

GIS Workshop

Statistics in Space

Chieko Maene
c-maene@northwestern.edu
 Maps and State Documents Librarian
 Government and Geographic Information and Data Services
 University Library (Evanston)
 Northwestern University

Statistics in Space

Focus: Three Approaches

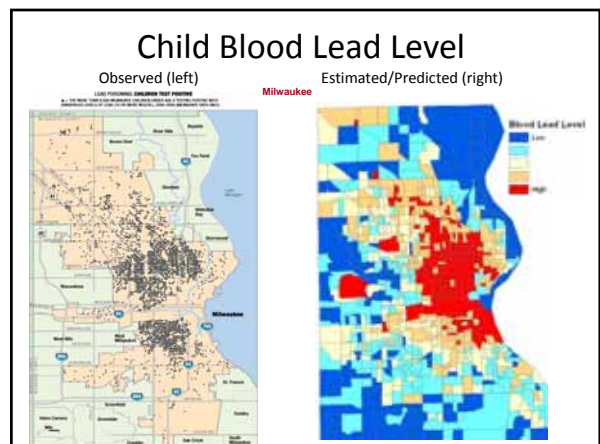
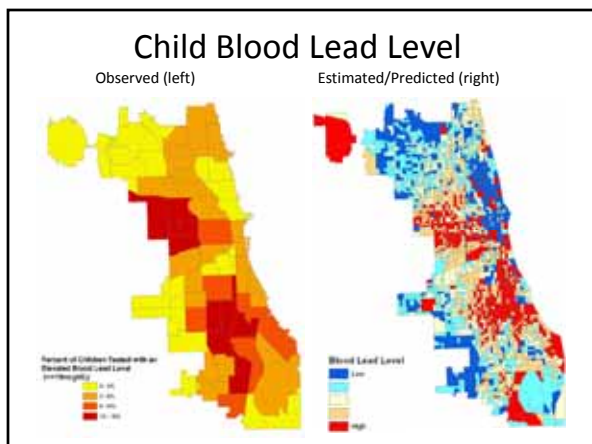
1. Use of **ecological analysis**
 - In short, visualization of statistical values calculated in statistical packages (or in ArcGIS) for geographic entities (i.e. counties, tracts, etc.)
 - ArcGIS is required only for visualization
2. Use of **spatial statistics (or tools)**
 - There are many space-related statistical measurements and methods, but our topic is limited to tools available in ArcGIS
 - Definition (in ArcGIS)
 - Software-based tools, methods and techniques for describing and modeling spatial distribution, patterns, processes and relationships
3. Use of **proximity/distance** as an additional variable

Ecological Regression Analysis

Steps
 (Note: be aware of potential "ecological fallacy" problem – we are dealing with spatially aggregated entities, **not** individuals)

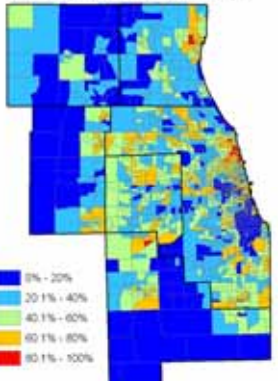
1. Create a regression model using smaller geographic entity as a unit of analysis (i.e. census tract, blocks)
2. Calculate predicted values for all the entities (i.e. tracts)
3. Visualize the values in GIS

- Ex: Miranda, Dolinoy & Overstreet (2002) *Mapping for Prevention: GIS Models for Directing Childhood Lead Poisoning Prevention Programs*. *Env. Health Perspectives* 110(9), pp.947-953
 - Investigated correlations among several variables
 - Case - children with blood lead levels (BLL) ≥ 10 mg/dL, Single parent, Renter-occupied, Minority population (Hispanic & Afro-Americans), Poverty, Income, House year built
 - Refined the model by eliminated insignificant factors
 - Multivariate regression model
 - Estimated BLL = 10 (varies by area) + (-0.0044 x year built) + (-4.42 x 0.000010 x median income) + (0.002 x percent African American)



Another Ecological Analysis example

Diversity Index

$$E_i = \frac{\sum_{k=1}^5 [P_{ik} * \log(P_{ik})]}{-1.609}$$


Ecological Diversity Index Analysis

- (Based on Shannon's) Diversity Index
 - A measure of how diverse a community is
 - Values between **0** (no diversity, completely homogeneous) and **1** (very diverse)
 - Calculates the probability that two groups picked at random will be of a different race and ethnicity

$$E_i = \frac{\sum_{k=1}^5 [P_{ik} * \log(P_{ik})]}{-1.609}$$

Where:
 E_i = Entropy, or Diversity Index, by Neighborhood of Residence (i)
 P_{ik} = Proportion of Population in Race (k) by Neighborhood of Residence (i)
 -1.609 = constant, based on five racial/ethnic categories.

