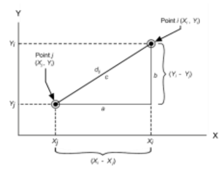
Spatial Statistics

GEOG 419: Lembo

Point Pattern Analysis

- Global methods to analyze point patterns across entire study region (or a map)
  - Quantitative tools for examining a spatial arrangement of point locations on the landscape
- Two common types of analysis
  - spacing of individual points – nearest neighbor analysis
    - Ex. fire stations locations – random or dispersed
      - Goal: equitable service throughout region
      - Design new configuration (e.g., relocating, new stations)
      - More or less dispersed than original configuration
  - nature of overall point pattern – are locations dispersed or clustered
    - Ex. diseased trees in a national forest
      - Widespread aerial spraying versus concentrated ground treatment

Center Point



Pythagorean theorem:

$a^2 + b^2 = c^2$

In a right-angled triangle, the square of the hypotenuse (c) is equal to the sum of the squares of the other two sides (a and b)

Euclidean (straight-line) distance:

$d_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}$

FIGURE 4.2
Calculation of Euclidean Distance (d_{ij}) from Point i to Point j

TABLE 4.2

Worktable for Calculating Central Point

Distance matrix between points in figure 4.1

	A	B	C	D	E	F	G	Total Distance
A	0	2.59	1.93	1.68	1.55	2.56	2.90	13.21
B	2.59	0	1.96	3.33	3.82	3.86	3.31	18.87
C	1.93	1.96	0	1.58	2.34	1.92	1.41	11.14
D	1.68	3.33	1.58	0	0.91	0.89	1.58	9.97
E	1.55	3.82	2.34	0.91	0	1.58	2.47	12.67
F	2.56	3.86	1.92	0.89	1.58	0	1.14	11.95
G	2.90	3.31	1.41	1.58	2.47	1.14	0	12.81

Point "D" is the lowest total distance from all other points: 9.97

Euclidean distance from point A to point B:

$$d_{AB} = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2}$$

$$= \sqrt{(2.8 - 1.6)^2 + (1.5 - 3.8)^2}$$

$$= 2.59$$

Locational coordinates: A (2.8, 1.5), B (1.6, 3.8)

Euclidean (straight-line) distance

- Total distance from all other points is lowest

Center Point

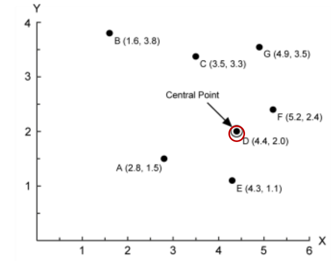


FIGURE 4.1
Graph of Locational Coordinates and Central Point

Salisbury UNIVERSITY

Mean Center

- mean center – average location of a set of points
  - Center of gravity of point pattern (spatial distribution)
  - average X, Y values
  - equal weights

$$\overline{X}_c = \frac{\sum X_i}{n} \text{ and } \overline{Y}_c = \frac{\sum Y_i}{n}$$

where:

$\overline{X}_c$ = mean center of X

$\overline{Y}_c$ = mean center of Y

X_i = X coordinate of point i

Y_i = Y coordinate of point i

n = number of points in the distribution

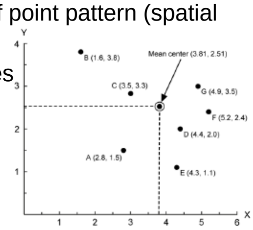


FIGURE 4.1
Graph of Locational Coordinates and Mean Center

Mean Center

- Outliers….
  - add point (15, 13)
  - Average location but…

TABLE 4.2

Worktable for Calculating Mean Center

Point	Locational coordinates *	
	X_i	Y_i
A	2.8	1.5
B	1.6	3.8
C	3.5	3.3
D	4.4	2.0
E	4.3	1.1
F	5.2	2.4
G	4.9	3.5

$n = 7$ $\sum X_i = 26.7$ $\sum Y_i = 17.6$

$\bar{X}_c = \frac{\sum X_i}{n} = \frac{26.7}{7} = 3.81$ $\bar{Y}_c = \frac{\sum Y_i}{n} = \frac{17.6}{7} = 2.51$

Mean center coordinates: (3.81, 2.51) *

* See figure 4.1 for graph of locational coordinates and mean center.

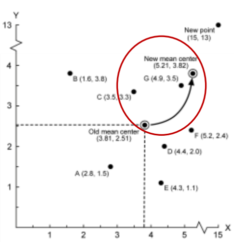


FIGURE 4.4
How an Outlier Might Affect Mean Center Location

Salisbury UNIVERSITY

Mean Center

- geographic "center of population" – point where a rigid map of the country would balance if equal weights (i.e., location of each person) were situated over it

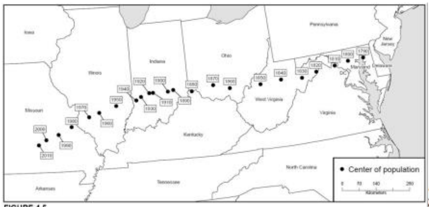


FIGURE 4.5
Geographic Center of U.S. Population, 1790 to 2010
Source: United States Bureau of the Census, 2010

