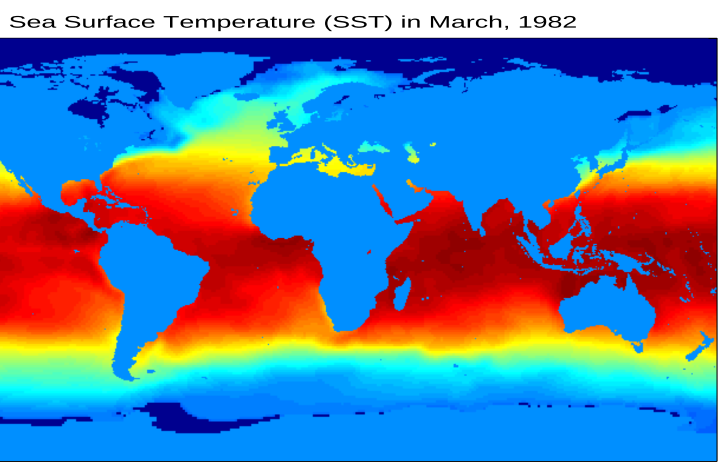
What's Special about Spatial Data Mining?

Shashi Shekhar
Pusheng Zhang
Yan Huang
Ranga Raju Vatsavai

Department of Computer Science and Engineering
University of Minnesota



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

⋆ 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?

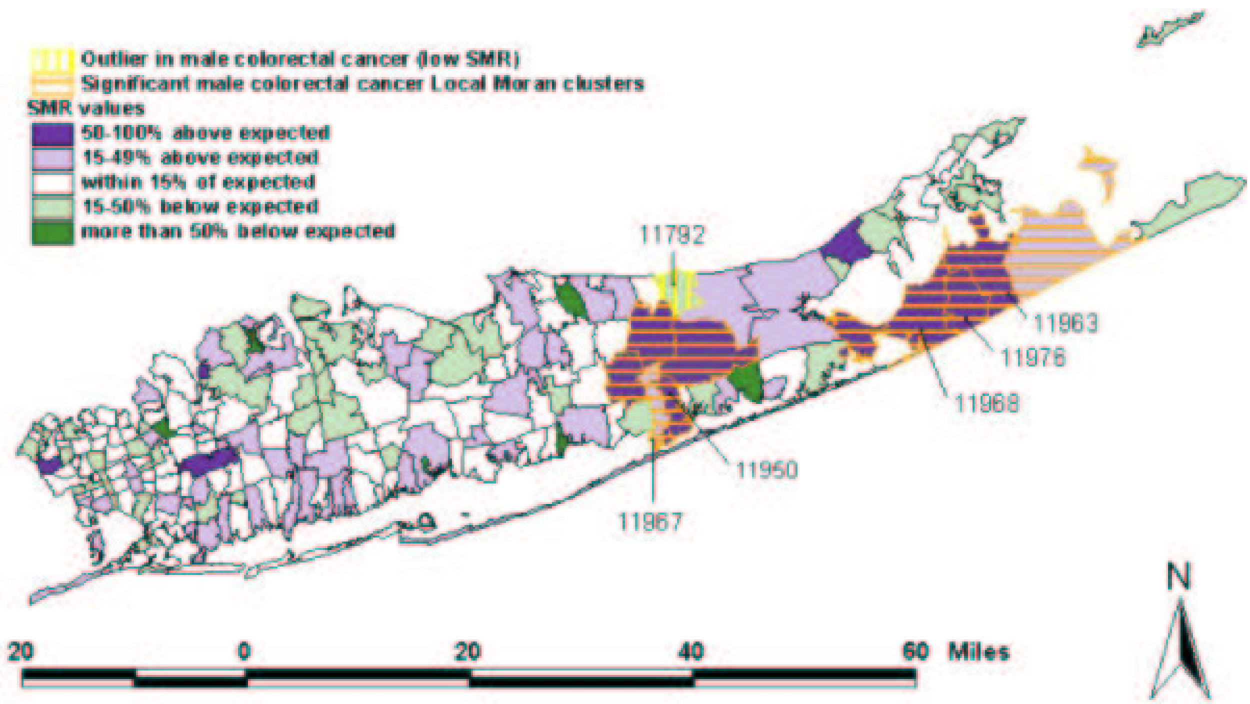
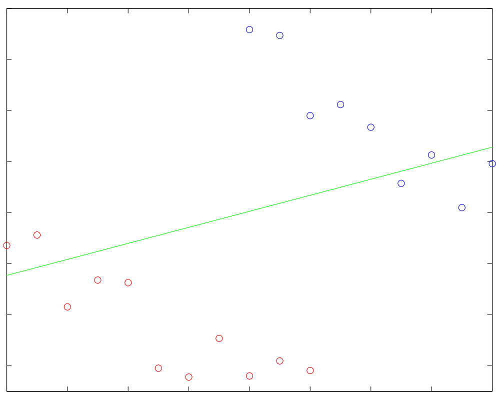


