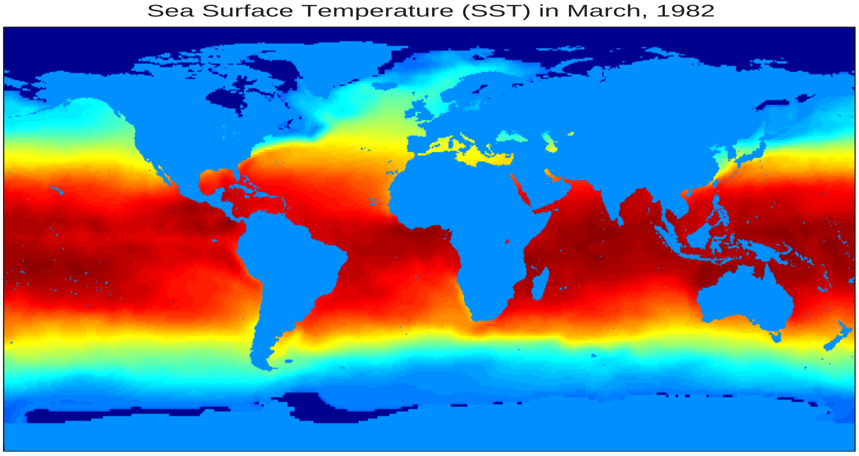
Spatial Data Mining: Accomplishments and Research Needs

Shashi Shekhar

Department of Computer Science and Engineering

University of Minnesota

Sea Surface Temperature (SST) in March, 1982



Why Data Mining?

- Holy Grail – Informed Decision Making
- Lots of Data are Being Collected
  - Business Applications:
    - Transactions: retail, bank ATM, air travel, etc
    - Web logs, e-commerce, GPS-track
  - Scientific Applications:
    - Remote sensing: e.g., NASA's Earth Observing System
    - Sky survey
    - Microarrays generating gene expression data
- Challenges:
  - Volume (data) $\gg$ number of human analysts
  - Some automation needed
- Data Mining may help!
  - Provide better and custmized insights for business
  - Help scientists for hypothesis generation

Spatial Data

⋆ Location-based Services

- Ex: MapQuest, Yahoo Maps, Google Maps, MapPoint

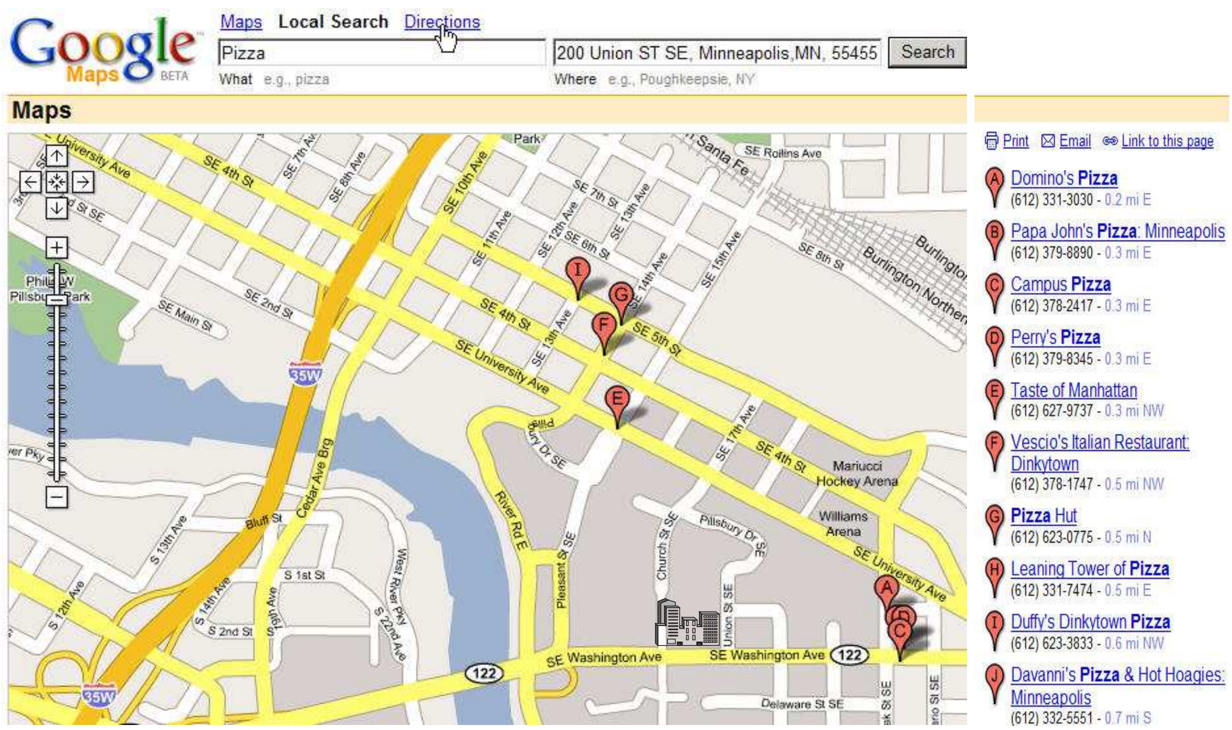


Figure 1: Google Local Search (http://maps.google.com)

⋆ In-car Navigation Device

Figure 2: Emerson In-Car Navigation System (In Coutesy of Amazon.com)

Spatial Data Mining (SDM)

⋆ The process of discovering

- interesting, useful, non-trivial patterns
  - patterns: non-specialist
  - exception to patterns: specialist
- from large spatial datasets

⋆ Spatial patterns

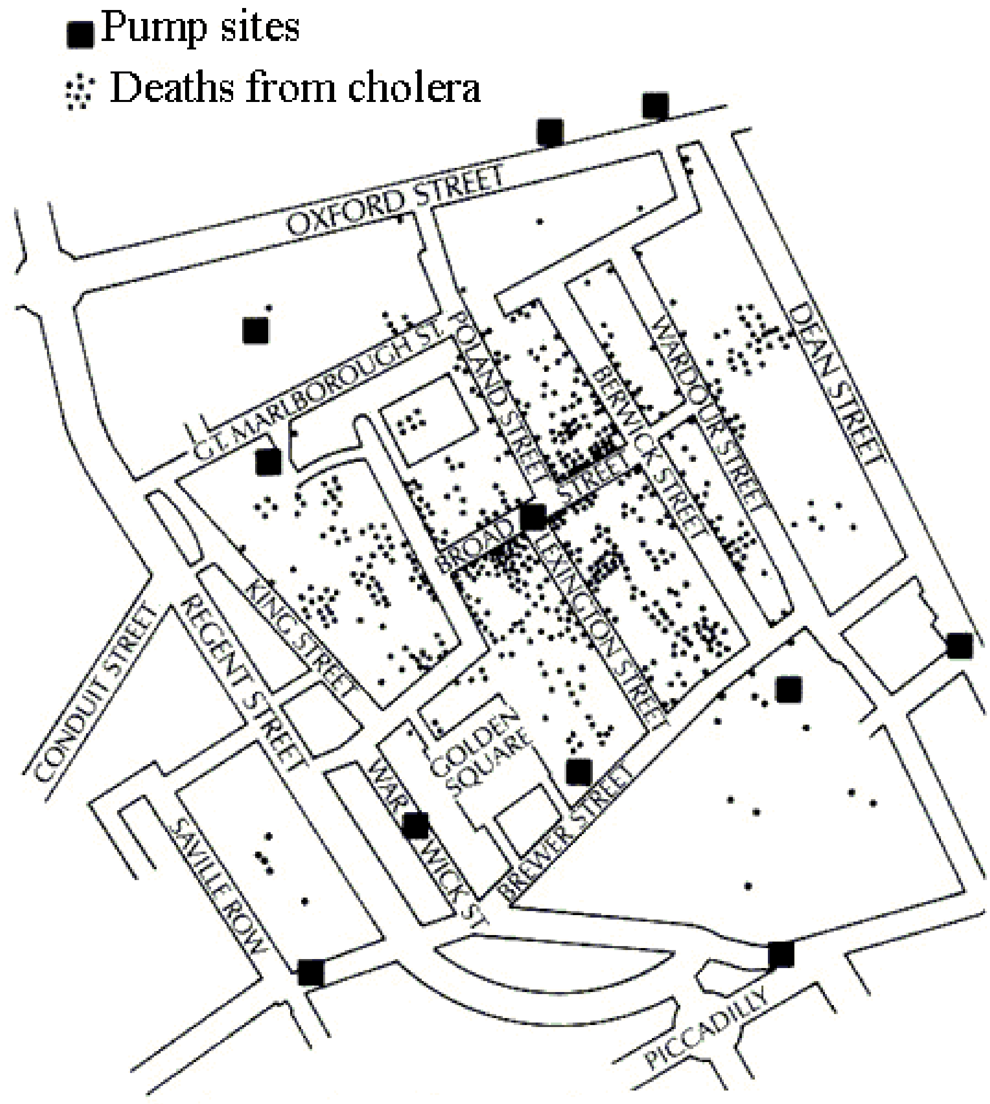
- Spatial outlier, discontinuities
  - bad traffic sensors on highways (DOT)
- Location prediction models
  - model to identify habitat of endangered species
- Spatial clusters
  - crime hot-spots (NIJ), cancer clusters (CDC)
- Co-location patterns
  - predator-prey species, symbiosis
  - Dental health and fluoride

Location As Attribute

- ⋆ Location as attribute in spatial data mining
- ⋆ What value is location as an explanatory variable?
  - most events are associated with space and time
  - surrogate variable
  - critical to data analyses for many application domains
    - physical science
    - social science
- ⋆ Location helps bring rich contexts
  - Physical: e.g., rainfall, temperature, and wind
  - Demographical: e.g., age group, gender, and income type
  - Problem-specific
- ⋆ Location helps bring relationships
  - e.g., distance to open water

