AMCS 313: Spatial Statistics

Prof. Marc G. Genton, CEMSE, KAUST

(joint notes with Prof. Mikyoung Jun, TAMU)

Before we start

- Prerequisites
- Homeworks and paper presentation
- R programming
- Project and final presentation/report
- Office hours
- Textbook(s)

Types of Spatial Data

- Regularly spaced data vs irregularly spaced data
- Point measurement vs block averages (or areal data)
- Point patterns
- Directional data
- Data from moving stations

- Temporal: $\{Z(t); t \geq 0\}$
- Spatial: $\{Z(\mathbf{s}); \mathbf{s} \in D\}$
- Spatio-temporal: $\{Z(\mathbf{s}, t); \mathbf{s} \in D, t \geq 0\}$
- Multivariate: $\mathbf{Z} \in \mathbb{R}^p$, e.g. $\{\mathbf{Z}(\mathbf{s}); \mathbf{s} \in D\}$
- On the sphere: use latitude/longitude

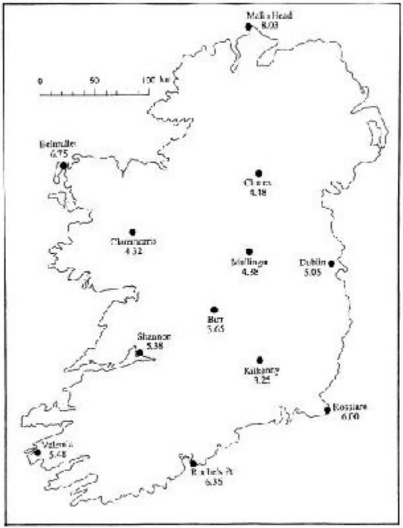
Law of Geography:
nearby things tend to be more alike than those far apart

- $D \in \mathbb{R}^d$, $d = 1, 2, 3$, "continuous" spatial index
- Mining: coal ash
- Pollution: soil, air (nuclear, chemical; ozone)
- Rainfall
- Temperature, pressure
- Wind speed and direction
- Remote sensing (satellite)

- D is a fixed collection of points, "discrete" spatial index
- Sudden Infant Death Syndrom (SIDS) in each county of the state of North Carolina, USA
- Cancer rates per city (or per state, or per country)
- Crime rates
- Census data
- Remote sensing (satellite)
- Agriculture: yield in a plot

- D is a random set, i.e. random locations
- Lansing wood trees in Michigan: hickory, maple
- Earth quake locations (e.g. San Francisco Bay)
- Mine fields
- Object recognition
- Nanoparticules
- Wildfires
- Lightning
- Blue spotted ribbon tail rays
- Spatial randomness? Clusters? Regularity?