Figure 1: Geographic distribution of male colorectal cancer in Long Island, New York(in courtesy of TerraSeer)

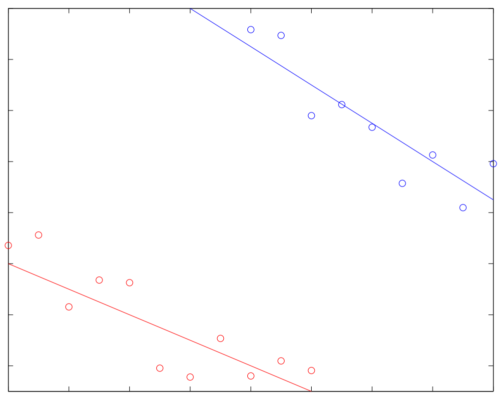
Spatial Slicing

⋆ Spatial heterogeneity

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



(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

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
  - space is an important **surrogate variable**
  - critical to hypothesis formation about relationships among variables

Domain	Spatial Observations	Hypothesis	Science
Social Science	central places, e.g., cities	power law	observed in social networks
Animal Behavior	co-occurrence(pant-hoot, food-bout) in space and time	chimpanzees use pant-hoot to share abundant food sources	observed in Gombe dataset
Physical Science	co-location(water in Colorado Springs, dental health)	water carries elements related to dental health	fluoride and dental health
Physical Science	1854, London: co-location(water pump, cholera)	water carries cholera agents	1883: germ theory

Spatial Data Mining (SDM)

⋆ The process of discovering

- interesting,useful, non-trivial patterns
- 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

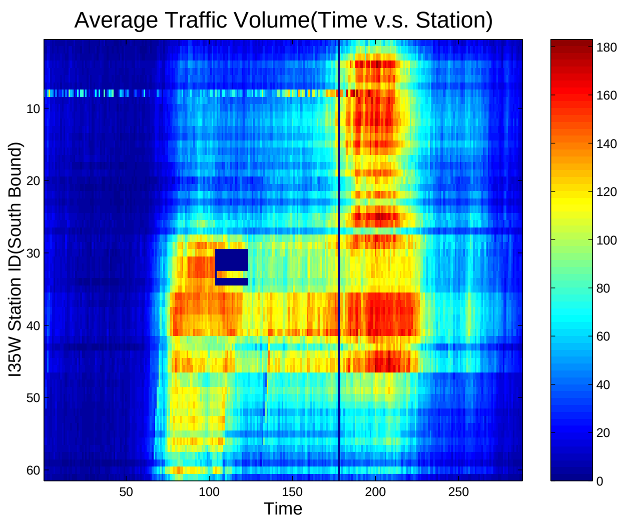
Example Spatial Pattern: Spatial Cluster

