Mapping and Geovisualization

Luc Anselin

http://spatial.uchicago.edu
from mapping to geovisualization to ESDA

map design primer

statistical maps

mapping rates
From Mapping to Geovisualization to ESDA
• Definitions

a map is “a collection of spatially defined objects” (Monmonier)

beyond mapping

• map as analysis vs map as presentation

geovisualization

geospatial visual analytics

exploratory spatial data analysis (ESDA)
Geovisualization

“the creation and use of visual representations to facilitate thinking, understanding and knowledge construction” (MacEachren)

exploration, synthesis, presentation, analysis
• Geospatial Visual Analytics

computer science perspective

visual analytics (Thomas and Cook 2005)

detect the expected and discover the unexpected (Kielman et al 2009) = facilitate analytical reasoning

bring in space = geospatial visual analytics
• Exploratory Spatial Data Analysis

“a collection of techniques to describe and visualize spatial distributions, identify atypical locations or spatial outliers, discover patterns of spatial association, clusters or hot spots and suggest spatial regimes or other forms of spatial heterogeneity” (Anselin 1999)
• Traditional Knowledge Discovery
  
deductive approach
  
hypothesis first, data later
  
inductive approach
  
data first, hypothesis later
• Alternative Knowledge Discovery

• abductive approach

  pattern discovered along with hypothesis

interaction between data exploration and human perception

visual popout = aha moment
• Geovisual Analytics

  leverages both geovisualization and visual analytics

  interactive mapping

  animation

  linking and brushing
How to Lie with Maps (Monmonier)

- manipulate map design parameters
  - legends, colors, intervals, scale
- choice of projection
  - manipulate area through choice of projection
    (political maps)
  - larger areas seem more important
- human perception can be tricked
Where is it the worst?

More than 1 percent of all U.S. households were in some stage of foreclosure last year, nearly double the 2006 figure, and foreclosures soared to an all-time high in the final quarter of last year. Chicago has fared a bit better but has been stung by the real estate crisis nonetheless, with foreclosures growing by 45 percent in 2007.

CHICAGO FORECLOSURES

2006:
9,788

2007:
14,250

http://xefer.com//2008/04/maps
Map Design Primer
- Choropleth Map

choro from region, NOT chloro

visualizing a spatial distribution

values at locations

map counterpart of histogram

values for discrete spatial units
• Map Design Parameters

  datum and coordinate system (geodesy)

  scale

  projection

    shape, area, distance, direction

  classification

  color

  legend
• Representing Value

  discrete

  selection of intervals

  all data points in same interval obtain the same color or shading

  continuous

  color ramp

  cross-hatch density

• symbol

  doesn’t really work for large data sets
Classification
• Map Classification

  choice of intervals

  selection of cut points

    equal interval, natural breaks (Jencks), manual

• statistical criteria

  equal share (quantile), standard deviation

  extreme values
histogram and equal intervals choropleth map
NYC Sub-boroughs - % rental
Natural Breaks
GeoDa Custom Category Editor
NYC Sub-boroughs - % rental
Custom Intervals
Colors
• Color Choice

perception of value

perception of pattern

• reds = hot, blues = cold

red = danger

colorbrewer2.org (Cynthia Brewer)
Pick a color scheme:
Multi-hue:

Single hue:

Color Brewer recommended color schemes
Legends
• Sequential Legend

ordered data

low to high

not appropriate for categorical data
ColorBrewer sequential legend
• Diverging Legend

equal emphasis on mid-range and extremes in either direction

stresses difference from central tendency, rather than ordering of data
ColorBrewer diverging legend
• Qualitative Legend

for categorical data

no ordering, no high or low

stress discrete categories, not values
ColorBrewer qualitative legend
Statistical Maps
• Quantile Map

data sorted from low to high

equal number of observations in each interval

examples

  quartile map (4 categories)

  quintile map (5 categories)

possible issues with ties
quintile map (NYC % rental units)
• **Box Map**

identifying outliers

same principle as in box plot

fence = median + 1.5 IQR or + 3 IQR

IQR = inter quartile range, 25% to 75%

six intervals

same principle as quartile map

outliers identified as a separate category
upper outliers in box plot and box map
(NYC median rent 2008)
lower outliers in box plot and box map
(NYC median rent 2008)
• Standard Deviational Map

based on standardized data values

mean = 0, standard deviation = 1

intervals correspond to one standard deviation

outliers are more than 2 standard deviations from the mean
standard deviational map
(NYC median rent 2008)
• Cartogram

areal unit proportional to variable of interest

avoid misleading effect of area

use transformed shapes

circular cartogram

contiguous cartogram
box map and circular cartogram
contiguous cartogram
area = number of votes in electoral college
source: Sarah Williams
• Conditional Maps

cc maps, conditioned choropleth maps (Carr)