Weighted mean center

- Unequal weights applied to points
  - Ex. retail store volume, city populations, etc.
  - Weights analogous to frequencies

$$\bar{X}_{wc} = \frac{\Sigma f_i X_i}{\Sigma f_i} \text{ and } \bar{Y}_{wc} = \frac{\Sigma f_i Y_i}{\Sigma f_i}$$

where $\bar{X}_{wc}$ = weighted mean center of X
$\bar{Y}_{wc}$ = weighted mean center of Y
f_i = frequency (weight) of point i

Weighted mean center

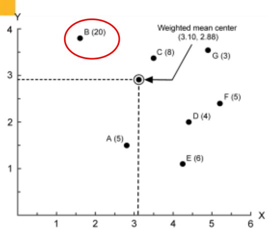


FIGURE 4.6
Graph of Point Locations, Weights (in Parentheses) and Weighted Mean Center

TABLE 4.4

Worktable for Calculating Weighted Mean Center

Point	Locational coordinates X_i	Locational coordinates Y_i	Weight f_i	Weighted coordinates $f_i X_i$	Weighted coordinates $f_i Y_i$
A	2.8	1.5	5	14.0	7.5
B	1.6	3.8	20	32.0	76.0
C	3.5	3.3	8	28.0	26.4
D	4.4	2.0	4	17.6	8.0
E	4.3	1.1	6	25.8	6.6
F	5.2	2.4	5	26.0	12.0
G	4.9	3.5	3	14.7	10.5

$n = 7$ $\quad \sum f_i = 51$ $\quad \sum f_i X_i = 158.1$ $\quad \sum f_i Y_i = 147.0$

$$\bar{X}_{wc} = \frac{\Sigma f_i X_i}{\Sigma f_i} = \frac{158.1}{51} = 3.10$$

$$\bar{Y}_{wc} = \frac{\Sigma f_i Y_i}{\Sigma f_i} = \frac{147.0}{51} = 2.88$$

Weighted mean center coordinates: (3.10, 2.88)

Spatial measures of dispersion

- standard distance – measures the amount of absolute dispersion in a point distribution
  - spatial equivalent to standard deviation
  - calculate Euclidean distance from each point to mean center

$$S_D = \sqrt{\frac{\Sigma (X_i - \overline{X}_c)^2 + \Sigma (Y_i - \overline{Y}_c)^2}{n}}$$

$$S_D = \sqrt{\left(\frac{\Sigma X_i^2}{n} - \overline{X}_c^2\right) + \left(\frac{\Sigma Y_i^2}{n} - \overline{Y}_c^2\right)}$$

Standard distance

TABLE 4.5

Worktable for Calculating Standard Distance

Point	X_i (Locational coordinates)	Y_i (Locational coordinates)	X_i^2	Y_i^2
A	2.8	1.5	7.84	2.25
B	1.6	3.8	2.56	14.44
C	3.5	3.3	12.25	10.89
D	4.4	2.0	19.36	4.00
E	4.3	1.1	18.49	1.21
F	5.2	2.4	27.04	5.76
G	4.9	3.5	24.01	12.25

$\overline{X}_c = 3.81$ $\overline{Y}_c = 2.51$ $\overline{X}_c^2 = 14.52$ $\overline{Y}_c^2 = 6.30$

$n = 7$ $\Sigma X_i^2 = 111.50$ $\Sigma Y_i^2 = 50.80$

$$S_D = \sqrt{\left(\frac{\Sigma X_i^2}{n} - \overline{X}_c^2\right) + \left(\frac{\Sigma Y_i^2}{n} - \overline{Y}_c^2\right)}$$

$$= \sqrt{\left(\frac{111.50}{7} - 14.52\right) + \left(\frac{50.80}{7} - 6.30\right)}$$

$$= 1.54$$

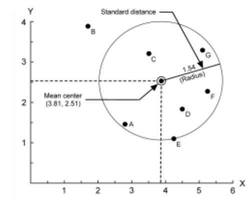


FIGURE 4.10
Graph of Point Locations, Mean Center, and Standard Distance

Relative Measure….

Weighted standard distance

- Used with weighted mean center

$$S_{WD} = \sqrt{\left(\frac{\Sigma f_i (X_i^2)}{\Sigma f_i} - \overline{X}_{wc}^2\right) + \left(\frac{\Sigma f_i (Y_i^2)}{\Sigma f_i} - \overline{Y}_{wc}^2\right)}$$

$$S_{WD} = \sqrt{\frac{\Sigma f_i (X_i - \overline{X}_{wc})^2 + \Sigma f_i (Y_i - \overline{Y}_{wc})^2}{\Sigma f_i}}$$

  - Difference
    - 1.54 vs. 1.70

TABLE 4.6

Worktable for Calculating Weighted Standard Distance

Point	f_i	X_i	X_i^2	$f_i(X_i)^2$	Y_i	Y_i^2	$f_i(Y_i)^2$
A	5	2.8	7.84	39.20	1.5	2.25	11.25
B	20	1.6	2.56	51.20	3.8	14.44	288.80
C	8	3.5	12.25	98.00	3.3	10.89	87.12
D	4	4.4	19.36	77.44	2.0	4.00	16.00
E	6	4.3	18.49	110.94	1.1	1.21	7.26
F	5	5.2	27.04	135.20	2.4	5.76	28.80
G	3	4.9	24.01	72.03	3.5	12.25	36.75

