Aug 23-25, 2001
Fudan University

Integrating Spatial Attribute Data and CHGIS for Spatial Analysis

Shuming Bao
sbao@umich.edu

China Data Center
University of Michigan

Topics

- Introduction
- Spatial Data Process
- Spatial Analysis
- Applications
- Tools for spatial analysis
- Research Issues

Introduc

- Some background about Chi (CITAS) and China Data Cer
  - CITAS project
  - Robert Hartware's CHGIS
  - **The missions of CDC**
- New opportunities provided by CHGIS project for scholars from different disciplinarians
- New challenges
  - Theories
  - Methodologies
  - Tools (stand alone and online tools)

– integrate historical, social and natural science data into a geographic information system (GIS)
– support research in the human and natural components of local, regional and global change
– promote quantitative research on China studies
– promote collaborative research in spatial studies
– promote the use of data on China in teaching
– promote data sharing

Types of Spatial Data

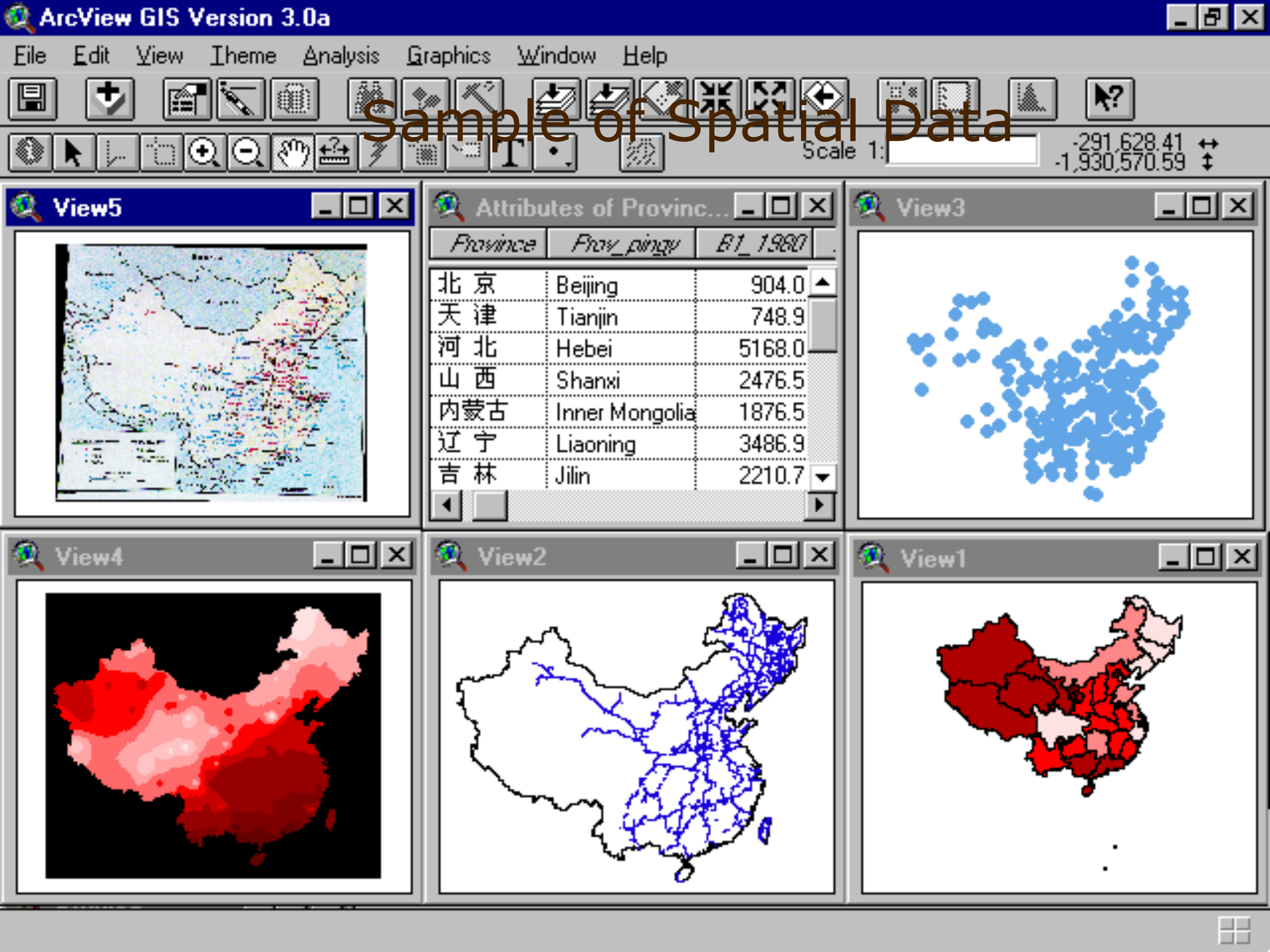
Types of Spatial Data:

