

Introductory Methods for Area Data

Bailey and Gatrell Chapter 7

Lecture 18

November 12, 2009

1

Exploring second order effects

A variogram or covariogram could be used if we assume that the attribute value is located at a point and use Euclidean distances between points

For area data, measures of **correlation** are used more commonly than covariance.

Autocorrelation

- Correlation of a random variable with itself
- In time domain: correlation between value at time t and at time $t - h$
- In spatial domain: correlation between value at a location i and a set of neighboring locations j

2

Spatial Autocorrelation Measures

Compare match between locational similarity and value (attribute) similarity

■ Locational similarity

Spatial weights - W

■ Types of measures for value similarity

Cross product: $x_i x_j$

Squared difference $(x_i - x_j)^2$

Absolute difference $|x_i - x_j|$

3

Spatial Randomness

- Values observed at locations do not depend on values observed at neighboring locations
- Observed spatial pattern of values is equally likely as any other pattern
- The location of values may be altered without affecting the information content of the data

Positive Spatial Autocorrelation

Like values tend to cluster in space

Negative Spatial Autocorrelation

Dissimilar values tend to cluster in space

4

Spatial Autocorrelation Statistics

Area data do not vary simply by location, but are functions of the fixed sub-regions into which they are divided.

Autocorrelation or variation must be measured using the proximity matrix **W**.

Same basic notion applies - characterize the similarity or difference of the increments of the function separated by a certain lag.

Measures of Spatial Autocorrelation

- Join counts
- Moran's I
- Geary's C

5

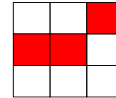
Join Count Statistics for Binary Data

For data measured on a nominal scale

Match value-location

Count joins in value that match joins in space

$x_i = 1$ or 0 White or Black(red)



Types of joins

BB	0-0	$(1 - x_i)(1 - x_j)$
WW	1-1	$x_i x_j$
BW	0-1	$(x_i - x_j)^2$

6

Join Count Statistics for Binary Data

$$WW = (1/2) \sum_i \sum_j w_{ij} x_i x_j$$

There are n_1 units coded 1 (W) and n_2 units coded 0 (B) and $n_1 + n_2 = n$

$$BB = (1/2) \sum_i \sum_j w_{ij} (1 - x_i)(1 - x_j)$$

$$BW = (1/2) \sum_i \sum_j w_{ij} (x_i - x_j)^2 \quad w_{ij} \text{ is proximity matrix}$$

Under random sampling number of map patterns is 2^n

$$E(WW) = J n_1^2 / n^2 \quad E(BW) = 2 J n_1 n_2 / n^2$$

$$E(BB) = J n_2^2 / n^2 \quad \text{Where } J = BB + WW + BW$$

7

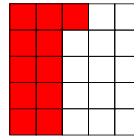
Join Count Statistics for Binary Data

When similarly coded units are next to each other

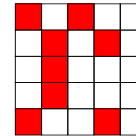
$BW \rightarrow 0$ and either BB increases, WW increases, or both

When dissimilarly coded units are next to each other

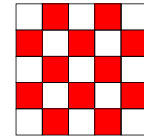
$BB \rightarrow 0$ $WW \rightarrow 0$ $BW \rightarrow J$



BB=14 WW=20 BW=6



BB=2 WW=14 BW=23



BB=0 WW=0 BW=40

8

Moran's I

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i=1}^n (y_i - \bar{y})^2 \right) \left(\sum_{i \neq j} \sum w_{ij} \right)}$$

Moran's I typically ranges from -1 to 1. An uncorrelated process has an expected $I \sim 0$. Negative values of I indicate negative autocorrelation. Positive values indicate positive autocorrelation.