From earlier calculation of weighted mean center:

$\overline{X}_{wc} = 3.10$ $\overline{Y}_{wc} = 2.88$ $\overline{X}_{wc}^2 = 9.61$ $\overline{Y}_{wc}^2 = 8.29$

$\Sigma f_i = 51$ $\Sigma f_i(X_i)^2 = 584.01$ $\Sigma f_i(Y_i)^2 = 475.98$

$$S_{WD} = \sqrt{\left(\frac{\Sigma f_i(X_i)^2}{\Sigma f_i} - \overline{X}_{WC}^2\right) + \left(\frac{\Sigma f_i(Y_i)^2}{\Sigma f_i} - \overline{Y}_{WC}^2\right)}$$

$$= \sqrt{\left(\frac{584.01}{51} - 9.61\right) + \left(\frac{475.98}{51} - 8.29\right)}$$

$$= 1.70$$

Standard Deviational Ellipse

- Extends standard distance to include orientation of the point pattern
  - Calculated separately for X and Y
    - Average distance points vary from mean center on X and average distance points vary

$$SD_x = \sqrt{\frac{\Sigma\left(X_i - \bar{X}_c\right)^2}{n}} \text{ and } SD_y = \sqrt{\frac{\Sigma\left(Y_i - \bar{Y}_c\right)^2}{n}} \quad \textbf{(4.16)}$$

Salisbury UNIVERSITY

Standard Deviational Ellipse

- cross of dispersion
- trigonometric function – angle of rotation
  - Rotated about mean center to minimize distance between both arms and points

TABLE 4.7

Summary Statistics for Standard Deviational Ellipse: Anti-Shipping Activity off the East African Coast

Year	# of incidents	$\bar{X}_c$	$\bar{Y}_c$	SD_x (miles)	SD_y (miles)	Rotation
2009	114	52.53	-1.00	379.7	604.7	45.4
2010	230	52.06	-1.58	415.9	956.3	56.3

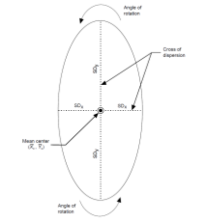
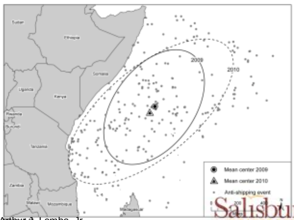


FIGURE 4.11
Standard Deviational Ellipse



Standard Deviational Ellipse: Anti-shipping Activity off the East African Coast, 2009 and 2010

Salisbury UNIVERSITY

Nearest Neighbor Analysis – (NNA)

- Distance of each point to its nearest neighbor is measured and mean distance for all points is determined
  - Objective: describe the pattern of points in a study region and make inferences about the underlying process

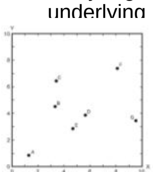


FIGURE 14.1
Location of Points for Nearest Neighbor Problem

Salisbury UNIVERSITY

Nearest Neighbor analysis – (NNA)

- Compare calculated value from point data to theoretical point distributions
  - Outcomes: random, clustered, dispersed
  - average nearest neighbor distance is an absolute index
    - Dependent on distance measure (ex. miles, km, meters, etc.)
    - Minimum = 0 (clustered), maximum is function of point density
  - standardized nearest neighbor index (R) is often used
    - Comparison of data to random

$$R = \frac{\overline{NND}}{\overline{NND}_R} \qquad (14.5)$$

Salisbury UNIVERSITY

Salisbury University

Nearest Neighbor analysis – (NNA)

$$\overline{NND} = \frac{\sum NND}{n} \qquad (14.1)$$

where n = number of points.

$$\overline{NND}_R = \frac{1}{2\sqrt{Density}} \qquad (14.2)$$

where $\overline{NND}_R$ = mean nearest neighbor distance in a random pattern

$Density$ = number of points (n)/Area

$$\overline{NND}_R = \frac{1}{2\sqrt{.07}} = 1.89$$

$$R = \frac{\overline{NND}}{\overline{NND}_R} \qquad (14.5)$$

TABLE 14.1

Coordinates and Nearest Neighbor Information for Example*

Point	X	Y	NN	NND
A	1.3	0.9	E	3.94
B	3.2	4.4	C	2.00
C	3.3	6.4	B	2.00
D	5.6	3.8	E	1.36
E	4.8	2.7	D	1.36
F	8.1	7.4	G	4.21
G	9.4	3.4	D	3.82
SUM				18.69

where NN = nearest neighbor

NND = nearest neighbor distance

$$\overline{NND} = \frac{\sum NND}{n} = \frac{18.69}{7} = 2.67$$

* See figure 14.1 for graph of point locations.

$$R = \frac{2.67}{1.89} = 1.41$$

Salisbury UNIVERSITY

NNA – R values

- Continuum…
  - Result?
  - Descriptive test.

$$R = \frac{2.67}{1.89} = 1.41$$

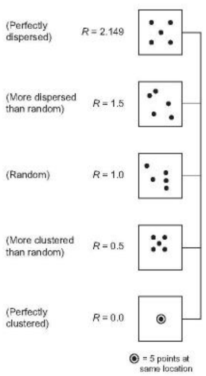
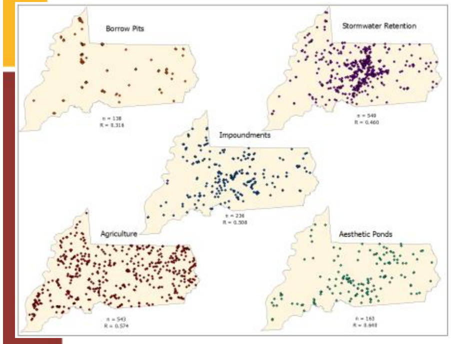


FIGURE 14.2
Continuum of R Values in Nearest Neighbor Analysis



Functional SWB

Results?