Example Spatial Pattern: Spatial Cluster

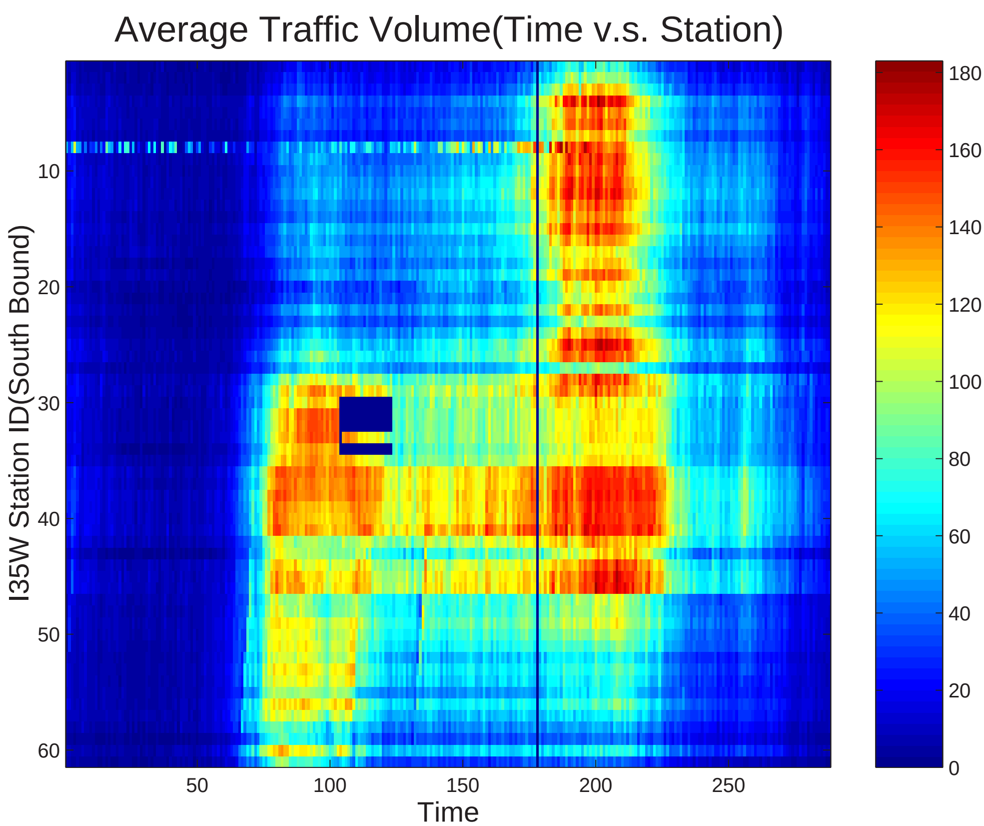
⋆ The 1854 Asiatic Cholera in London



Example Spatial Pattern: Spatial Outliers

⋆ Spatial Outliers

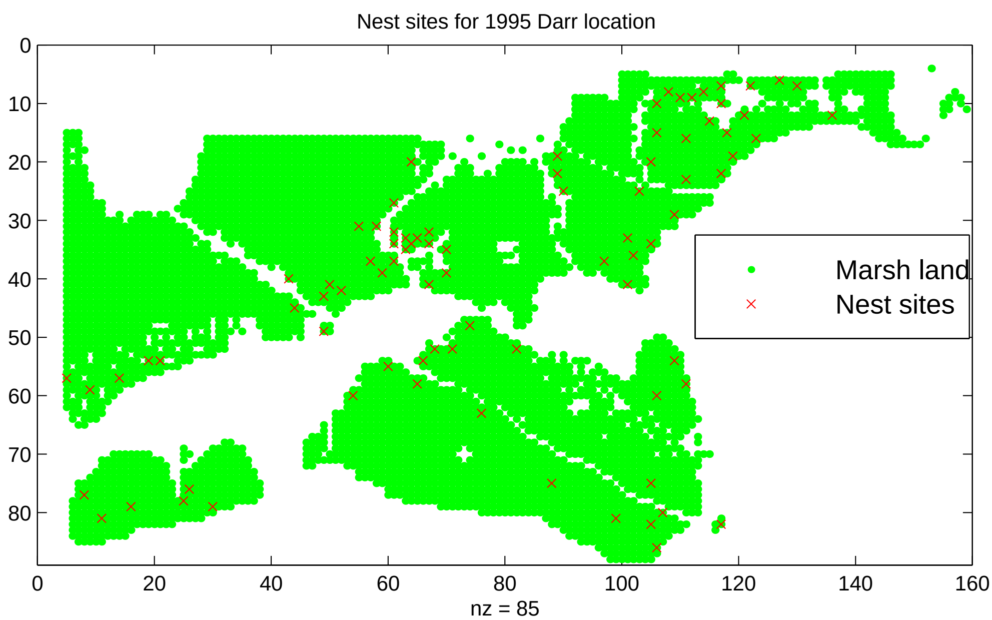
- Traffic Data in Twin Cities
- Abnormal Sensor Detections
- Spatial and Temporal Outliers



Example Spatial Pattern: Predictive Models

⋆ Location Prediction: Bird Habitat Prediction

- Given training data
- Predictive model building
- Predict new data

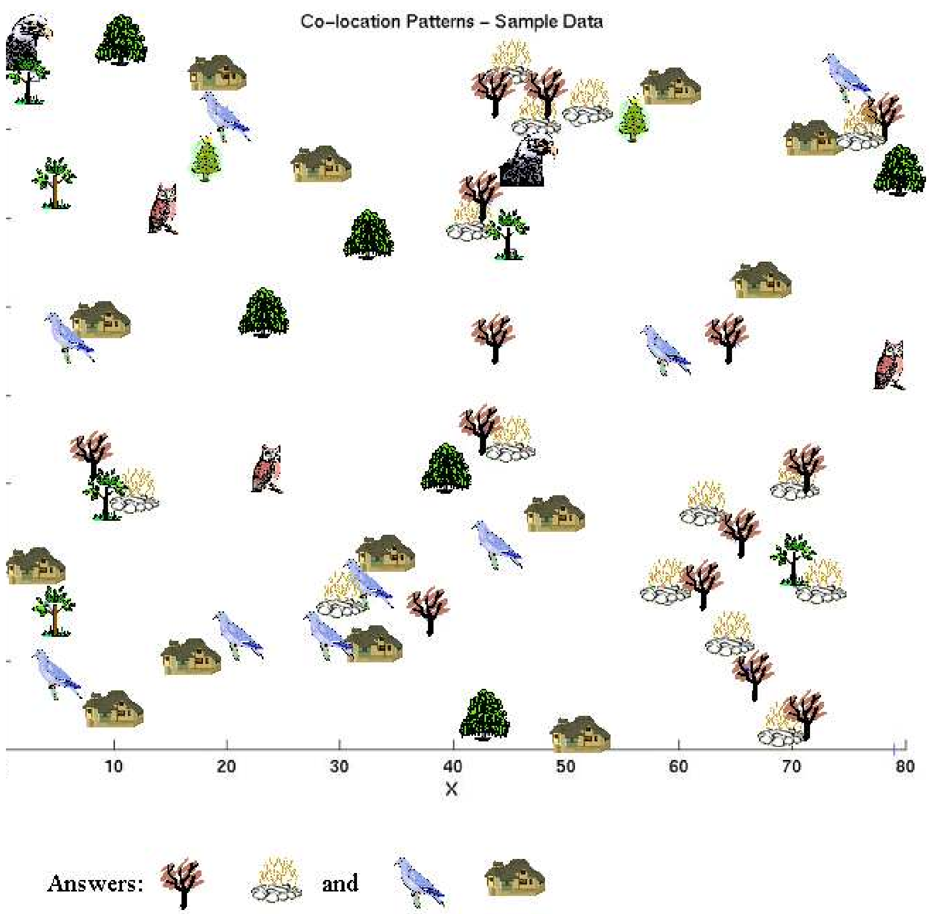


Example Spatial Pattern: Co-locations (backup)

⋆ Given:

- A collection of different types of spatial events

⋆ Illustration



⋆ Find: Co-located subsets of event types

What's NOT Spatial Data Mining

⋆ Simple Querying of Spatial Data

- Find neighbors of Canada given names and boundaries of all countries
- Find shortest path from Boston to Houston in a freeway map
- Search space is not large (not exponential)

⋆ Testing a hypothesis via a primary data analysis

- Ex. Female chimpanzee territories are smaller than male territories
- Search space is not large !
- SDM: secondary data analysis to generate multiple plausible hypotheses

⋆ Uninteresting or obvious patterns in spatial data

- Heavy rainfall in Minneapolis is correlated with heavy rainfall in St. Paul, Given that the two cities are 10 miles apart.
- Common knowledge: Nearby places have similar rainfall

⋆ Mining of non-spatial data

- Diaper sales and beer sales are correlated in evening

Application Domains

⋆ Spatial data mining is used in

- NASA Earth Observing System (EOS): Earth science data
- National Inst. of Justice: crime mapping
- Census Bureau, Dept. of Commerce: census data
- Dept. of Transportation (DOT): traffic data
- National Inst. of Health(NIH): cancer clusters
- Commerce, e.g. Retail Analysis

⋆ Sample Global Questions from Earth Science

- How is the global Earth system changing?
- What are the primary forcings of the Earth system?
- How does the Earth system respond to natural and human-included changes?
- What are the consequences of changes in the Earth system for human civilization?
- How well can we predict future changes in the Earth system

Example of Application Domains

⋆ Sample Local Questions from Epidemiology[TerraSeer]

- What's overall pattern of colorectal cancer?
- Is there clustering of high colorectal cancer incidence anywhere in the study area?
- Where is colorectal cancer risk significantly elevated?
- Where are zones of rapid change in colorectal cancer incidence?

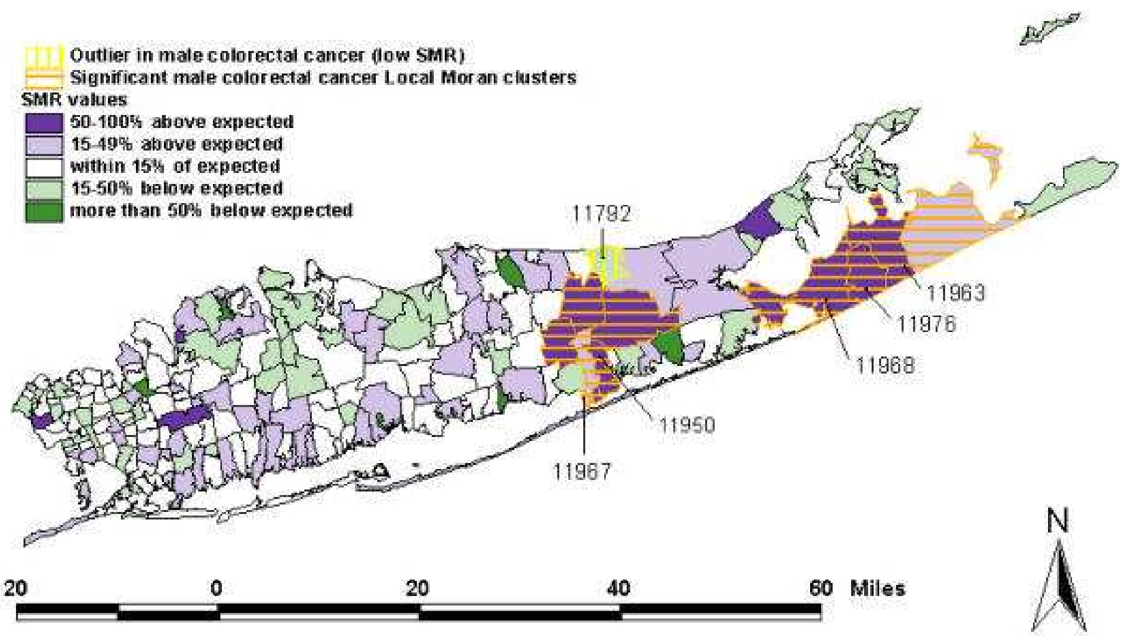


Figure 3: Geographic distribution of male colorectal cancer in Long Island, New York(in courtesy of TerraSeer)

Business Applications

⋆ Sample Questions:

- What happens if a new store is added?
- How much business a new store will divert from existing stores
- Other "what if" questions:
  - changes in population, ethic-mix, and transportation network
  - changes in retail space of a store
  - changes in choices and communication with customers

⋆ Retail analysis: Huff model [Huff, 1963]

- A spatial interaction model
  - Given a person p and a set S of choices
  - Pr[person p selects choice c] $\propto$ perceived_utility($\forall c \in S$, p)
  - perceived_utility(store c, person p) = f (square-footage(c), distance (c, p), parameters)
- Connection to SDM
  - Parameter estimation, e.g., via regression
- For example:
  - Predicting consumer spatial behaviors
  - Delineating trade areas
  - Locating retail and service facilities
  - Analyzing market performance

Map Construction

⋆ Sample Questions

- Which features are anomalous?
- Which layers are related?
- How can the gaps be filled?

⋆ Korea Data

- Latitude 37deg15min to 37deg30min
- Longitude 128deg23min51sec to 128deg23min52sec

⋆ Layers

- Obstacles (Cut, embankment, depression)
- Surface drainage (Canal, river/stream, island, common open water, ford, dam)
- Slope
- Soils (Poorly graded gravel, clayey sand, organic silt,disturbed soil)
- Vegetation (Land subject to inundation, cropland, rice field, evergreen trees, mixed trees)
- Transport (Roads, cart tracks, railways)

Colocation in Example Data

- ⋆ Road: river/stream
- ⋆ Crop land/rice fields: ends of roads/cart roads
- ⋆ Obstacles, dams and islands: river/streams
- ⋆ Embankment obstacles and river/stream: clayey soils
- ⋆ Rice, cropland, evergreen trees and deciduous trees :river/stream
- ⋆ Rice: clayey soil, wet soil and terraced fields
- ⋆ Crooked roads: steep slope