⋆ The 1854 Asiatic Cholera in London

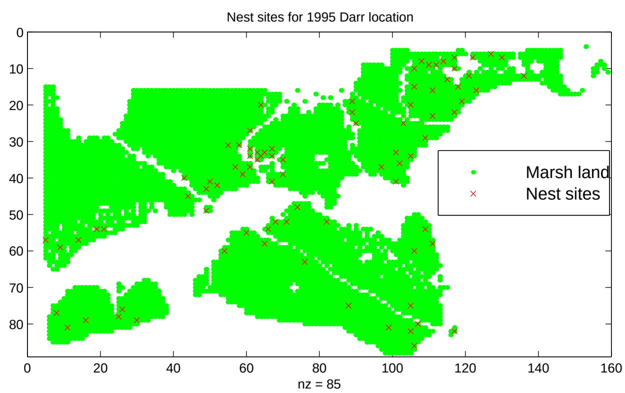
Example Spatial Pattern: Spatial Outliers and Predictive Models

---

* Spatial Outliers



* Predictive Models

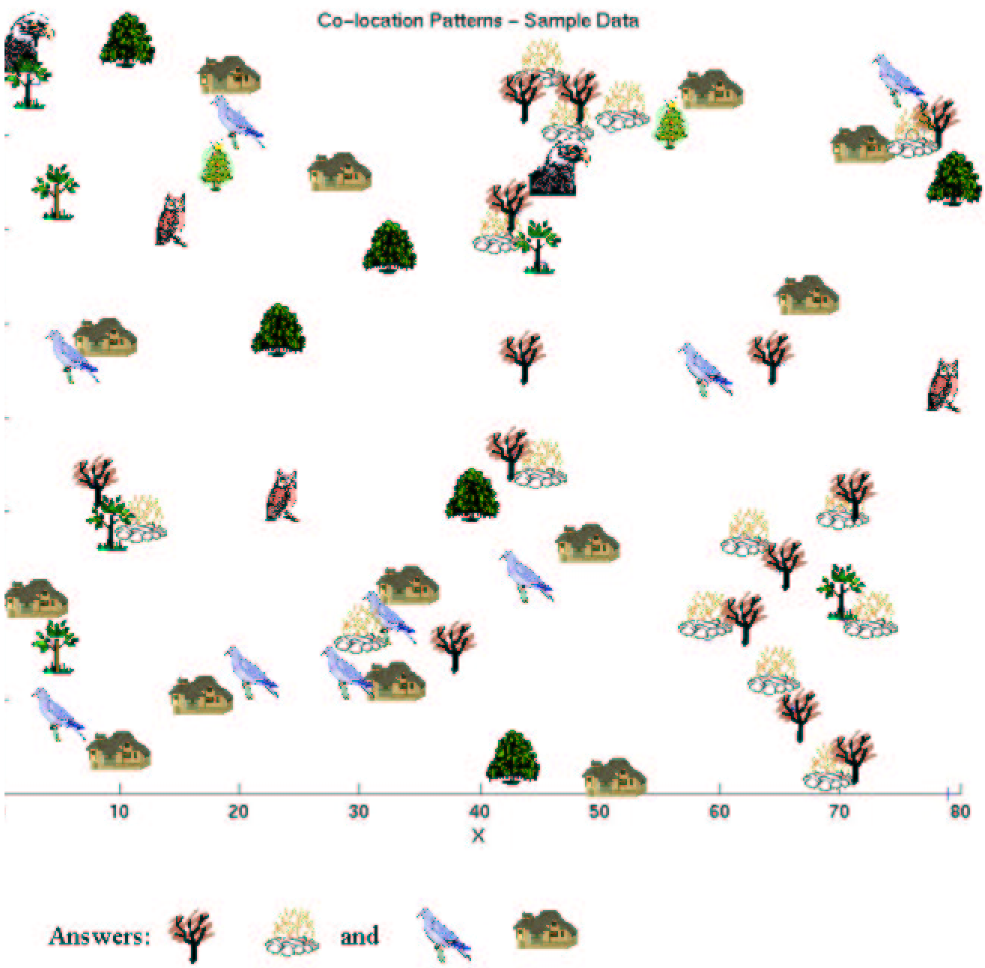


Example Spatial Pattern: Co-locations (backup)

⋆ Given:

- A collection of different types of spatial events

⋆ Illustration



⋆ Find: Co-located subsets of event types

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

General Approaches in SDM

⋆ Materializing spatial features

- e.g., spatial association rule mining[Koperski, Han, 1995]
- commercial tools: e.g., Arc/Info family

⋆ Spatial slicing

- e.g., association rule with support map[P. Tan et al]

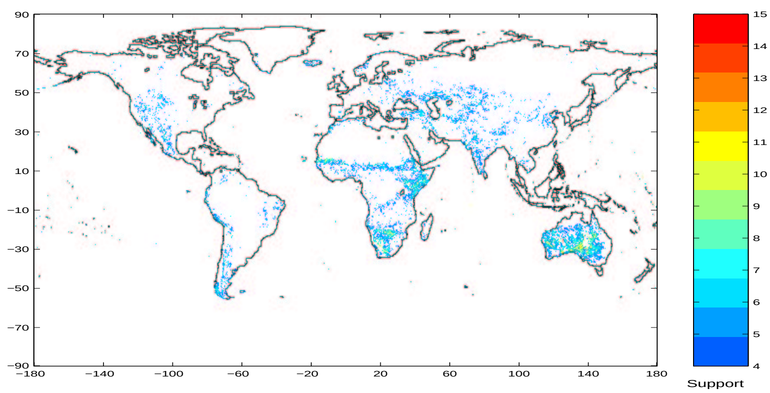


Figure 2: Association rule with support map(FPAR-high → NPP-high)

- commercial tools: e.g.,Matlab, SAS, R, Splus

⋆ Customized spatial techniques

- e.g., MRF-based Bayesian Classifier
- commercial tools
  - e.g.,Splus spatial/R spatial/terraseer + customized codes

Overview

$\Rightarrow$ Input

$\star$ Statistical Foundation

$\star$ Output

$\star$ Computational process

Overview of Input

⋆ Data

- Table with many columns(attributes)

tid	f_1	f_2	...	f_n

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 3: 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

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 3: 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

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]

- 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

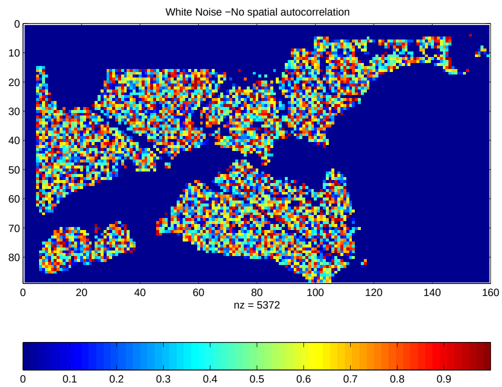
		Point Process	Lattice	Geostatistics
raster			$\surd$	$\surd$
vector	point	$\surd$	$\surd$	$\surd$
	line			$\surd$
	polygon		$\surd$	$\surd$
graph				

Table 4: Data Types and Statistical Models

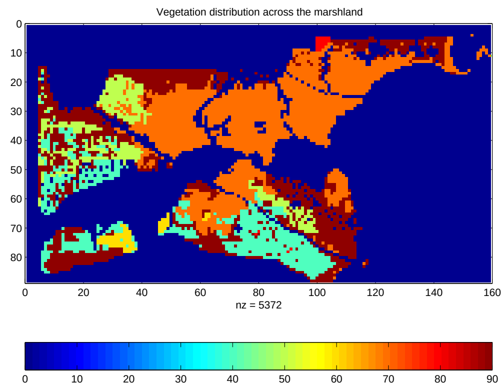
Spatial Autocorrelation(SA)

⋆ 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 4: 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, $\cdots$