Salisbury UNIVERSITY

Nearest neighbor analysis (nna)

- A difference test can be used to determine if the observed nearest neighbor index $(\overline{NNA})$ differs significantly from the theoretical norm $(\overline{NNA}_R)$
  - H_0: There is no difference between our distribution and a random distribution (Poisson)

$$Z_n = \frac{\overline{NND} - \overline{NND}_R}{\sigma_{\overline{NND}}} \quad (14.6)$$

where $\sigma_{\overline{NND}}$ = standard error of the mean nearest neighbor distances

The standard error for the nearest neighbor test can be estimated with the following formula:

$$\sigma_{\overline{NND}} = \frac{.26136}{\sqrt{n(Density)}} \quad (14.7)$$

where: n = number of points $Density$ = number of points (n)/Area

Nearest Neighbor Analysis

Primary Objective: Determine whether a random (Poisson) process has generated a point pattern

Requirements and Assumptions:

1. Random sample of points from a population
2. Sample points are independently selected

Hypotheses:

$H_0 : \overline{NND} = \overline{NND}_R$ (point pattern is random)

$H_A : \overline{NND} \neq \overline{NND}_R$ (point pattern is not random)

$H_A : \overline{NND} > \overline{NND}_R$ (point pattern is more dispersed than random)

$H_A : \overline{NND} < \overline{NND}_R$ (point pattern is more clustered than random)

Test Statistic:

$$Z_n = \frac{\overline{NND} - \overline{NND}_R}{\sigma_{\overline{NND}}}$$

Nearest neighbor analysis (nna) Example: Community Services in Toronto

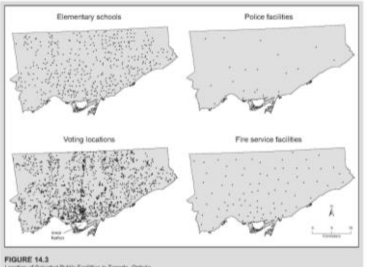


FIGURE 14.3
Location of Selected Public Facilities in Toronto, Ontario
Source: City of Toronto Planning Division

- Emergency services: fire and police
  - Seek dispersion to provide services equally
- Nonemergency services: polling sites and elementary schools
  - Seek clustering…why?

Salisbury UNIVERSITY

Nearest neighbor analysis (nna) Example: Community Services in Toronto

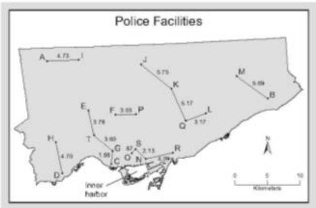


FIGURE 14.4
Nearest Neighbor Distances for Police Service Facilities in Toronto, Ontario

TABLE 14.2

Worktable for Nearest Neighbor Analysis: Police Service Facilities in Toronto, Ontario

$H_0: \overline{NND} = \overline{NND}_R$ (point pattern is random)
$H_A: \overline{NND} \neq \overline{NND}_R$ (point pattern is not random)

Calculate mean nearest neighbor distance:

$$\overline{NND} = \frac{\sum NND}{n} = \frac{72.60}{20} = 3.63$$

where: n = number of points

Calculate random nearest neighbor distance:

$$\overline{NND}_R = \frac{1}{2\sqrt{Density}}$$

where: Density = n/Area

$$\overline{NND}_R = \frac{1}{2\sqrt{\frac{20}{641.45}}} = \frac{1}{2\sqrt{0.0312}} = \frac{1}{0.1766} = 2.83$$

Calculate standardized nearest neighbor distance:

$$R = \frac{\overline{NND}}{\overline{NND}_R} = \frac{3.63}{2.83} = 1.28$$

Calculate test statistic:

$$Z_n = \frac{\overline{NND} - \overline{NND}_R}{\sigma_{\overline{NND}}}$$

where: Density = n/Area

$$\sigma_{\overline{NND}} = \frac{0.26136}{\sqrt{20(0.0312)}} = \frac{0.26136}{0.7899} = 0.3309$$

$$Z_n = \frac{3.63 - 2.83}{0.3309} = \frac{0.80}{0.3309} = 2.42 \quad (p = 0.0078)$$

- Result?

