National Cancer Institute

Statistical Perspectives on Spatial Social Science presented by Michael Goodchild

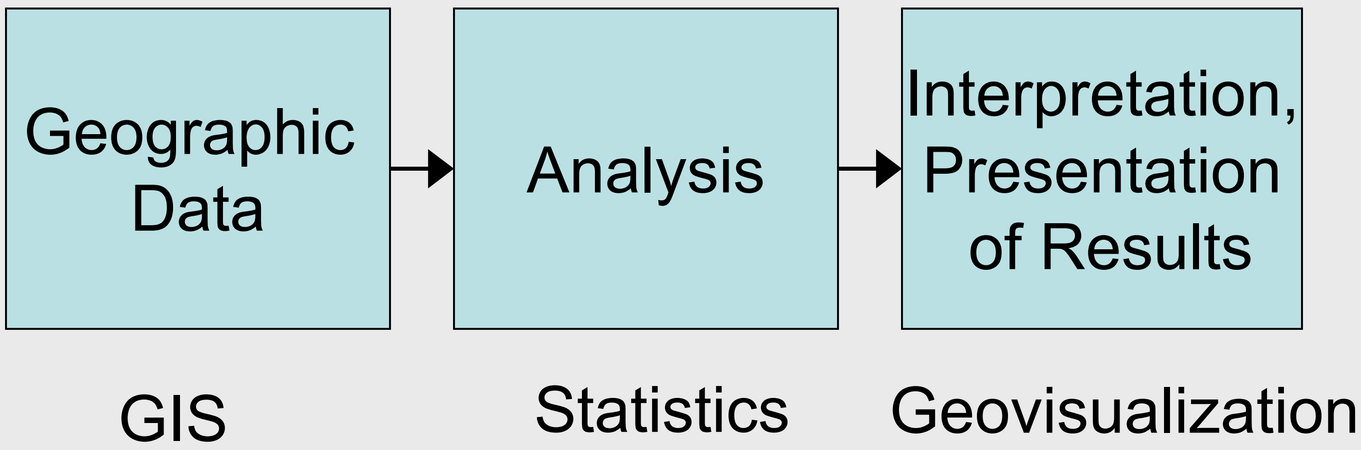
Discussant: Linda Williams Pickle

The 2006 Morris Hansen Lecture
November 6, 2006

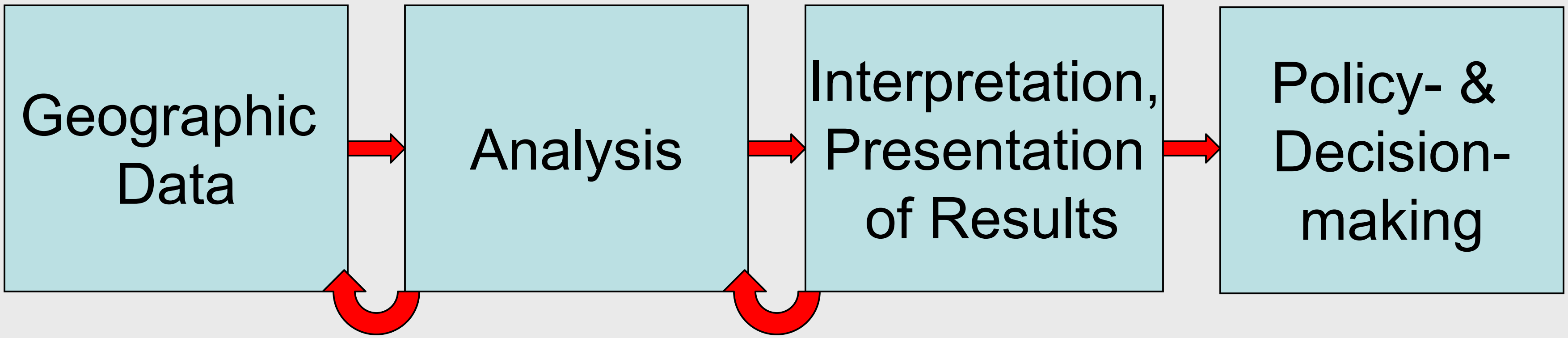
Outline of remarks

- A process for spatial data analysis
- Data
  - Examples of place-based analysis and policy formulation
  - Adding value to a geographic dataset by a GIS
- Spatial statistical analysis
  - Characteristics of geographic data impacting ability to apply statistical methods (uncertainty, required assumptions)
  - Improvements in statistical models for spatial data
- Future directions
  - Increasing familiarity with geographic information by the public
  - Social science applications in cancer control