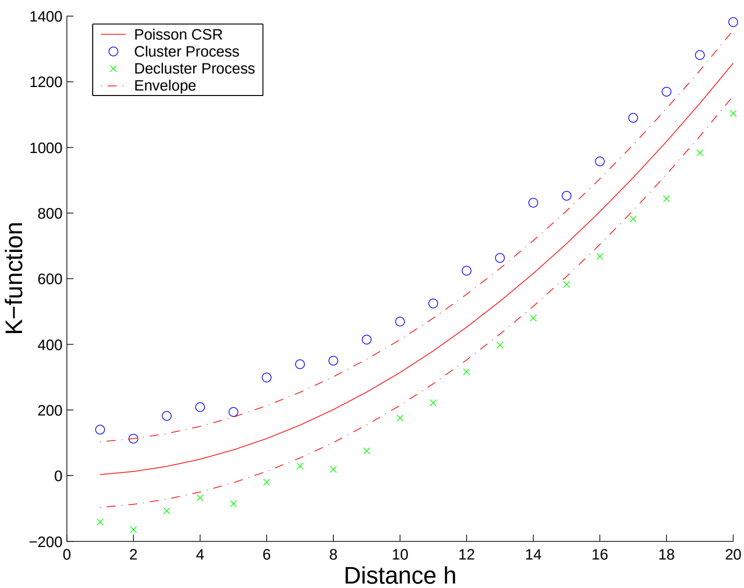
Spatial Autocorrelation: Distance-based Measure

⋆ 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

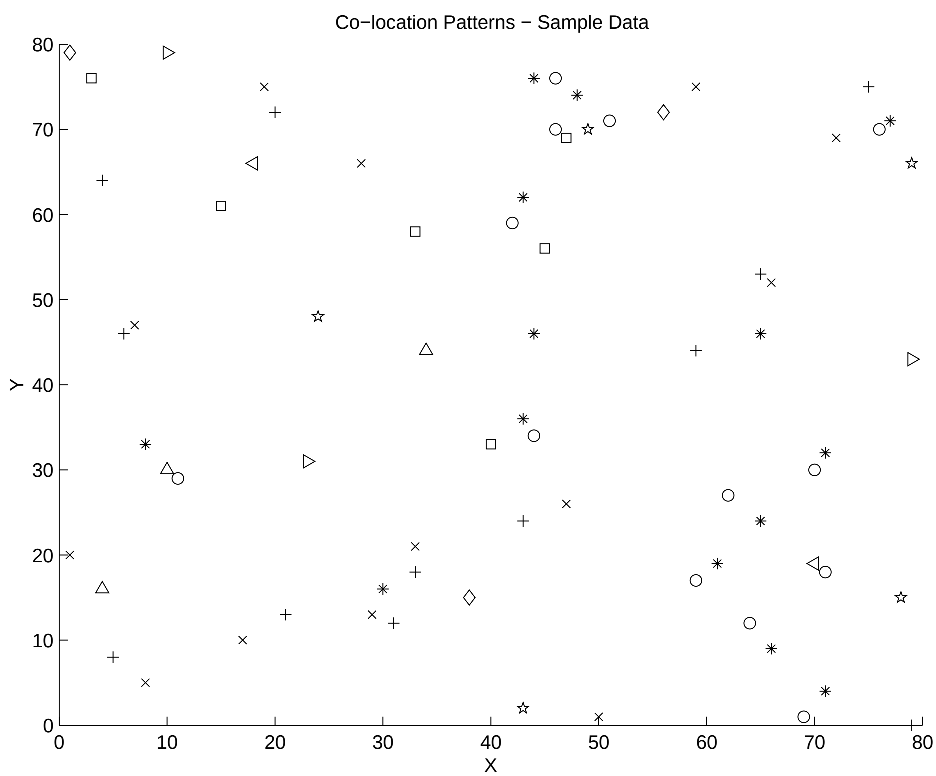


Figure 5: Example Data (o and * ; x and +)

