

# Introductory Methods for Area Data

Bailey and Gatrell Chapter 7

Lecture 17

November 11, 2009

1

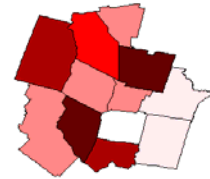
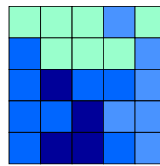
## Analysis Methods for Area Data

Values are associated with a fixed set of areal units covering the study region

We assume a value has been observed for all areas

Assume the areas are the only locations at which the attributes can be measured.

The areal units may take the form of a regular lattice or irregular units



2

## Analysis Methods for Area Data

### Objectives

- Not prediction - there are typically no unobserved values, the attribute has been measured for every unit.
- Model spatial patterns in the values associated with fixed areas - determine possible explanations for such patterns

3

## Analysis Methods for Area Data

For continuous data models we were more concerned with explanation of patterns in terms of locations

In this case explanation is in terms of covariates measured over the same units as well as in terms of the spatial arrangement of the areal units

### Examples

Relationship of disease rates and socio-economic variables

Relationship of election results and demographic variables

Relationship of forest biomass to elevation, precipitation

4

## Analysis Methods for Area Data

$$\{Y(s), s \in R\}$$

### Continuous case

Random variable Y indexed by locations

### Area case

$$\{Y(A_i), A_i \in A_1 \dots A_n\}$$

Random variable Y indexed by a fixed set of areal units

$$A_1 \cup \dots \cup A_n = R$$

The set of areal units covers the study region R

Conceive of the sample as a sample from a super population – all realizations of the process over these areas that might ever occur.

The sample observations are one possible realization

5

## Analysis Methods for Area Data

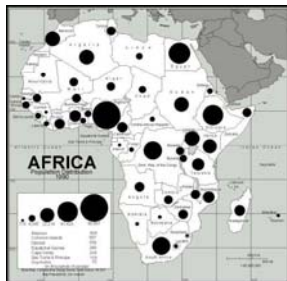
Explore these attribute values in the context of global trend or first order variation and second order variation – pattern in values given the spatial arrangement of the set of areas

- First order variation as variation in the of mean,  $\mu_i$  of  $Y_i$
- Second order variation as spatial dependence or  $COV(Y_i, Y_j)$

6

## Visualizing Area Data

Can use proportional symbols superimposed over the areal units  
Symbols are proportionate to the attribute value of the area



7

## Visualizing Area Data

Can use *choropleth* maps – each area is shaded according to the attribute value associated with the areal unit

- Attribute of interest is scaled to a set of discrete ranges or classes
- Each zone is shaded or colored according to its attribute value



8

## Visualizing Area Data

Class intervals compress attribute ranges into a relatively small number of discrete values.

### Number of classes

Should be a function of the range of data variability

Limited by human perception to ~8 classes

Rule of thumb: # classes =  $1 + 3.3 \log n$

5 observations = 3 classes

20 observations = 6 classes

40 observations = 7 classes

200 observations = 8 classes

1500 observations = 11 classes

9

## Visualizing Area Data

Same class intervals as applied to continuous data

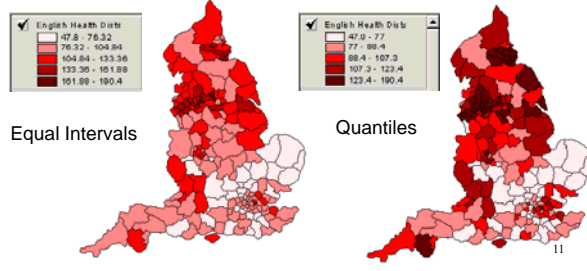
- Equal Intervals - for fairly uniformly distributed data
- Trimmed equal intervals - handles a few outliers from a uniform distribution
- Percentiles - rank data and get  $x$  evenly distributed classes of width  $1/x$
- Quartile map - 4 classes, lowest quartile, second quartile, third, and highest
- Standard Deviates - divide data into units of standard deviation around the mean