A process for spatial data analysis



A process for spatial data analysis: This process is really non-linear



GIS for data storage, Creation of new measures

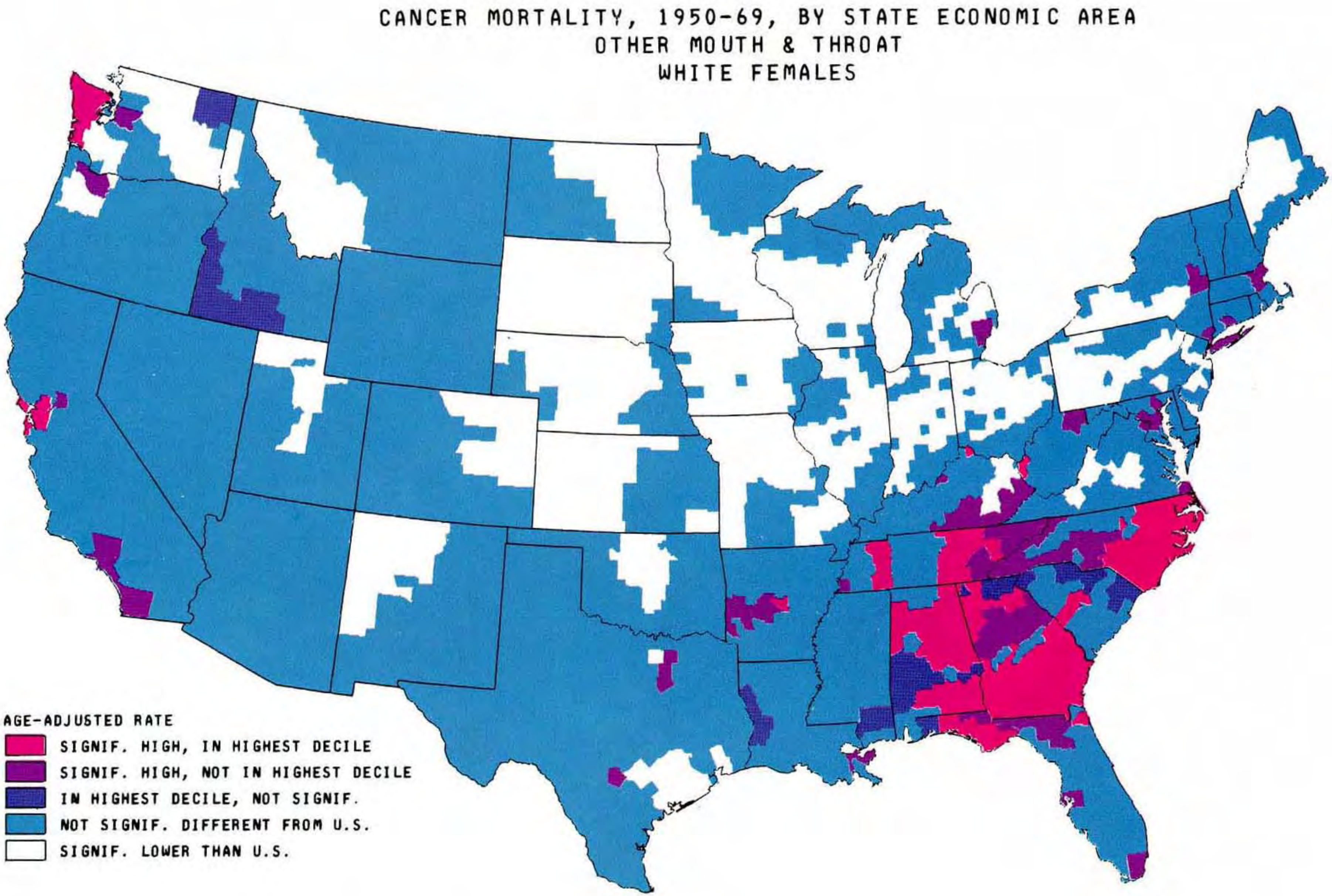
Exploratory, Confirmatory, Prediction

Map original values, Plot/map residuals, Plot/map results

DATA: Examples of place-based analysis and policy formulation

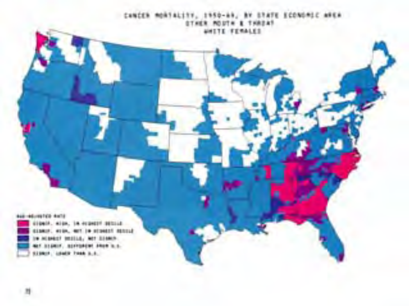
- Nomothetic vs idiographic science: Can we generalize from knowledge at distinct locations, or is every place unique?
- Are descriptive methods useful in the analytic process?
- How can results of spatial statistical analyses inform policy making?