Illustration of Cross-Correlation

⋆ Illustration of Cross K-Function for Example Data

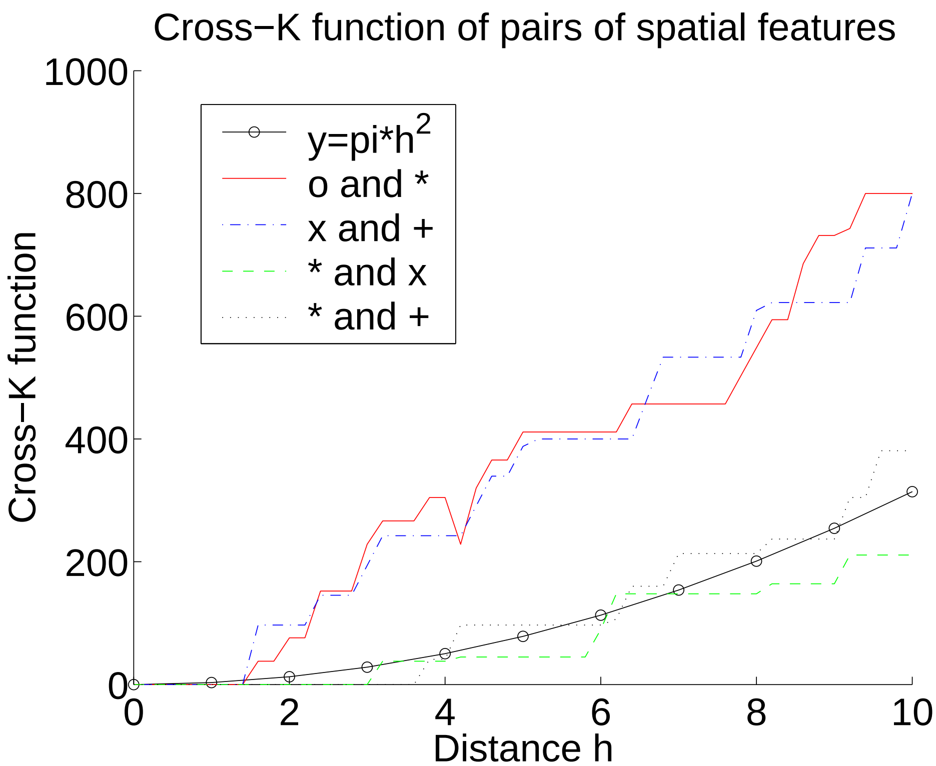


Figure 6: Cross K-function for Example Data

Overview

$\surd$ Input

$\surd$ Statistical Foundation

$\Rightarrow$ Output

$\star$ Computational process

Overview of Data Mining Output

⋆ Supervised Learning: Prediction

- Classification
- Trend

⋆ Unsupervised Learning:

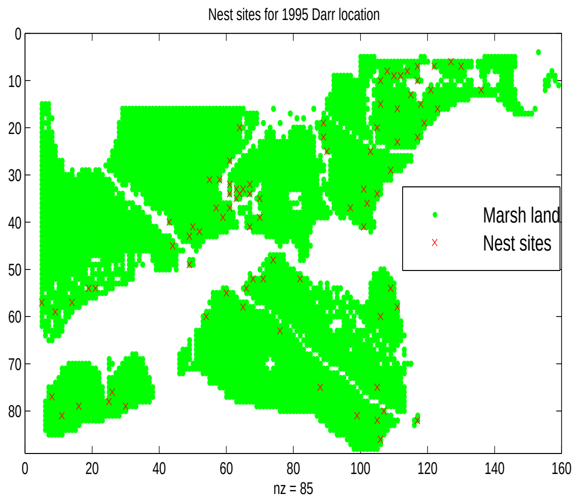
- Clustering
- Outlier Detection
- Association

⋆ Input Data Types vs. Output Patterns

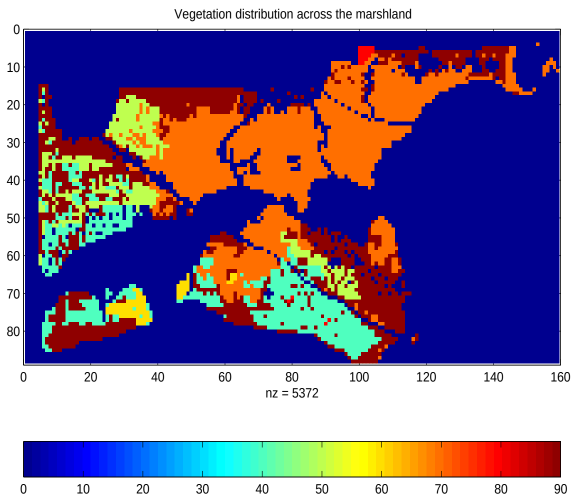
Patterns	Point Process	Lattice	Geostatistics
Prediction	$\surd$	$\surd$	
Trend			$\surd$
Clustering	$\surd$	$\surd$	
Outliers	$\surd$	$\surd$	$\surd$
Associations	$\surd$	$\surd$	