- Geospatial data
  - Polygons
  - Points
  - Lines
  - Images/Grid
- Socioeconomic data
  - County/Province statistics
  - Census data
  - Social surveys

Spatial Data Sources:

- Geographic data (polygons, points and lines)
- Arc/Info data
- Shape files (*.shp, *.shx, and *.dbf)
- Grid
- Image data (ERDAS Image, JPEG, TIFF, BMP and Arc/Info Image)
- Tabular data (dBASE, INFO and TEXT)
- SQL
- SDE (Spatial Data Engine)

Sample of Spatial Data



Elevation and Major Cities of China

The Integration of HGIS data with other data

Geographical data
- River
- Roads
- Elevation

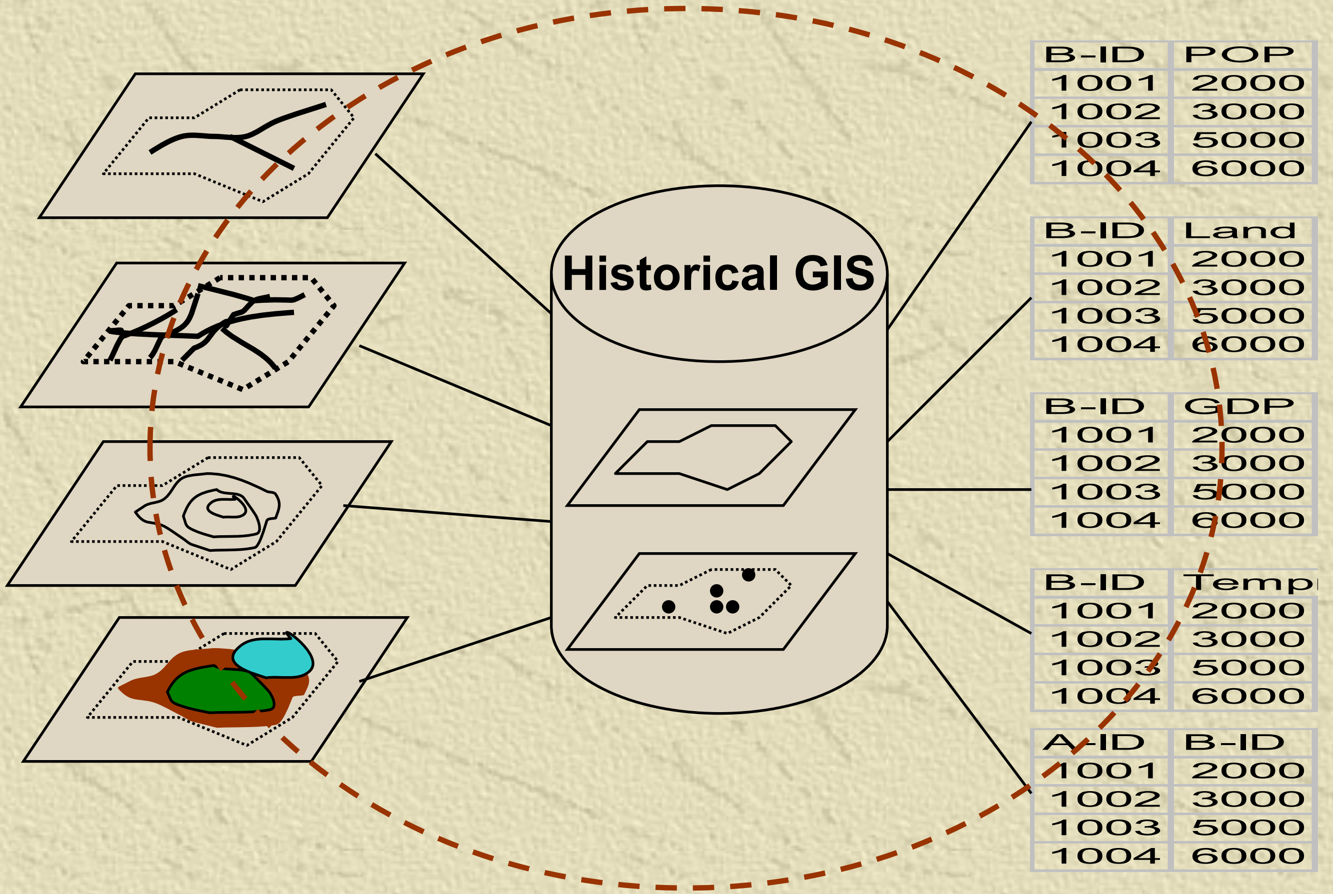
Local attributes
- Climate
- Culture
- Education
- Languages
- Agriculture
- Business