- Applications from cancer epidemiology

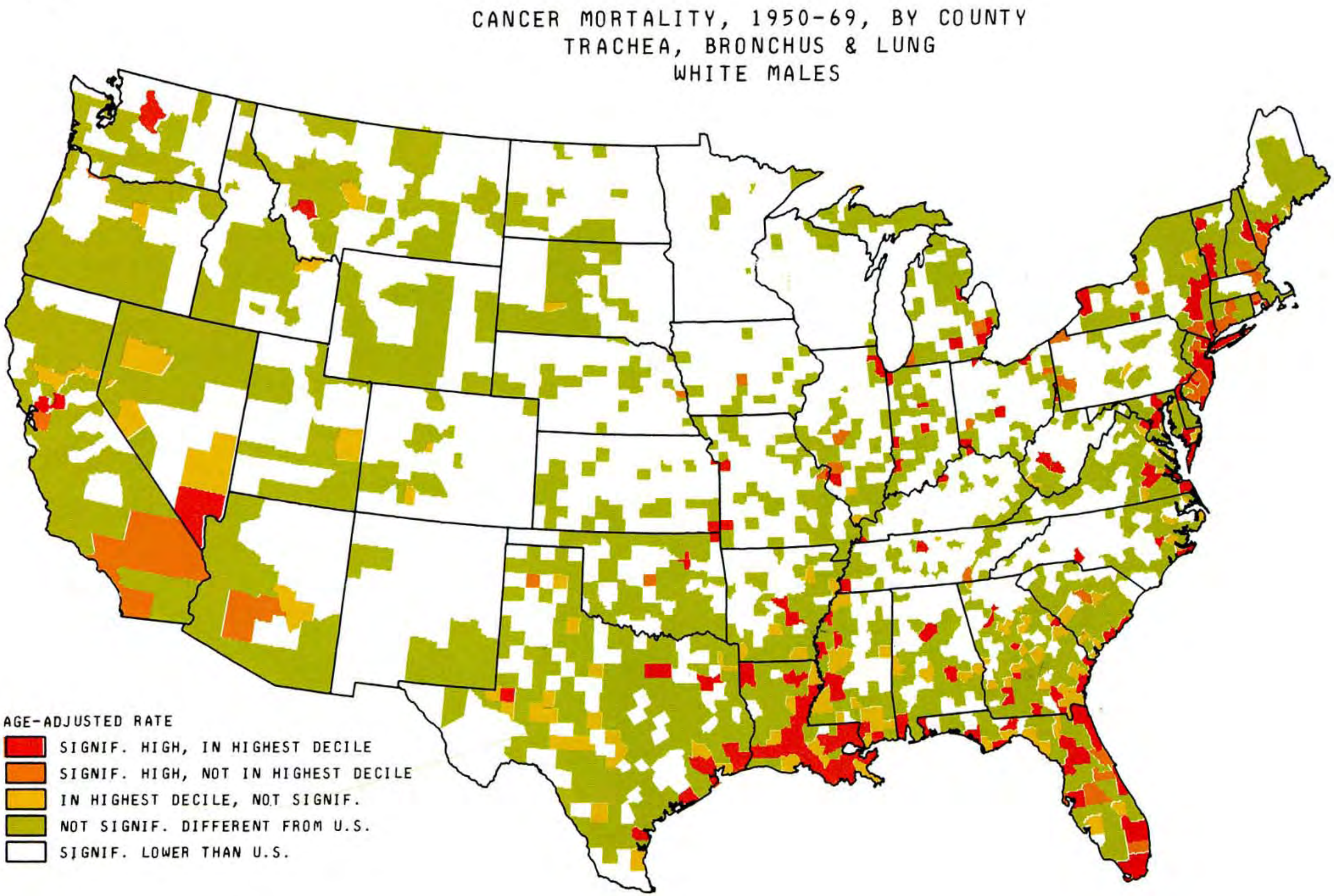


Source: Mason et al., *Atlas of Cancer Mortality for U.S. Counties*, NCI, 1975.

Oral cancer & snuff dipping

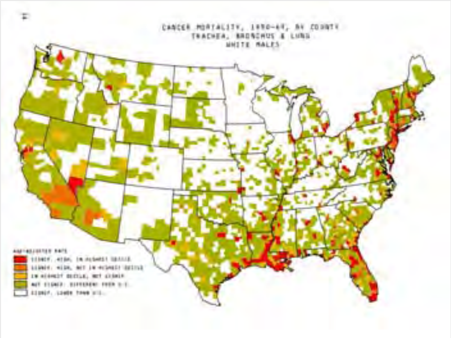


- **Uniqueness**: strong cluster in Southeast
  - Can we generalize to the entire US from findings only in SE?
- **Descriptive analysis** (mortality map with significance tests) identified specific areas where NCI epidemiologists could conduct interview studies with oral cancer cases & controls
  - Working hypothesis of the study: exposure to textile mill dust
- **Resulting generalization**: carcinogenic components of smokeless tobacco (snuff) cause extremely high risk of oral cancer at the exact site where the tobacco was in contact with gum tissue
- **Policy changes**
  - Ban of sales of smokeless tobacco to minors
  - Campaigns to stop smokeless tobacco use among role models for young people, e.g., baseball players



Source: Mason et al., *Atlas of Cancer Mortality for U.S. Counties*, NCI, 1975

Lung cancer & asbestos



- **Uniqueness**: Attribute cluster in coastal cities
- **Descriptive analysis** (mortality map with significance tests) identified specific areas for study
  - Working hypothesis: community or occupational exposure to airborne pollutants from the petrochemical industry
- **Resulting generalization**: Occupational exposure to asbestos in tasks requiring installation or removal of asbestos-containing insulation is sufficient to cause lung cancer (& mesothelioma) about 20 years later
- **Policy changes**: asbestos containment & abatement laws

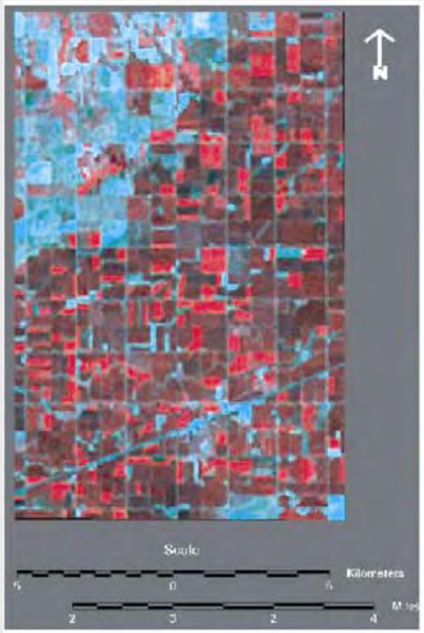
Adding value to data by GIS

- GIS can provide information about potential exposures that cannot be obtained through traditional epidemiologic methods, e.g., personal interviews
- Examples
  - Use of satellite imagery to reconstruct historical crop patterns for environmental exposure assessment

    (Ward et al. Env Health Perspectives,2000)
  - Roadway characteristics influencing walking behavior in Los Angeles