Table 5: Output Patterns vs. Statistical Models

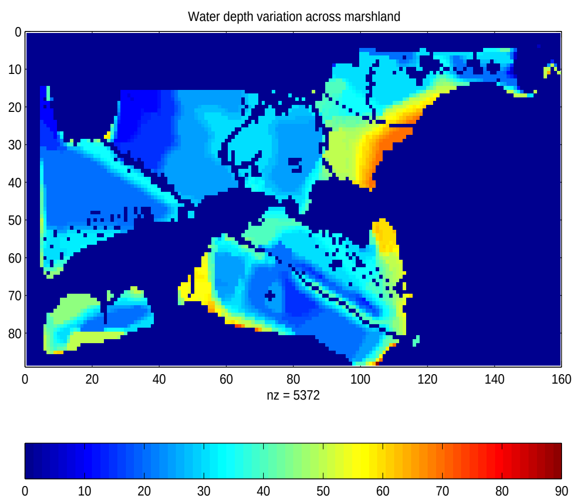
Illustrative Application to Location Prediction (Backup)



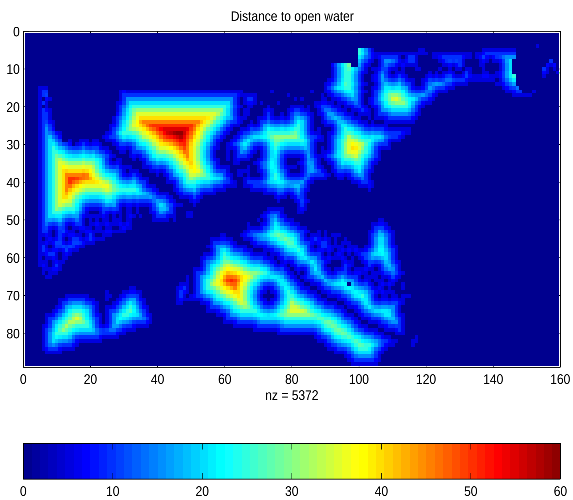
(a) Nest Locations



(b) Vegetation



(c) Water Depth



(d) Distance to Open Water

Prediction and Trend

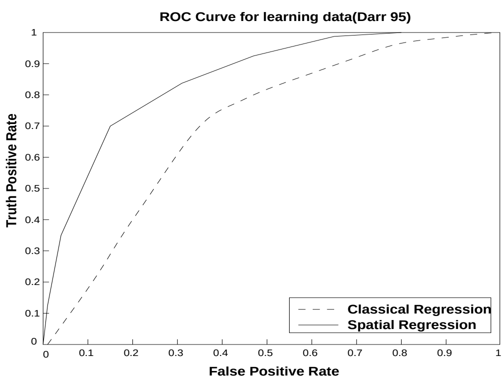
⋆ Prediction

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

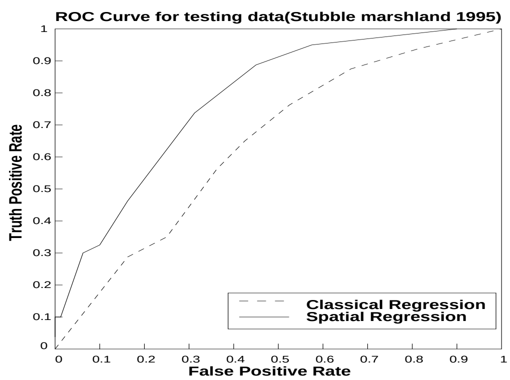
Classical	Spatial
$\mathbf{y} = \mathbf{X}\beta + \epsilon$	$y = \rho W y + X\beta + \epsilon$
$Pr(C_i \mid X) = \frac{Pr(X \mid C_i) Pr(C_i)}{Pr(X)}$	$Pr(c_i \mid X, C_N) = \frac{Pr(c_i) * Pr(X, C_N \mid c_i)}{Pr(X, C_N)}$