special case of trellis graphs

micromap matrix

conditioning variables on the axes

matrix of mini maps for the variable of interest conditioned by the values on the axes
child malnutrition cc map conditioned on poverty index and per capita income (Nepal districts)
• Map Animation

map movie

highlight observations in increasing or decreasing order

one at a time

cumulative

visual impression of patterning/clustering
Mapping Rates
Risk and Rates
• Concept of Risk

  many meanings

  risk = probability that an event may occur

  actual risk is not observed

  only events are observed
• Risk Estimate

raw rate or crude rate

number of events / population at risk

typically expressed in per 1,000 or some multiple

crude rate is maximum likelihood estimate
• Rate Maps

focus on spatial heterogeneity

risk is not uniform across space

interest in identifying areas of elevated risk

associate elevated risk with causal factors
Foreclosure Count in Franklin county (OH), 2007
Housing Units in Franklin county (OH), 2007
Foreclosure Rate Outliers in Franklin county (OH), 2007
Excess Risk
• **What is Excess Risk?**
  
elevated risk = higher than some standard

what is the standard

rate computed for a reference group

e.g., foreclosure for the whole city
• **Average Risk**

not the average of the rates

total number of events / total population

e.g., all the foreclosed homes in the metro area over all the homes in the metro area

weighted average of the rates, weighted by their population share
• Average Risk Computation

\( O_i \) : observed number of events

\( P_i \) : “population at risk”

\( r_i \) : rate for i, \( r_i = O_i / P_i \)

average risk \( r = ( \sum_i O_i ) / (\sum_i P_i) \)
• Expected Events

\[ E_i = r \times P_i \]

expected events = average rate times the “population” in area \( i \)

e.g., if average risk of an event is 1 per 10,000, then a county of 30,000 would have 3 expected events
• **Relative Risk**

compare observed to expected

observed = number of events, \( O_i \)

expected = number of events if average were applied to population, \( E_i \)

relative risk

\[
\text{observed} / \text{expected}, \ O_i / E_i
\]
• Excess Risk Map

compare relative risk to unity

> 1: more events than on average

higher (excess) risk

< 1: fewer events than on average

choropleth map of excess risk
Franklin County 2007 foreclosures excess risk
Smoothing Rates
number of events as draws from a binomial distribution

\[ \text{Prob}[O = x] = \binom{P}{x} \pi^x (1 - \pi)^{P-x}, \text{ for } x = 0, 1, \ldots, P. \]

probability of \( x \) events given risk of \( \pi \)

mean: \( \mathbb{E}[O] = \pi P \)

variance: \( \text{Var}[O] = \pi (1 - \pi) P \)
• Moments of the Rate Estimate

O is random variable, P is not

\[ r = \frac{O}{P} \quad \text{O events, P population} \]

\[ E[r] = \frac{E[O]}{P} = \pi \frac{P}{P} = \pi \]

\[ \text{Var}[r] = \frac{\text{Var}[O]}{P^2} = \pi (1 - \pi) \frac{P}{P^2} = \pi (1 - \pi) \frac{1}{P} \]
• Variance Instability

P in denominator

smaller areas have larger variance = less precision

Example

with true (unknown) $\pi = 0.1$

Pop 1 = 500 and Pop 2 = 100,000

$SE1 = 0.013$ and $SE2 = 0.0009$
• Why Smooth?

rate estimates have variable reliability

less precision for smaller areas

why trust (e.g., no events = zero risk?)

borrow strength

use additional information to improve estimate
• Shrinkage (James-Stein)

new estimate that improves overall precision while sacrificing some bias

bias-variance tradeoff

shrink crude rate towards overall mean

overall mean contains useful information

shrink (smooth) rate as inverse function of variance


Effect of Shrinkage

Smoothing depends on variance:

- Small population → large variance → a lot of smoothing
- Large population → small variance → little smoothing
• Empirical Bayes Smoothing

shrinkage estimate as a weighted average of the crude rate and a prior

prior estimated from the data (reference rate)

hence “empirical” Bayes

$$\pi_i = w_i r_i + (1 - w_i) \theta$$ with $\theta$ as reference rate

weights inversely proportional to variance
Foreclosure Raw Rate in Franklin county (OH), 2007
Foreclosure EB Rate in Franklin county (OH), 2007
• Effect of Empirical Bayes Smoothing

position changes in cumulative distribution

small outlier areas move towards the overall mean

large high/low value areas may become outliers

spurious outliers are removed
To Smooth or Not To Smooth
• Pros of Smoothing

better estimates in MSE sense

adjusts for variance instability

removes spurious outliers

better estimates of true extremes
• Cons of Smoothing

degree of arbitrariness

sensitive to smoothing method

oversmoothing hides interesting “outliers”
• Smoothing in Practice

  some healthy debate

  use of original data vs transformed data

  many options

  need for sensitivity analysis