Using GIS to calculate a new risk measure

Landsat image

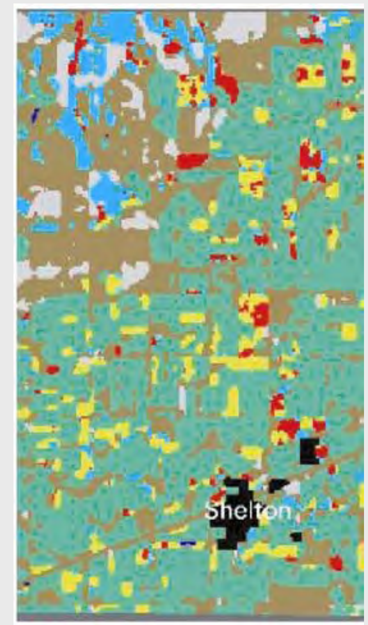


+

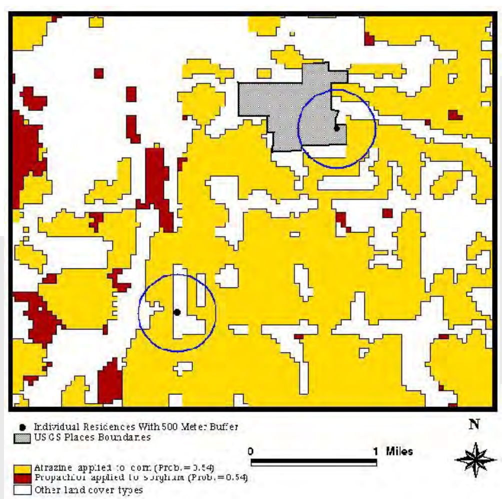
Farmers' crop reports
(ground truth)

=

Classified land cover



RESULTS: an estimate of likelihood of exposure to particular pesticides at each location (assumes each farmer uses same type & "dose" of pesticide for each crop)



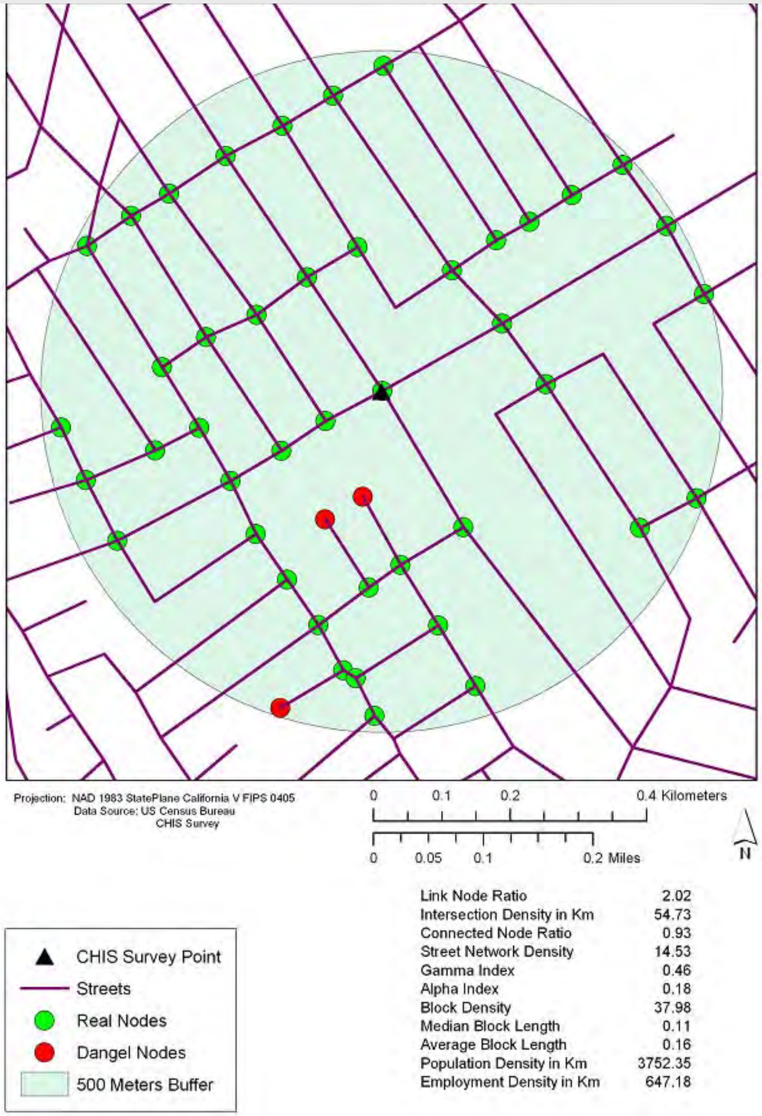
Yellow = likely atrazine exposure

Dot + buffer around homes of cases & controls (non-Hodgkin's lymphoma)

