Spatial statistical analysis at the National Cancer Institute

NATIONAL CANCER INSTITUTE®

Linda Williams Pickle, Ph.D.

ESRI Health GIS Conference

October 26, 2005

Steps in any statistical analysis project

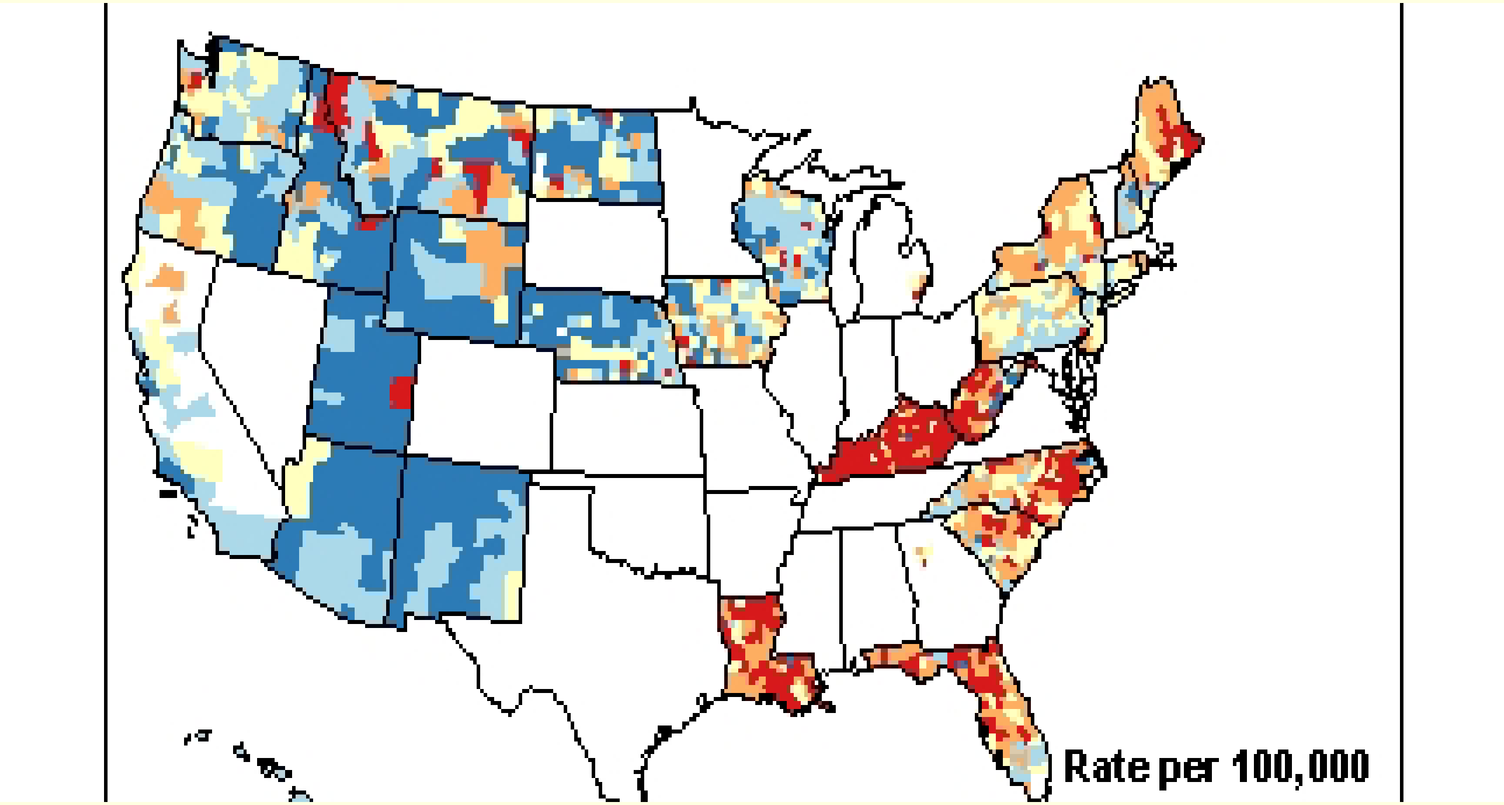
- Data exploration
  - Quality control – errors in data?
  - Distribution of variables
  - Associations among predictor variables? between outcome & predictors?
  - Establish hypotheses to be tested, goals
- Statistical modeling (or hypothesis testing)
  - Verify model assumptions
  - Apply model; estimate parameters
  - Assess fit of the model, modify until satisfactory
- Communicate results to client, public, etc.

- Will illustrate application to spatial data using cancer rate examples

Spatial data exploration

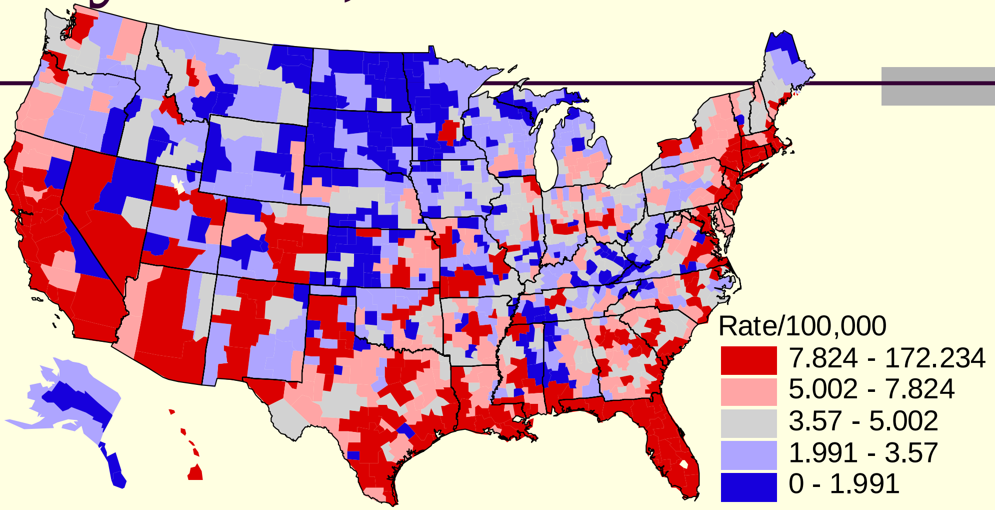
- Visualization of patterns in original spatial data
  - QC: any anomalies? Holes in map, extreme outliers?
  - Evaluate general pattern
    - Smoothing
    - Clustering
    - Outliers
    - Measures of spatial correlation
- Hypothesis generation
  - What risk factor maps look similar to cancer rate map? (How to measure similarity of patterns on multiple maps?)
- Select potential predictors for model (covariates)

Quality control: Missing some CA data

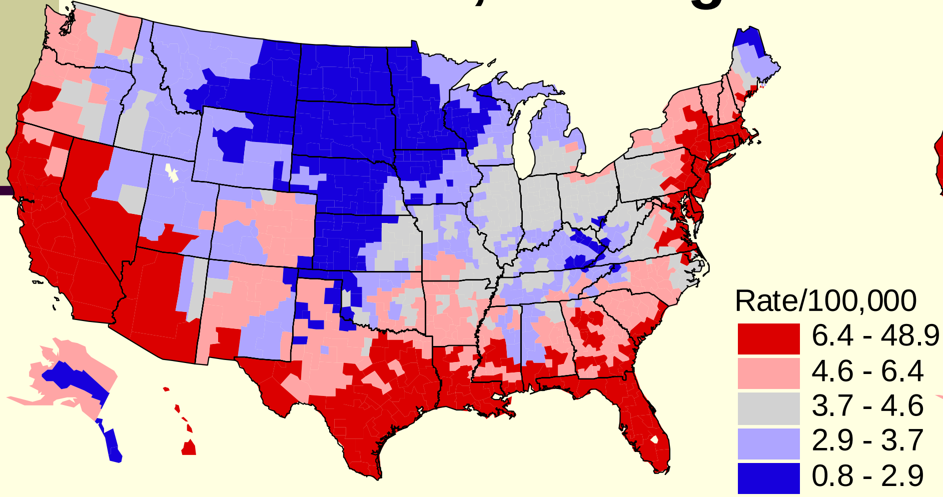


Smoothing:
HIV mortality rates, 1988-92

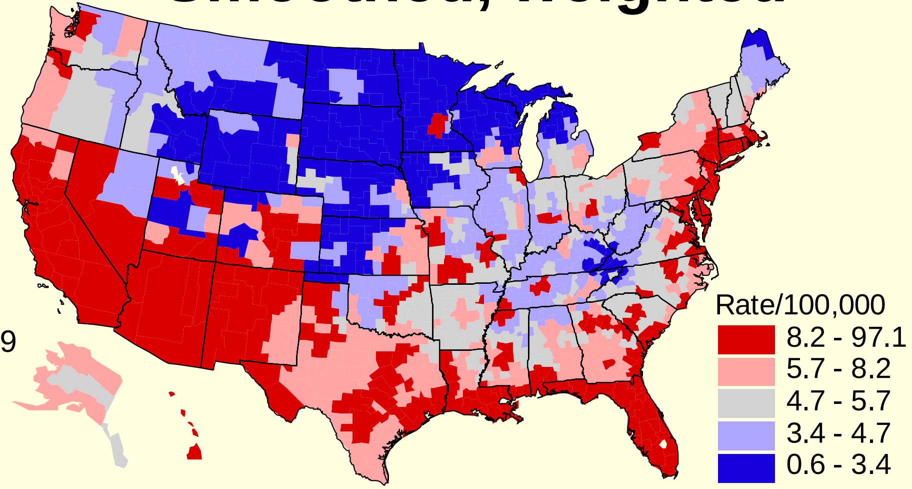
Original data:



Smoothed, unweighted



Smoothed, weighted

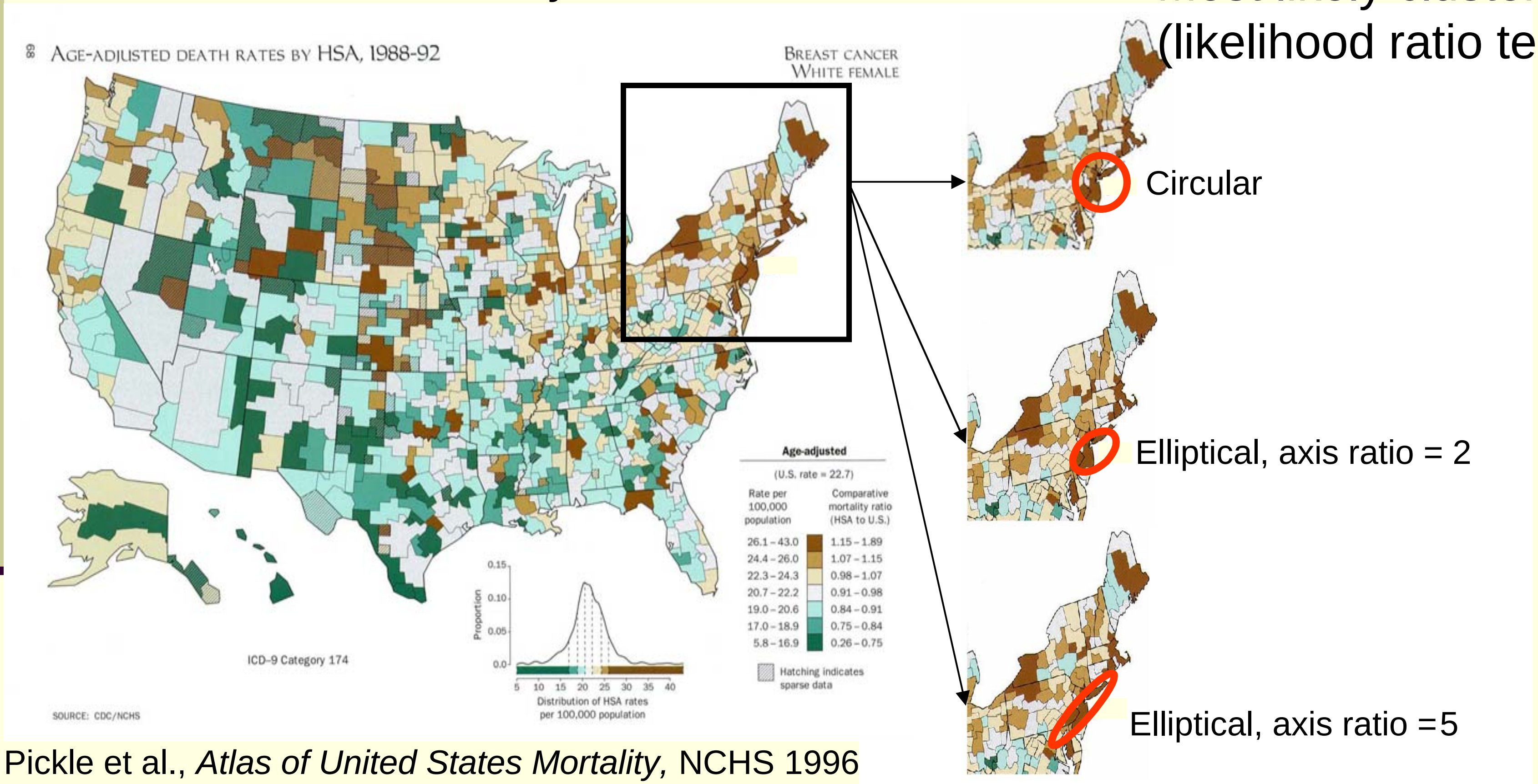


Source: Pickle et al., *Atlas of United States Mortality*, NCHS, 1996.

Cluster identification using SaTScan: Breast cancer mortality rates

Breast cancer mortality rates

Most likely cluster (likelihood ratio test)



Circular

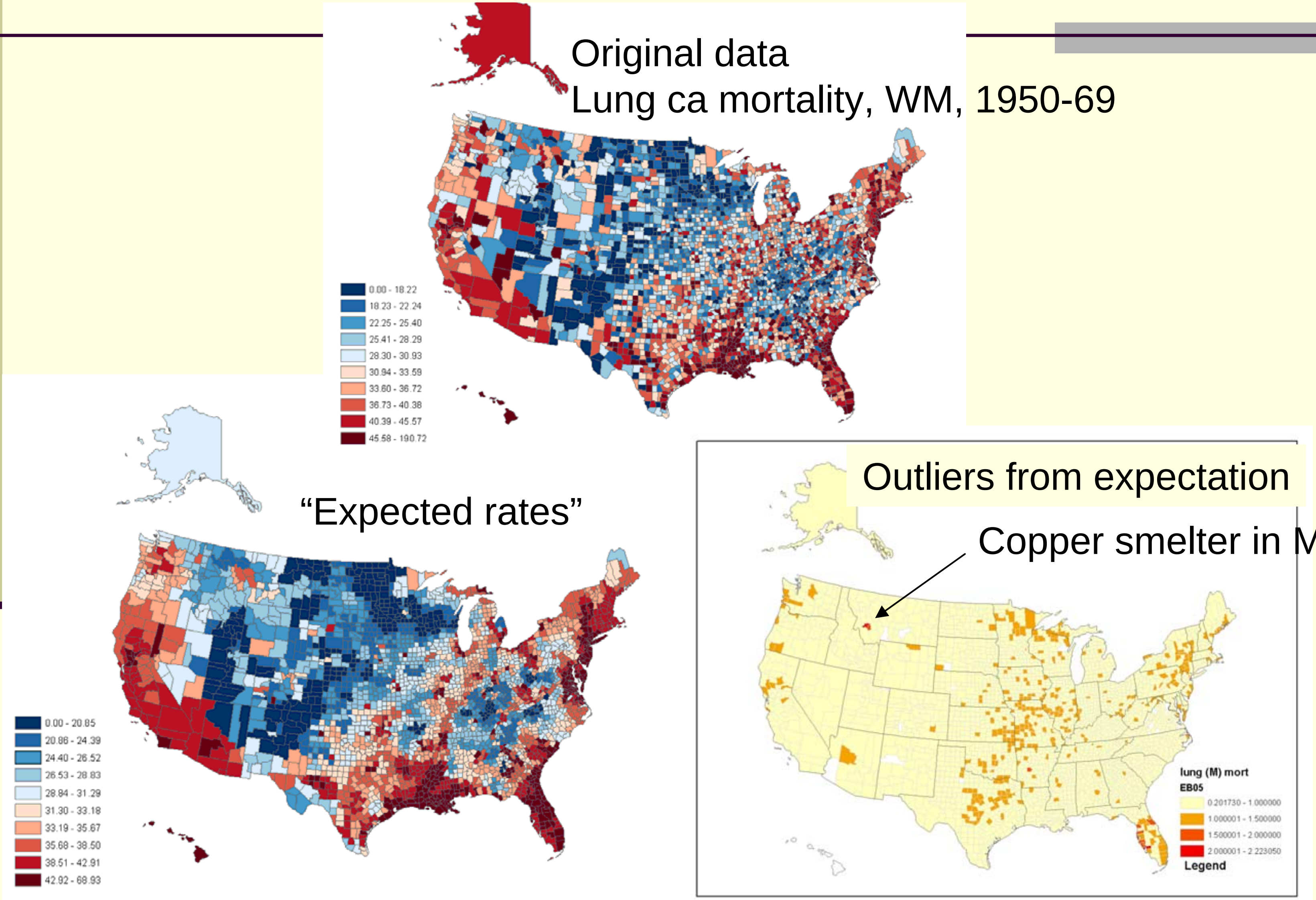
Elliptical, axis ratio = 2

Elliptical, axis ratio =5

Pickle et al., *Atlas of United States Mortality,* NCHS 1996

SaTScan by Martin Kulldorff, available at *www.satscan.org*

Outlier identification:
Values statistically different from expectation



Original data
Lung ca mortality, WM, 1950-69

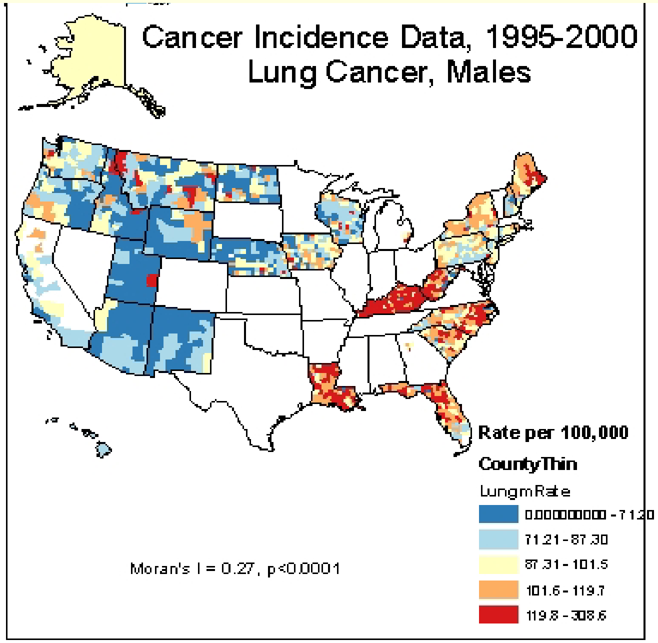
“Expected rates”