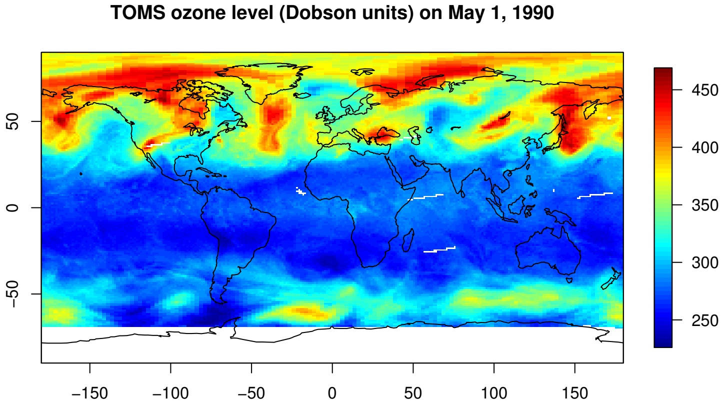
Irish wind data



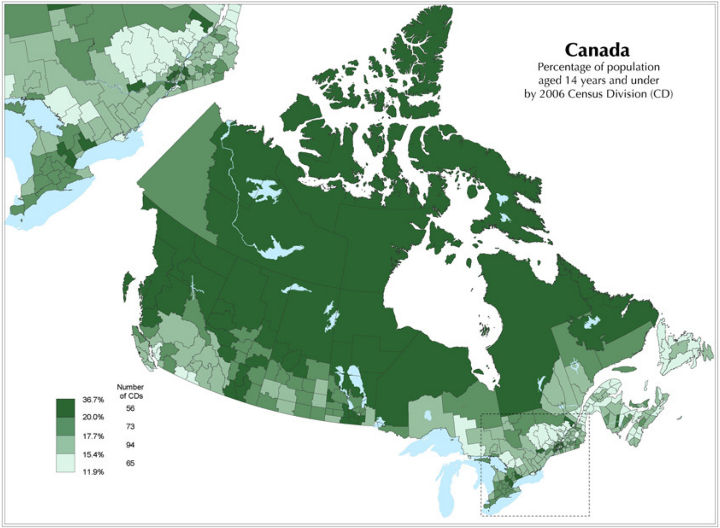
source: Haslett and Raftery (1989, Applied Statistics)

Total column ozone levels (TOMS data)

From satellite: Total Ozone Mapping Spectrometer instrument

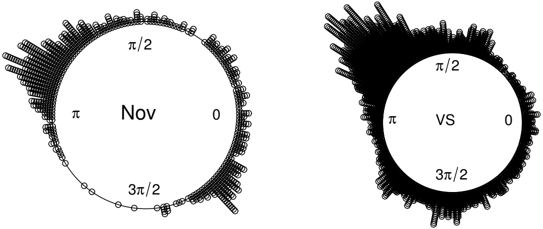


Level 3 data (NASA-produced) on regular grid
(1 degree latitude by 1.25 degrees longitude)
source: Jun and Stein (2008, AOAS)



source: www.statcan.ca

Wind speed and direction data

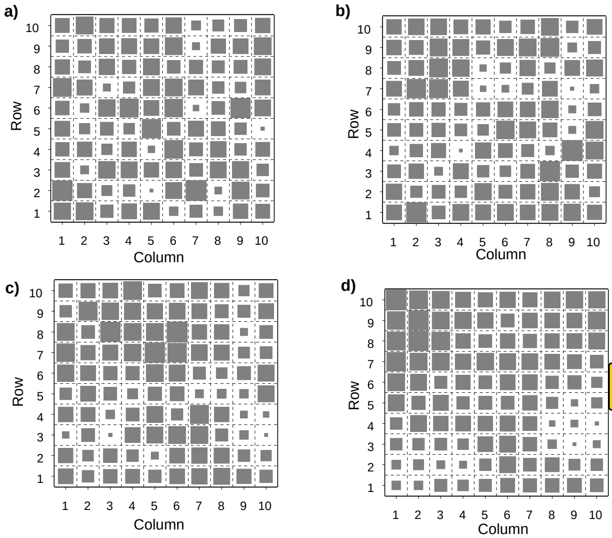


source: Hering and Genton (2010, JASA)

