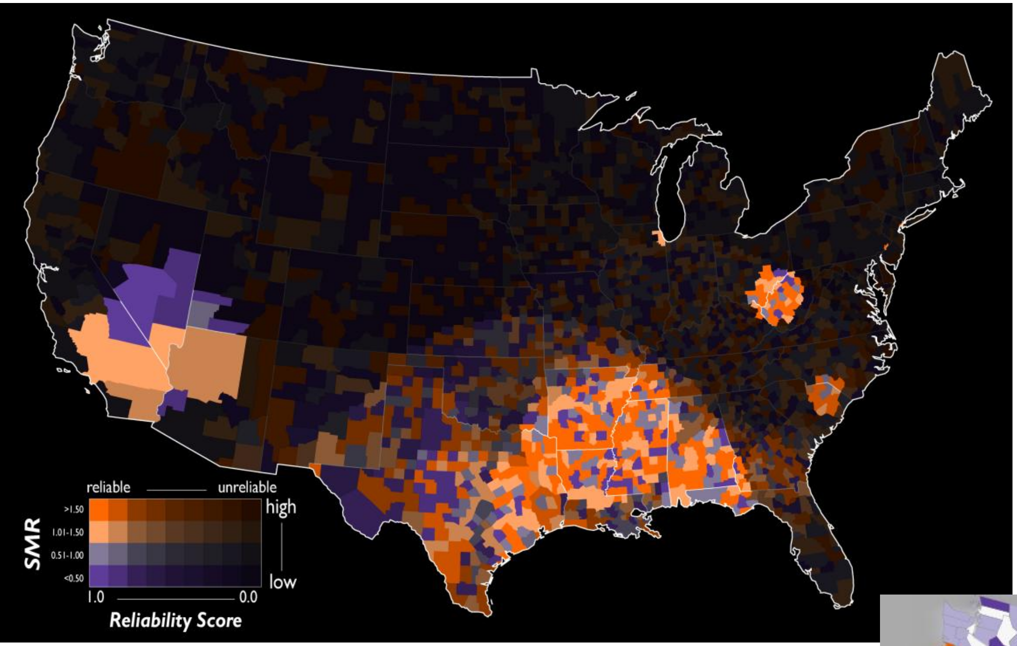


Concurrently visualize SMR and reliability score

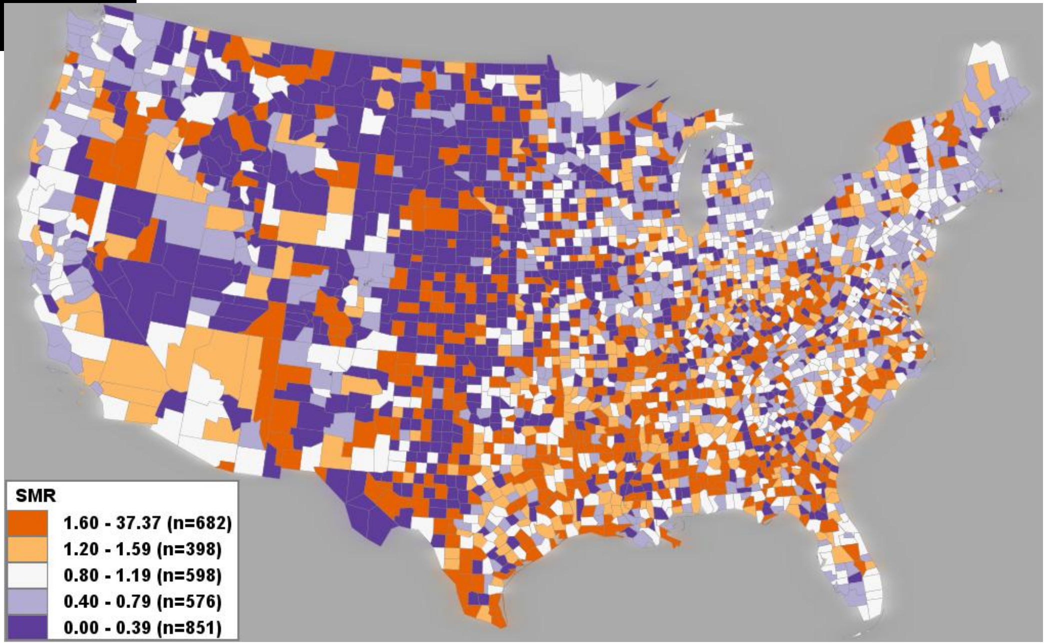
Reliable, high risk regions of US cervical cancer mortality 2000-2004: southern CA, New Mexico, NV, Deep South, Appalachia, south Carolina, Chicago area.