Ecological Diversity Index Analysis

- How to calculate
 - Sum of ([Proportion (% frequency) of each groups] multiplied by Natural Log ([Proportion (% frequency) of each groups])) divided by a constant (-1.609 for 5 categories)

Example

Groups	Proportion	(Prop) LN (Prop)	(Even Prop) LN (Even Prop)
White	0.235	-0.3403	-0.3219
Afro-American	0.351	-0.3675	-0.3219
Asian	0.156	-0.2898	-0.3219
Hispanic	0.219	-0.3326	-0.3219
Other	0.039	-0.1265	-0.3219
Total	1	-1.4567	-1.609

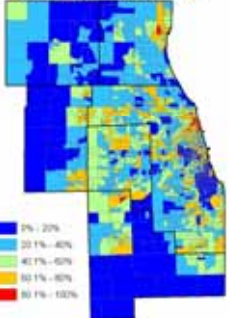
$\frac{-1.4567}{-1.609} = 0.905$

Exercise 19

Diversity Index - Racial Groups

```

((([NH_WHITE] / [TOT_POP]) * Log ( ([NH_WHITE] / [TOT_POP] ) ) )
+ ([NH_BLACK] / [TOT_POP]) * Log ( ([NH_BLACK] / [TOT_POP] ) ) )
+ ([NH_ASIAN] / [TOT_POP]) * Log ( ([NH_ASIAN] / [TOT_POP] ) ) )
+ ([NH_OTHER] / [TOT_POP]) * Log ( ([NH_OTHER] / [TOT_POP] ) ) )
+ ([HISPANIC] / [TOT_POP]) * Log ( ([HISPANIC] / [TOT_POP] ) ) )
) / ([PLNP_W] + [PLNP_B] + [PLNP_A] + [PLNP_O] + [PLNP_H]) / -1.609
    
```

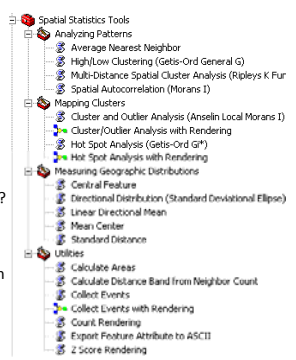


Spatial Statistics

- What is Spatial Statistics?
 - Statistics that incorporate space (area, length, proximity, orientation, and/or spatial relationship) directly into their mathematics
- Why Spatial Statistics?
 - Sometimes visualization isn't enough
 - Ambiguity in color, classes – what are low or high values?
 - Maps can be misleading
 - Give us reasoning and concrete **proofs** or **evidence**, which help us make decisions with confidence! (We can say: it's statistically significant!)

Spatial Statistics Tools

- In ArcGIS
 - Analyzing patterns
 - If clusters (spatial autocorrelation) exist?
 - Mapping clusters
 - Does the clusters has patterns (hot/cold spots)?
 - Measuring geographic distribution
 - Are there any distribution patterns?
 - Utilities



Mean Center

- Where is the center?
 - Simple
 - Weighted
- What feature is the most central?
 - Simple
 - Weighted
- Or track changes in the distribution..



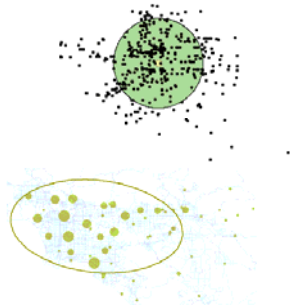
Ex. 20 : Mean Center

- Let's replicate John Snow's famous mapping analysis using GIS
- He found the source of the Cholera outbreak in the mid 19th century by mapping Cholera death cases



Distribution - Distance

- Standard Distance
 - Measures distribution of features around the mean
- Directional Distance
 - Identify spatial trends in the distribution of features



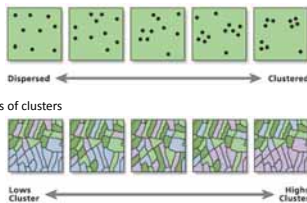
Ex. 20 : Directional Distribution

- Examining different time period of crimes
 - Does the time of crime show different distribution patterns?



Spatial Autocorrelation

- Finding clusters/dispersion using **First of Law of Geography**
 - "Near things are more related than distant things"
- Though visualization can help us identify "clusters", maps can be deceiving – we may want **statistics to prove** if features are indeed clustered, and measure (how much?) the **clusterness!**
- Methods in ArcGIS
 - Global calculations
 - To see patterns or trends
 - Local calculations
 - To see the extent and locations of clusters



Hot Spot Analysis (Getis-Ord Gi*)

- Hot spots are clusters of high values (instances)
 - ... and cold spots are cluster of low values..
- Based on local calculation of spatial autocorrelation
- Local **G-statistic** and its **Z score** tell us whether high values or low values tend to cluster in a study area, and identify where clustering occurs in both high and low values.
 - High Z-Score – hot spot
 - Low Z-Score – cold spot
- Use Z-scores for visualization
 - To visualize statistical significance



Ex. 20 : Hot Spot Analysis

- Do crime clusters exist?
 - Would it be different by type of crimes?