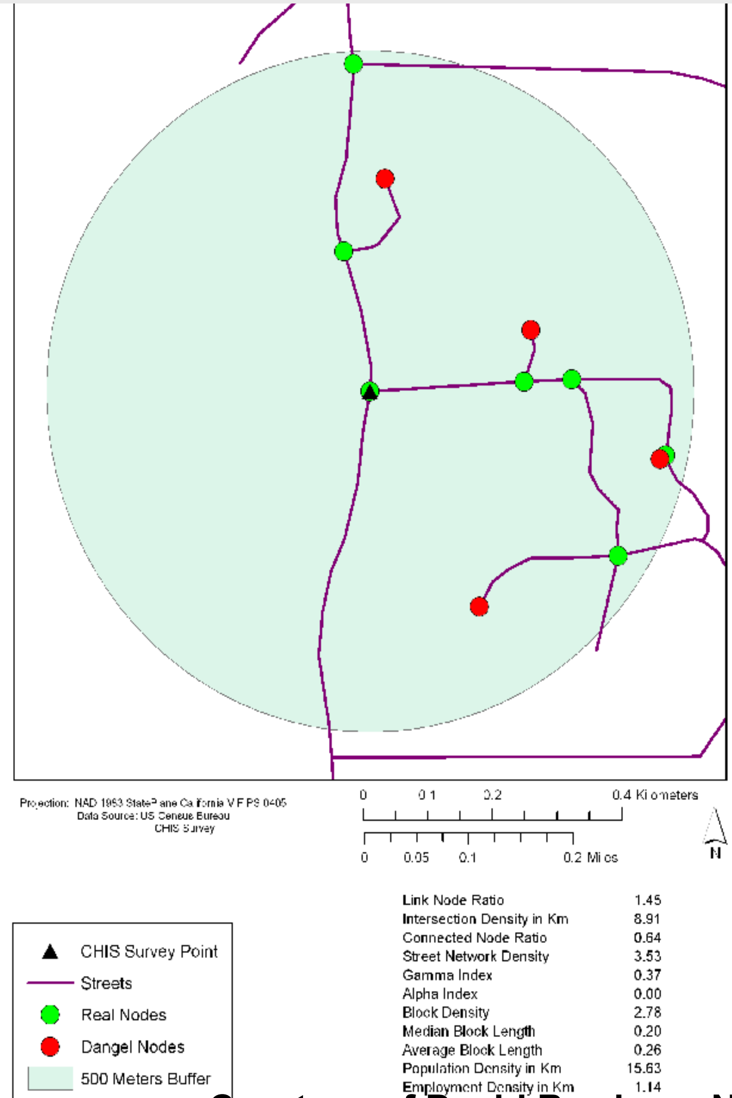
Source: Ward, EHP, 2000

Defining potential risk factors using a GIS: High & low street connectivity buffers

(Los Angeles County, California Health Interview Survey, 2001)



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



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

Courtesy of David Berrigan, NCI

Characteristics of geographic data that impact statistical analysis

- Sources of uncertainty
  - Measurement error in mapped values (outcome variable)
  - Imprecise boundary definitions
  - Lack of replicability in defining classes (interrater disagreement)
  - Location variability due to earth's axis wobble, tectonic movement
- Additional sources:
  - Random error (statistical error)
  - Model choice
  - Measurement errors in covariates
  - Location errors due to geocoding problems