Historical GIS
- Boundaries
- Settlements

Remotely sensed data
- Images
- Grid

Statistical data
- Socioeconomic data
- Survey data
- Census data

The Integration of HGIS data with other data (b)



B-ID	POP
1001	2000
1002	3000
1003	5000
1004	6000

B-ID	Land
1001	2000
1002	3000
1003	5000
1004	6000

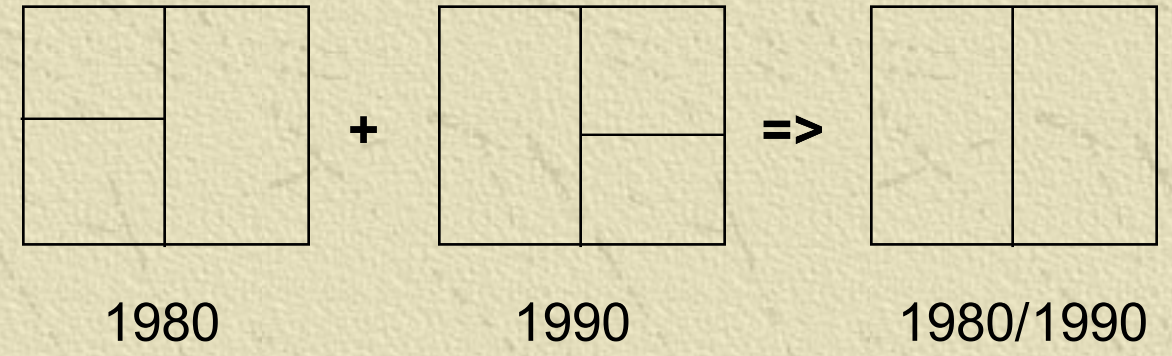
B-ID	GDP
1001	2000
1002	3000
1003	5000
1004	6000

B-ID	Temp
1001	2000
1002	3000
1003	5000
1004	6000

A-ID	B-ID
1001	2000
1002	3000
1003	5000
1004	6000

Integration of Data: Spatial Data Process

Space-Time Information
=> Comparable base map

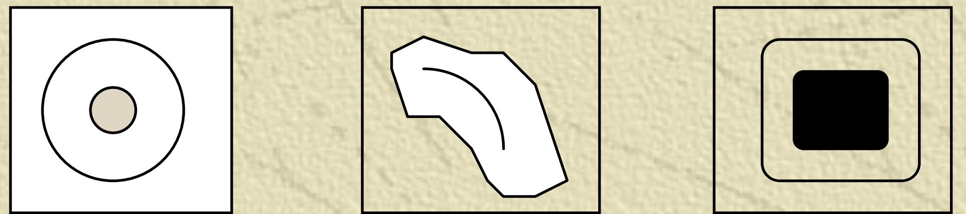


Multilayers Information
=> Joint table

B-ID	POP	LAND	WATER
1001	2000	U	30
1002	3000	U	20
1003	5000	R	40
1004	6000	R	10

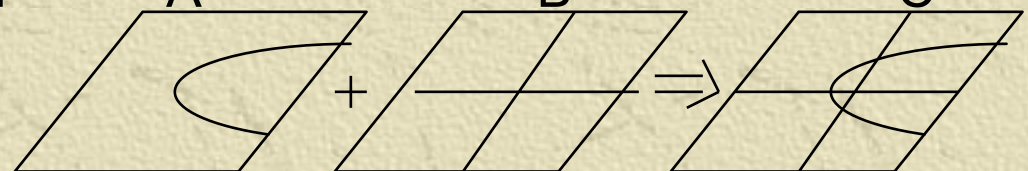
Integration of Data: Spatial Operations

Buffer:



A B C

Overlay:



Join:

A-ID	U/R
10	R
20	U

+

B-ID	POP
1001	2000
1002	3000
1003	5000
1004	6000

⇒

C-ID	B-ID	POP	A-ID	U/R
1	1001	2000	10	U
2	1001	2000	20	R
3	1002	3000	10	U
4	1002	3000	20	R
5	1003	5000	10	U
6	1003	5000	20	R
7	1004	6000	10	U
8	1004	6000	20	R

Questions

- Is there any spatial cluster over space?
- Are spatial observations distributed randomly over space?
- Are spatial observations correlated ?
- Is there any spatial outlier?
- Is there any spatial trend?
- What is the interaction (statistically and theoretically) between different factors?
- How to predict an unknown spatial value at a specific location ?