Colocation Example

⋆ Interestingness

- Patterns to Non-Specialist vs. Exceptions to Specialist

⋆ Road-River/Stream Colocation

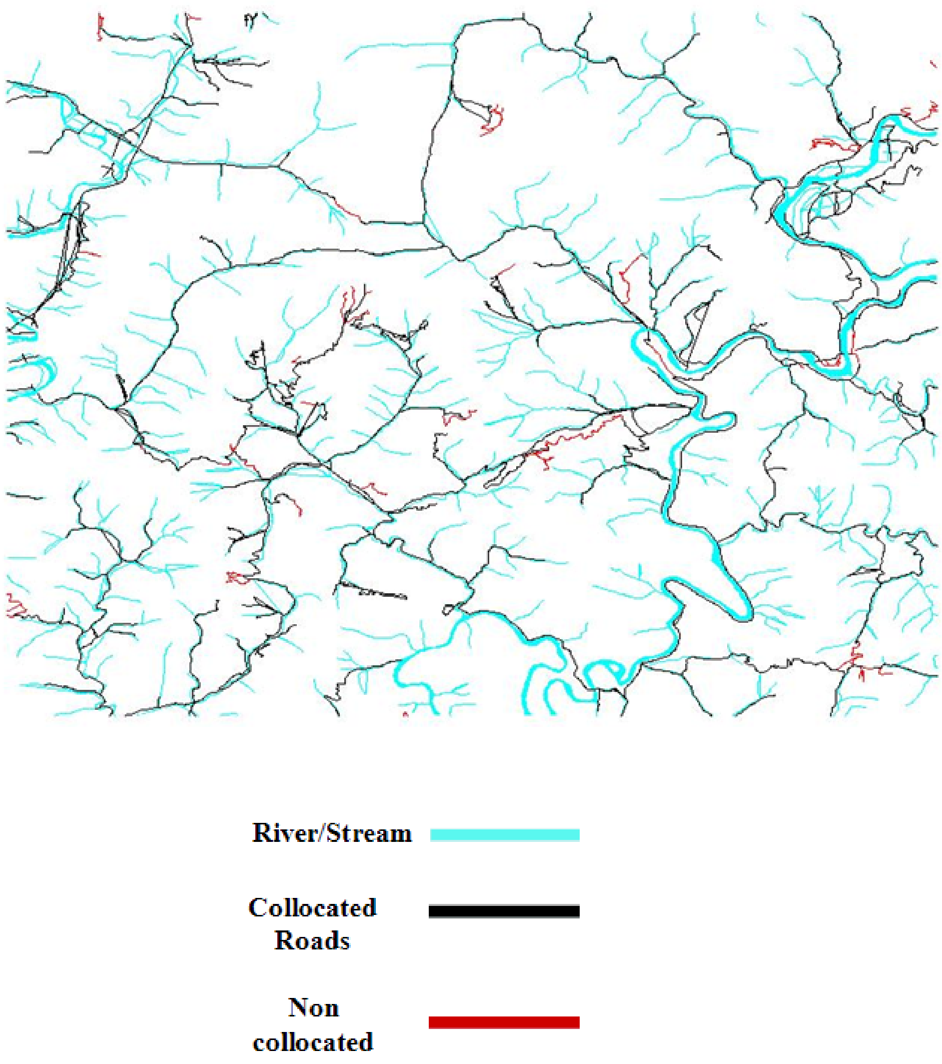


Figure 4: Road-River Colocation Example(Korea dataset)

SQL Example for Colocation Query

- ⋆ SQL3/OGC (Postgres/Postgis)
- ⋆ Detecting Road River Colocation Pattern:
  - Spatial Query Fragment

```
CREATE TABLE Road-River-Colocation AS
        SELECT DISTINCT R.*
        FROM River-Area-Table T, Road-Line-Table R
        WHERE distance ( T.geom, R.geom ) < 0.001;

CREATE TABLE Road-Stream-Colocation AS
        SELECT DISTINCT R.*
        FROM Stream-Line-Table T, Road-Line-Table R
        WHERE distance ( T.geom, R.geom ) < 0.001;

CREATE TABLE Cartroad-River-Colocation AS
        SELECT DISTINCT R.*
        FROM River-Area-Table T, Cartroad-Line-Table R
        WHERE distance ( T.geom, R.geom ) < 0.001;

CREATE TABLE Cartroad-Stream-Colocation AS
        SELECT DISTINCT R.*
        FROM Stream-Line-Table T, Cartroad-Line-Table R
        WHERE distance ( T.geom, R.geom ) < 0.001;
```

Colocation: Road-River

* 375 road features
* Center-line to center-line distance threshold = 0.001 units (about 100 meters)
* 77 % of all roads colocated with river

Colocation Pattern	Number of Colocated Features	**Interest Measure (%)** (Colocated roads / Total roads) * 100
Road with stream	153 of 239	64 %
Road with river	96 of 239	40 %
Road with stream or river	176 of 239	74 %
Cartroad with stream	97 of 136	71 %
Cartroad with river	44 of 136	32 %
Cartroad with stream or river	111 of 136	82 %
All roads with river or stream	287 of 375	**77 %**

Figure 5: Road-River Colocation Example(Korea dataset)

More Complex Colocation Examples

⋆ Complex Colocation/Outlier Example:

- Cropland colocated with river, stream or road

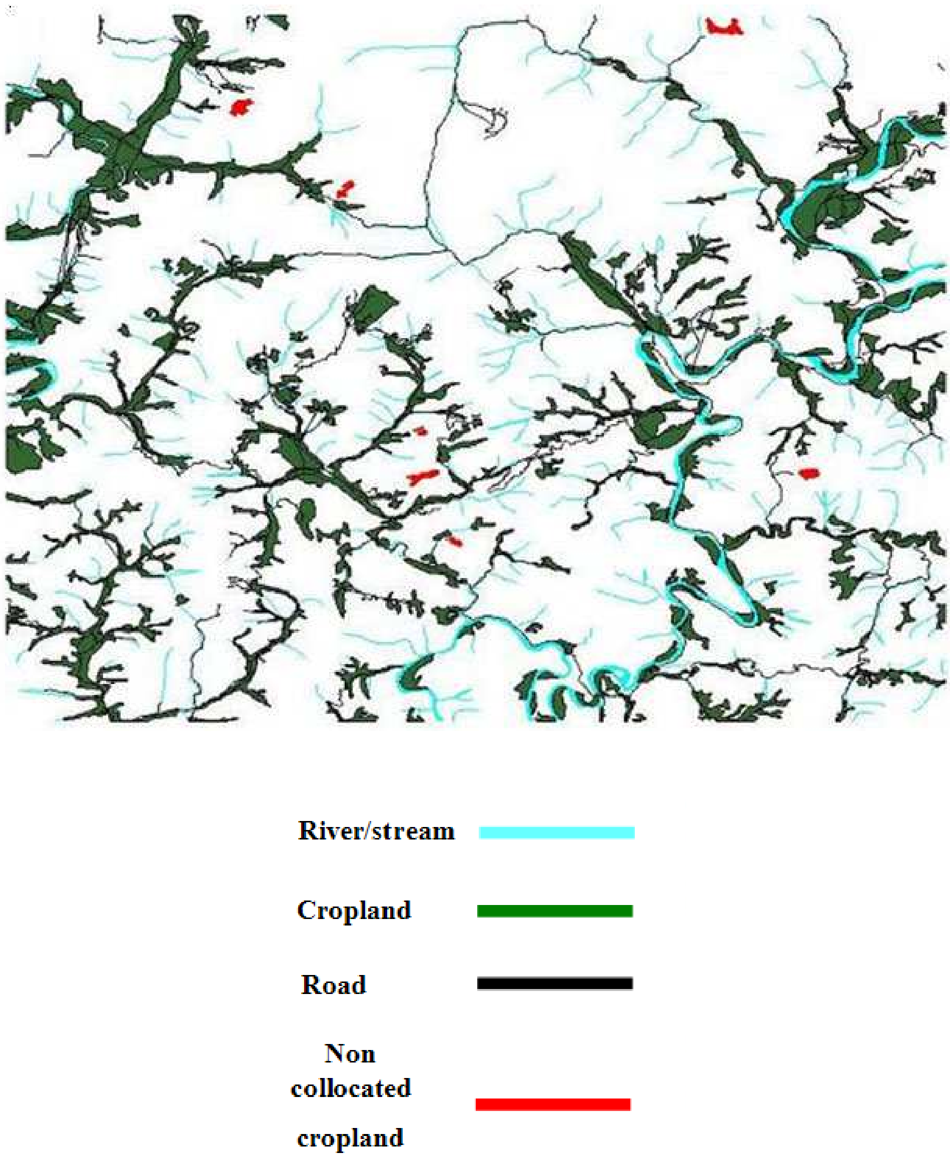


Figure 6: Complex Colocation Example

Outliers in Example Data

- ⋆ Outlier detection
  - Extra/erroneous features
  - Positional accuracy of features
  - Predict mislabeled/misclassified features
- ⋆ Overlapping road and river
- ⋆ Road crossing river and disconnected road Stream mislabeled as river
- ⋆ Cropland close to river and road
- ⋆ Cropland outliers on edges

Outliers in Example

⋆ Map production

- Identifying errors
  - e.g., expected colocation : (bridge, $\cap$(road, river))
  - violations illustrated below:

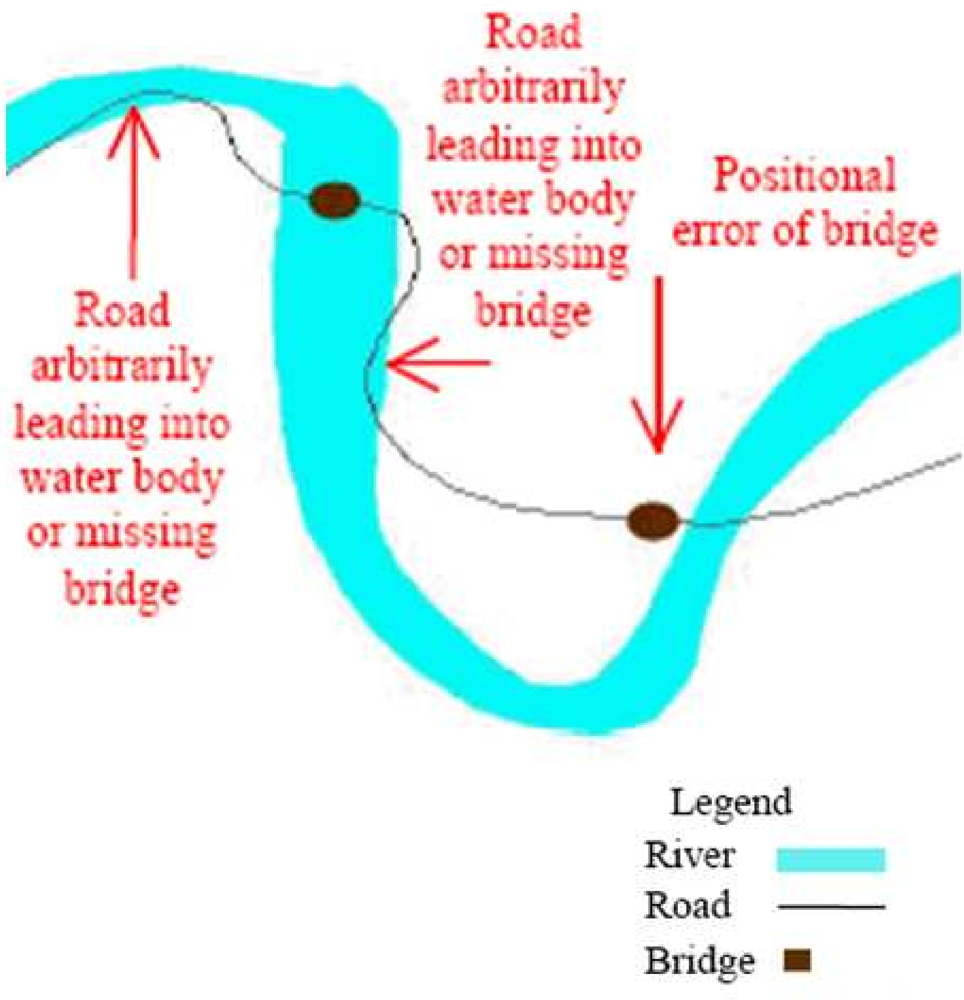


Figure 7: Finding errors in maps having road, river and bridges (Korea dataset)

Overview

⋆ Spatial Data Mining

- Find interesting, potentially useful, non-trivial patterns from spatial data

⋆ Components of Data Mining:

- Input: table with many columns, domain(column)
- Statistical Foundation
- Output: patterns and interest measures
  - e.g., predictive models, clusters, outliers, associations
- Computational process: algorithms

Overview

⇒ Input

⋆ Statistical Foundation

⋆ Output

⋆ Computational process

Overview of Input

⋆ Data

- Table with many columns(attributes)

tid	f_1	f_2	$\ldots$	f_n
0001	3.5	120	$\ldots$	Yes
0002	4.0	121	$\ldots$	No

Table 1: Example of Input Table

  - e.g., tid: tuple id; f_i: attributes

- Spatial attribute: geographically referenced
- Non-spatial attribute: traditional

⋆ Relationships among Data

- Non-spatial
- Spatial

Data in Spatial Data Mining

⋆ Non-spatial Information

- Same as data in traditional data mining
- Numerical, categorical, ordinal, boolean, etc
- e.g., city name, city population

⋆ Spatial Information

- Spatial attribute: geographically referenced
  - Neighborhood and extent
  - Location, e.g., longitude, latitude, elevation
- Spatial data representations
  - Raster: gridded space
  - Vector: point, line, polygon
  - Graph: node, edge, path

Figure 8: Raster and Vector Data for UMN Campus (in courtesy of UMN, MapQuest)

Relationships on Data in Spatial Data Mining

* Relationships on non-spatial data
  - Explicit
  - Arithmetic, ranking(ordering), etc.
  - Object is_instance_of a class, class is a subclass_of another class, object is part_of another object, object is a membership_of a set

* Relationships on Spatial Data
  - Many are **implicit**
  - Relationship Categories
    - Set-oriented: union, intersection, and membership, etc
    - Topological: meet, within, overlap, etc
    - Directional: North, NE, left, above, behind, etc
    - Metric: e.g., Euclidean: distance, area, perimeter
    - Dynamic: update, create, destroy, etc
    - Shape-based and visibility
  - Granularity

Granularity	Elevation Example	Road Example
local	elevation	on_road?
focal	slope	adjacent_to_road?
zonal	highest elevation in a zone	distance to nearest road

Table 2: Examples of Granularity

OGC Model

⋆ Open GIS Consortium Model

- Support spatial data types: e.g. point, line, polygons
- Support spatial operations as follows:

Operator Type	Operator Name
Basic Function	SpatialReference()
	Envelope()
	Export()
	IsEmpty()
	IsSimple()
	Boundary()
Topological/Set Operators	Equal
	Disjoint
	Intersect
	Touch
	Cross
	Within
	Contains
	Overlap
Spatial Analysis	Distance
	Buffer
	ConvexHull
	Intersection
	Union
	Difference
	SymmDiff

Table 3: Examples of Operations in OGC Model

Mining Implicit Spatial Relationships

⋆ Choices:

- Materialize spatial info + classical data mining
- Customized spatial data mining techniques

Relationships		Materialization	Customized SDM Tech.
Topological	Neighbor, Inside, Outside	Classical Data Mining	NEM, co-location
Euclidean	Distance,	can be used	K-means
	density		DBSCAN
Directional	North, Left, Above		Clustering on sphere
Others	Shape, visibility		

Table 4: Mining Implicit Spatial Relationships

⋆ What spatial info is to be materialized?