An additional source of uncertainty: geocoding errors

All cases, by county

Cases geocodable to census tract

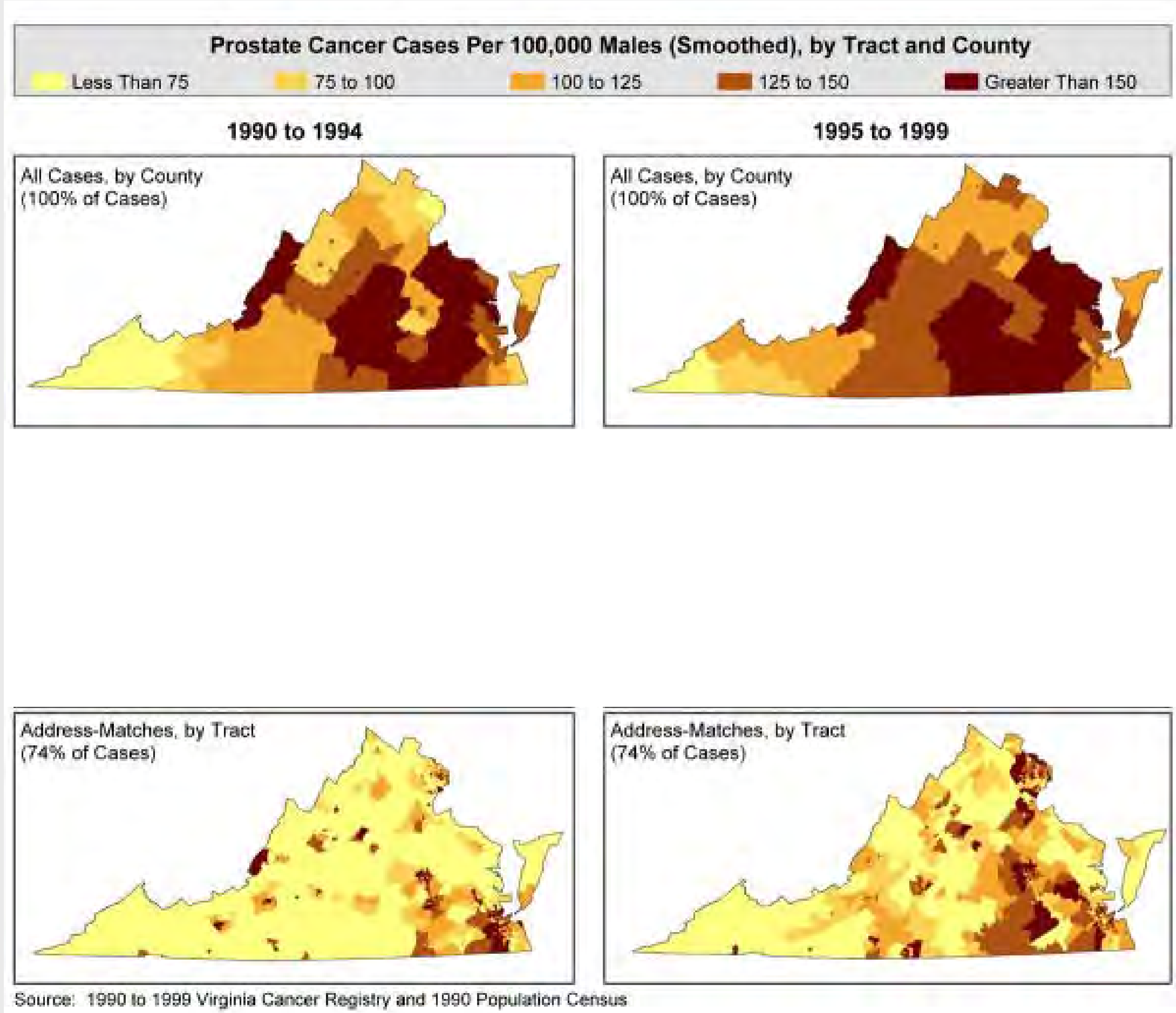


Fig 1. Annualized, age-adjusted prostate cancer incidence rates in VA, 1990-99
Source: Oliver, Matthews, Siadaty, Hauck, Pickle, *Int J of Health Geographics* 4:29, 2005

An additional source of uncertainty: geocoding errors

All cases, by county

County maps include only cases with geocodable addresses

Cases geocodable to census tract

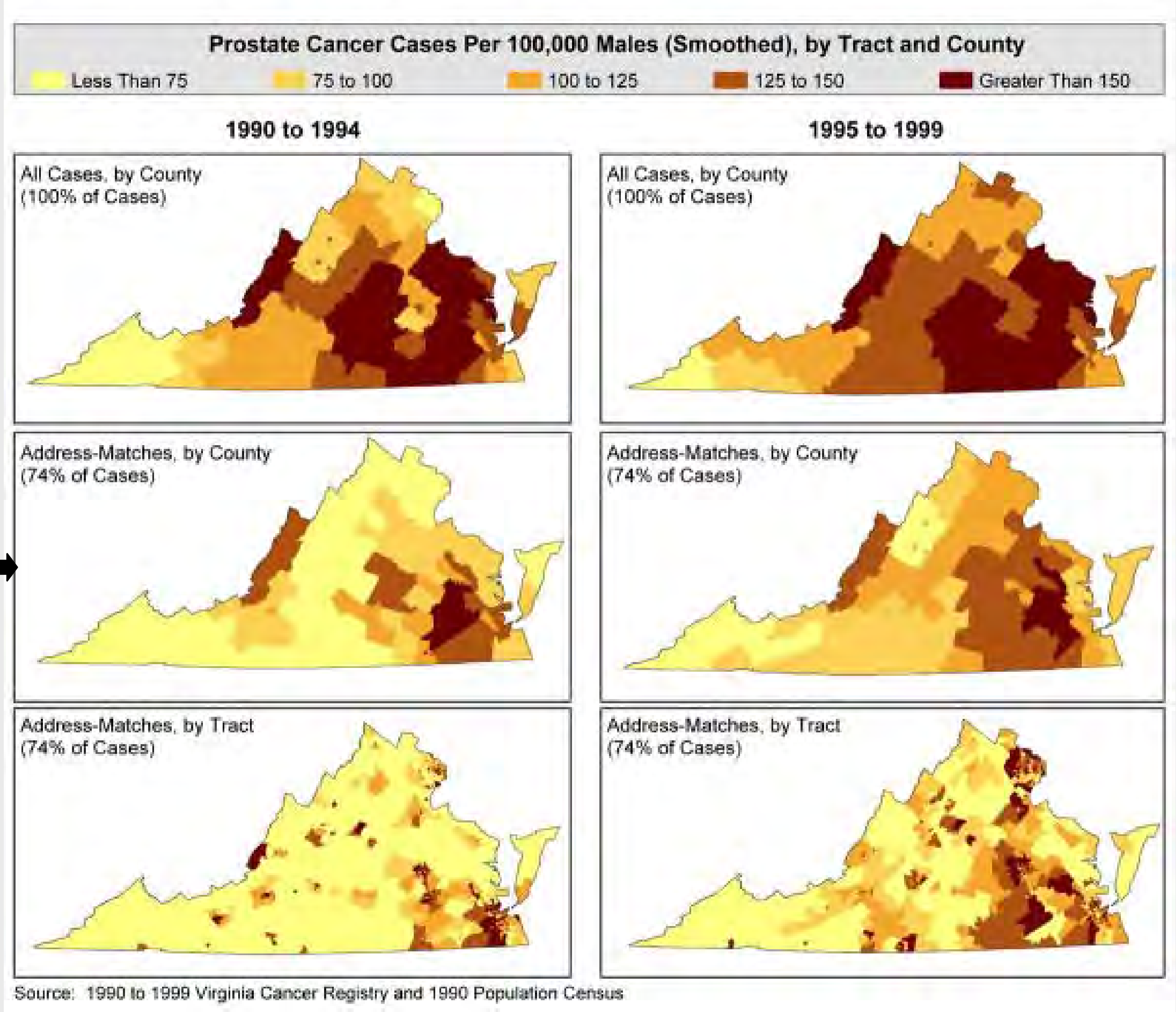


Fig 1. Annualized, age-adjusted prostate cancer incidence rates in VA, 1990-99
Source: Oliver, Matthews, Siadaty, Hauck, Pickle, *Int J of Health Geographics* 4:29, 2005

Characteristics of geographic data that impact statistical analysis, continued

- Spatial dependence (autocorrelation)
- Strong non-stationarity (spatial heterogeneity)
- Fractal behavior, e.g., of coastline (a scale issue)
- Inability to do random sampling