10

## Visualizing Area Data

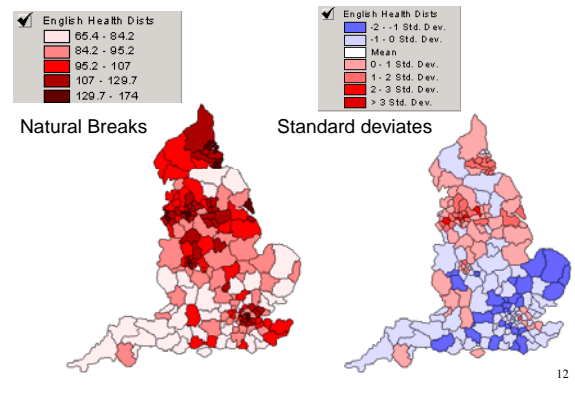
Visual outcome depends heavily on class interval choice, color and shading

Use of discrete classes to assign colors can give false impressions of both uniformity (within units) and discontinuity (between units)



11

## Visualizing Area Data



12

## Visualizing Area Data

### Problems with choropleth maps

Arbitrariness of the areal units

What was the origin of areal unit boundaries?

Designed for convenience or efficiency rather than reflection of the underlying spatial pattern – most enumeration units

Typically referred to as modifiable area unit problem - MAUP

13

## Modifiable Area Unit Problem

Different zones will produce virtually any numbers from the same underlying distribution

### Original data (individuals living in households)

Mean 3.75 var 2.6

2	4	6	1
3	6	3	5
1	5	4	2
5	4	5	4

mean 3.75 var 0.50

3	3.5
4.5	4
3	3
4.5	4.5

mean 3.75 var 0.00

3.75	3.75
3.75	3.75

mean 3.75 var 0.93

2.5	5	3
	4.5	
3	4.5	3
	4.5	

mean 3.17 var 2.11

4	1
4	3.67

Source Jelenski and Wu (1996) "The modifiable area unit problem and implications for landscape ecology". *Landscape Ecology* 11(3) p. 129-140.

14

## Visualizing Area Data

### Problems with choropleth maps

- Large areas tend to dominate the information content of the map
- Area shading of attribute values means two visual variables apply to the data – color and area
- Should not map absolute numbers with a choropleth map
- With absolute counts there is no relation of the attribute variable to the absolute and/or relative size of the spatial units
- Choropleth maps are not appropriate for counts unless the data are corrected for area, population or some other measure that factors out the size of the enumeration districts.

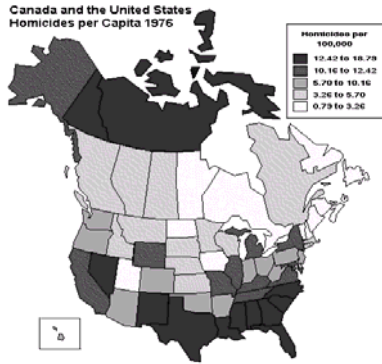
15

## Visualizing Area Data

- The problem of class limits
  - Use of continuous shading
- Area dependence problem
  - Correct for population or area
  - Transform areas to be proportionate to the attribute value