Related to covariogram $\hat{C}(h) = \frac{1}{N(h)} \sum \{Z(s_i) - \hat{\mu}\} \{Z(s_j) - \hat{\mu}\}_9$

Moran's I

Example

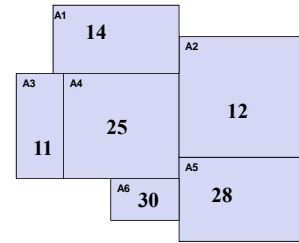
$$n = 6$$

$$\sum_i (y_i - \bar{y})^2 = 370$$

$$\bar{y} = \sum_i y_i / 6 = 20$$

$$\sum_i \sum_j w_{ij} = 18$$

$$W = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$



10

Moran's I Example

$$\begin{aligned} \sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y}) &= (14-20)(12-20) + (14-20)(11-20) \\ &+ (14-20)(25-20) + (12-20)(14-20) \\ &+ (12-20)(25-20) + (12-20)(28-20) \\ &+ (11-20)(14-20) + (11-20)(25-20) \\ &+ (25-20)(14-20) + (25-20)(12-20) \\ &+ (25-20)(11-20) + (25-20)(28-20) \\ &+ (25-20)(30-20) + (28-20)(12-20) \\ &+ (28-20)(25-20) + (28-20)(30-20) \\ &+ (30-20)(25-20) + (30-20)(28-20) \\ &= 186 \end{aligned}$$

$$I = \frac{n \sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_i \sum_j w_{ij} \right) \sum_i (y_i - \bar{y})^2} = \frac{6 * 186}{18 * 370} = 0.1676$$

11

Geary's C

$$C = \frac{(n-1) \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - y_j)^2}{2 \left(\sum_{i=1}^n (y_i - \bar{y})^2 \right) \left(\sum_{i \neq j} \sum w_{ij} \right)}$$

Geary's C typically ranges from 0 to 3. It cannot be negative. An uncorrelated process has an expected $C = 1$. Values less than 1 indicate positive spatial autocorrelation, while values greater than 1 indicate negative autocorrelation.

Related to variogram $2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{i < j} [z(s_i) - z(s_j)]^2$

12

Geary's C Example

$$\begin{aligned} \sum_i \sum_j w_{ij} (y_i - y_j)^2 &= (14-12)^2 + (14-11)^2 + (14-25)^2 + (12-14)^2 \\ &\quad + (12-25)^2 + (12-28)^2 + (11-14)^2 + (11-25)^2 \\ &\quad + (25-14)^2 + (25-12)^2 + (25-11)^2 + (25-28)^2 \\ &\quad + (25-30)^2 + (28-12)^2 + (28-25)^2 + (28-30)^2 \\ &\quad + (30-25)^2 + (30-28)^2 \\ &= 1586 \end{aligned}$$

$$C = \frac{(n-1) \sum_i \sum_j w_{ij} (y_i - y_j)^2}{2 \left(\sum_i \sum_{j \neq i} w_{ij} \right) \sum_i (y_i - \bar{y})^2} = \frac{6-1 * 1586}{2 * 18 * 370} = 0.5953$$

13

Relationship of Moran's I and Geary's C

- C approaches 0 and I approaches 1 when similar values are clustered
- C approaches 3 and I approaches -1 when dissimilar values tend to cluster
- High values of C measures correspond to low values of I
- So the two measures are inversely related

14

Local Moran's I

A global spatial autocorrelation measure provides only one statistic to summarize the whole study area.

Local Moran's I is used to determine if local autocorrelation exists around a specified subregion i ($i=1, \dots, n$).

$$I_i = \frac{n(y_i - \bar{y})}{\sum_j (y_j - \bar{y})^2} \sum_j w_{ij} (y_j - \bar{y})$$