- Distance measure:
  - Point: Euclidean
  - Extended objects: buffer-based
  - Graph: shortest path
- Transactions: i.e., space partitions
  - Circles centered at reference features
  - Gridded cells
  - Min-cut partitions
  - Voronoi diagram

Research Needs for Data

⋆ Limitations of OGC Model

- Aggregate functions - e.g. mapcube
- Direction predicates - e.g. absolute, ego-centric
- 3D and visibility
- Network analysis
- Raster operations

⋆ Needs for New Research

- Modeling semantically rich spatial properties
- Moving objects
- Spatial time series data

Overview

- $\surd$ Input
- $\Rightarrow$ Statistical Foundation
- $\star$ Output
- $\star$ Computational process

Statistics in Spatial Data Mining

⋆ Classical Data Mining

- Learning samples are independently distributed
- Cross-correlation measures, e.g., χ^2, Pearson

⋆ Spatial Data Mining

- Learning sample are **not independent**
- Spatial Autocorrelation
  - Measures:
    * distance-based(e.g., K-function)
    * neighbor-based(e.g., Moran's I)
- Spatial Cross-Correlation
  - Measures: distance-based, e.g., cross K-function
- Spatial Heterogeneity

Overview of Statistical Foundation

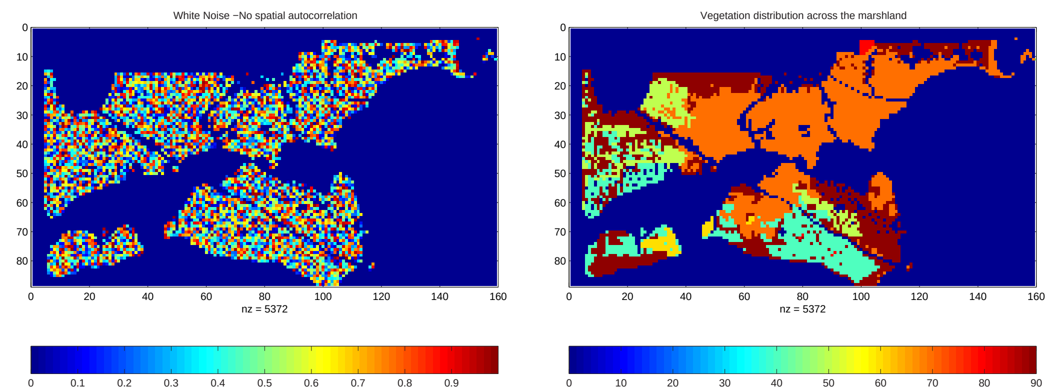
⋆ Spatial Statistics[Cressie, 1991][Hanning, 2003]

- Geostatistics
  - Continuous
  - Variogram: measure how similarity decreases with distance
  - Spatial prediction: spatial autocorrelation
- Lattice-based statistics
  - Discrete location, neighbor relationship graph
  - Spatial Gaussian models
    - Conditionally specified spatial Gaussian model
    - Simultaneously specified spatial Gaussian model
  - Markov Random Fields, Spatial Autoregressive Model
- Point process
  - Discrete
  - Complete spatial randomness (CSR): Poisson process in space
  - K-function: test of CSR

Spatial Autocorrelation(SA)

⋆ First Law of Geography

- "All things are related, but nearby things are more related than distant things. [Tobler, 1970]"



(a) Pixel property with independent identical distribution

(b) Vegetation Durability with SA

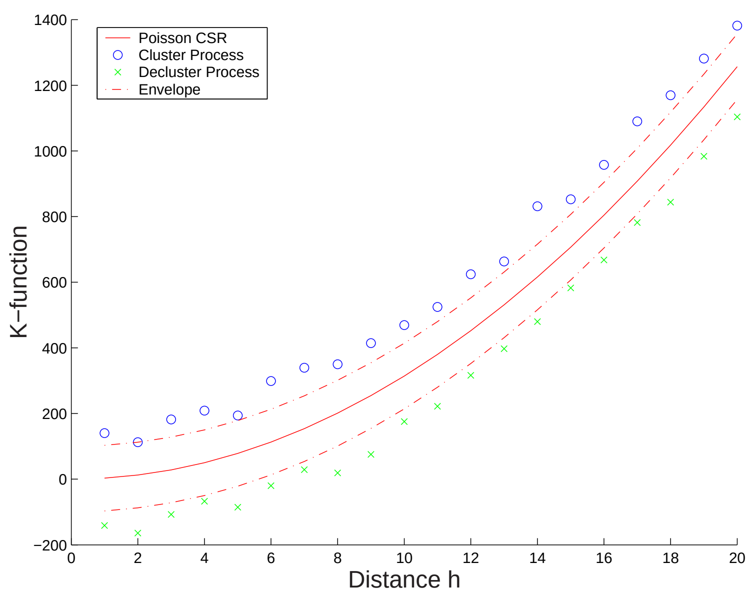
Figure 9: Spatial Randomness vs. Autocorrelation

⋆ Spatial autocorrelation

- Nearby things are more similar than distant things
- Traditional i.i.d. assumption is not valid
- Measures: K-function, Moran's I, Variogram, $\cdots$

Spatial Autocorrelation: Distance-based Measure

- K-function Definition:
  - Test against randomness for point pattern
  - $K(h) = \lambda^{-1} E$[number of events within distance h of an arbitrary event]
    - λ is intensity of event
  - Model departure from randomness in a wide range of scales
- Inference
  - For Poisson complete spatial randomness(csr): $K(h) = \pi h^2$
  - Plot Khat(h) against h, compare to Poisson csr
    - $>$: cluster
    - $<$: decluster/regularity



Spatial Autocorrelation: Topological Measure

- ⋆ Moran's I Measure Definition:

$$MI = \frac{zWz^t}{zz^t}$$

- $z = \{x_1 - \bar{x}, \ldots, x_n - \bar{x}\}$
  - x_i : data values
  - $\bar{x}$**:** mean of x
  - n**:** number of data
- W**:** the contiguity matrix

- ⋆ Ranges between -1 and +1
  - higher positive value ⇒ high SA, Cluster, Attract
  - lower negative value ⇒ interspersed, de-clustered, repel
  - e.g., spatial randomness ⇒ MI = 0
  - e.g., distribution of vegetation durability ⇒ MI = 0.7
  - e.g., checker board ⇒ MI = -1

Cross-Correlation

⋆ Cross K-Function Definition

- $K_{ij}(h) = \lambda_j^{-1} E$ [number of type j event within distance h of a randomly chosen type i event]
- Cross K-function of some pair of spatial feature types
- Example
  - Which pairs are frequently co-located?
  - Statistical significance

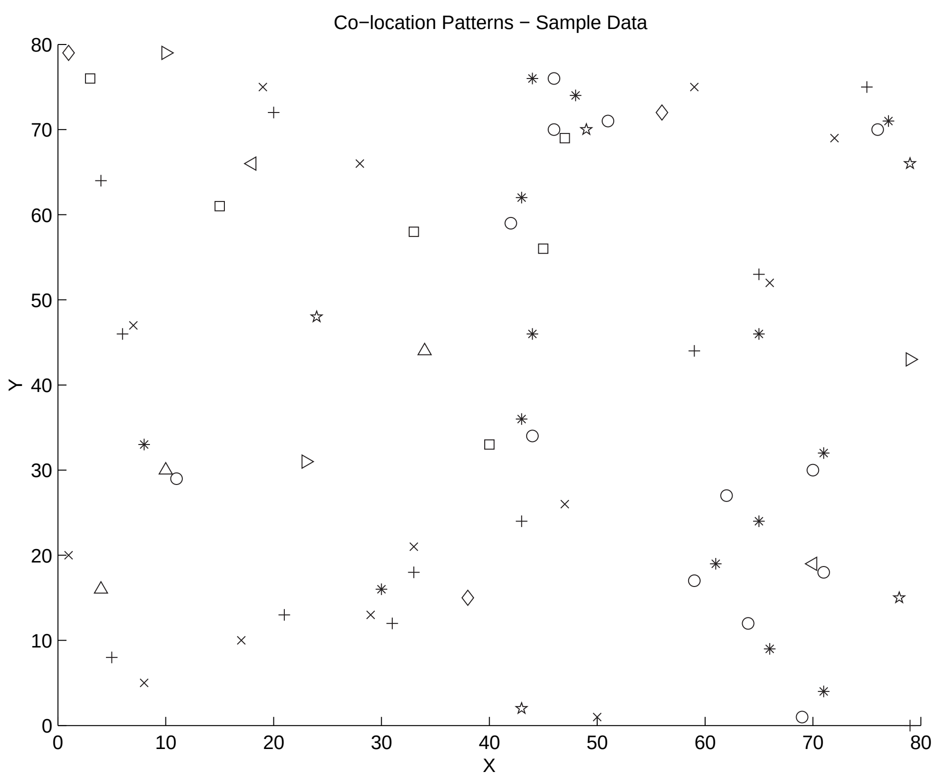


Figure 10: Example Data (o and * ; x and +)

Illustration of Cross-Correlation

⋆ Illustration of Cross K-Function for Example Data

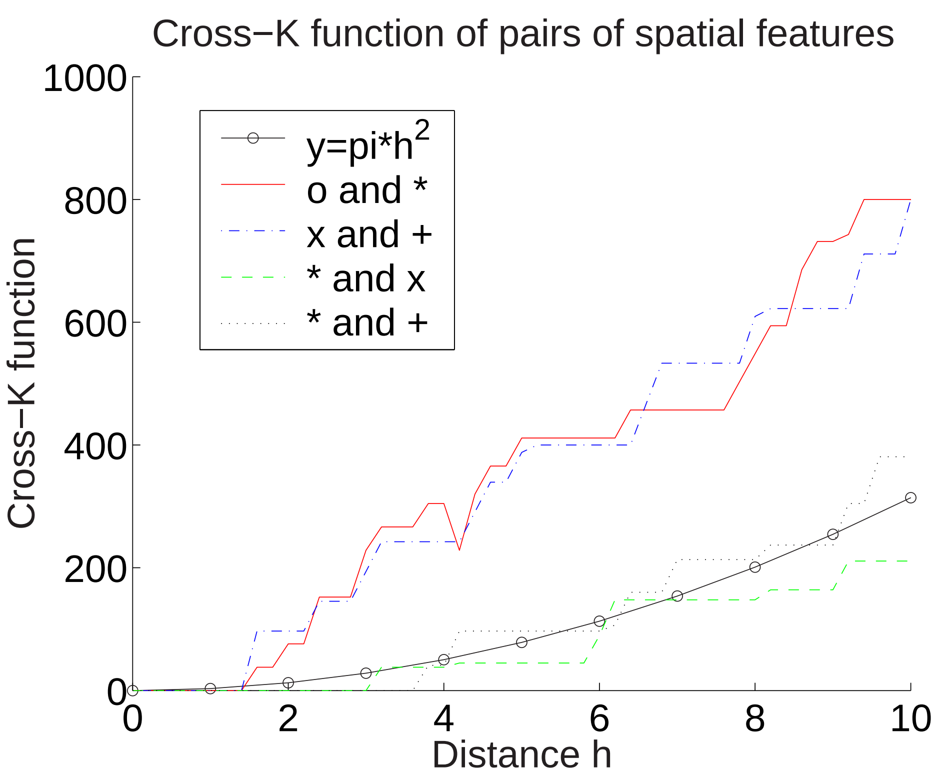
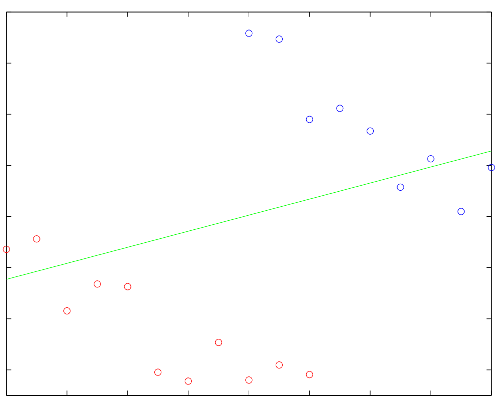