Nearest neighbor analysis (nna) Example: Community Services in Toronto

TABLE 14.3

Nearest Neighbor Values for Selected Public Facilities in Toronto, Ontario

Public Facility	$\overline{NND}$	Density	$\overline{NND}_R$	R	Z_n	p
Fire	2.06	.1278	1.40	1.47	8.21	.0000
Police	3.63	.0312	2.83	1.28	2.42	.0153
Elementary schools	0.59	.7358	0.58	1.02	0.95	.3439
Voting locations	0.27	2.4351	0.32	0.84	−12.23	.0000

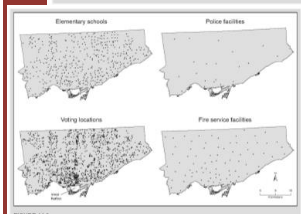


FIGURE 14.3
Location of Selected Public Facilities in Toronto, Ontario

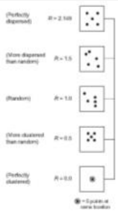


FIGURE 14.2
Continuum of R Values in Nearest Neighbor Analysis

Emergency services – more dispersed

Voting locations – more clustered

Elementary schools - random

Nearest neighbor analysis (nna) Example: Community Services in Toronto

- Issues to consider…
  - Study area boundaries – political boundary or research delimited
    - Doesn't impact NNA distances but does impact area (point density function)
  - Nearest feature – may be outside study area!
    - Problem with using political boundaries
  - More advanced techniques available – *Ripley's K*
    - Evaluates more than one nearest neighbor
    - Can define distances – How many police stations within 1km? 2km?

$$\sigma_{\overline{NND}} = \frac{.26136}{\sqrt{n(Density)}} \qquad \textbf{(14.7)}$$

where: n = number of points

General Issues in Inferential Spatial Statistics

- Geographers are interested in spatial patterns produced by physical or cultural processes
  - Explain patterns of points and areas
    - "global" overall arrangement
      - Random vs. Nonrandom spatial processes
    - "local" concentrations or absences
      - Clusters – points or areas within larger area
        - » Groups of high values – "hot spots"
        - » Groups of low values – "low spots"

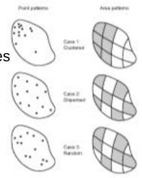


FIGURE 13.1
Types of Point and Area Patterns

Types of Spatial Patterns

- Compare existing pattern to theoretical pattern
- Clustered
  - Density of points varies significantly from one part of study area to another
    - Points: retail locations near highway interchange
    - Areas: registered majority political party affiliation
  - Patterns result from nonrandom factors
    - Accessibility, income, race, etc.

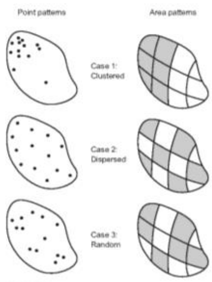


FIGURE 13.1
Types of Point and Area Patterns

Types of Spatial Patterns

- Dispersed
  - Uniformly distributed across study area
    - Suggests systematic spatial process
    - Area example: Central Place Theory – settlements are uniformly distributed across landscape to best serve needs of a dispersed rural population

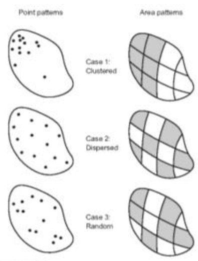


FIGURE 13.1
Types of Point and Area Patterns

Types of Spatial Patterns

- Random
  - No dominant trend toward clustering or dispersion
    - Suggests spatially random process (Poisson)
    - Ex. lightning strikes
- Geographic problems
  - Patterns typically appear as some combination of these three patterns
    - Along continuum…

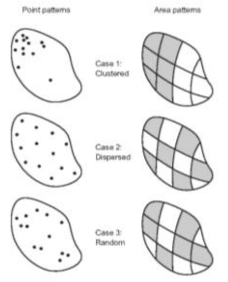


FIGURE 13.1
Types of Point and Area Patterns

Spatial Autocorrelation

- Tobler's Law – "Everything is related to everything else but near things are more related than distant things"
- spatial autocorrelation: measures the degree to which a geographic variable is correlated with itself through space
  - Positive, negative or non-existent
    - Positive spatial autocorrelation: objects near one another tend to be similar
      - Features with high values are near other features with high values, features with medium values are near other features with medium values, etc.
    - Negative spatial autocorrelation: objects near one another tend to have sharply contrasting values
      - Features with high values near features with low values
- Most geographic phenomena exhibit positive spatial autocorrelation
  - Examples: rainfall amounts, home values, etc.

Variogram

- Visualization of spatial autocorrelation
- variogram: scatterplot that display the differences in values between geographic locations against the differences in distances between the geographic locations
  - Y-axis: average *variance* (really half the variance) in *values* for a set of geographic objects
  - X-axis: *distance* between objects
  - Use plot to determine average difference in values at specific distances
    - Ex. 100 miles, 500 miles

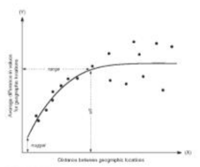


FIGURE 13.2
An Example of a Variogram

Geographic locations near one another tend to have smaller differences than geographic locations at greater distances (positive autocorrelation)!