Outliers from expectation

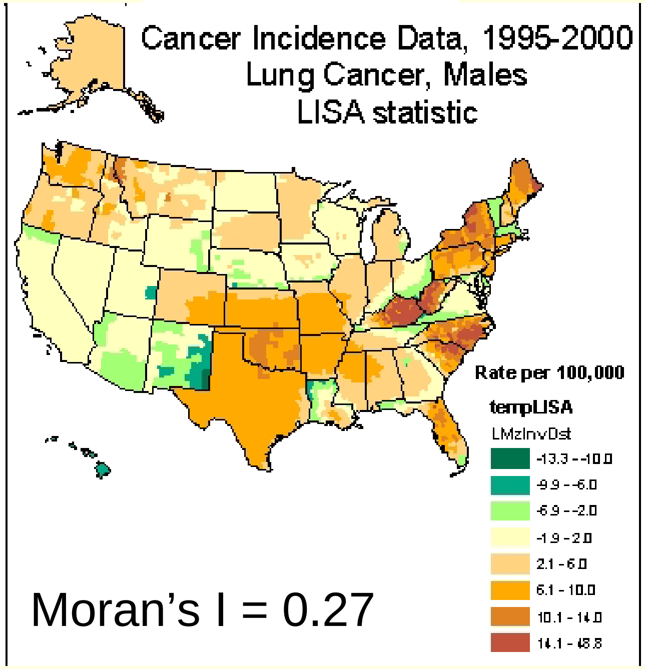
Copper smelter in MT

Method: *(DuMouchel & Pregibon, Proc KDD, 2001; Lincoln Technologies, Inc)*

Cluster detection in ArcMap



LISA statistic

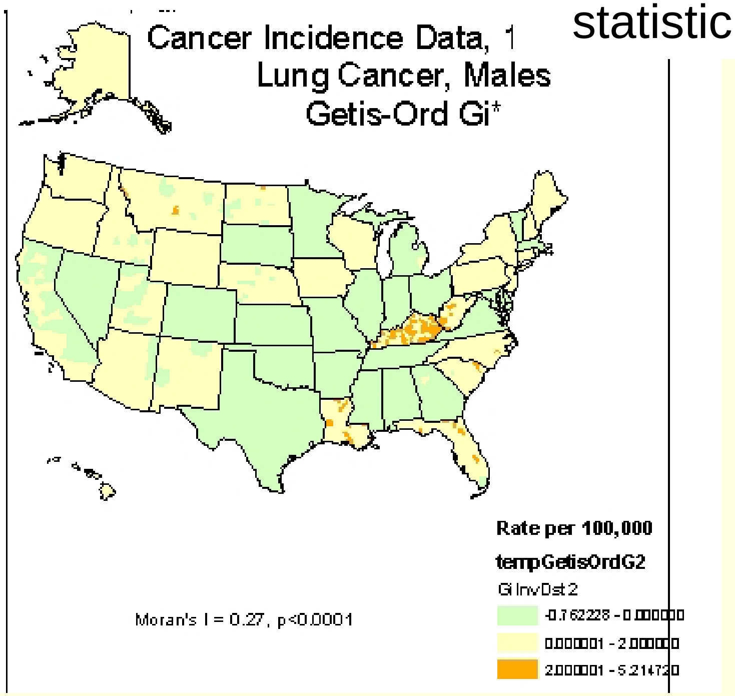


Moran's I = 0.27

Color Breaks For Z:

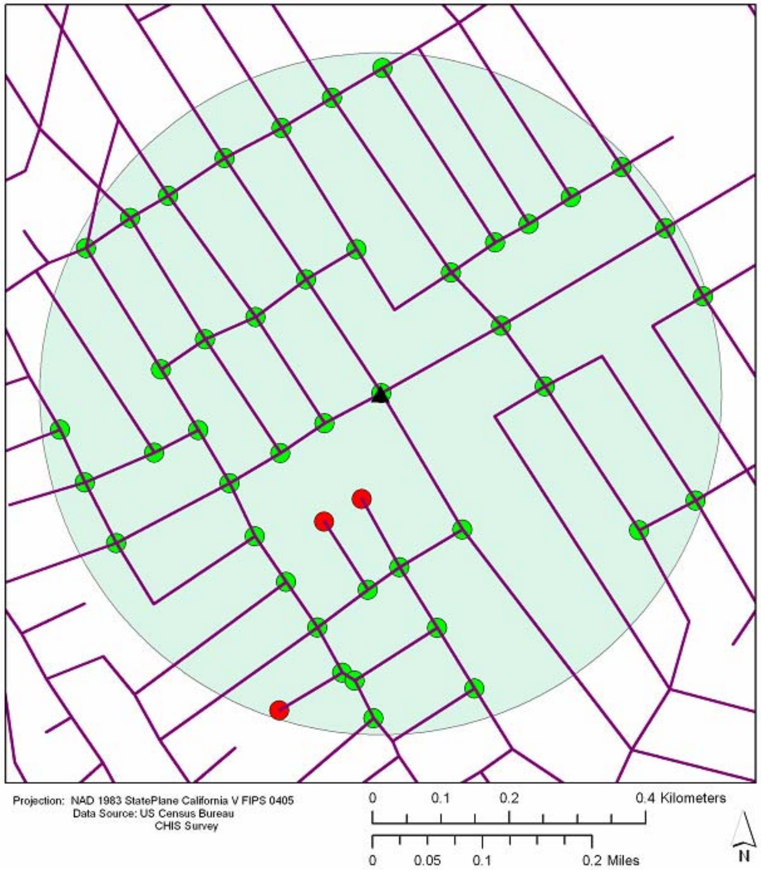
min
-10
-6
-2
+2 (NS)
+6
+10
+14
max

Getis-Ord Gi* statistic



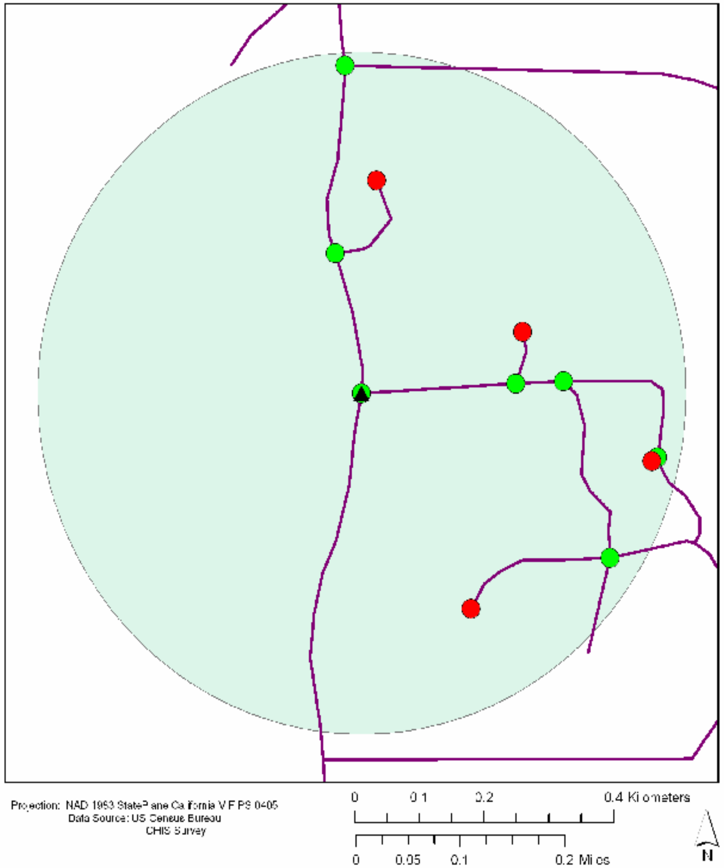
Defining potential risk factors using a GIS:
High and Low Connectivity Buffers in Los Angeles

(From The California Health Interview Survey, 2001)


Projection: NAD 1983 StatePlane California V FIPS 0405
Data Source: US Census Bureau
CHIS Survey
0
0.1
0.2
0.4 Kilometers
0
0.05
0.1
0.2 Miles
N

Link Node Ratio	2.02
Intersection Density in Km	54.73
Connected Node Ratio	0.93
Street Network Density	14.53
Gamma Index	0.46
Alpha Index	0.18
Block Density	37.98
Median Block Length	0.11
Average Block Length	0.16
Population Density in Km	3752.35
Employment Density in Km	647.18