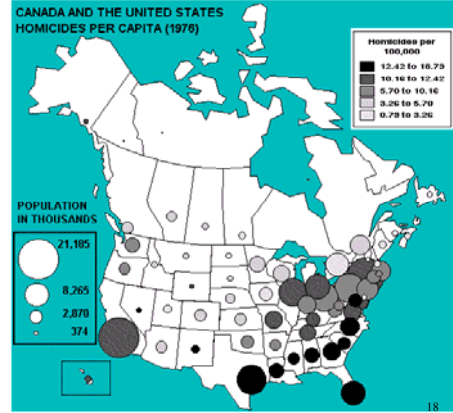
16

Map of murder rates per 100,000 US and Canada



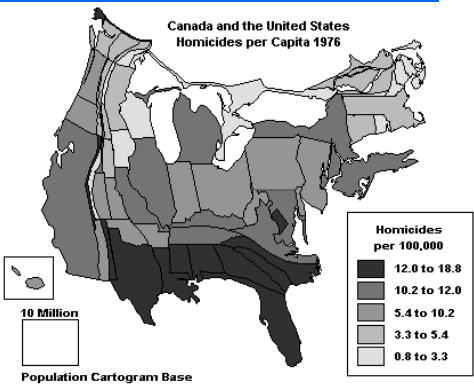
17

Corrects for population

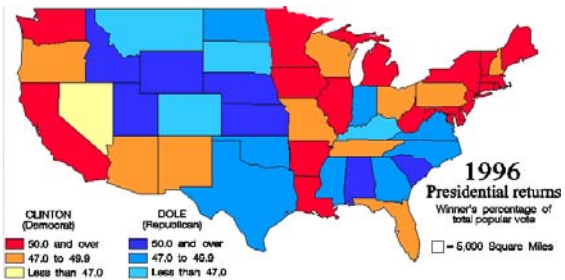


18

Cartogram - Attempt to correct for area or population



19



20

## Cartograms

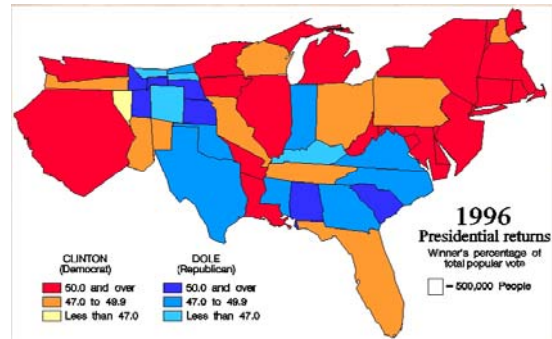


There are two distinct and conflicting goals in the construction of cartograms:

- adjusting region sizes
- retaining region shapes

21

## Cartograms

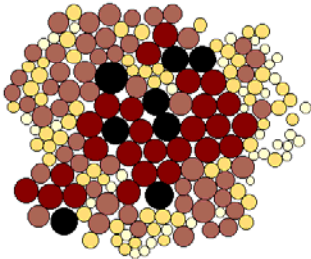


22

## Cartograms

Relaxation of contiguity and shape constraints

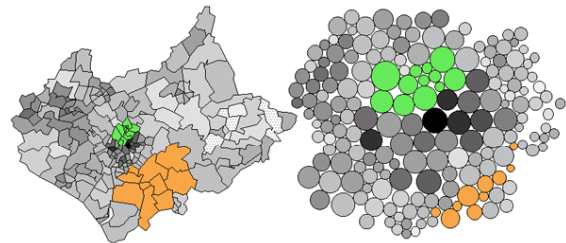
A non-continuous population cartogram based on the population values for the census zones of Leicestershire created using Dorling's 'circle growing' algorithm.



Both the circle areas and shades are proportionate to the zone populations.

## Cartograms

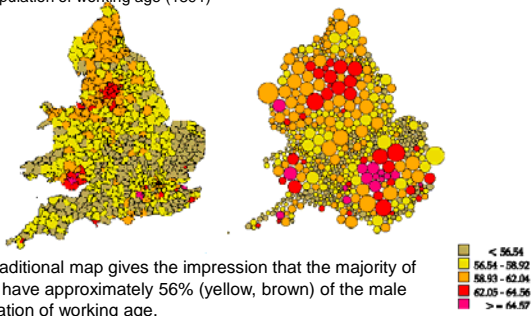
Cartograms represent areas in relation to their population size.



Patterns are displayed in relation to the number of people involved instead of the size of the area involved.

24

Comparison of traditional choropleth map with cartogram showing percent of male population of working age (1891)



The traditional map gives the impression that the majority of areas have approximately 56% (yellow, brown) of the male population of working age.

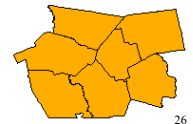
The cartogram gives a different overall view - the majority of areas have a male population of working age over 58% (orange, red, and purple on the map).



## Exploring Area Data

- Approaches for examining mean values over the areal units
- Techniques for exploring spatial dependence

What was essential information for exploration of these with point and continuous data?