Why Spatial is Special ?

- Why spatial data is different from non-spatial data ? (spatial neighborhood)
- Statistical property for spatial data:
  - Spatial dependence (autocorrelation)
  - Heterogeneity
  - Spatial trend (non-stationarity)
- Sensitive to spatial boundaries and spatial unit (Country, County, Tract) Lat / Long grid

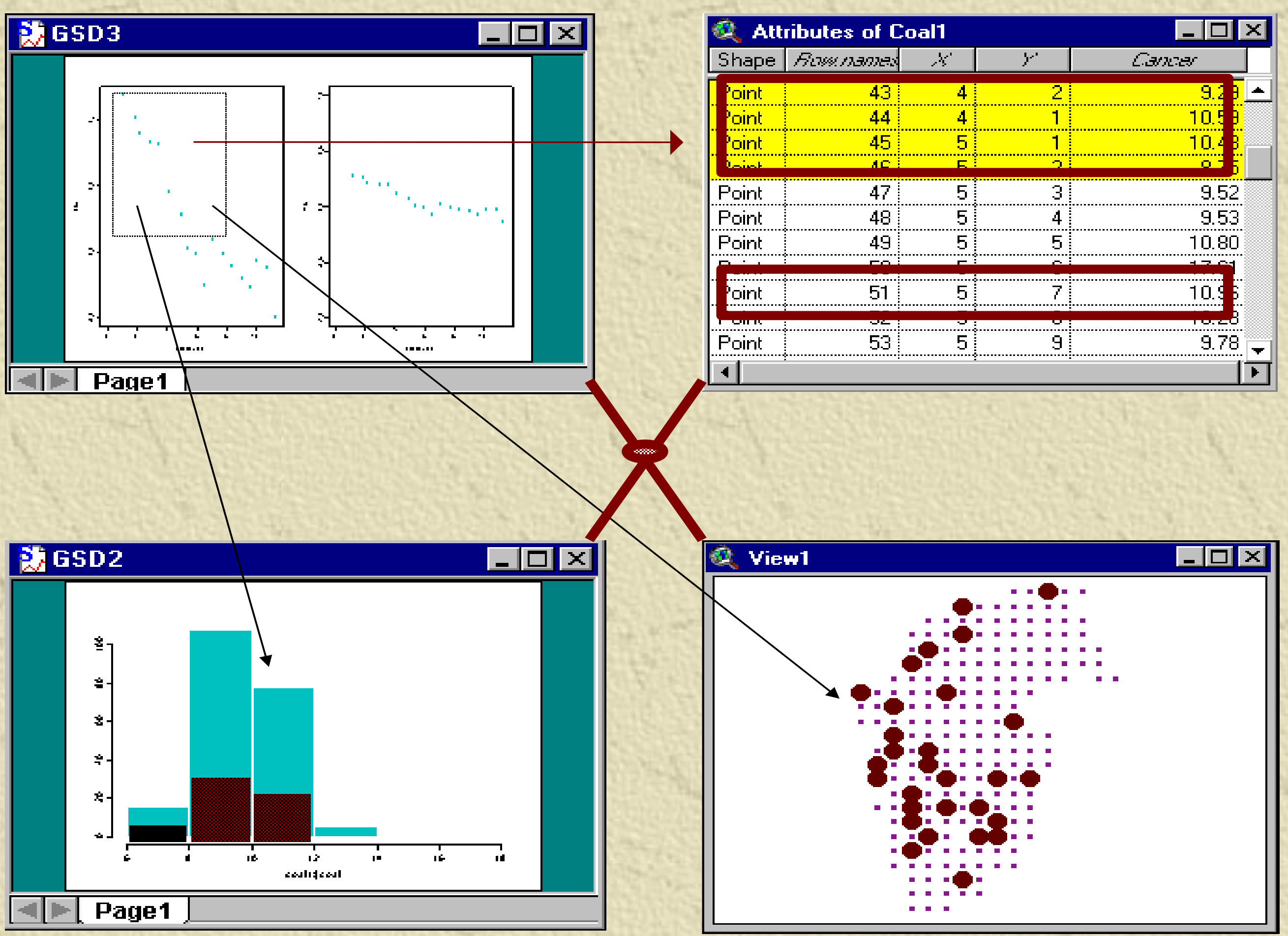
Spatial Analysis

- Tests on spatial patterns:
  - Tests on spatial non-stationarity
  - Tests on spatial autocorrelation
  - Tests on Spatial stationarity and non-stationarity