▲ CHIS Survey Point
— Streets
🟢 Real Nodes
🔴 Dangel Nodes
🟩 500 Meters Buffer

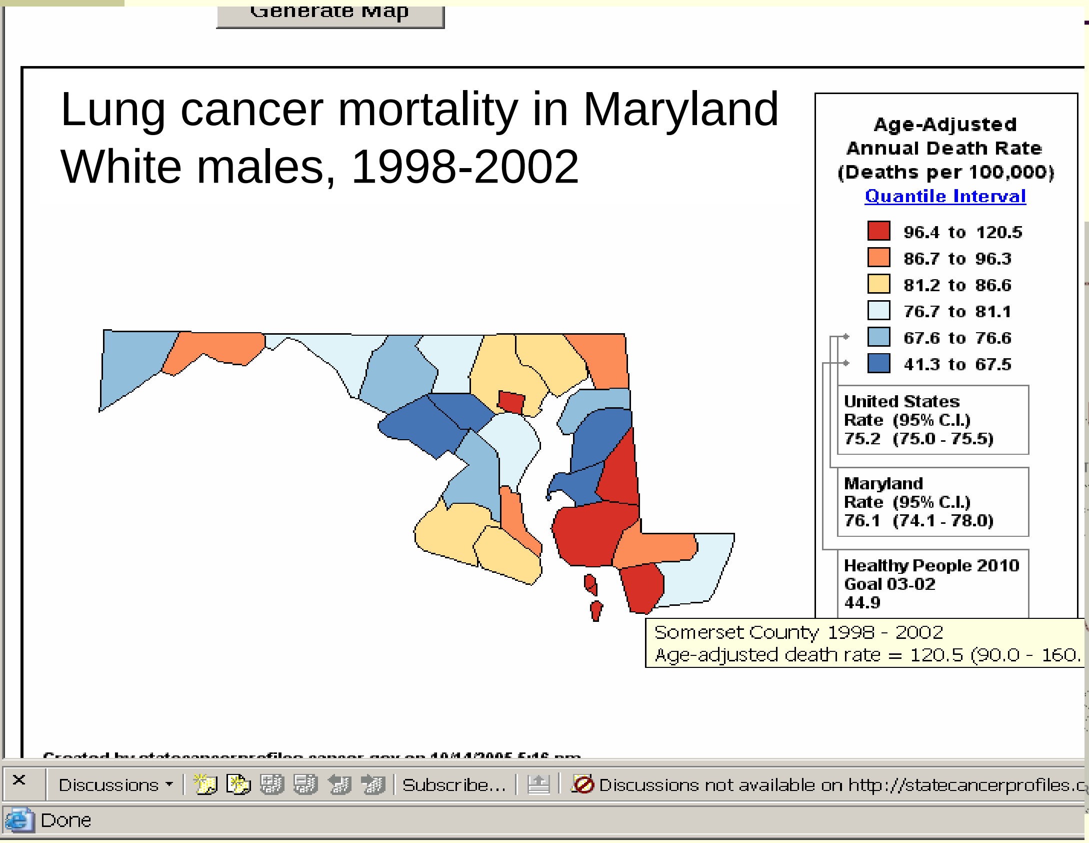

Projection: NAD 1983 StatePlane California V FIPS 0405
Data Source: US Census Bureau
CHIS Survey
0
0.1
0.2
0.4 Kilometers
0
0.05
0.1
0.2 Miles
N

Link Node Ratio	1.45
Intersection Density in Km	8.91
Connected Node Ratio	0.64
Street Network Density	3.53
Gamma Index	0.37
Alpha Index	0.00
Block Density	2.78
Median Block Length	0.20
Average Block Length	0.26
Population Density in Km	15.63
Employment Density in Km	1.14

▲ CHIS Survey Point
— Streets
🟢 Real Nodes
🔴 Dangel Nodes
🟩 500 Meters Buffer

Courtesy of David Berrigan, NCI

Choosing covariates for the analysis: Comparison of patterns of outcome and potential risk factors



Lung cancer mortality in Maryland
White males, 1998-2002

Age-Adjusted Annual Death Rate (Deaths per 100,000)
Quantile Interval

- 🟥 96.4 to 120.5
- 🟧 86.7 to 96.3
- 🟨 81.2 to 86.6
- ⬜ 76.7 to 81.1
- 🟦 67.6 to 76.6
- 🟦 41.3 to 67.5

United States Rate (95% C.I.)
75.2 (75.0 - 75.5)

Maryland Rate (95% C.I.)
76.1 (74.1 - 78.0)

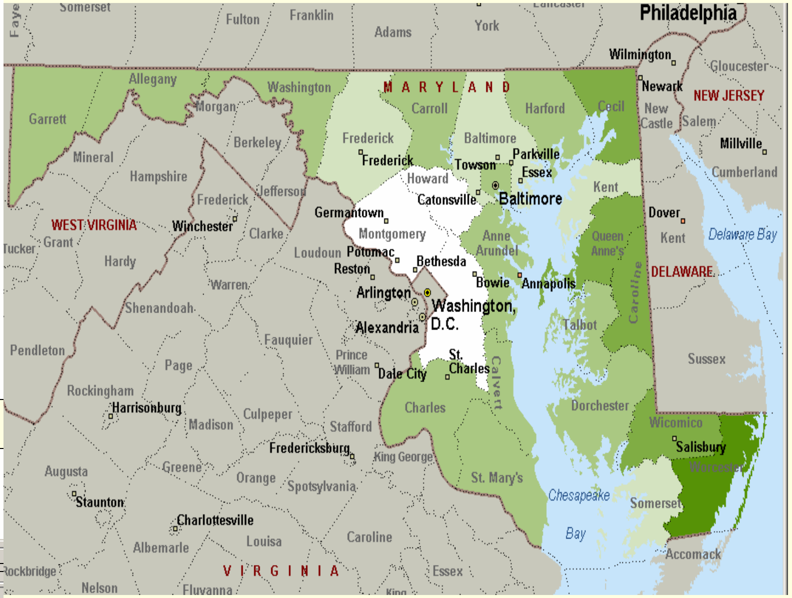
Healthy People 2010 Goal 03-02
44.9

Somerset County 1998 - 2002
Age-adjusted death rate = 120.5 (90.0 - 160.

Created by statecancerprofiles.cancer.gov on 10/14/2005 5:46 pm

URL: statecancerprofiles.cancer.gov

% men ever smoked cigarettes



Data source: BRFSS, CDC

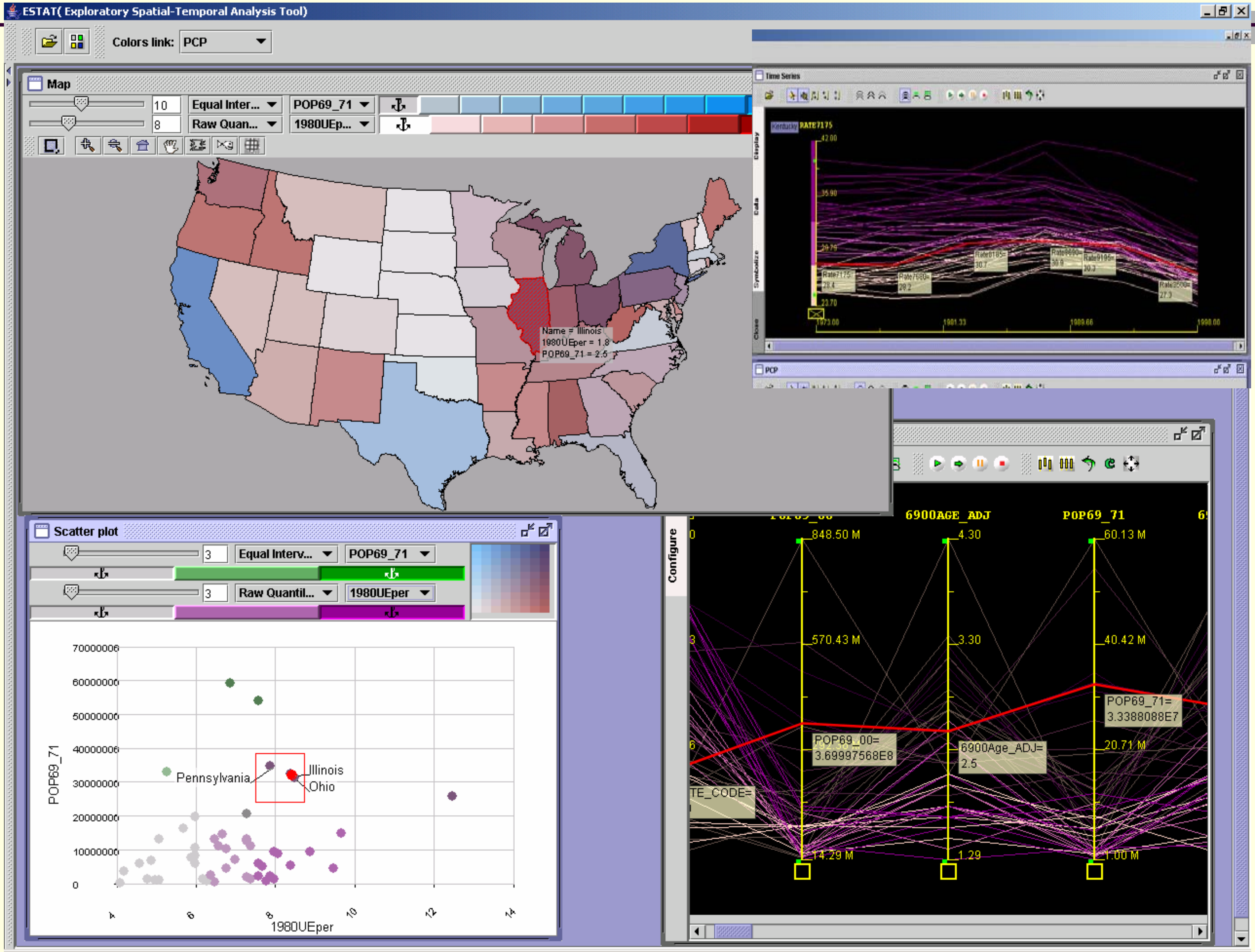
Exploratory Spatio-Temporal Analysis Tool (ESTAT)

Map

Rate Time Series Plot

Scatter plot

Covariate PCP plot



ESTAT(Exploratory Spatial-Temporal Analysis Tool)