Figure 11: Cross K-function for Example Data

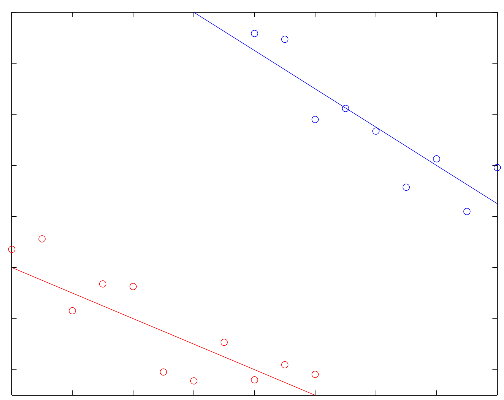
Spatial Slicing

⋆ Spatial heterogeneity

- "Second law of geography" [M. Goodchild, UCGIS 2003]
- Global model might be inconsistent with regional models
  - spatial Simpson's Paradox (or Ecological Inference)



(a) Global Model



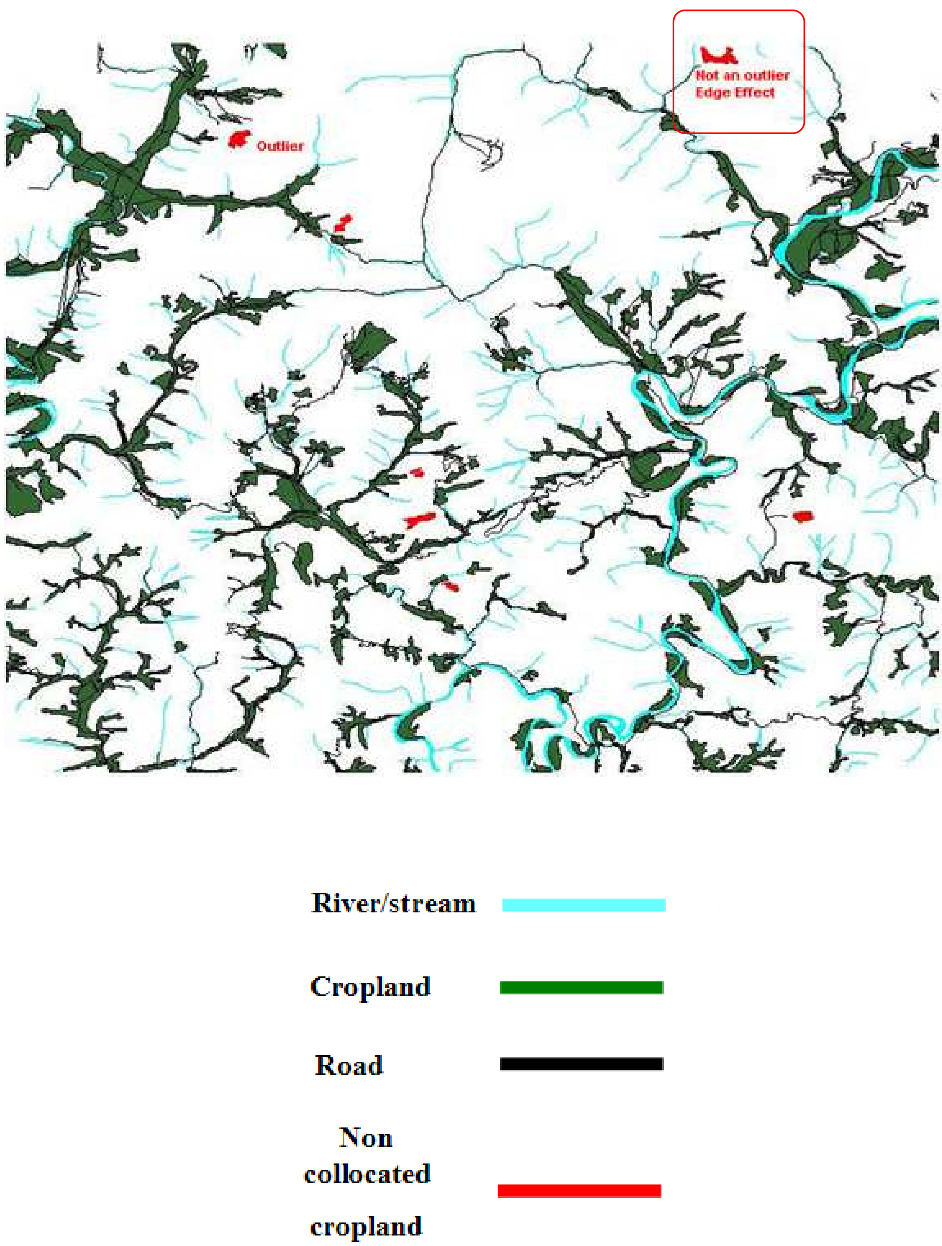
(b) Regional Models

⋆ Spatial Slicing

- Slicing inputs can improve the effectiveness of SDM
- Slicing output can illustrate support regions of a pattern
  - e.g., association rule with support map

Edge Effect

- Cropland on edges may not be classified as outliers
- No concept of spatial edges in classical data mining



Research Challenges of Spatial Statistics

⋆ State-of-the-art of Spatial Statistics

		Point Process	Lattice	Geostatistics
raster			√	√
vector	point	√	√	√
	line			√
	polygon		√	√
graph				

Table 5: Data Types and Statistical Models

⋆ Research Needs

- Correlating extended features:
  - Example data: Korea data
  - e.g. road, river (line strings)
  - e.g. cropland (polygon), road, river
- Edge effect
- Relationship to classical statistics
  - Ex. SVM with spatial basis function vs. SAR

Overview

- $\surd$ Input
- $\surd$ Statistical Foundation
- $\Rightarrow$ Output
- $\star$ Computational process

General Approaches in SDM

- ⋆ Materializing spatial features, use classical DM
  - Ex. Huff's model - distance(customer, store)
  - Ex. spatial association rule mining[Koperski, Han, 1995]
  - Ex: wavelet and fourier transformations
  - commercial tools: e.g., SAS-ESRI bridge
- ⋆ Spatial slicing, use classical DM
  - Ex. association rule with support map[P. Tan et al]

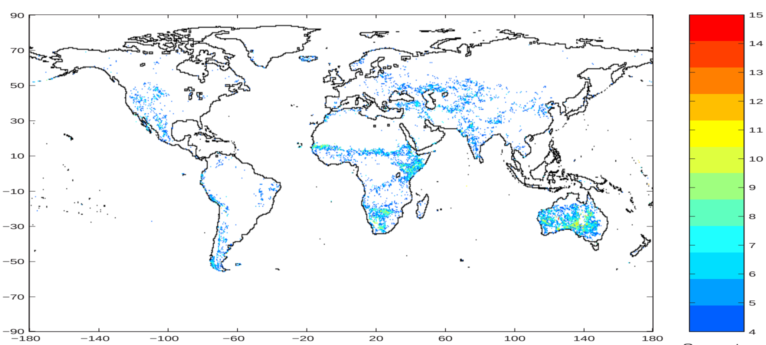


Figure 12: Association rule with support map(FPAR-high → NPP-high)

  - commercial tools: e.g.,Matlab, SAS, R, Splus
- ⋆ Customized spatial techniques
  - Ex. geographically weighted regression: parameter = f(loc)
  - e.g., MRF-based Bayesian Classifier (MRF-BC)
  - commercial tools
    - e.g.,Splus spatial/R spatial/terraseer + customized codes

Overview of Data Mining Output

⋆ Supervised Learning: Prediction

- Classification
- Trend

⋆ Unsupervised Learning:

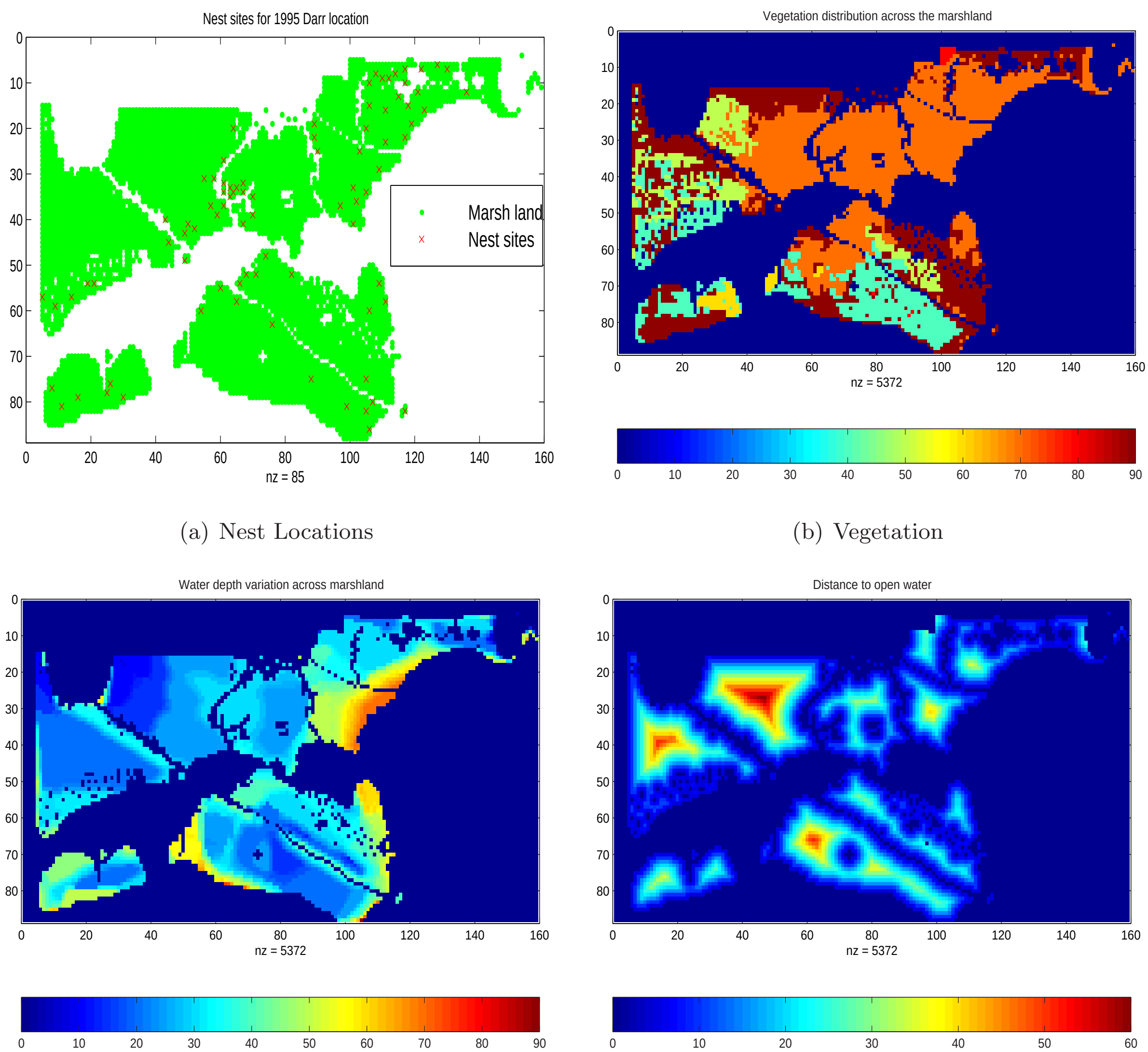
- Clustering
- Outlier Detection
- Association

⋆ Input Data Types vs. Output Patterns

Patterns	Point Process	Lattice	Geostatistics
Prediction	$\surd$	$\surd$	
Trend			$\surd$
Clustering	$\surd$	$\surd$	
Outliers	$\surd$	$\surd$	$\surd$
Associations	$\surd$	$\surd$	

Table 6: Output Patterns vs. Statistical Models

Illustrative Application to Location Prediction (Backup)