Compare to simple rate mapping



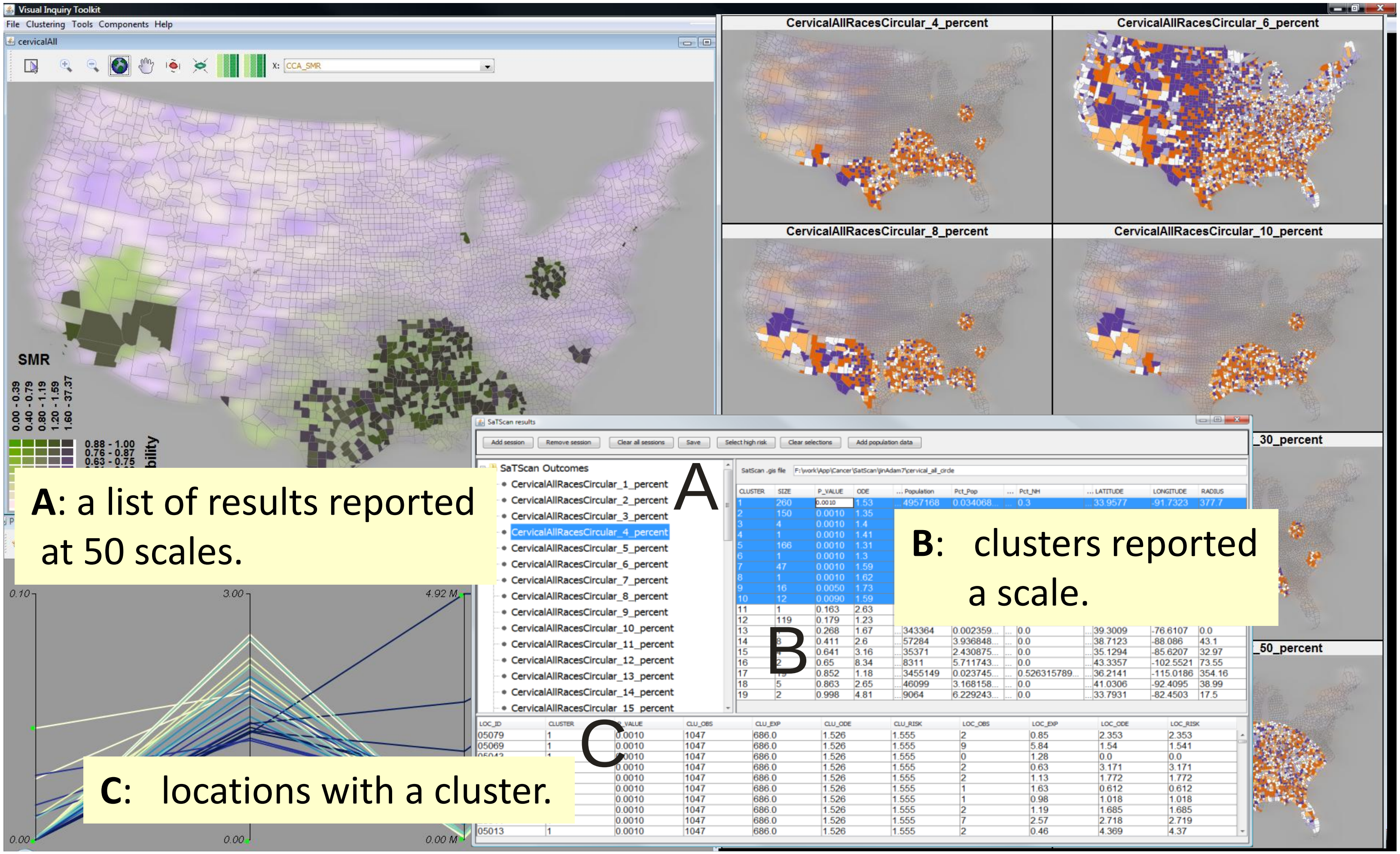
SMR + reliability



SMR map

Visual Inquiry Toolkit (VIT)

- VIT is coupled with spatial scan statistic to present clusters on geographic maps.
- Allow interactive exploration of spatial clusters at multiple scales



Conclusion

- A single "optimized" scale is hardly found, multi-scale cluster analysis is necessary.
- Reliability visualization can alleviate scale sensitivity of spatial scan statistics methods
- See more in the paper:

INTERNATIONAL JOURNAL OF HEALTH GEOGRAPHICS

UNOFFICIAL IMPACT FACTOR 1.61

Methodology

Highly accessed Open Access

Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality

Jin Chen[1], Robert E Roth[1], Adam T Naito[1], Eugene J Lengerich[2] and Alan M MacEachren[1]

1 GeoVISTA Center, Department of Geography, the Pennsylvania State University, University Park, USA
2 Department of Public Health Sciences, the Pennsylvania State University, Hershey, USA

author email corresponding author email

International Journal of Health Geographics 2008, **7**:57   doi:10.1186/1476-072X-7-57

The electronic version of this article is the complete one and can be found online at: http://www.ij-healthgeographics.com/content/7/1/57

Received: 29 July 2008
Accepted: 7 November 2008
Published: 7 November 2008

Thank you very much for your attention.

- Related information can be found at: http://www.personal.psu.edu/users/j/x/jxc93/ (or simply google: **Jin CHEN PSU**)

- Questions?