Colors link: PCP

Name = Illinois
1980UEper = 1.8
POP69_71 = 2.5

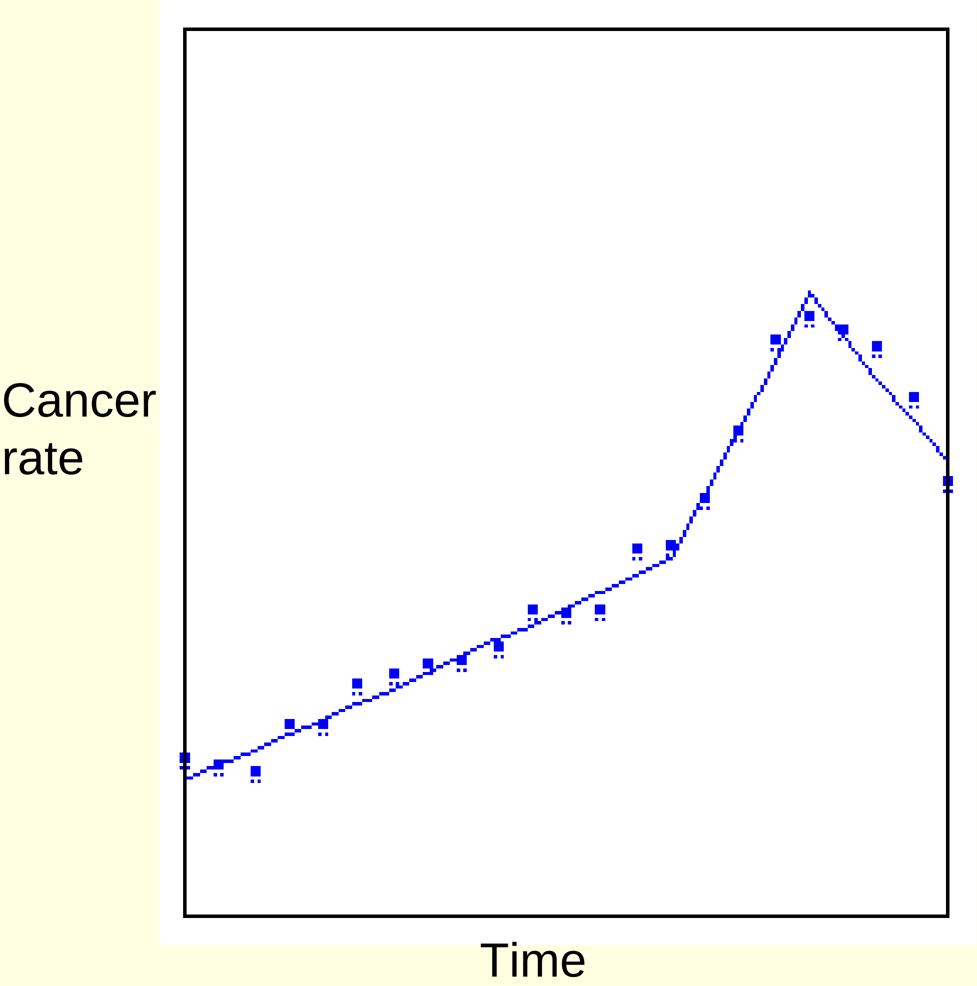
POP69_71=
3.3388088E7

POP69_00=
3.69997568E8

6900Age_ADJ=
2.5

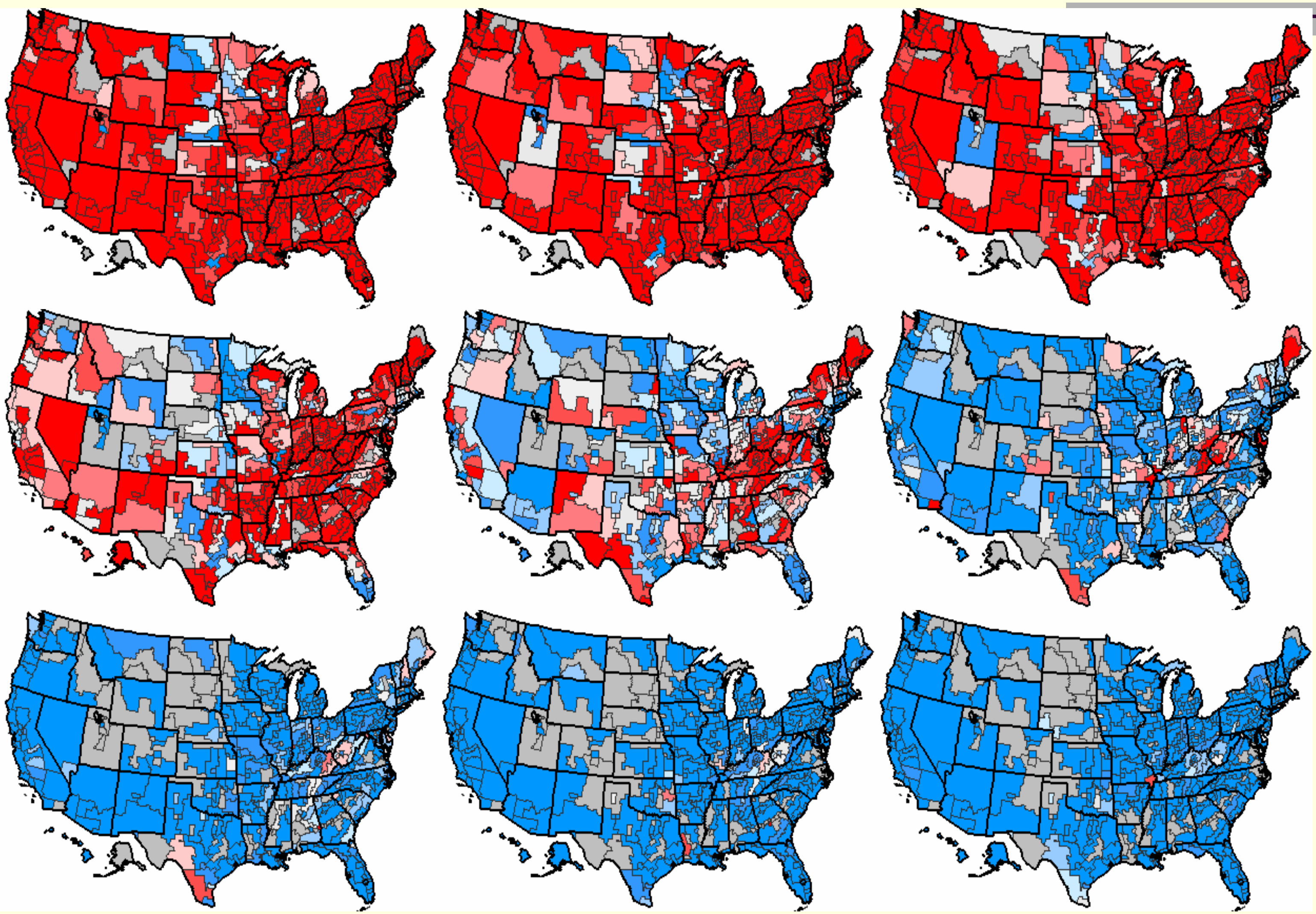
Developed by Alan MacEachren & GeoVista staff, Penn State University

Cancer surveillance:
Has the cancer trend changed? If so, where?



Joinpoint software determines when time trend changed significantly

Spatio-temporal patterns
Cervical cancer mortality 1950-94 by 5 years



URL: www.cancer.gov/atlasplus

Goals of statistical modeling

- Inference
  - Explain the patterns, calculate risks
  - Emphasis is on interpretation of model results
- Prediction
  - Either smooth observed data or fill in gaps
  - Emphasis is on fit of the model
  - Further complication: project ahead in time
- Approaches to modeling
  - Include as many covariates & interactions as possible to explain spatial patterns & correlation
  - Use a simpler model and let the spatial correlation “soak up” some of the spatial variation
- Spatial correlation must be included if necessary, otherwise variance estimates will be wrong

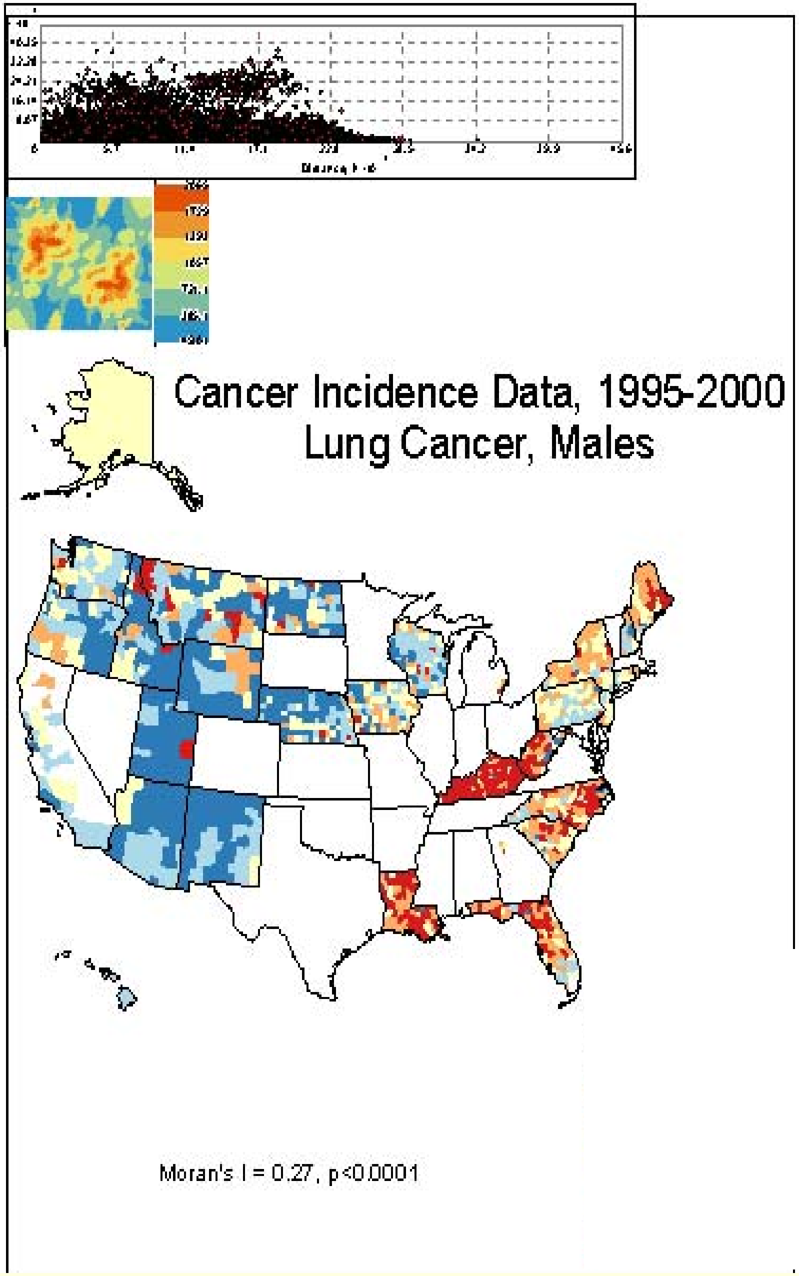
The process of statistical modeling

SAS,
S+,
Etc.