- Data-driven approaches (Exploratory Spatial Data Analysis)
  - Global Statistics
  - Local statistics

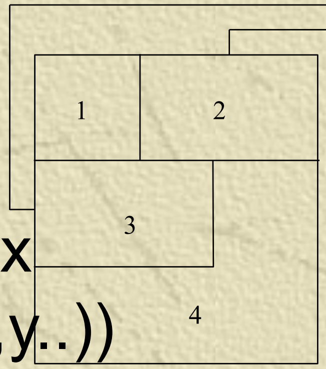
- Model-driven approaches
  - Spatial linear and non-linear models
  - Space-temporal models

Visualization of Spatial Data



Defining Spatial Linkage

- **Criteria:** theoretical and empirical
  - •Accessibility (roads, rivers, railways, airlines and Internet)
  - •Economic linkage (commuter flows, migrations, trade flows)
  - •Social linkage (college admission, language)
  - •Locational linkage (neighborhood, geographical distance)
- **Methodology:**
  - •Binary matrix
  - •Row standardized matrix
  - •Weight function (wij=f(x,y..))



ROW.ID	COL.ID	WEIGHTA	WEIGHTB
1	2	1	0.5
1	3	1	0.5
2	1	1	0.33
2	3	1	0.33
2	4	1	0.33
3	1	1	0.33
3	2	1	0.33
3	4	1	0.33
4	2	1	0.5
4	3	1	0.5

Defining Spatial Weight Matrices

Adjacency criterion:

$$w_{ij} = \begin{cases} 1 & \text{if location j is adjacent to i,} \\ 0 & \text{if location j is not adjacent to i.} \end{cases}$$

Distance criterion:

$$w_{ij}(d) = \begin{cases} 1 & \text{if location j is within distance d from i,} \\ 0 & \text{otherwise.} \end{cases}$$

A general spatial distance weight matrices:

$$w_{ij}(\mathrm{d}) = d_{ij}^{-a} \cdot \beta^{b}$$

Identifying Spatial Outliers

- Mapping
- Table analysis
- Exploratory spatial data analysis
- Statistical analysis

Identifying Spatial Trend

Theoretical Variogram: $$\gamma(h) = \frac{1}{2} E[(Z(x) - Z(x'))^2]$$

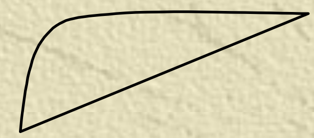
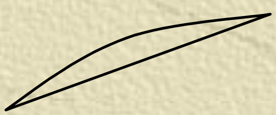
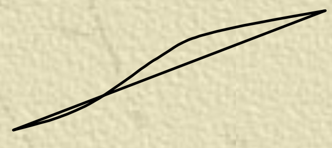
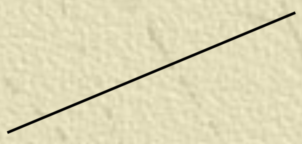
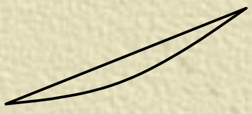
Experimental Variogram: $$\hat{\gamma}(h_k) = \frac{1}{2|N(h_k)|} \sum_{i=1}^{Nk} [z(x_i) - z(x_i{}')]^2$$

$$h_k{}^l \leq || x_i - x_i{}' || < h_k{}^u, \; h_k = \frac{1}{N_k} \sum_{i=1}^{Nk} || x_i - x_i{}' || \qquad h_k = \frac{1}{2} | h_k{}^u - h_k{}^l |$$

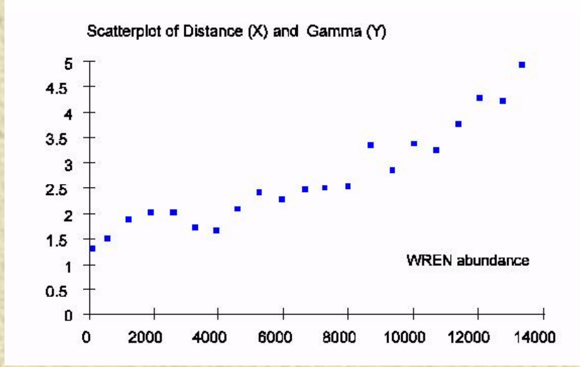
where $N(h_k)=\{(i,j): x_i - x_{i'} = h\}$, $|N(h_k)|$ is the number of distinct elements of $N(h_k)$.