(a) Nest Locations

(b) Vegetation

(c) Water Depth

(d) Distance to Open Water

Prediction and Trend

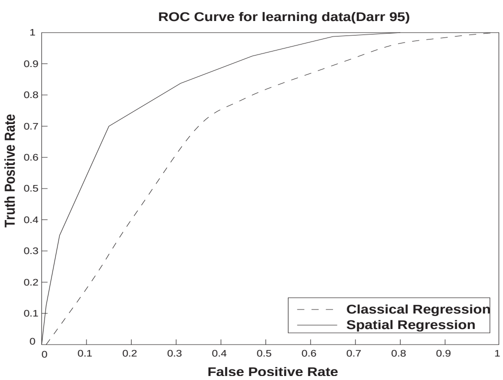
⋆ Prediction

- Continuous: trend, e.g., regression
  - Location aware: spatial autoregressive model(SAR)
- Discrete: classification, e.g., Bayesian classifier
  - Location aware: Markov random fields(MRF)

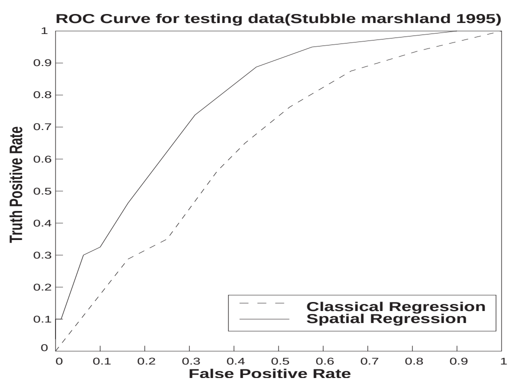
Classical	Spatial
$\mathbf{y} = \mathbf{X}\beta + \epsilon$	$y = \rho W y + X\beta + \epsilon$
$Pr(C_i\|X) = \frac{Pr(X\|C_i)Pr(C_i)}{Pr(X)}$	$Pr(c_i\|X, C_N) = \frac{Pr(c_i)*Pr(X,C_N\|c_i)}{Pr(X,C_N)}$

Table 7: Prediction Models

- e.g., ROC curve for SAR and regression



(e) ROC curves for learning



(f) ROC curves for testing

Figure 13: (a) Comparison of the classical regression model with the spatial autoregression model on the Darr learning data. (b) Comparison of the models on the Stubble testing data.

Spatial Contextual Model: SAR

⋆ Spatial Autoregressive Model (SAR)

$$y = \rho W y + X\beta + \epsilon.$$

- Assume that dependent values y_i' are related to each other

$$y_i = f(y_j) \; i \neq j.$$

- Directly model spatial autocorrelation using W

⋆ Geographically Weighted Regression (GWR)

- A method of analyzing spatially varying relationships
  - parameter estimates vary locally
- Models with Gaussian, logistic or Poisson forms can be fitted
- Example:

$$y = X\beta' + \epsilon'.$$

- where β' **and** ϵ' **are location dependent**

Spatial Contextual Model: MRF

⋆ Markov Random Fields Gaussian Mixture Model (MRF-GMM)

- Undirected graph to represent the interdependency relationship of random variables
- A variable depends only on neighbors
- Independent of all other variables
- $f_C(S_i)$ independent of $f_C(S_j)$, if $W(s_i, s_j) = 0$
- Predict $f_C(s_i)$, given feature value X and neighborhood class label C_N

$$Pr(c_i|X, C_N) = \frac{Pr(c_i) * Pr(X, C_N|c_i)}{Pr(X, C_N)}$$

  – Assume: $Pr(c_i), Pr(X, C_N|c_i), and Pr(X, C_N)$ are mixture of Gaussian distributions.

Research Needs for Spatial Classification

⋆ Open Problems

- Estimate W for SAR and MRF-BC
- Scaling issue in SAR
  - scale difference: $\rho \mathbf{Wy}$ **vs.** $\mathbf{X}\beta$
- Spatial interest measure: e.g., avg dist(actual, predicted)

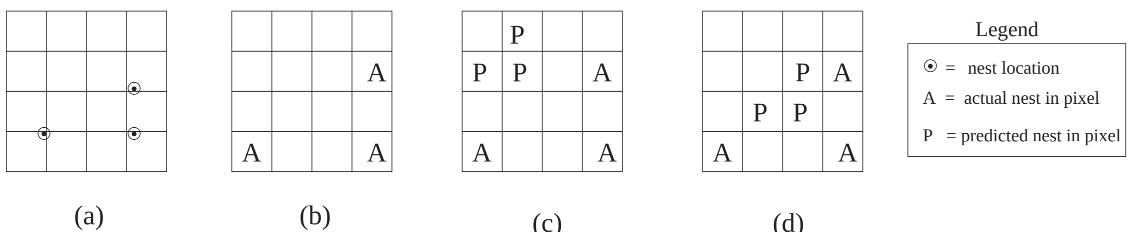


Figure 14: An example showing different predictions: (a)The actual sites, (b)Pixels with actual sites, (c)Prediction 1, (d)Prediction 2. Prediction 2 is spatially more accurate than 1.

Clustering

⋆ Clustering: Find groups of tuples

⋆ Statistical Significance

- Complete spatial randomness, cluster, and decluster

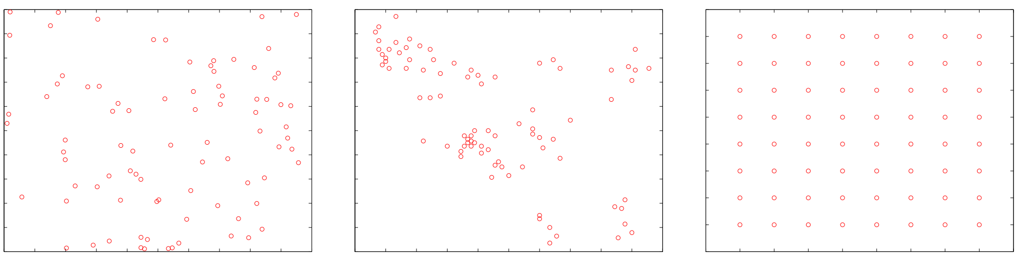


Figure 15: Inputs: Complete Spatial Random (CSR), Cluster, and Decluster

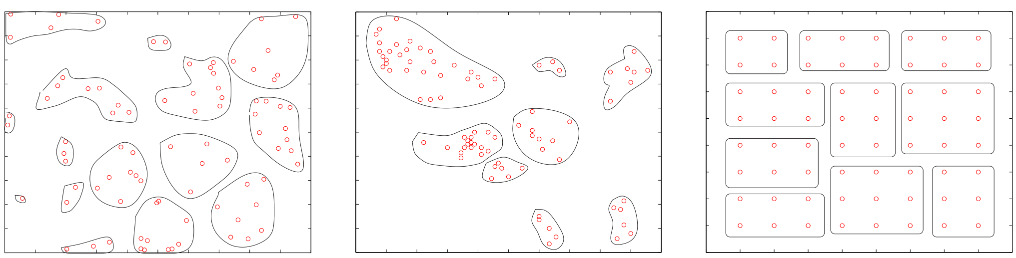


Figure 16: Classical Clustering

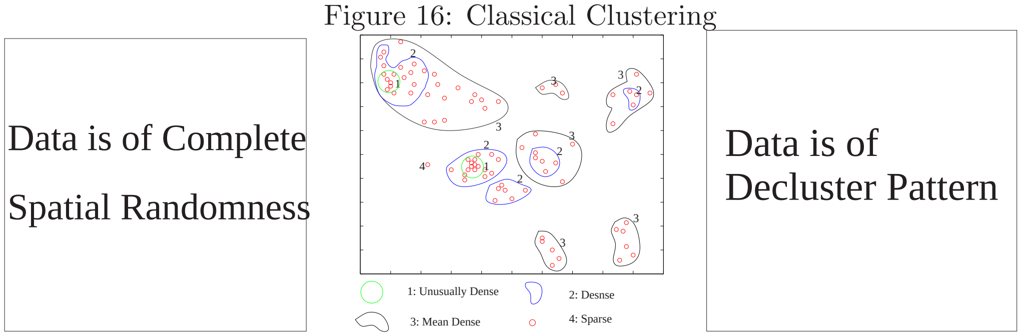


Figure 17: Spatial Clustering

Clustering

- Similarity Measures
  - Non-spatial: e.g., soundex
  - Classical clustering: Euclidean, metric, graph-based
  - Topological: **neighborhood EM**(NEM)
    - seeks a partition that is both well clustered in feature space and spatially regular
    - Implicitly based on locations
  - Interest measure:
    - spatial continuity
    - cartographic generalization
    - unusual density
    - keep nearest neighbors in common cluster
- Challenges
  - Spatial constraints in algorithmic design
    - Clusters should obey obstacles
    - Ex. rivers, mountain ranges, etc

Semi-Supervised Bayesian Classification

⋆ Motivation: high cost of collecting labeled samples

⋆ Semi-supervised MRF

- Idea: use unlabeled samples to improve classification
  - Ex. reduce salt-N-pepper noise
- Effects on land-use data - smoothing

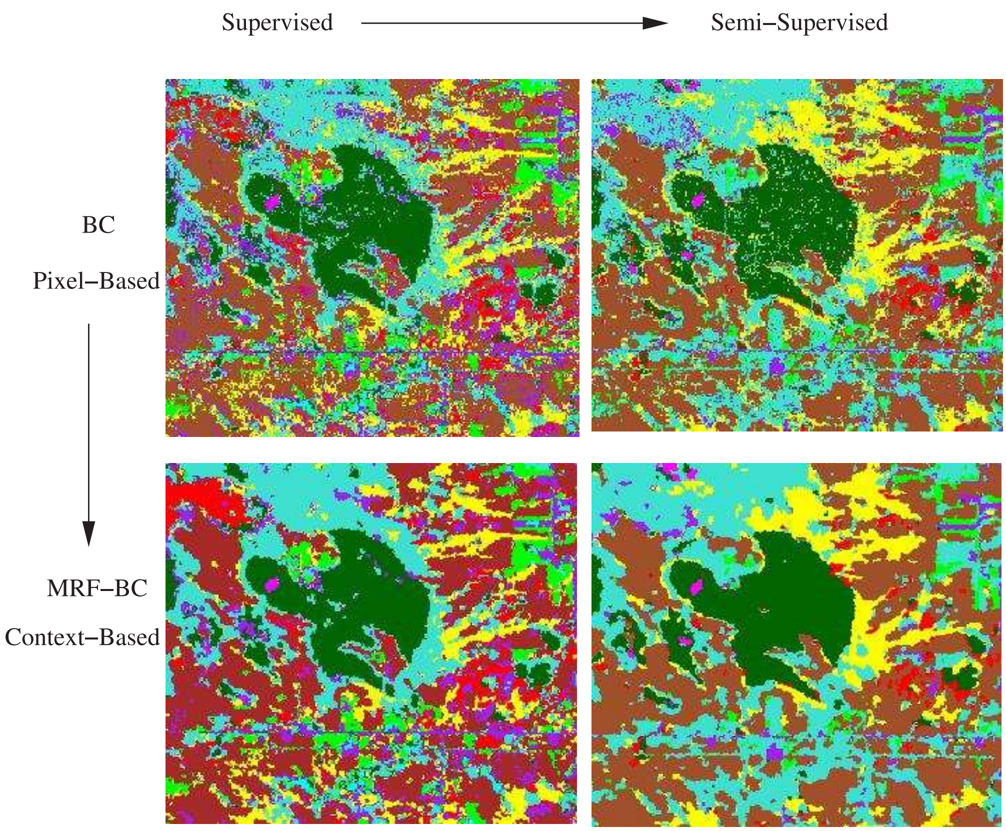
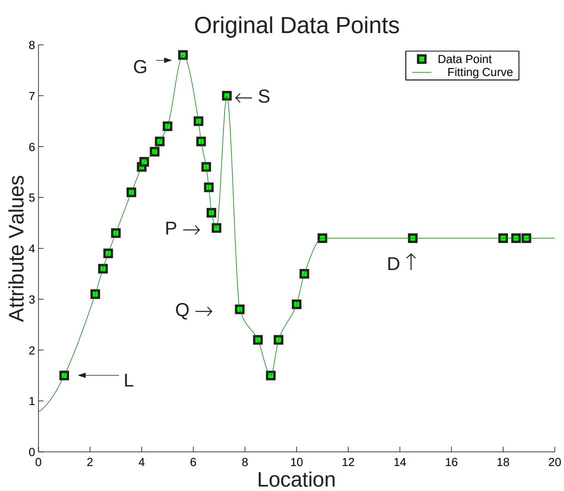