Variogram

- Displayed as best-fitting curve (function)
  - Differences in values with distance noted and then diminishes
    - range - distance at which the difference in values are no longer correlated
    - sill – average difference in value where there is no relationship between location and value
    - nugget – degree of uncertainty when measuring values for geographic locations that are very close to one another
      - Effect of sampling, measurement error, etc.
      - Unlikely that two samples near each other will have the exact same value

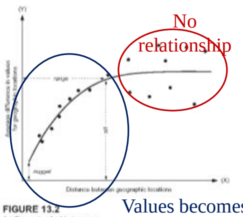
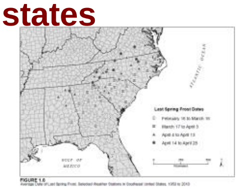


FIGURE 13.2
An Example of a Variogram

Variogram Example: Last Spring Frost IN SE United states

- Two nearby stations, LSF dates should be similar
  - 0 to 400 miles: distances between stations are large, dates are different
  - Beyond 400 miles, no longer spatially autocorrelated…



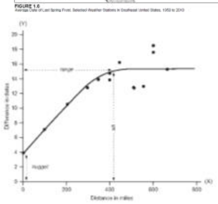


FIGURE 13.3

Spatial Autocorrelation: Importance in Geographic Research

- GIS – push of a button
  - Calculates relationship for any distances…
    - Is the test appropriate for any distance?
- Presence of spatial autocorrelation
  - Inferential statistics assume independent observations
    - Example: last spring frost dates are spatially correlated!
    - Impact: sample locations close together, just like taking the same sample
      - Sample size impacts size of standard error
        - » Smaller standard error than warranted
      - Standard deviation calculation impacted
        - » Even smaller standard error
- Global or local measurement
  - global – examine a distribution of subset (ex. ethnic group) across entire area (ex. city)
    - One group more clustered, dispersed or random than another
  - local – compares each geographic object (ex. all group members) with its surrounding neighbors
    - Is area (ex. neighborhood) more clustered, dispersed or random than another?

Spatial Autocorrelation: Neighbor Definitions

- Measure of interaction between geographic features
  - Defining neighbor…
    - adjacency – share common border
      - Binary: yes or no
        - » Ex. New York and Pennsylvania, New York and California
    - distance threshold – cut-off distance
      - Salisbury, MD – neighbor definition 60 miles…Easton, Wilmington, DE?
    - inverse-distance – strength of "neighborliness" between two objects as a function of distance separating them (1/distance)
        - » New York City and Boston: 1/189 miles or .005,
        - » NYC and LA: 1/2588 miles or .0004
          - » Interaction measure ("neighborliness") is 12 times stronger between NYC and Boston versus NYC and LA
  - In equations/modeling, takes the form of weights
    - w_{ij} : weight between geographic object *i* and *j*
      - Binary: 0 or 1
      - Inverse-distance: continuous value …

Spatial Autocorrelation

- First law of geography: "everything is related to everything else, but near things are more related than distant things" – Waldo Tobler
- Many geographers would say "I don't understand spatial autocorrelation" Actually, they don't understand the mechanics, they do understand the concept.

Spatial Autocorrelation

- Spatial Autocorrelation – correlation of a variable with itself through space.
  - If there is any systematic pattern in the spatial distribution of a variable, it is said to be spatially autocorrelated
  - If nearby or neighboring areas are more alike, this is *positive spatial autocorrelation*
  - *Negative autocorrelation* describes patterns in which neighboring areas are unlike
  - Random patterns exhibit *no spatial autocorrelation*

Why spatial autocorrelation is important

- Most statistics are based on the assumption that the values of observations in each sample are independent of one another
- Positive spatial autocorrelation may violate this, if the samples were taken from nearby areas
- Goals of spatial autocorrelation
  - Measure the strength of spatial autocorrelation in a map
  - test the assumption of independence or randomness

Salisbury UNIVERSITY

Spatial Autocorrelation

- Spatial Autocorrelation is, conceptually as well as empirically, the two-dimensional equivalent of redundancy
- It measures the extent to which the occurrence of an event in an areal unit constrains, or makes more probable, the occurrence of an event in a neighboring areal unit.

Salisbury UNIVERSITY

Spatial Autocorrelation

- Non-spatial independence suggests many statistical tools and inferences are inappropriate.
  - Correlation coefficients or ordinary least squares regressions (OLS) to predict a consequence assumes that the observations have been selected randomly.
  - If the observations, however, are spatially clustered in some way, the estimates obtained from the correlation coefficient or OLS estimator will be biased and overly precise.
  - They are biased because the areas with higher concentration of events will have a greater impact on the model estimate and they will overestimate precision because, since events tend to be concentrated, there are actually fewer number of independent observations than are being assumed.

Salisbury UNIVERSITY

Indices of Spatial Autocorrelation

- Moran's I
- Geary's C
- Ripley's K

Salisbury University

Moran's *I* Index (Global)

- Popular technique for quantifying level of spatial autocorrelation in a set of geographic areas
- Moran's *I* Index takes into account geographic locations (points or areas) as well as attribute values (ordinal or interval/ratio) to determine if areas are clustered, randomly located or dispersed
  - Positive : clustered – nearby locations have similar attribute values
  - Negative: dispersed – nearby locations have dissimilar attribute values
  - Near zero: attribute values are randomly dispersed throughout study area

Salisbury University

Moran's *I* Index (Global)

The general form of Moran's Index for areas is shown in equation 15.4, and the mathematical equivalent, as presented in Ebdon (1988), appears in equation 15.5:

$$I = \frac{\begin{pmatrix}\text{number}\\ \text{of areas}\end{pmatrix}\begin{pmatrix}\text{sum of cross-products}\\ \text{for all contiguous pairs } (i,j)\end{pmatrix}}{\begin{pmatrix}\text{number}\\ \text{of joins}\end{pmatrix}\begin{pmatrix}\text{variance of the}\\ \text{area attribute values}\end{pmatrix}} \quad (15.4)$$

$$I = \frac{n\sum(x_i - \bar{x})(x_j - \bar{x})}{J\sum(x - \bar{x})^2} \quad (15.5)$$

where:

n = the number of areas
x = an area attribute value
$\bar{x}$ = the mean of all area attribute values
x_i and x_j = the values of contiguous pairs

$\sum(x_i - \bar{x})(x_j - \bar{x})$ = the sum of all contiguous pairs
J = the number of joins
$\sum(x - \bar{x})^2$ = the variance of the attribute values.