Need for spatial statistics

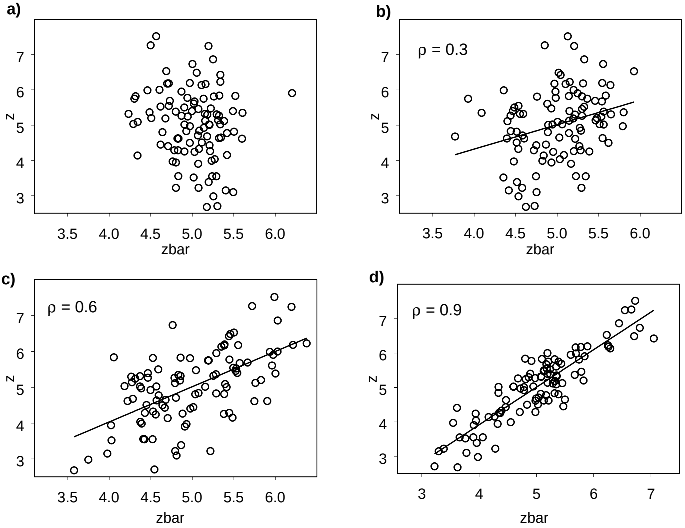
- Roots: geology (mining), geography, meteorology, environmetrics
- Classical statistics: $X_1, \dots, X_n$ iid $\sim F$, e.g. F is normal (Gaussian) distribution
- Spatial data: measurements/observations taken at specific locations or within specific regions
- Key features of spatial data: autocorrelation of observations in space, i.e. observations spatially close tend to be more similar
- Example: simulate data on 10×10 lattice, iid from normal N(5,1)
- a) observations assigned randomly to lattice coordinates
- b)-d) data rearranged: each value surrounded by more similar values (by simulated annealing algorithm)
- Define nearest neighbors: move queen piece on chess board
- $(\mathbf{s}_i, \bar{Z}_i)$, $i = 1, \dots, 100$, $\bar{Z}_i$=average of neighboring sites of $\mathbf{s}_i$ (note: edge effect!)
- Plot: $(\bar{Z}_i, Z(\mathbf{s}_i))$

Different autocorrelations



source: Schabenberger & Gotway (2005)

Different autocorrelations

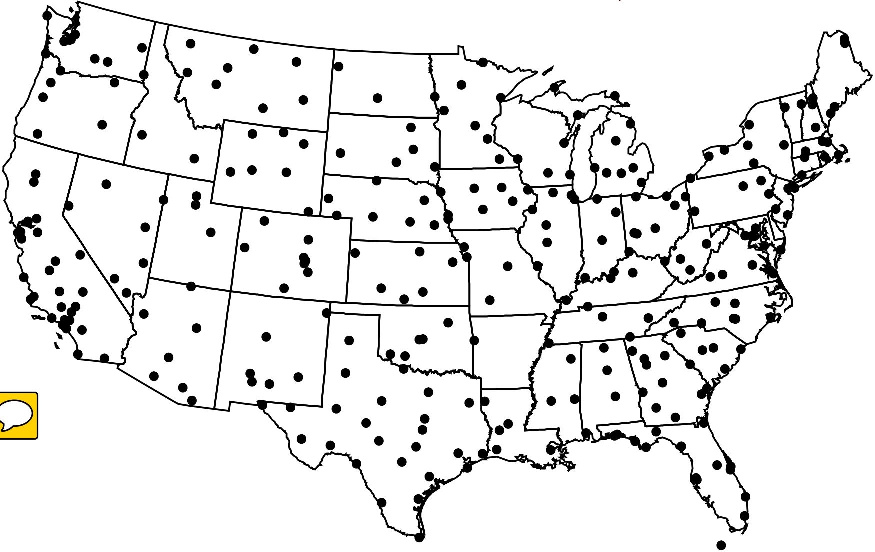


source: Schabenberger & Gotway (2005)

- $Y_1, \ldots, Y_n$: $Y_i \sim N(\mu_Y, \sigma^2)$, $\mathrm{Cov}(Y_i, Y_j) = \sigma^2\rho$, $i \neq j$ (equicorrelation)
- $X_1, \ldots, X_n$: $X_i \sim N(\mu_X, \sigma^2)$, $\mathrm{Cov}(X_i, X_j) = \sigma^2\rho$, $i \neq j$ (equicorrelation)
- Y_i's independent of X_j's
- Effect of ignoring correlation:
- $\hat{\mu} = \bar{Y}$ is "natural" estimator for μ_Y
- $\mathrm{Var}(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}(Y_i, Y_j) = \frac{1}{n^2}\{n\sigma^2 + n(n-1)\sigma^2\rho\} = \frac{\sigma^2}{n}\{1 + (n-1)\rho\}$. So if $\rho > 0$ then $\mathrm{Var}(\bar{Y}) > \sigma^2/n$, i.e. more dispersed than in random sample
- $\mathrm{E}(\bar{Y}) = \mu_Y$ and $\lim_{n\to\infty} \mathrm{Var}(\bar{Y}) = \sigma^2\rho$, so $\bar{Y}$ is not a consistent estimator of μ_Y
- Effective sample size: $n' = \frac{n}{1+(n-1)\rho}$, e.g. $n' = \frac{5}{1+4*0.25} = 2.5$