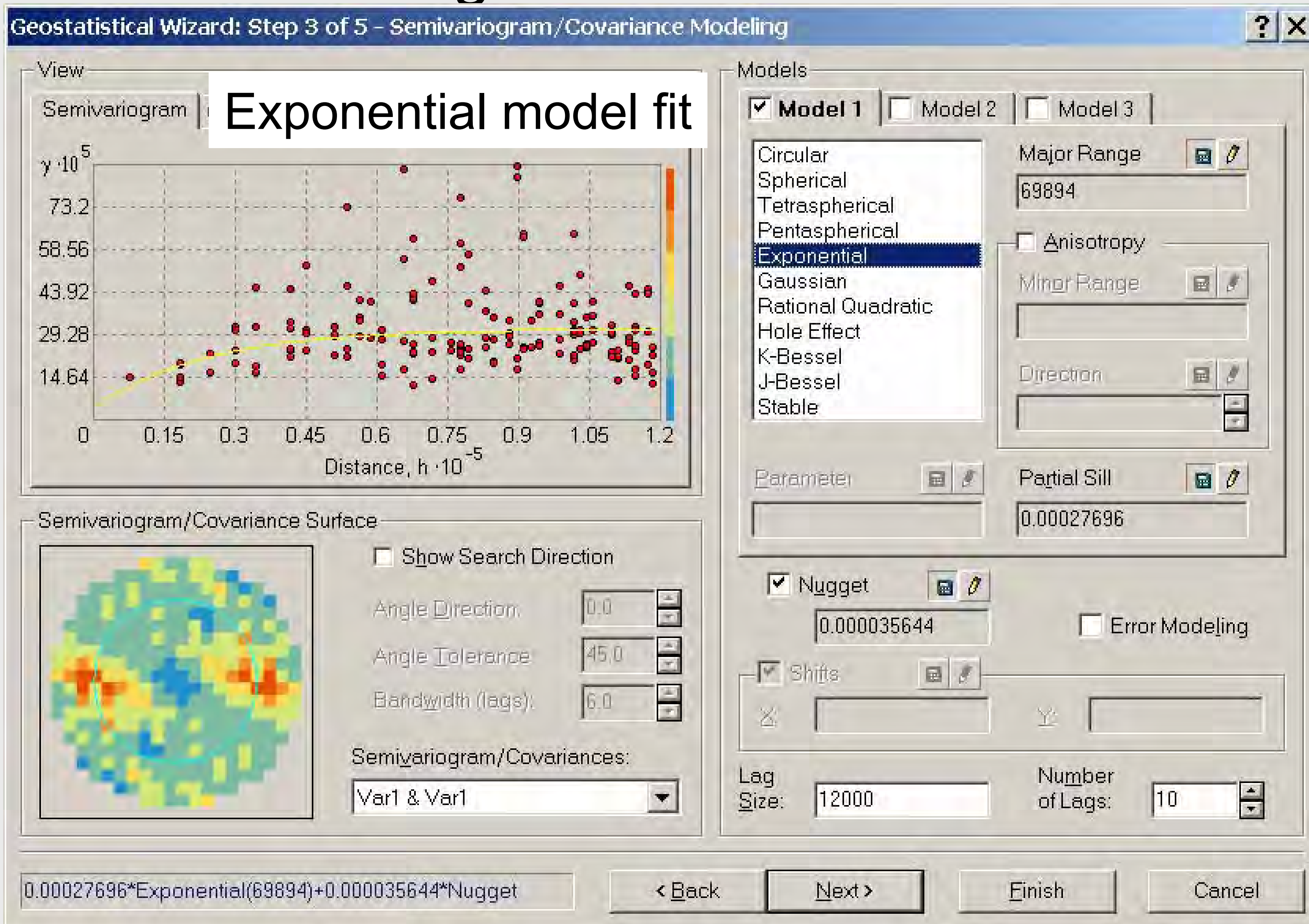
Stationarity & spatial heterogeneity

- **Strong stationarity**: joint distribution of process only depends on relative, not actual, locations of observations
- **Weak stationarity**:
  - Constant mean over all locations {s}
  - Covariances and variances of pairs of observations only depend on distances between them, not on actual locations

    Var[Y(s+h)-Y(s)] = 2γh ➡ semi-variogram plot of pairwise (variances/2) vs. binned distances h
- Weak stationarity assumption needed for many statistical methods, but for model-based analysis assumption usually satisfied (or close) for <u>residuals</u>, even if not for original outcome variable
- **Spatial heterogeneity** due to population variation is usually not of interest; can adjust or weight to remove

  e.g., $d_i \sim \text{Pois}(\lambda_i)$ is not stationary, but d_i/n_i probably is

ESRI Geostatistical Analyst®
Semivariogram of CA Ozone Data



Improved spatial statistical models

- Consider simple fixed effects regression model:

  $$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \,, \quad \varepsilon_i \sim \text{iid } N(0,\sigma^2)$$

- If errors are spatially correlated, then $\varepsilon_i \sim N(0,\Sigma)$

  where Σ is a variance-covariance matrix that describes their spatial dependencies in terms of distances or neighbors
  - Alternatively, can write residual ε_i as sum of spatially-dependent and spatially-independent errors
  - Common goal: add covariates sufficient to remove autocorrelation

- Adding uncertainty via random effects, e.g., errors in covariates

  $$y_i = \beta_0 + \beta_1 X_{1i} + b_2 X_{2i} + \varepsilon_i \,, \quad b_2 \sim N(\beta_2,\Omega)$$