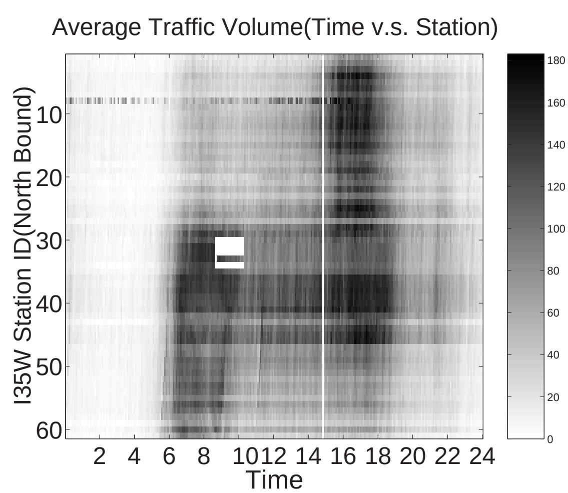
Figure 18: Bayesian Classifier (Top Left); Semi-Supervised BC (Top Right);BC-MRF (Bottom Left); BC-EM-MRF (Bottom Right)

Spatial Outlier Detection

- ⋆ Spatial Outlier Detection
  - Finding anomalous tuples
  - Global vs. Spatial outlier
  - Detection Approaches
    - Graph-based outlier detection: Variogram, Moran Scatter Plot
    - Quantitative outlier detection: Scatter Plot, Z-score
- ⋆ Location-awareness



(a) Outliers in Example Data



(b) Outliers in Traffic Data

An Example of Spatial Outlier Detection (Backup)

⋆ Consider Scatter Plot

⋆ Model Building

- Neighborhood aggregate function $f^N_{aggr} : E(x) = \frac{1}{k} \sum_{y \in N(x)} f(y)$
- Distributive aggregate functions
  - $\sum f(x), \sum E(x), \sum f(x)E(x), \sum f^2(x), \sum E^2(x)$
- Algebraic aggregate functions
  - $m = \frac{N \sum f(x)E(x) - \sum f(x) \sum E(x)}{N \sum f^2(x) - (\sum f(x)^2}$
  - $b = \frac{\sum f(x) \sum E^2(x) - \sum f(x) \sum f(x)E(x)}{N \sum f^2(x) - (\sum f(x))^2}$
  - $\sigma_\epsilon = \sqrt{\frac{S_{yy} - (m^2 S_{xx})}{(n-2)}}$,
  - where $S_{xx} = \sum f^2(x) - [\frac{(\sum f(x))^2}{n}]$
  - and $S_{yy} = \sum E^2(x) - [\frac{(\sum E(x))^2}{n}]$

⋆ Testing

- Difference function F_{diff}
  - $\epsilon = E(x) - (m * f(x) + b)$
  - where $E(x) = \frac{1}{k} \sum_{y \in N(x)} f(y)$
- Statistic test function ST
  - $|\frac{\epsilon - \mu_\epsilon}{\sigma_\epsilon}| > \theta$

Spatial Outlier Detection

⋆ Separate two phases

- Model Building
- Testing: test a node (or a set of nodes)

⋆ Computation Structure of Model Building

- Key insights:
  - Spatial self join using $N(x)$ relationship
  - Algebraic aggregate function can be computed in one disk scan of spatial join

⋆ Computation Structure of Testing

- Single node: spatial range query
  - Get_All_Neighbors(x) operation
- A given set of nodes
  - Sequence of Get_All_Neighbor(x)

Research Needs in Spatial Outlier Detection

⋆ Multiple spatial outlier detection

- Eliminating the influence of neighboring outliers
- Incremental

⋆ Multi-attribute spatial outlier detection

- Use multiple attributes as features

⋆ Design of spatial statistical tests

⋆ Scale up for large data

Association Rules - An Analogy

⋆ Association rule e.g. (Diaper in T ⇒ Beer in T)

rans.	Items Bought
	{socks,  milk, , beef, egg, … }
	{ pillow,  , toothbrush, ice-cream, muffin, … }
	{  ,  , pacifier, formula, blanket, … }
.	…
	{battery, juice, beef, egg, chicken, … }

- Support: probability(Diaper and Beer in T) = 2/5
- Confidence: probability(Beer in T|Diaper in T)= 2/2

⋆ Algorithm Apriori [Agrawal, Srikant, VLDB94]

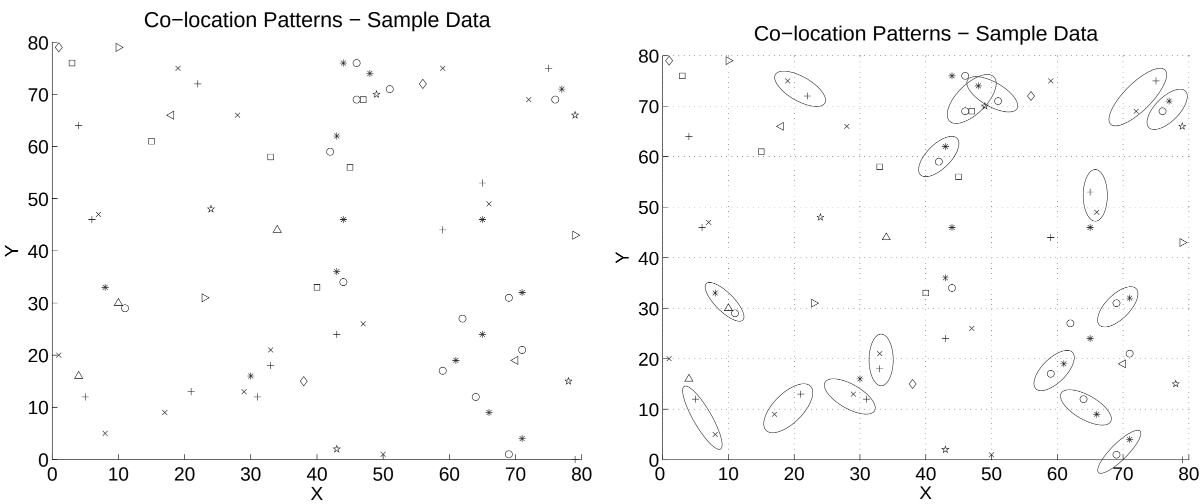
- Support based pruning using monotonicity

⋆ Note: **Transaction is a core concept!**

Spatial Colocation

⋆ Association

- Domain(f_i) = union { **any, domain**(f_i)}
- Finding frequent itemsets from f_i
- Co-location
  - Effect of transactionizing: **loss of info**
  - Alternative: use spatial join, statistics



(d)

Figure 19: a) A spatial dataset. Shapes represent different spatial feature types. (b) Transactionazing continuous space splits circled instances of colocation patterns into separated transactions

Spatial Colocation Approaches

* Approaches
  - Spatial Join-based Approaches
    - Join based on map overlay, e.g. [Estivill-Castro and Lee, 1001]
    - Join using K-function, e.g. [Shekhar and Huang, 2001]
  - Transaction-based Approaches
    - e.g., [Koperski and Han, 1995] and [Morimoto,2001]
* Challenges
  - Neighborhood definition
  - "Right" transactionazation
  - Statistical interpretation
  - Computational complexity
    - large number of joins
    - join predicate is a conjunction of:
      - neighbor
      - distinct item types

Overview

- $\surd$ Input
- $\surd$ Statistical Foundation
- $\surd$ Output
- $\Rightarrow$ Computational process

Computational Process

⋆ Most algorithmic strategies are applicable

⋆ Algorithmic Strategies in Spatial Data Mining:

Classical Algorithms	Algorithmic Strategies in SDM	Comments
Divide-and-Conquer	Space Partitioning	possible info loss
Filter-and-Refine	Minimum-Bounding-Rectangle(MBR), Predicate Approximation	
Ordering	Plane Sweeping, Space Filling Curves	possible info loss
Hierarchical Structures	Spatial Index, Tree Matching	
Parameter Estimation	Parameter estimation with spatial autocorrelation	

Table 8: Algorithmic Strategies in Spatial Data Mining

⋆ Challenges

- Does spatial domain provide computational efficiency?
  - Low dimensionality: 2-3
  - Spatial autocorrelation
  - Spatial indexing methods
- Generalize to solve spatial problems
  - Linear regression vs SAR
    - Continuity matrix W is assumed known for SAR, however, **estimation of anisotropic W** is non-trivial
  - Spatial outlier detection: spatial join
  - Co-location: bunch of joins

Example of Computational Process

⋆ Teleconnection

- Find locations with climate correlation over θ
  - e.g., El Nino affects global climate

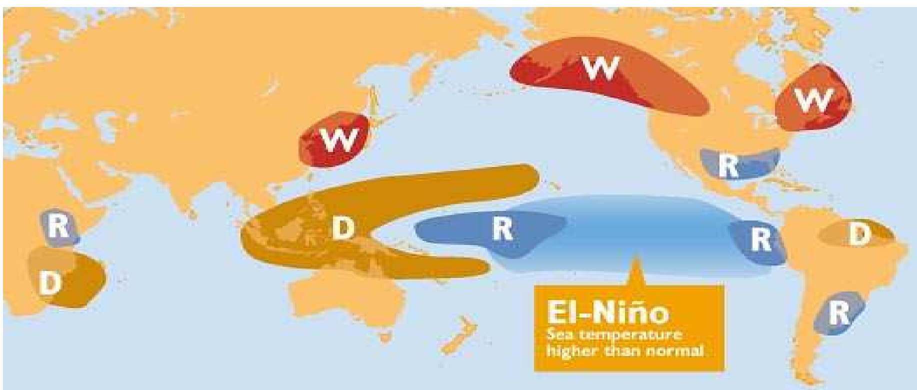


Figure 20: Global Influence of El Nino during the Northern Hemisphere Winter(D: Dry; W:Warm; R:Rainfall)

Example: Teleconnection (Cont')

⋆ Challenge:

- high dim(e.g., 600) feature space
- 67k land locations and 100k ocean locations
- 50-year monthly data

⋆ Computational Efficiency

- Spatial autocorrelation:
  - Reduce Computational Complexity
- Spatial indexing to organize locations
  - Top-down tree traversal is a strong filter
  - Spatial join query: filter-and-refine
    - save 40% to 98% computational cost at $\theta = 0.3$ to 0.9

Parameter estimation for SAR

⋆ Spatial Auto-Regression Model

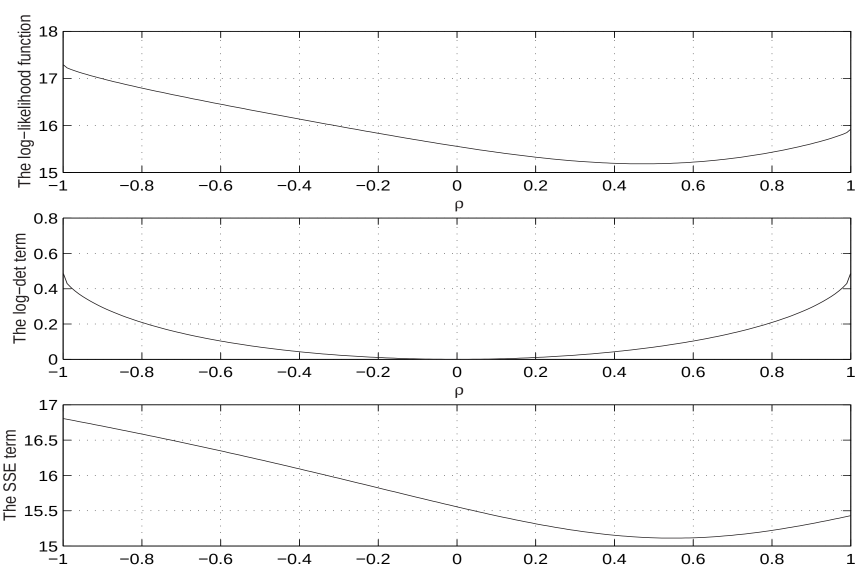
- Estimate ρ and β for $y = \rho W y + X\beta + \epsilon$
- The estimation uses maximum-likelihood (ML) theory

⋆ Log-likelihood function $\mathbf{LLF}$ = log-det + SSE + const

- log-det $= \ln|\mathbf{I} - \rho\mathbf{W}|$
- SSE $= \frac{1}{2\sigma^2}\{\mathbf{y}^T(\mathbf{I} - \rho\mathbf{W})^T\mathbf{M}^T\mathbf{M}(\mathbf{I} - \rho\mathbf{W})\mathbf{y}\}$

⋆ Computational Insight:

- LLF is uni-modal [Kazar et al., 2005]: breakthrough result
- Optimal ρ found by Golden Section Search or Binary Search