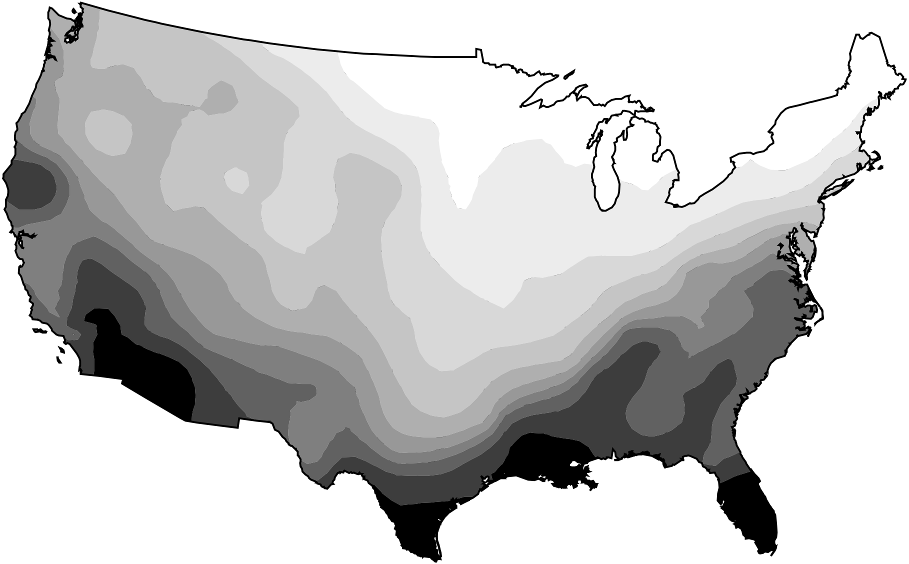
- Hypothesis test: $H_0 : \mu_X = \mu_Y$
- Ignoring correlation: $Z^*_{obs} = \frac{\bar{Y}-\bar{X}}{\sigma\sqrt{2/n}}$
- Correct test statistics: $Z_{obs} = \frac{\bar{Y}-\bar{X}}{\sigma\sqrt{2\{1+(n-1)\rho\}/n}}$
- Z^*_{obs} too large, so p-values too small, i.e. test rejects more often than it should
- n correlated observations contain less information than n uncorrelated observations
- Generalized least squares estimator of μ_Y: $\hat{\mu}_Y = \frac{\mathbf{1}^T\Sigma^{-1}\mathbf{Y}}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}$ where $\Sigma = \sigma^2\{(1-\rho)I_n + \rho\mathbf{1}\mathbf{1}^T\}$
- Can derive similar results for AR(1) structure: $\text{Cov}(Y_i, Y_j) = \sigma^2\rho^{|i-j|}$
- Later: effect of autocorrelation on prediction
- Exercise 1: Study the effect of AR(1) autocorrelation structure on classical statistical inference

US weather stations: air temperature (National Climatic Data Center)

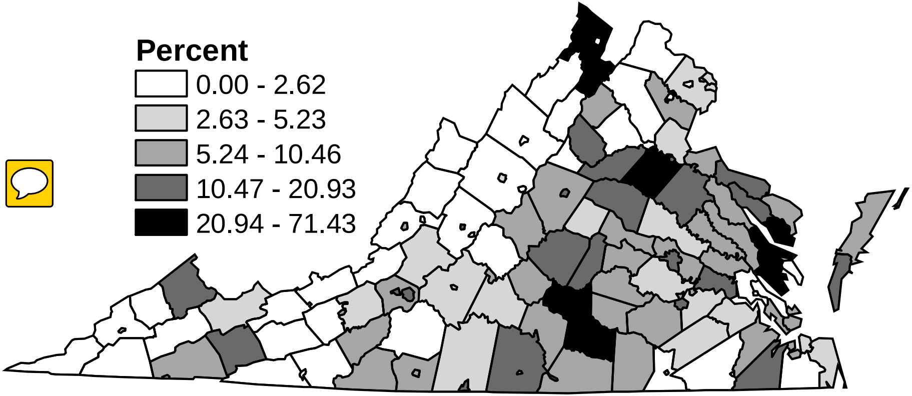


source: Schabenberger & Gotway (2005)