## Proximity Measures

Need measures of proximity for irregular areal units

$$w_{ij} \begin{cases} 1 & \text{centroid of } A_j \text{ is one of } k \text{ nearest centroids to that of } A_i \\ 0 & \text{otherwise} \end{cases}$$

$$w_{ij} \begin{cases} 1 & \text{centroid of } A_j \text{ is within distance } d \text{ of centroids of } A_i \\ 0 & \text{otherwise} \end{cases}$$

$$w_{ij} \begin{cases} 1 & A_j \text{ shares a boundary with } A_i \\ 0 & \text{otherwise} \end{cases}$$

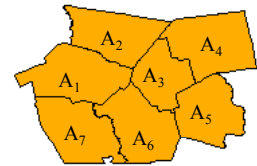
$$w_{ij} \begin{cases} d_{ij}^\gamma & \text{if intercentroid distance } d_{ij} < d \\ 0 & \text{otherwise} \end{cases}$$

27

## Proximity Measures

$$\mathbf{W} = \begin{matrix} & \begin{matrix} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 \end{matrix} \\ \begin{matrix} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \\ A_7 \end{matrix} & \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

Adjacency



The spatial proximity matrix  $\mathbf{W}$  with elements  $w_{ij}$  represents a measure of proximity of  $A_i$  to  $A_j$

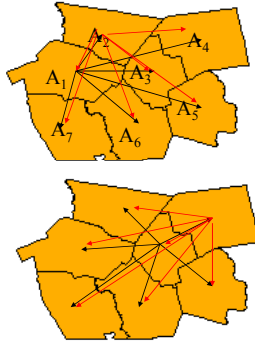
28

## Proximity Measures

$$W = \begin{matrix} & A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix}$$

k nearest centroids

Not symmetric



29

## Proximity Measures

Proximity measures can be specified as measures of different orders – spatial lags

These can be represented as different proximity matrices for different lags

For example  $W_1$  represents first spatial lag,  $W_2$  the second spatial lag, etc

30

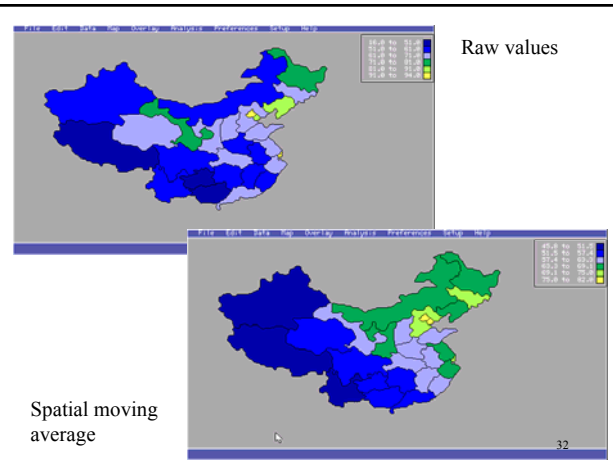
## Spatial Moving Averages

One way to estimate the mean is by the average of the values in “neighboring” areas

The proximity matrix  $W$  represents the neighbors

$$\hat{\mu}_i = \frac{\sum_{j=1}^n w_{ij} y_j}{\sum_{j=1}^n w_{ij}}$$

31



## Median Polish

If areal units are a regular grid can use median polish to estimate the trend

Median is more resistant to outliers in the data than the mean

Each value is decomposed into:

$$y_{ij} = \mu + \mu_i + \mu_j + \varepsilon_{ij}$$

where  $\mu$  is a fixed overall effect

$\mu_i$  and  $\mu_j$  represent fixed row and column effects

$\varepsilon_{ij}$  is a random error term

The overall mean is  $\mu_{ij} = \mu + \mu_i + \mu_j$

33

## Median Polish algorithm

1. Take the median of each row and record the value to the side of the row – subtract the row median from each value in that row
2. Compute the median of the row medians, and record the value as the overall effect, Subtract the overall effect from each of the row medians
3. Take the median of each column and record the value beneath the column, Subtract the column median from each value in that particular column
4. Compute the median of the column medians, and add the value to the current overall effect. Subtract this addition to the overall effect from each of the column medians.