15

Local Moran's I

Example using region 6 from previous example

$$n = 6$$

$$\bar{y} = \sum_i y_i / 6 = 20 \quad \sum_i (y_i - \bar{y})^2 = 370$$

$$\text{For region 6} \quad y_6 - \bar{y} = 30 - 20 = 10$$

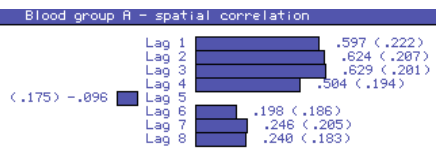
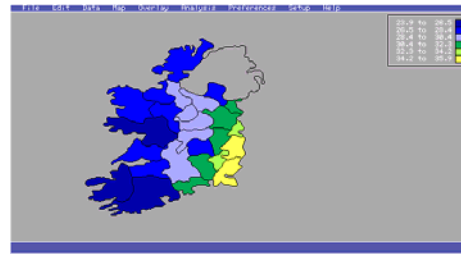
$$I_6 = \frac{n(y_6 - \bar{y})}{\sum_j (y_j - \bar{y})^2} \sum_j w_{6j} (y_j - \bar{y}) = \frac{6 * 10}{370} * 13 = 2.108$$

16

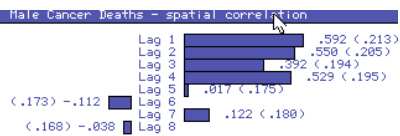
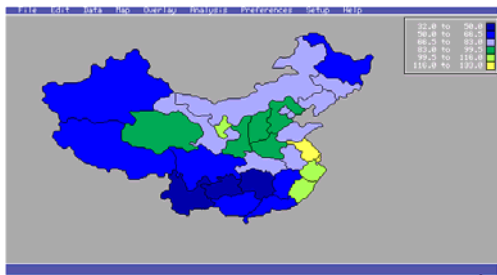
Correlograms

Correlograms are constructed by calculating spatial autocorrelation at different spatial lags and plotting the correlation values against the lag distances

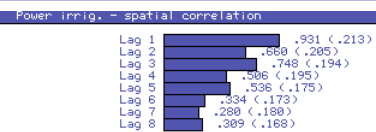
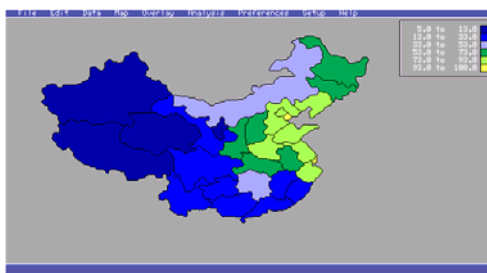
17



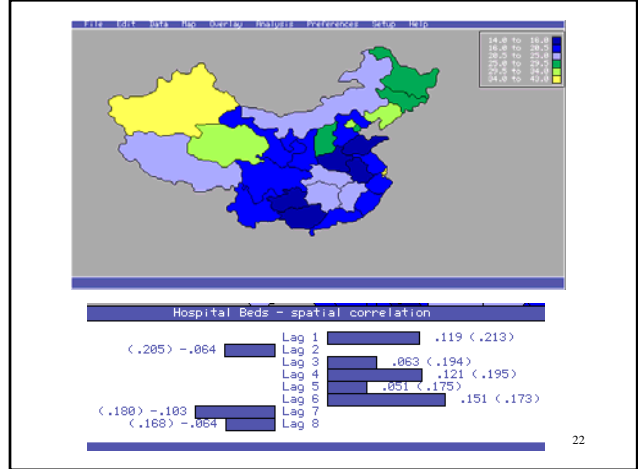
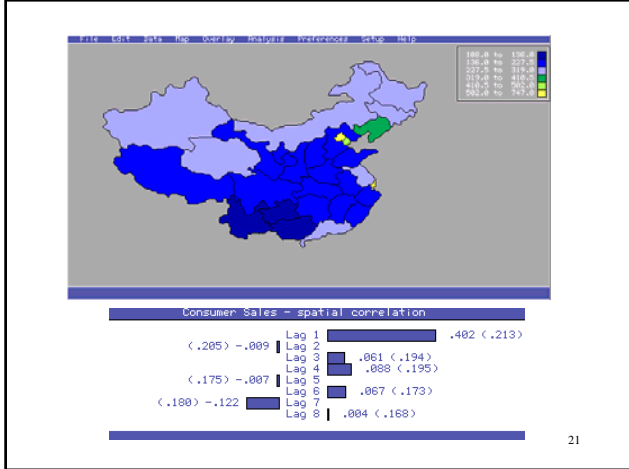
18