Temperature surface

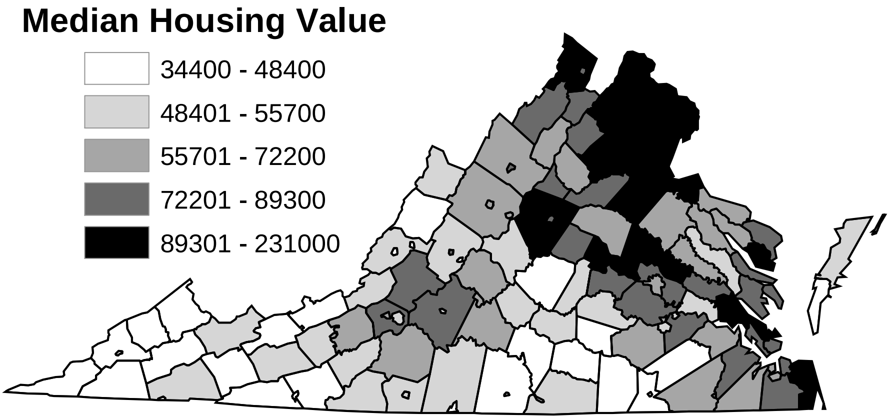


source: Schabenberger & Gotway (2005)



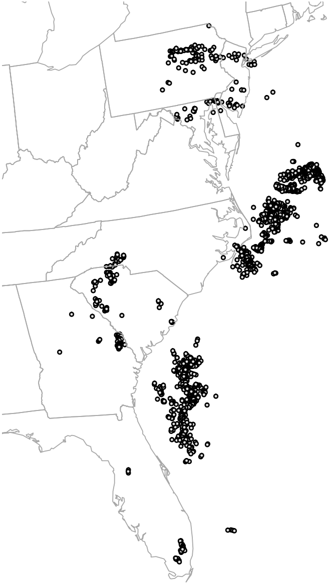
Percent of children under the age of 72 months with elevated blood lead levels in Virginia in 2000 (133 counties).
source: Schabenberger & Gotway (2005)

Median housing value per county in Virginia in 2000



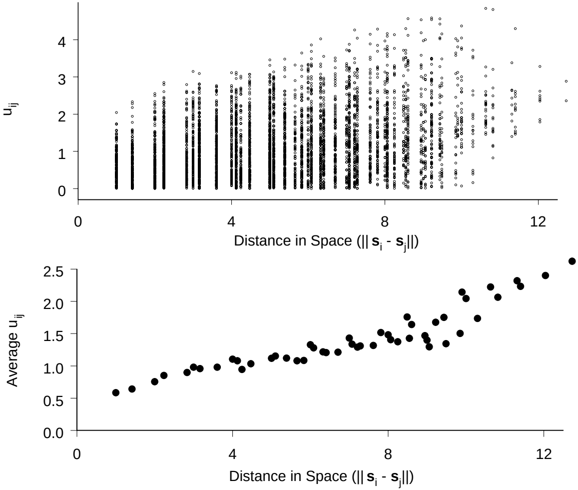
source: Schabenberger & Gotway (2005)