34

## Median Polish algorithm

5. Repeat steps 1-4 until no changes occur with the row or column medians

3	4	5
5	4	6
5	6	5

	1	2	3	s+1
1	3	4	5	0
2	5	4	6	0
3	5	6	5	0
r+1	0	0	0	0

1)

			s+1	
	-1	0	1	4
	0	-1	1	5
	0	1	0	5
r+1	0	0	0	0

3a)

			s+1	
	-1	0	1	-1
	0	-1	1	0
	0	1	0	0
r+1	0	0	1	5

2a)

			s+1	
	-1	0	1	4
	0	-1	1	5
	0	1	0	5
r+1	0	0	0	5

3b)

			s+1	
	-1	0	0	-1
	0	-1	0	0
	0	1	-1	0
r+1	0	0	1	5

2b)

			s+1	
	-1	0	1	-1
	0	-1	1	0
	0	1	0	0
r+1	0	0	0	5

4a)

			s+1	
	-1	0	0	0
	0	-1	0	0
	0	1	-1	0
r+1	0	0	0	5

## Median Polish algorithm

Resulting cell values  $\hat{\mu}_{ij} = \hat{\mu} + \hat{\mu}_i + \hat{\mu}_j$

Represents a global trend in the data

Also provides a method to remove trend – leaving a set of residuals to analyze

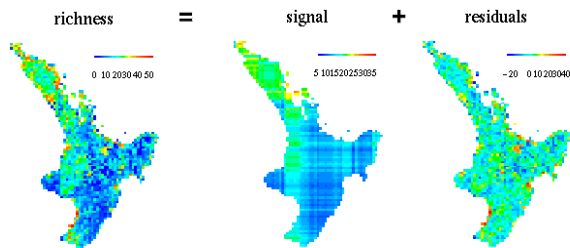
Can produce banding effects since it decomposes trend according to the directions of the grid

These direction may not correspond to the direction of trend in the data

No control over the degree of smoothing

36

## Median polish example



Source:

<http://www.rem.sfu.ca/gis/Projects/Eh/Nzbirds/geostats/polish.htm>

## Kernel Estimation

Kernel estimation was employed to explore intensity variations for point pattern data and to describe changes in first order trend in continuous data.

It may be used in the area case, as well.

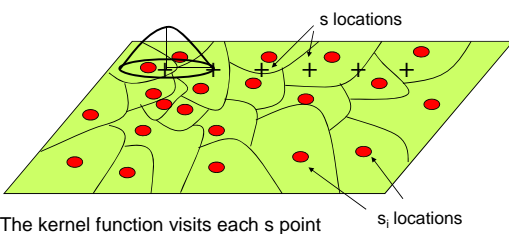
Kernel methods are really only applicable to distance, and therefore do not use the **W** matrix.

$$\hat{\mu}_\tau(s) = \frac{\sum_{i=1}^n k\left(\frac{(s-s_i)}{\tau}\right) y_i}{\sum_{i=1}^n k\left(\frac{(s-s_i)}{\tau}\right)}$$

use simple distance from the zone centroids to a set of points (small areas) **s**

38

## Kernel Estimation



The kernel function visits each **s** point

39

## Kernel Estimation

When observations are counts the previous approach is not appropriate - need an estimate of density

$$\hat{\lambda}_\tau(s) = \sum_{i=1}^n \frac{1}{\tau^2} k\left(\frac{(s-s_i)}{\tau}\right) y_i$$

Represents the total per unit area (density)

Count estimate derived from density estimate times area of region or integrating density estimate over the region

One approach to cross areal interpolation

40

## Cross Areal Interpolation

With areal data there is often a need to interpolate values from one set of areas to another

Kernel estimation has been used to convert count data from set of irregular units to set of finer grid units.

The data can then be re-aggregated to other sets of areas and used with data from an alternate set of areas

41

## Cross Areal Interpolation

### Other methods

#### Assign data to nearest centroid

Data are interpolated on the basis of assigning all data associated with an area in one set to that area in the other set that is closest in terms of smallest centroid-centroid distance

#### Point in polygon interpolation

Assign data in one set of areas to a set in which the centroid lies

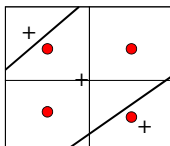
#### Area weighting

Interpolate data based on weighted average of data in areas in one set by which the area in second set overlaps

42

## Cross Areal Interpolation

### Example



43