- Verify model assumptions (statistical & spatial)
- Apply model to data, estimate parameters
- Assess fit of the model
  - Chi-square goodness-of-fit, deviance statistics, etc
  - Scatter plots of observed vs. predicted values, leverage (observations that most influence results),…
  - Map the results
    - Does predicted value map look like observed value map?
    - Do residuals appear to be spatially random?
    - Are residuals still spatially correlated? (variogram)
- If necessary (almost always!), modify model & rerun

Use of Geostatistical Analyst to check spatial assumptions

- Stationarity
- Isotropy
- Functional form of spatial correlation (variogram models)

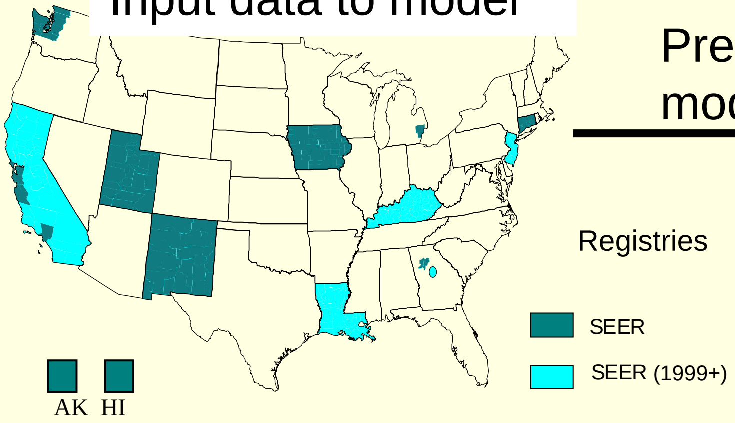


Cancer Incidence Data, 1995-2000
Lung Cancer, Males

Moran's I = 0.27, p<0.0001

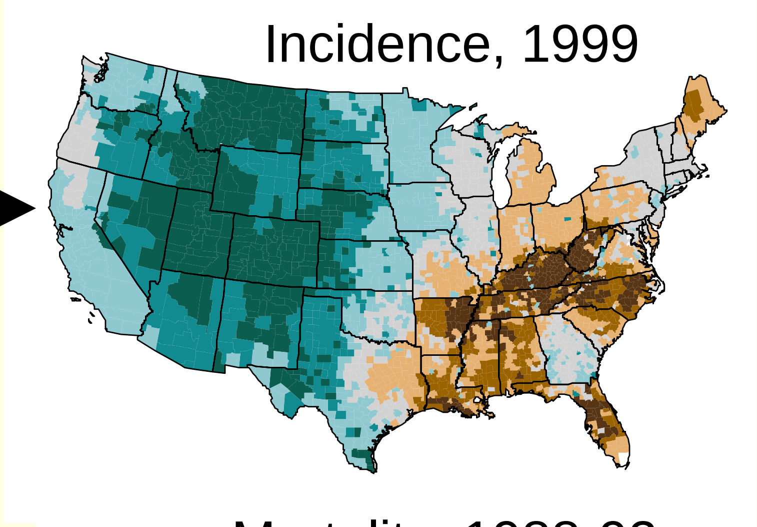
Map comparison as a goodness-of-fit tool

NCI cancer registries
Input data to model

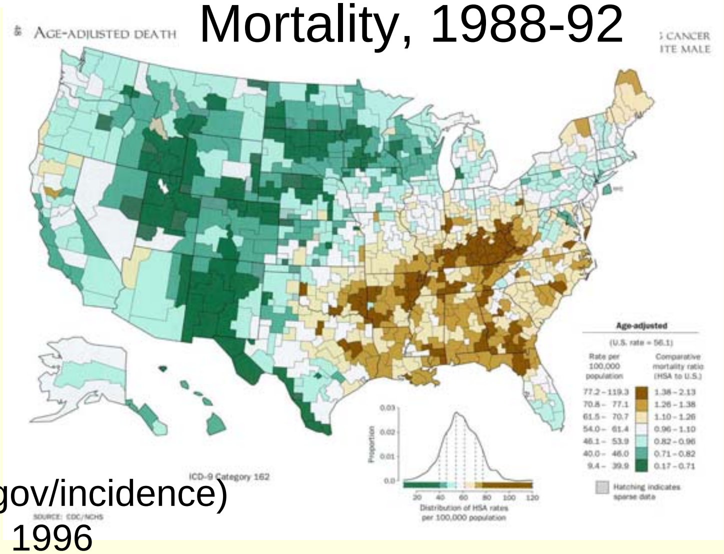


Prediction model

Incidence, 1999



Mortality, 1988-92



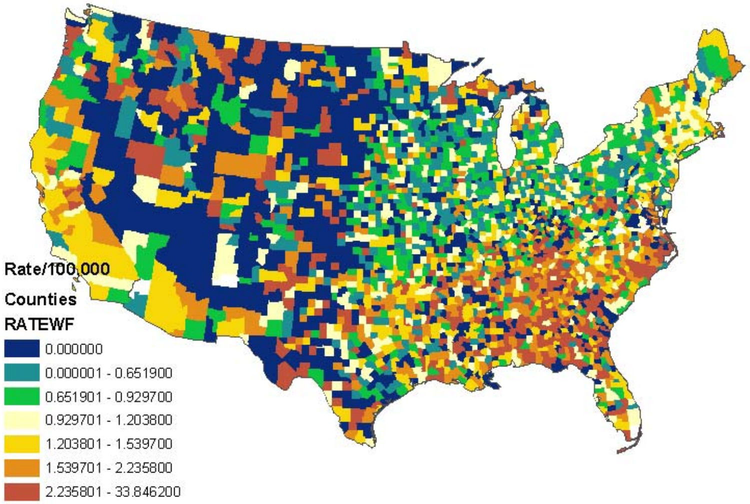
Top: Pickle et al., NCI monograph 2003 (URL: srab.cancer.gov/incidence)
Bottom: Pickle et al., *Atlas of United States Mortality,* NCHS 1996

Communicating results of spatial analysis

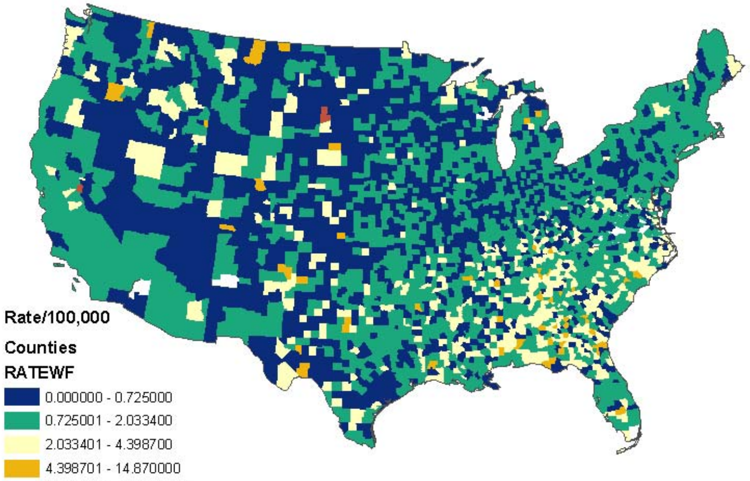
- Important for gov’t agencies to disseminate information
  - Back to original data collectors (cancer registries, states)
  - To researchers in the subject area
  - To the public
- Information needs to be accurate and clear to diverse audience, dissemination tools need to be user friendly
- Accuracy on a map
  - Tension between precision (narrow rate categories) & readability (not too many categories)
  - Cartographic choices can impact visual impressions
- Uncertainty of statistic must be communicated
- Often need multiple maps for multiple purposes

Oral cancer mortality, white females, 1950-69

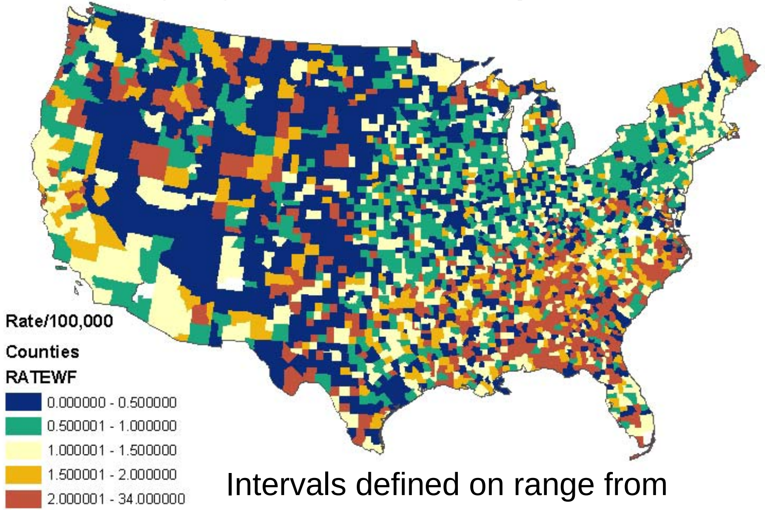
7 category quantile



Jenks (4 categories)