Locations of lightning strikes: April 17-20, 2003



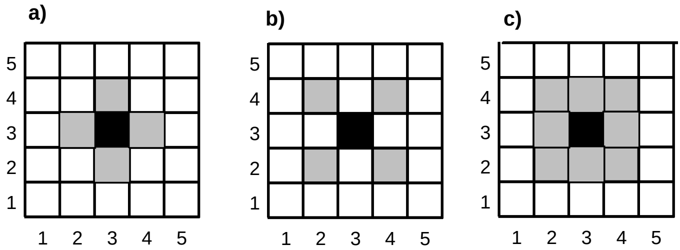
source: Schabenberger & Gotway (2005)

Scatter plot of $u_{ij} = |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|$ for simulated lattice data D)



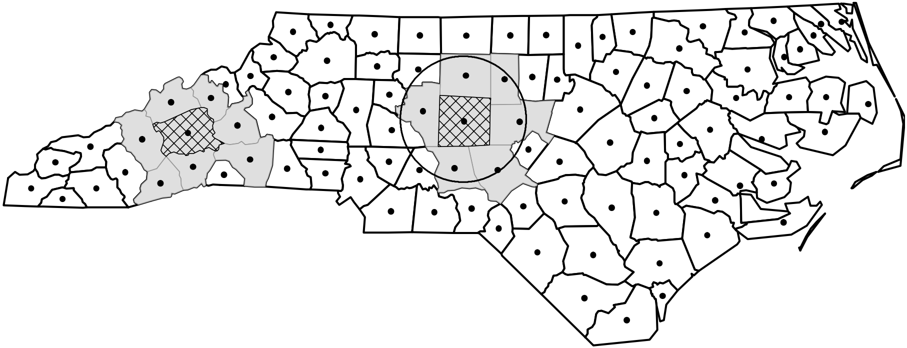
source: Schabenberger & Gotway (2005)

Possible definitions of spatial connectedness (contiguity) for a regular lattice



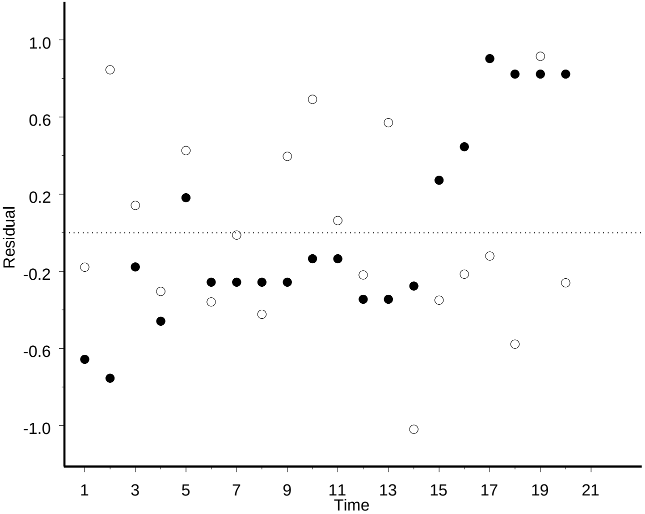
rook; bishop; queen
source: Schabenberger & Gotway (2005)

Two definitions of spatial connectedness (contiguity) for an irregular lattice



Circle of radius 35 miles
source: Schabenberger & Gotway (2005)

Independent vs AR(1) observations (var is 0.3)



source: Schabenberger & Gotway (2005)

- R is a free statistical package similar to S-plus (`http://r-project.org`)
- Even if you are a first time R user, there are more than enough resources for you to get started:
  - `http://dist.stat.tamu.edu/pub/rvideos/`
  - `http://www.statmethods.net/index.html`
  - `http://zoonek2.free.fr/UNIX/48_R/all.html`

- R packages for spatial statistics
  - fields (http://www.image.ucar.edu/Software/)
  - geoR (http://www.leg.ufpr.br/geoR)
  - RandomFields (https://cran.r-project.org/web/packages/RandomFields/index.html)
  - useful link:
    - Pages 1-15 can be useful (http://www.unc.edu/~rls/s890/ShortCourseMalta.pdf)
- R package for point patterns
  - spatstat (http://cran.r-project.org/web/packages/spatstat/index.html)