Hot spot analysis works often with aggregated data (such as block/tract, etc) because these tools need a weight, which one-record-per-incident type of data don't have.

More Information

- Extend Crime Analysis with ArcGIS Spatial Statistics Tools http://www.esri.com/news/arcuser/0405/ss_crimestats1of2.html
- ArcInfo Using the Spatial Statistics Tools <http://vid01.esri.com/winmedia/ArcGIS9/SpatialStats.wmv>
- ArcGIS 9.3 Desktop Help, Modeling Spatial Relationships http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Modeling_spatial_relationships
- Andy Mitchell (2005) "The ESRI Guide to GIS Analysis: Volume 2, Spatial Measurements & Statistics" ESRI Press. Redland, CA.
- Luc Anselin (2005) Spatial Statistical Modeling in a GIS Environment in "GIS, Spatial Analysis, and Modeling" ESRI Press. Redland, CA.

Proximity/distance Analysis

- GIS can answer almost (almost) any questions that have something to do with **spatial relationships**
 - Questions about **proximity/distance** are very common
 - Many useful proximity tools available in ArcGIS
 - Some tools require the highest level of licenses (ArcInfo workstation) but there is almost (almost) always another way to get around to do stuff without expensive licenses

Proximity Analysis – Buffer

- How many people live within 1 miles from a station?
 - Buffer
 - Select or Intersect

Proximity Analysis - Distance

- How far is it between two points?
 - There is no special tool in ArcGIS to create direct lines but it is not difficult to draw direct lines
 - All you need are pairs of two points ("origin – destination" type)
 - These are also called **spider diagram!**
 - Try tools from support.ESRI.com

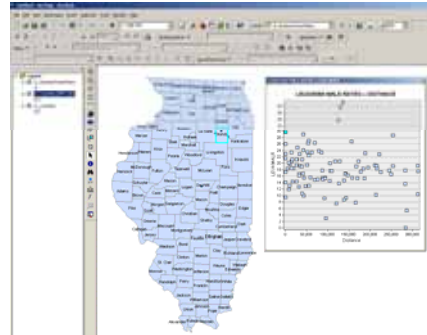
Proximity Analysis - Raster

- Show distance from a road
 - Spatial Analyst Tools > Distance > **Euclidean distance** tool

Exercise 21 : Proximity Analysis

- Use of **distance** as one of **analysis variables**
 - Is there a **relationship between pediatric cancer cases and a distance to a closest nuclear power plant?**
- **Note:**
- This is a very very **simplified** and not a good example with limited information resources
 - Cancer registry (SEER) exists but only serious researchers can access such detailed data
 - It has been clinically proved there was no such relationship..

Exercise 21: Proximity



Exercise 22 Introduction to Geodatabase, ModelBuilder and Python Script

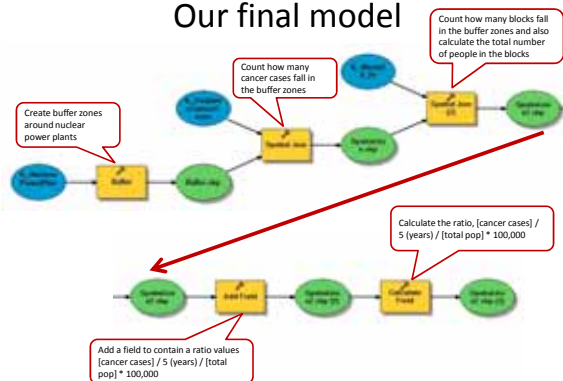
Disclaimer
This exercise is **NOT about statistics!** But this rather **focuses on the tools!**

Modeling

- Models are sets of tasks, how you automate your work, which let you modify, execute multiple times to improve the results
- ModelBuilder is an ArcGIS application to chain together tools using the output of one tools as the input to another tool

http://training.esri.com/acb2000/showdetl.cfm?DID=6&Product_ID=844

Our final model



Python Scripts

- Python
 - Another way of automating work flows
 - Object-oriented programming language
 - Let's access ArcObjects (tools, functions) and data without opening ArcMap or ArcCatalog
- To edit, debug & test, use PythonWin
 - PythonWin: open source Python IDE for Windows – download from SourceForge, or find an executable (.exe) from ArcGIS installation CD.
- Documentation & examples:
 - ArcGIS 9.3 Help: http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=About_getting_started_with_writing_geoprocessing_scripts
 - The library has a PDF document, Writing_Geoprocessing_Scripts.pdf – a bit old (2005) and said to include examples that don't work..

http://training.esri.com/acb2000/showdetl.cfm?DID=6&Product_ID=815