- Spatio-temporal models are extensions of spatial models, but with possible temporal autocorrelation

Comparison of results of various models for spatial data (Waller & Gotway, 2004, Chapter 9*)

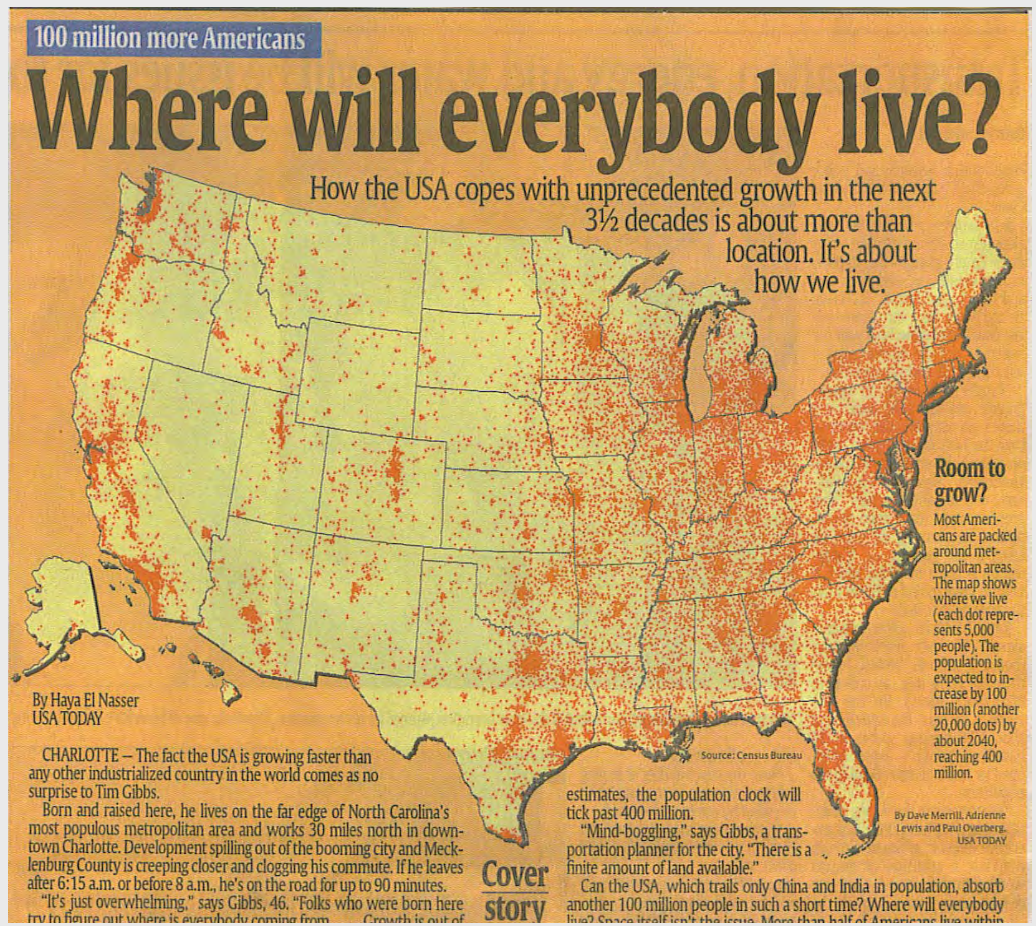
- NY leukemia data: # cases by census tract in 8 counties
- Methods compared
  - Linear regression assuming independent observations
  - Linear regression assuming spatially correlated observations
  - Simultaneous autoregressive (SAR) model
  - Conditional autoregressive (CAR) model
  - Each of these 4 applied to transformed tract rates with & without weights to account for population heterogeneity
  - Hierarchical Poisson regression, Bayesian implementation
- Conclusions
  - Better to use any method for modeling spatial autocorrelation than to assume independence (but choice can affect estimates)
  - Accounting for population heterogeneity is very important.

*Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*, Wiley, 2004.

Future directions: Increasing familiarity with geographic information by the public

- Weather maps
- Google Earth – being used, e.g., for traffic reports
- More sophisticated maps in newspapers, magazines

Cover Story, USA Today, Oct. 27, 2006: dot density map of population



100 million more Americans

Where will everybody live?

How the USA copes with unprecedented growth in the next 3½ decades is about more than location. It's about how we live.

By Haya El Nasser
USA TODAY

CHARLOTTE – The fact the USA is growing faster than any other industrialized country in the world comes as no surprise to Tim Gibbs.

Born and raised here, he lives on the far edge of North Carolina's most populous metropolitan area and works 30 miles north in downtown Charlotte. Development spilling out of the booming city and Mecklenburg County is creeping closer and clogging his commute. If he leaves after 6:15 a.m. or before 8 a.m., he's on the road for up to 90 minutes.

"It's just overwhelming," says Gibbs, 46. "Folks who were born here

Cover story

estimates, the population clock will tick past 400 million.

"Mind-boggling," says Gibbs, a transportation planner for the city. "There is a finite amount of land available."

Can the USA, which trails only China and India in population, absorb another 100 million people in such a short time? Where will everybody

Room to grow?

Most Americans are packed around metropolitan areas. The map shows where we live (each dot represents 5,000 people). The population is expected to increase by 100 million (another 20,000 dots) by about 2040, reaching 400 million.

Source: Census Bureau

By Dave Merrill, Adrienne Lewis and Paul Overberg, USA TODAY

1 dot = 5000 people

Social science applications in public health