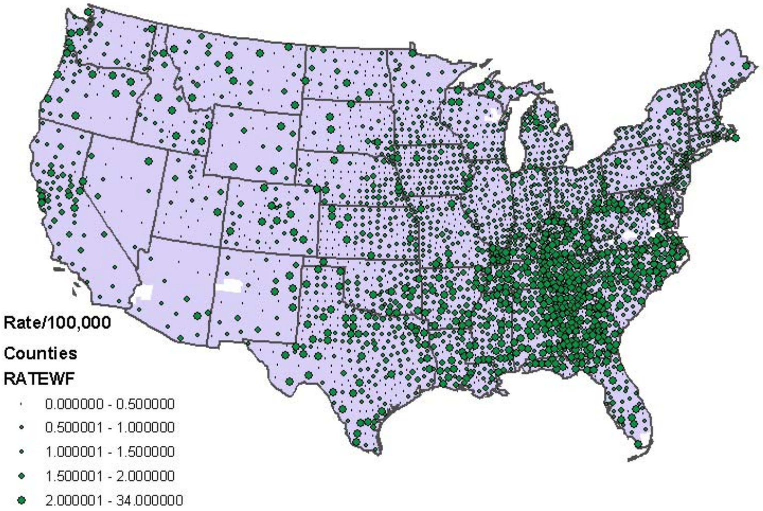


5 category truncated equal interval

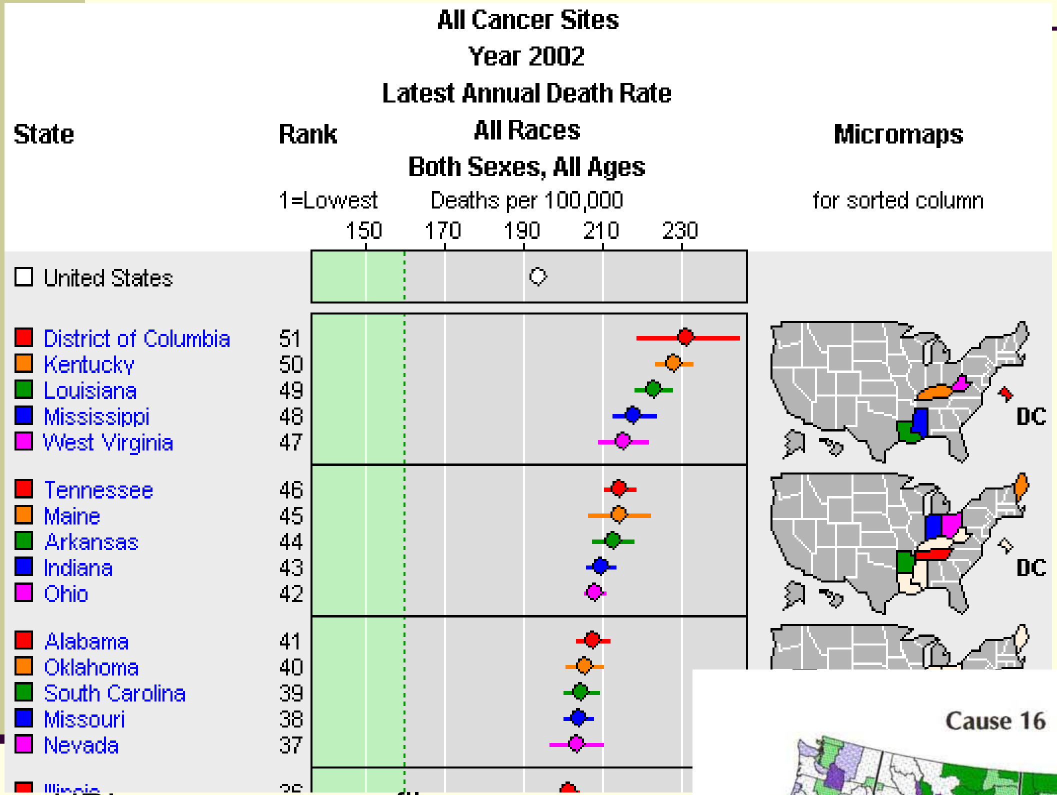


Intervals defined on range from 10%tile to 90%tile of distribution

Categorical symbols



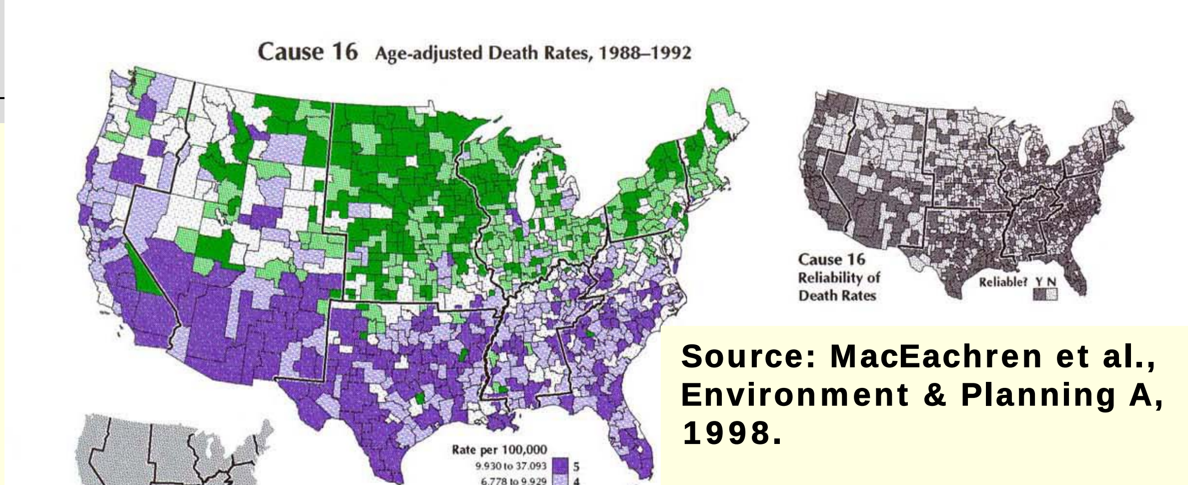
Methods of communicating uncertainty by separation of value and variance information



Confidence intervals on Graphic linked to map

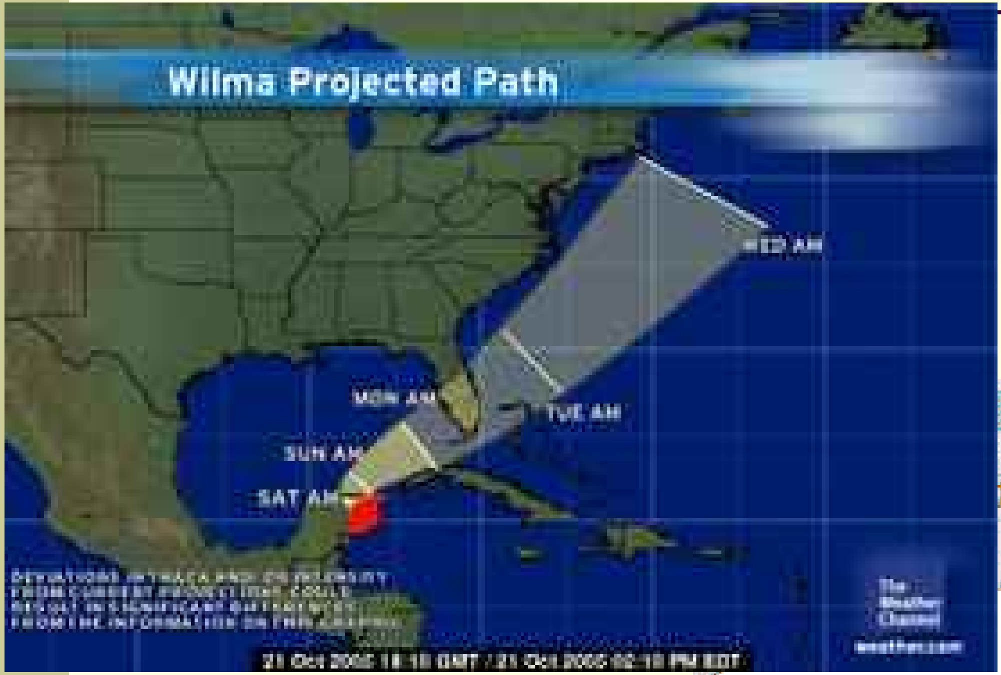
URL: statecancerprofiles.cancer.gov

Separate maps

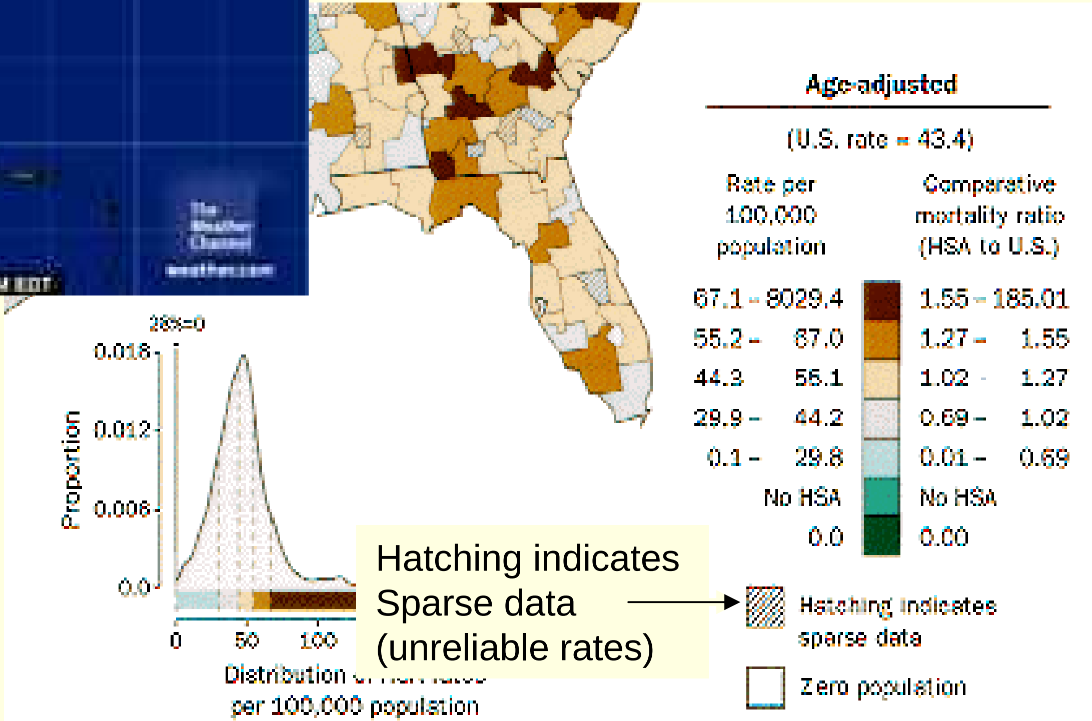


Source: MacEachren et al., Environment & Planning A, 1998.