19



20



Modeling Areal Data

In modeling we want to establish a relationship between attribute y_i and, the relative spatial arrangement of A_i , and other possible attributes x_j recorded for each A_i

First consider models to account for first order variation in the mean values

Non-spatial regression models

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

\mathbf{Y} is a $(n \times 1)$ vector of random variables – response

\mathbf{X} is a matrix of $(n \times p)$ explanatory variables

$\boldsymbol{\beta}$ is a $(p \times 1)$ vector of coefficients

$\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector of zero mean random variables ε_i

23

Modeling Areal Data

ε_i represent fluctuations about the mean $\mu_i = x_i^T \boldsymbol{\beta}$

The errors ε_i , are assumed to have constant variance and be independent – zero covariance, and normally distributed

Under these assumptions data can be fit by ordinary least squares to derive estimates for $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\text{VAR}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

24

Modeling Areal Data

Some of the explanatory variables ($x_{i1} \dots x_{ip}$) could be the centroids of the areas, their powers or cross products to add a trend surface to the model

The variance σ^2 given in $VAR(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ will most often be unknown and will need to be estimated from the residuals of the fitted models

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p} \quad \text{where } \hat{y}_i = x_i^T \hat{\beta}$$

Examination of the residuals is used to assess the model fit 25

Modeling Areal Data

Ordinary regression assumptions are likely to be violated

Residuals will not be independent

Variance is not likely to be constant

As a result confidence intervals for coefficients and assessment of significance are affected and may be invalid

26

Modeling Areal Data

Non-constant variance in the residuals can be corrected by:

Weighted regression

A special case of generalized least squares where C is a diagonal matrix. Each observation in the regression is weighted in inverse proportion to its variance – less weight to observations with large variance

Transformations of y_i to y_i'

Variance of transformed variable is constant

Linearizes the relationship between the mean and the covariates

27

Transformations of y_i to y_i'

$$y_i' = \frac{y_i^\lambda - 1}{\lambda} \quad \lambda \neq 0 \quad y_i' = \log(y_i) \quad \lambda = 0$$

For counts:

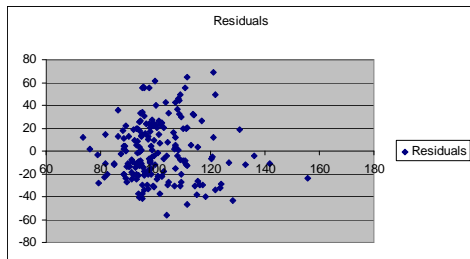
might expect a Poisson distribution

variance is proportional to mean

Square root transformation will convert variance to a constant

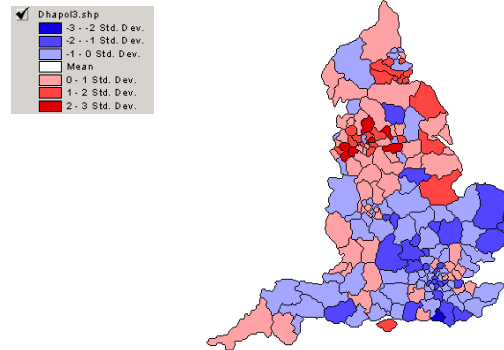
Logarithmic transformation often used to linearize relation of mean to covariates and to stabilize variance 28

Plot of residuals against fitted values



33

Residuals from OLS Regression Model



34

Spatial autocorrelation of residuals from OLS regression model – Jarman Score

Spatial Correlation Estimate
 Statistic = "moran" Sampling = "free"
 Correlation = **0.3921**
 Variance = 0.001937
 Std. Error = 0.04401
 Normal statistic = 9.029
 Normal p-value (2-sided) = 1.729e-19