Theoretical Variogram Models & Empirical Variogram

Theoretical variogram:

• 1) Exponenti • 2) Gaussian • 3) Spherical • 4) Linear • 5) Power	 a. Exponential	 b. Gaussian
 c. Spherical	 d. Linear	 e. Power

Empirical variogram:

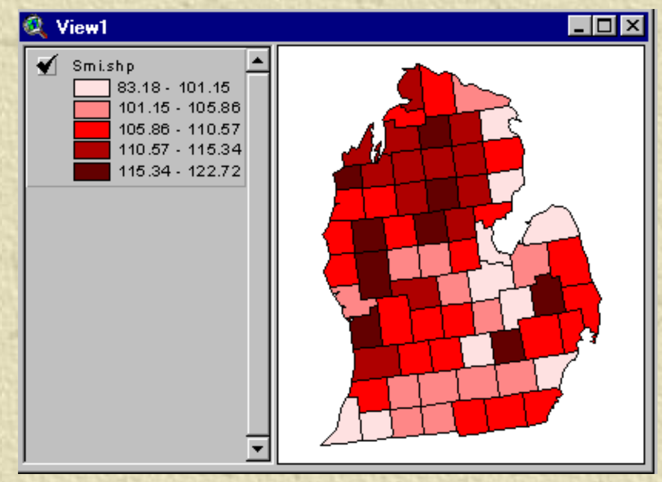


Identifying Global Pattern of Spatial Distribution

Moran I:

$$I(d) = \sum_{i}^{n} \sum_{j}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x}) \Big/ (S^2 \sum_{i}^{n} \sum_{j}^{n} w_{ij})$$

$$S^2 = \frac{1}{n} \sum_{i}^{n} (x_i - \bar{x})^2 \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$



Geary C:

$$C(d) = (n-1) \Big/ (2 \sum_{i}^{n} \sum_{j}^{n} w_{ij}) \{\sum_{i}^{n} \sum_{j}^{n} w_{ij}(x_i - x_j)^2 \Big/ \sum_{i}^{n} (x_i - \bar{x})^2\}$$

Moran I (Z value) is

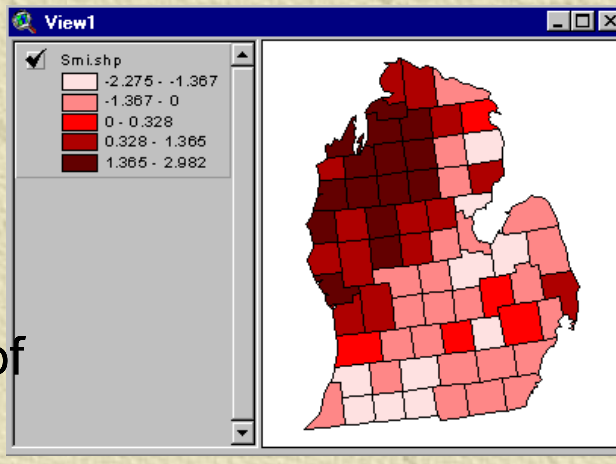
- positive: observations tend to be similar;
- negative: observations tend to be dissimilar;
- approximately zero: observations are arranged randomly over space.

Geary C:

- large C value (>>1): observations tend to be dissimilar;
- small C value (<<1) indicates that they tend to be similar.

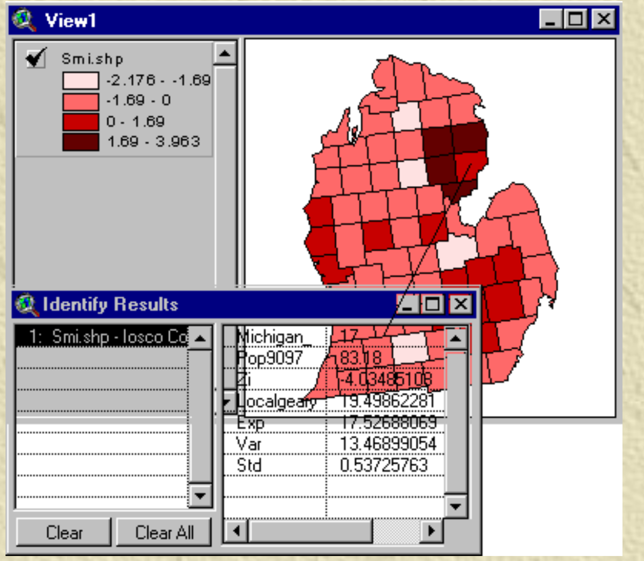
Identifying Local Patterns of Spatial Distribution

Local Moran: $I_i(d) = \sum_{j \neq i}^{n} w_{ij} Z_j$



- significant and negative if location i is associated with relatively low values in surrounding locations;
- significant and positive if location i is associated with relatively high values of the surrounding locations.