Reducing Computational Cost

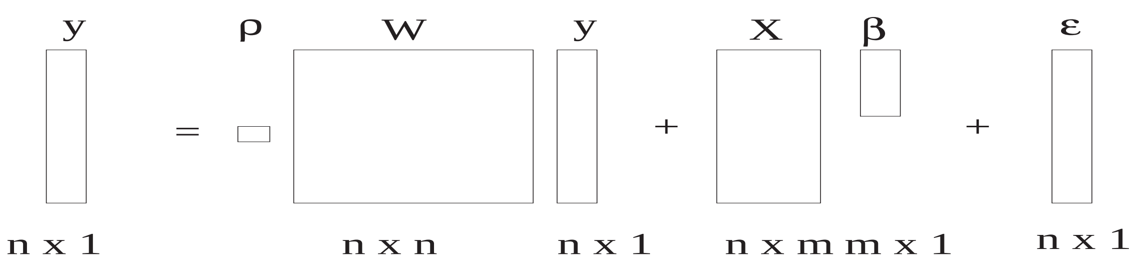
⋆ Exact Solution

- Bottleneck = evaluation of log-det
- Reduce cost by getting a seed for ρ minimizing SSE term [Kazar et.al., 2005]

⋆ Approximate Solution

- Reduce cost by approximating log-determinant term
- E.g., Chebyshev Polynomials, Taylor Series [LeSage and Pace, 2001]
- Comparison of Accuracy, e.g., Chebyshev Polynomials $\gg$ Taylor series [Kazar et.al., 2004]

⋆ Parallel Solution



⋆ Computational Challenges

- Eigenvalue + Least square + M. L.
- Computing all eigenvalues of a large matrix
- Memory requirement

Life Cycle of Data Mining

- ⋆ CRISP-DM (CRoss-Industry Standard Process for DM)
  - Application/Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment
  - [1] CRISP-DM URL: http://www.crisp-dm.org

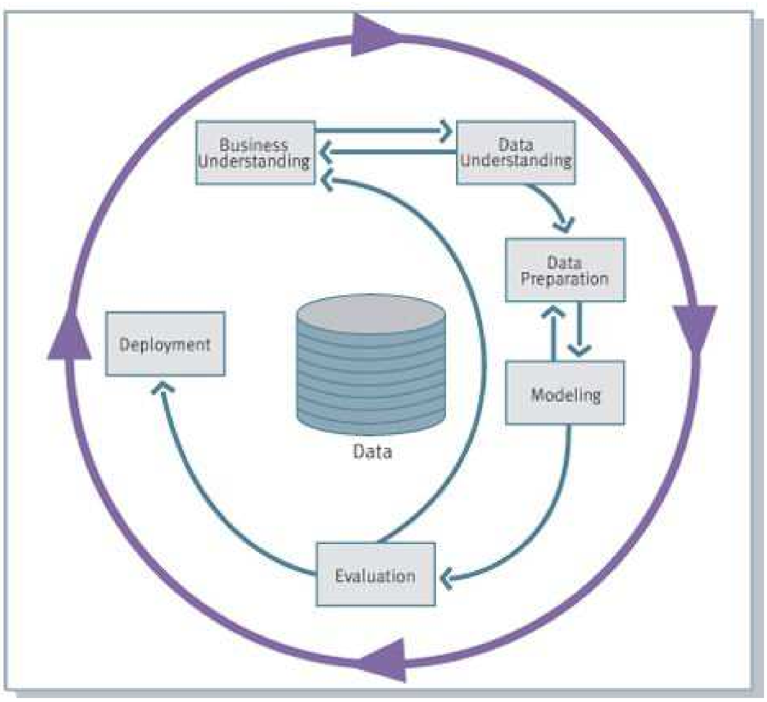


Figure 21: Phases of CRISP-DM [1]

- ⋆ Is CRISP-DM adequate for Spatial Data Mining?

Summary

⋆ What's Special About Spatial Data Mining?

- Input Data
- Statistical Foundation
- Output Patterns
- Computational Process

	Classical DM	Spatial DM
Input	All explicit, simple types and transactions	often Implicit relationships, complex types
Stat Foundation	Independence of samples	spatial autocorrelation
Output	Interest measures: set-based	Location-awareness
Computational Process	Combinatorial optimization Numerical alg.	Computational efficiency opportunity Spatial autocorrelation, plane-sweeping New complexity: SAR, co-location mining Estimation of anisotropic W is nontrivial
Objective Function	Max likelihood Min sum of squared errors	Map_Similarity(Actual, Predicted)
Constraints	Discrete space Support threshold Confidence threshold	Keep NN together Honor geo-boundaries
Other Issues		Edge effect, scale

Table 9: Summary of Spatial Data Mining

Book

http://www.cs.umn.edu/research/shashi-group

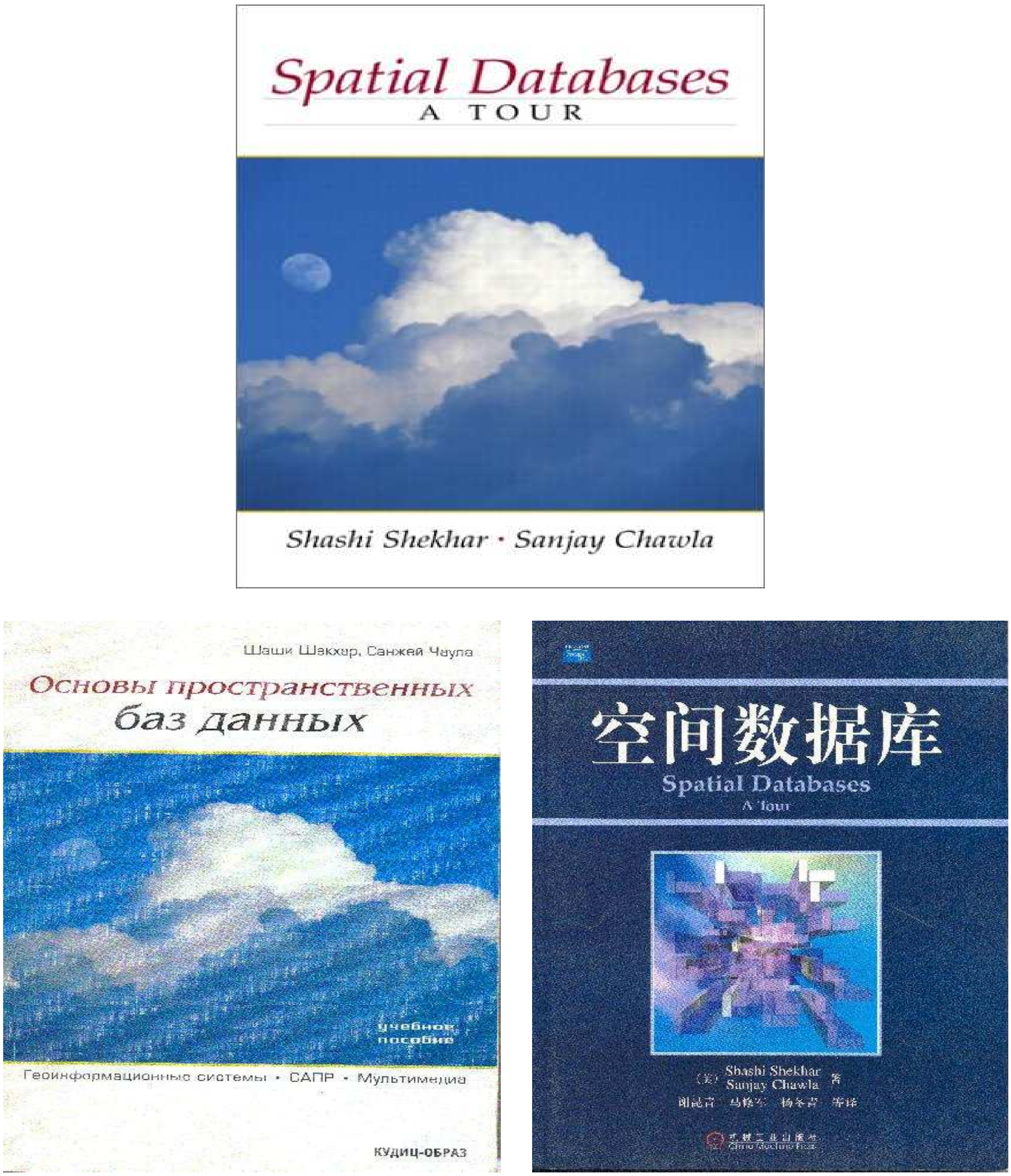


Figure 22: Spatial Databases: A Tour (a) English Version (b) Russian Version (c) Chinese Version

References

⋆ References

- [Cressie, 1991], N. Cressie, ***Statistics for Spatial Data***, John Wiley and Sons, 1991
- [Degroot, Schervish, 2002], M. Degroot and M. Schervish, ***Probability and Statistics (Third Ed.)***, Addison Wesley, 2002
- [Fotheringham et al, 2002], A. Fotheringham, C. Brunsdon, and M. Charlton, ***Geographically Weighted Regression : The Analysis of Spatially Varying Relationships***, John Wiley & Sons, 2002.
- [Goodchild, 2001], M. Goodchild, ***Spatial Analysis and GIS***, 2001 ESRI User Conference Pre-Conference Seminar
- [Hanning, 2003], R. Hanning, ***Spatial Data Analysis : Theory and Practice***, Cambridge University Press, 2003
- [Hastie et al, 2002], T. Hastie, R. Tibshirani, and J. Friedman, ***The Elements of Statistical Learning***, Springer-Verlag, 2001
- [Huff, 1963], D. Huff, ***A Probabilistic Analysis of Shopping Center Trade Areas***, Lan Economics, 1963
- [Kazar et al., 2004], B. M. Kazar, S. Shekhar, D. J. Lilja, R. R. Vatsavai, R. K. Pace, ***Comparing Exact and Approximate Spatial Auto-Regression Model Solutions for Spatial Data Analysis***, GIScience 2004
- [Kazar et al., 2005], B.M. Kazar, D. Boley, S. Shekhar, D.J. Lilja, R.K. Pace, J. LeSage, ***Parameter Estimation for the Spatial Autoregression Model: A Summary of Results***, submitted to KDD 2005

References

⋆ References

- [Koperski, Han, 1995], K. Kopperski and J. Han, ***Discovery of Spatial Association Rules in Geographic Information Database***, SSTD, 1995
- [Koperski et al, 1996], K. Kopperski, J. Adhikary, and J. Han, ***Spatial Data Mining: Progress and Challenges***, DMKD, 1996
- [LeSage and Pace, 2001], J. LeSage and R. K. Pace, ***Spatial Dependence in Data Mining***, in Data Mining for Scientific and Engineering Applications, R. L. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. R. Namburu (eds.), Kluwer Academic Publishing, p. 439-460, 2001.
- [Miller, Han, 2001], H. Miller and J. Han(eds), ***Geographic Data Mining and Knowledge Discovery***, Taylor and Francis, 2001
- [Roddick, 2001], J. Roddick, K. Hornsby and M. Spiliopoulou, ***Yet Another Bibliography of Temporal, Spatial Spatio-temporal Data Mining Research***, KDD Workshop, 2001
- [Shekhar et al, 2003], S. Shekhar, C. T. Lu, and P. Zhang, ***A Unified Approach to Detecting Spatial Outliers***, GeoInformatica, 7(2), Kluwer Academic Publishers, 2003
- [Shekhar, Chawla, 2003], S. Shekhar and S. Chawla, ***Spatial Databases: A Tour***, Prentice Hall, 2003
- [Shekhar et al, 2002], S. Shekhar, P. Schrater, R. Vatsavai, W. Wu, and S. Chawla, ***Spatial Contextual Classification and Prediction Models for Mining Geospatial Data***, IEEE Transactions on Multimedia (special issue on Multimedia Databases), 2002

References

⋆ References

- [Shekhar et al, 2001], S. Shekhar and Y. Huang, ***Discovering Spatial Co-location Patterns: A Summary of Results*** ,SSTD, 2001
- [Tan et al, 2001], P. Tan and M. Steinbach and V. Kumar and C. Potter and S. Klooster and A. Torregrosa, ***Finding Spatio-Temporal Patterns in Earth Science Data, KDD Workshop on Temporal Data Mining, 2001***
- [Tobler, 1970], W. Tobler, ***A Computer Movie Simulating Urban Growth of Detroit Region***, Economic Geography, 46:236-240, 1970
- [Zhang et al, 2003], P. Zhang, Y. Huang, S. Shekhar, and V. Kumar, ***Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries***, SSTD, 2003
- [Zhang et al., 2005], P. Zhang, M. Steinbach, V. Kumar, S. Shekhar, P. Tan, S. Klooster, C. Potter, ***Discovery of Patterns of Earth Science Data Using Data Mining***, to appear in Next Generation of Data Mining Applications, edited by Mehmed M. Kantardzic and Jozef Zurada, IEEE Press, 2005