Spatial Correlation Estimate
 Statistic = "geary" Sampling = "free"
 Correlation = **0.5858**
 Variance = 0.003013
 Std. Error = 0.05489
 Normal statistic = -7.546
 Normal p-value (2-sided) = 4.494e-14

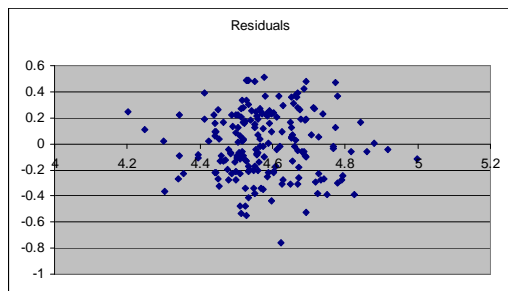
35

OLS regression of log Heart Attacks and log Jarman index

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.439004552					
R Square	0.192724996					
Adjusted R Sq	0.18843098					
Standard Error	0.248655421					
Observations	190					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	2.775046422	2.775046	44.88223	2.35904E-10	
Residual	188	11.62394942	0.06183			
Total	189	14.39899584				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	1.717468515	0.42713325	4.02092	8.39E-05	0.874878908	2.560058
Injarman	0.623107036	0.093009104	6.69942	2.36E-10	0.439631495	0.806583

36

Plot of residuals against fitted values



37

Spatial autocorrelation of residuals from OLS regression model – Jarman Score

Statistic = "moran" Sampling = "free"

Correlation = 0.4723

Variance = 0.001937

Std. Error = 0.04401

Normal statistic = 10.85

Normal p-value (2-sided) = 1.947e-27

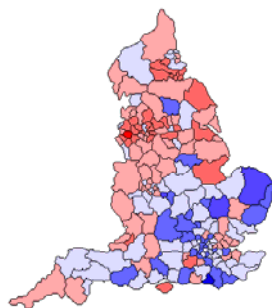
Null Hypothesis: No spatial autocorrelation

Summary of the permutation-correlations :

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
permutation p-value = 0	-0.1287	-0.03516	-0.006908	-0.005121	0.02254	0.1555

38

Residuals



Log Jarman score only

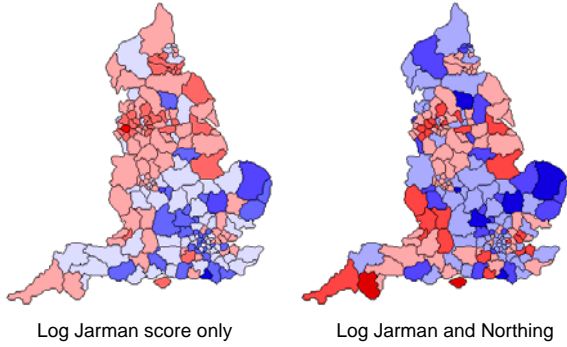
39

OLS regression of log heart attacks, Northing and log Jarman index

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.657648315				
R Square	0.432501306				
Adjusted R Square	0.426431801				
Standard Error	0.209039199				
Observations	190				
<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	6.227585	3.113792	71.25809	9.9E-24
Residual	187	8.171411	0.043697		
Total	189	14.399			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i> <i>Upper 95%</i>
Intercept	2.498893747	0.369686	6.759496	1.71E-10	1.769602 3.228185
Injarman	0.392225422	0.082392	4.760471	3.86E-06	0.229688 0.554763
NORTHINC	1.14473E-05	1.29E-06	8.888767	5.16E-16	8.91E-06 1.4E-05

40

Residuals



41

Spatial autocorrelation of residuals from OLS regression model – log Jarman Score and Northing

Statistic = "moran" Sampling = "free"

Correlation = 0.2981

Variance = 0.001937

Std. Error = 0.04401

Normal statistic = 6.893

Normal p-value (2-sided) = 5.472e-12

Summary of the permutation-correlations :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.1353	-0.03573	-0.00809	-0.00549	0.02365	0.1546

permutation p-value = 0

42