Weighted cross-products: deviation values for contiguous pairs multiplied together and summed

- Positive: neighboring areas with similar attribute values either large or small (clustered)
  - Larger deviation from mean, greater magnitude
- Negative: neighboring areas with dissimilar attribute values contiguous (dispersed)
  - Larger deviation from mean, greater magnitude
- Near zero: random…
- *I* ranges from -1.00 to 1.00

Salisbury University

Moran's *I* Index (Global): Significance test

The null hypothesis is that the area values are arranged in a completely random spatial pattern, and that spatial autocorrelation is not present within the study area. The Z score is computed as:

$$Z = \frac{I - E_I}{\sqrt{Var_I}} \qquad (15.6)$$

where: $E_I = \frac{-1}{(n-1)}$

Similar to the free and non-free sampling techniques mentioned with join count analysis and binary data, the variance for Moran's *I* may be calculated under the assumption of normality or randomization. *Normality* is similar to free sampling in that the observed values of a variable are taken randomly from a normally distributed population, whereas *randomization* is similar to non-free sampling in that the values are taken directly from the data set. The variance under the *normality* assumption is computed as:

$$Var_I = \frac{n^2J + 3J^2 - n\sum L^2}{J^2(n^2-1)}$$

where: n = the number of objects
J = the number of joins, and
$\sum L$ = the sum of the number of joins for each individual area

H_0: No spatial autocorrelation in the data (Values of areas are completely random)
H_A: Spatial autocorrelation in the data (Values of areas are not completely random)

- If p-value is not significant, then you should not reject the null hypothesis
  - The observed pattern is not different from complete spatial randomness
- p-value significant and *Z-score* ...itive ...stering
- ...ant and *Z-score* ...tive

Variance under the *randomization* assumption is computed as:

$$Var_I = \frac{n[J(n^2+3-3n)+3J^2-n\sum L^2]-k[J(n^2-n)+6J^2-2n\sum L^2]}{J^2(n-1)(n-2)(n-3)}$$

where: k represents the kurtosis for the variable x

Salisbury University

TABLE 15.6

Worktable for Moran's *I*: Five-area Example Pattern

H_0: no spatial autocorrelation in the data

H_A: spatial autocorrelation exists

Joins			Area values		
Area	L	L^2	x	$(x-\bar{x})$	$(x-\bar{x})^2$
A	2	4	70	25	625
B	3	9	80	35	1,225
C	3	9	20	−25	625
D	2	4	10	−35	1,225
Joins(J) = 5		$\sum L^2$ = 26	$\bar{x} = \frac{180}{4} = 45$		$\sum(x-\bar{x})^2$ = 3,700

Join number	x_i	$(x_i-\bar{x})$	x_j	$(x_j-\bar{x})$	$(x_i-\bar{x})(x_j-\bar{x})$
1	70	25	80	35	875
2	70	25	20	−25	−625
3	20	−25	80	35	−875
4	80	35	10	−35	−1,225
5	10	−35	20	−25	875
SUM					−975

$$E_I = \frac{-1}{(n-1)} = -\frac{1}{4-1} = -0.33$$

$$I = \frac{n\sum(x_i-\bar{x})(x_j-\bar{x})}{J\sum(x-\bar{x})^2} = \frac{4(-975)}{5(3{,}700)} = -0.21$$

$$Var_I = \frac{n^2J + 3J^2 - n\sum L^2}{J^2(n^2-1)} = \frac{(4^2\times 5)+(3\times 5^2)-(4\times 26)}{5^2(4^2-1)} = \frac{51}{375} = 0.136$$

$$Z = \frac{I-E_I}{\sqrt{Var_I}} = \frac{-0.21-(-0.33)}{\sqrt{0.136}} = 0.326$$

p = 0.74

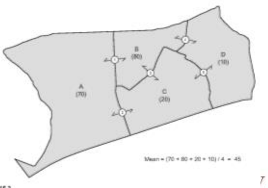


FIGURE 15.3
Areas and Attribute Values for Moran's *I*

Result?