Local Geary: $C_i(d) = \sum_{j \neq i}^{n} w_{ij} (Z_i - Z_j)^2$



- significant and small Local Geary (t<0) suggests a positive spatial association (similarity);
- significant and large Local Geary (t>0) suggests a negative spatial association (dissimilarity).

Identifying Factors for Spatial Changes

- Spatially autoregressive model
- Spatial moving average model
- Semi-parametric model
- Kriging

A Simple Spatial Autoregressive Model

$Y = \rho WY + \varepsilon$

where y is an observed variable over space D: $\{Y(s_i): s_{i \in D,\ i=1?n}\}$,
W is a spatial weight matrix (*nxn*),
ρ is the spatial autoregressive parameter, and $\varepsilon \sim N(0, \sigma^2)$.

OLS estimates are biased and inconsistent:

$$\hat{\rho} = \left[(Wy)'(Wy)\right]^{-1}(Wy)'y = \rho + \left[(Wy)'(Wy)\right]^{-1}(Wy)'\varepsilon$$

$$E(\hat{\rho}) \neq \rho$$

A General Form of Spatial Process Model

$$y = \rho W_1 y + X\beta + \varepsilon$$

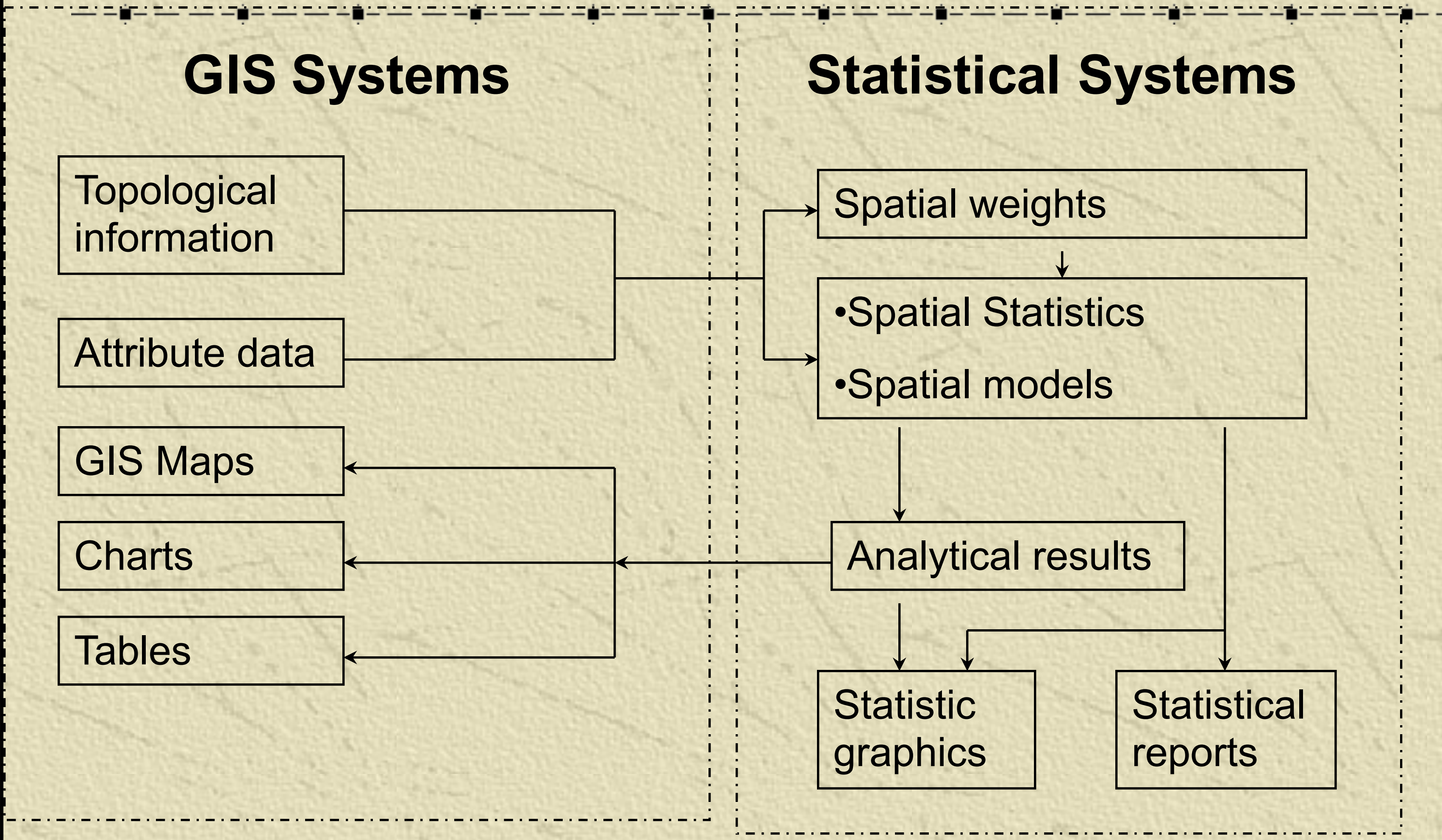
$$\varepsilon = \lambda W_2 \varepsilon + \mu$$

where W_1 and W_2 are spatial weight matrices, $\mu \sim N(0,\Omega)$.