Methods of communicating uncertainty by superimposing uncertainty layer on map


Wilma Projected Path
WED AM
MON AM
TUE AM
SUN AM
SAT AM
DEVIATIONS IN TRACK AND OR INTENSITY FROM CURRENT PROJECTIONS COULD RESULT IN SIGNIFICANT DIFFERENCES FROM THE INFORMATION ON THIS GRAPHIC
The Weather Channel
weather.com

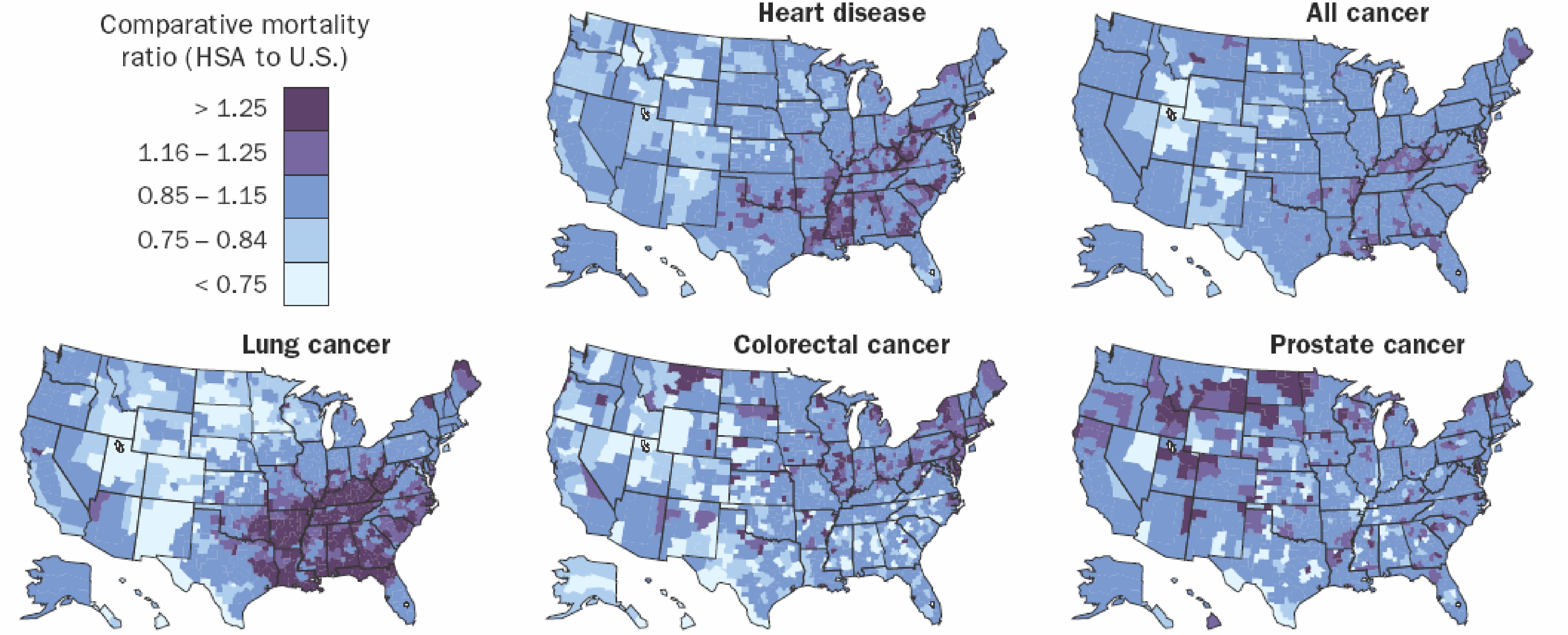
Meteorologist’s “cone of uncertainty”



Hatching indicates Sparse data (unreliable rates)

Source: Pickle et al., *Atlas of United States Mortality,* NCHS 1996

Small multiple maps useful for comparisons: Relative mortality ratios (area:US), 1988-92 Selected causes of death, white males



Pickle et al., *Atlas of United States Mortality,* NCHS 1996

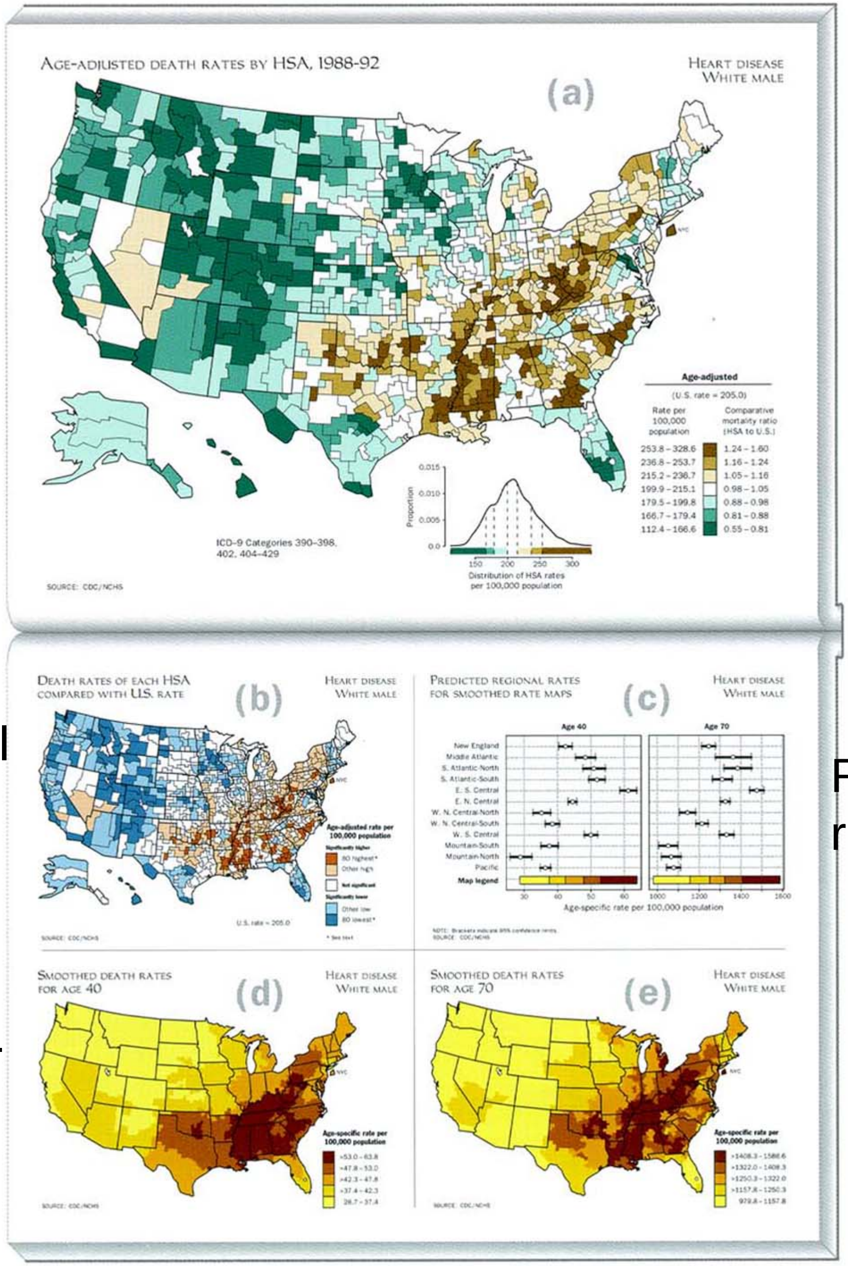
Multiple maps for multiple purposes

FIGURE 1. GRAPHICAL COMPONENTS OF THE TWO-PAGE ATLAS LAYOUT

Observed rates & reliability

Results of statistical significance test

Modeled age-specific rates



Regiona rates

Source: Pickle et al., *Atlas of United States Mortality*, NCHS, 1996.

Conclusions:
Methods important for spatial statistical analysis

- Flexible smoothing methods to explore patterns
- Inclusion of weights for smoothing, spatial statistics in order to account for population heterogeneity
- Output must include a measure of reliability (variance)
- Methods to examine spatial pattern characteristics: spatial correlation, stationarity, isotropy
- Methods to identify significant patterns, e.g., clusters & outliers – eye can be fooled
- Link to commonly-used statistical software packages
- Cartographic choices that present least biased view of the data (colors, categorization, unreliability, symbology)