© Arthu...
Salisbury University

Example: Cleveland Census Block Groups

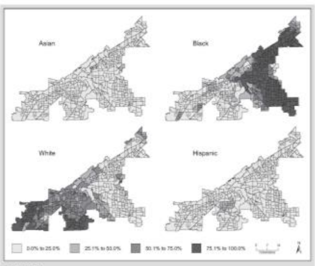


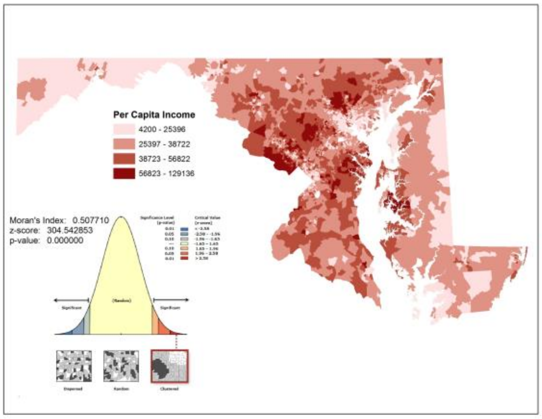
FIGURE 15.4
Distribution of Various Racial and Ethnic Groups by Census Block Group in Cleveland, Ohio

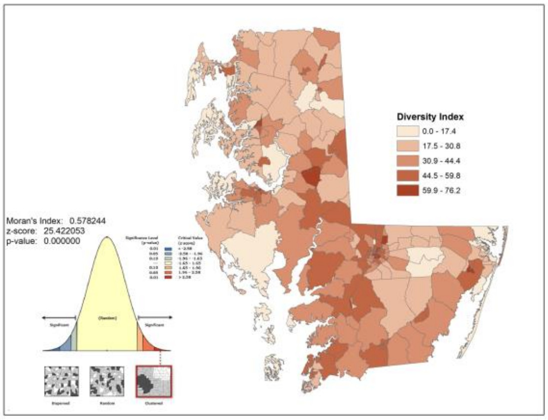
TABLE 15.7

Moran's *I* Values and Test Statistics for Selected Racial and Ethnic Groups in Cleveland, Ohio Census Block Groups

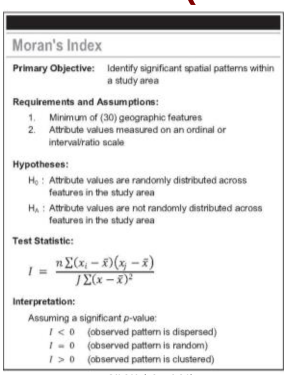
Group	Moran's Index	Z	p-value
Asian	0.35	18.39	0.000
Black	0.58	64.6	0.000
White	0.64	65.4	0.000
Hispanic	0.81	65.7	0.000

Salisbury University





Moran's *I* Index (Global)



Moran's Index

Primary Objective: Identify significant spatial patterns within a study area

Requirements and Assumptions:

1. Minimum of (30) geographic features
2. Attribute values measured on an ordinal or interval/ratio scale

Hypotheses:

H_0 : Attribute values are randomly distributed across features in the study area

H_A : Attribute values are not randomly distributed across features in the study area

Test Statistic:

$$I = \frac{n \sum (x_i - \bar{x})(x_j - \bar{x})}{J \sum (x - \bar{x})^2}$$

Interpretation:

Assuming a significant *p*-value:

$I < 0$ (observed pattern is dispersed)

$I = 0$ (observed pattern is random)

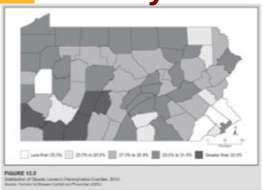
$I > 0$ (observed pattern is clustered)

Moran's I Index (local)

- Global spatial autocorrelation (Moran's *I*) may indicate a lack of spatial autocorrelation
  - Local pockets may exist– hotspots
  - LISA – Local Indicators of Spatial Association
    - Quantify similarity of each geographic observation with an identified group of geographic neighbors
      - Identifies local clusters – geographic locations where adjacent or nearby areas have similar values
      - Spatial outliers – geographic locations that are different from adjacent or nearby areas
    - Each geographic area receives individual measure

Salisbury UNIVERSITY

Moran's I Index (local): Example: Obesity in PA



$$I = \frac{(x_i - \bar{X})}{S_i^2} \sum_{j=1, j \neq i} w_{ij}(x_j - \bar{X}) \quad (15.7)$$

$$S_i^2 = \sum_{j=1, j \neq i} \frac{(x_j - \bar{X})^2}{n-1} - \bar{X}^2$$

where: x_i = the value for a particular geographic entity
x_j = the value for the neighboring geographic entity
$\bar{X}$ = the average of all attributes
w_{ij} = the spatial weights.

The overall computation of the local Moran index is beyond the scope of this book, but a short example will help illustrate its usefulness.

- Global Moran's *I* = .69, *p-value* = .25
- Local Moran's *I* for each county…

Positive values: similar levels in adjacent counties (clustering)

- Philly…
- Johnstown/Altoona

Negative values: dissimilar values – outlier

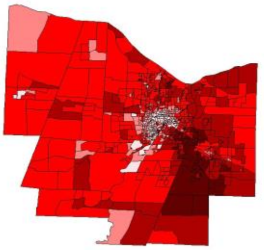
- Fayette County

Salisbury UNIVERSITY

Example of Moran's I – Per Capita Income in Monroe County

Using Polygons:
Morans I: .66
P: < .001

Using Points:
I: .12
Z: 65



UNIVERSITY

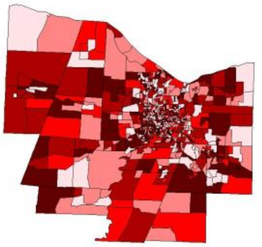
Example of Moran's I – Random Variable

Using Polygons:

Moran's I: .012
p: .515

Using Points:

Moran's I: .0091
Z: 1.36



Salisbury University