Applications

- Historical studies
- Socioeconomic development
- Environment
- Religion
- Anthropology studies
- Population studies
- Minority studies
- ….

Integration of Spatial Analysis with HGIS



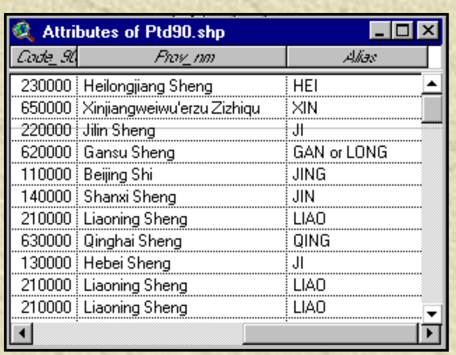
S-PLUS for ArcView GIS
http://www.mathsoft.com

- An enhanced version of S language specially for exploratory data analysis and statistics.
- An integrated suite for data manipulation, data analysis and graphical display.
- An interpreted language, in which individual language expressions are read and then immediately executed.
- Object-oriented programming(method, class, and object).
- ***S+SpatialStats*** for geostatistical data, polygon data and point data (2000+ analytical functions).

S-PLUS for ArcView

China Data

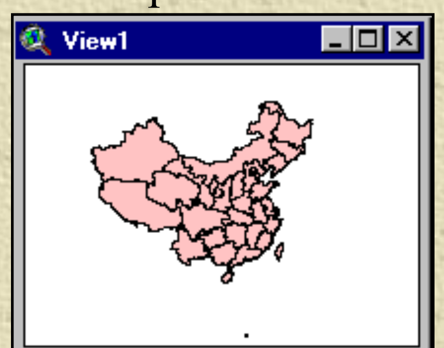
Attribute data:



Attributes of Ptd90.shp

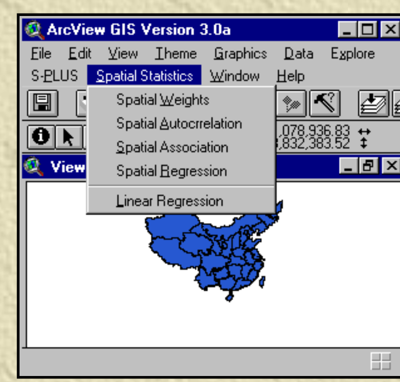
Code_90	Prov_nm	Alias
230000	Heilongjiang Sheng	HEI
650000	Xinjiangweiwu'erzu Zizhiqu	XIN
220000	Jilin Sheng	JI
620000	Gansu Sheng	GAN or LONG
110000	Beijing Shi	JING
140000	Shanxi Sheng	JIN
210000	Liaoning Sheng	LIAO
630000	Qinghai Sheng	QING
130000	Hebei Sheng	JI
210000	Liaoning Sheng	LIAO
210000	Liaoning Sheng	LIAO

GIS map data:

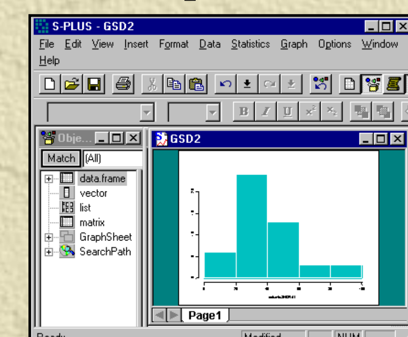


Application Interface

ArcView GIS



S-PLUS/SpatialStats



Maps

Analysis

Reports

Statistical Graphics

Research Issues

- ❖ Spatial data process (missing data, fuzzy data, large volume of data, space-time data structure, references)
- ❖Spatial data sharing and management (Metadata, GIS data, attribute data; distributed centers; update, search, online analysis)
- ❖ Integration of CHGIS with natural and social information
- ❖ Development of new methodology and tools for spatial data analysis (sampling, survey, clustering, autocorrelation, association, modeling, simulation, web tools)
- ❖ Applications of GIS, database, and new technology in historical and other studies