Table 6: Prediction Models

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



(e) ROC curves for learning



(f) ROC curves for testing

Figure 7: (a) Comparison of the classical regression model with the spatial autoregressive model on the Darr learning data. (b) Comparison of the models on the Stubble testing data.

Prediction and Trend

⋆ Open Problems

- Estimate W for SAR
- Spatial interest measure: e.g., avg dist(actual, predicted)

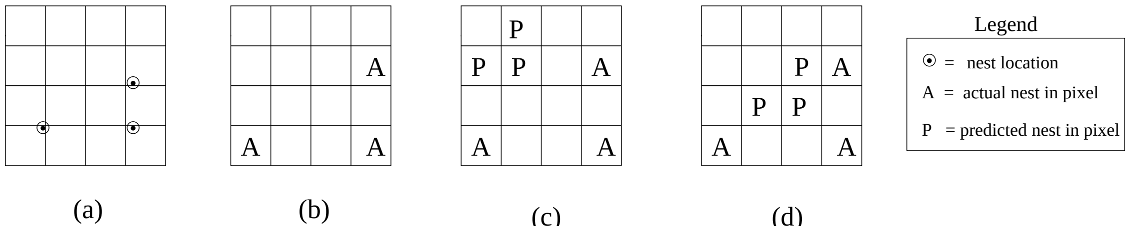


Figure 8: An example showing different predictions: (a)The actual sites, (b)Pixels with actual sites, (c)Prediction 1, (d)Prediction 2. Prediction 2 is spatially more accurate than 1.

Clustering

⋆ Clustering: Find groups of tuples

⋆ Statistical Significance

- Complete spatial randomness, cluster, and decluster

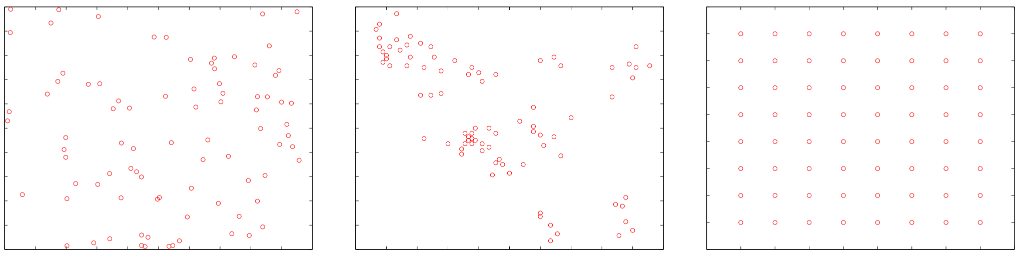


Figure 9: Inputs: Complete Spatial Random (CSR), Cluster, and Decluster

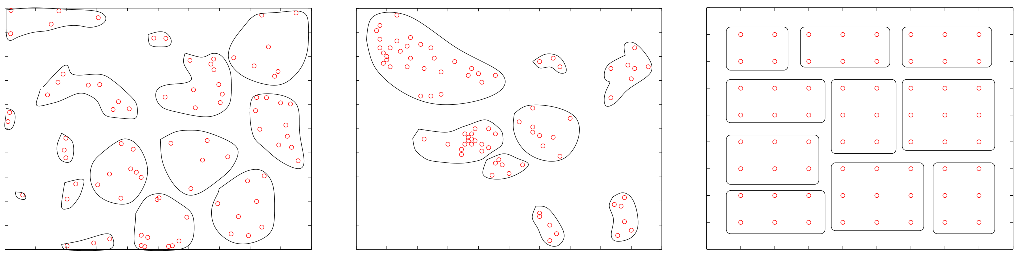


Figure 10: Classical Clustering

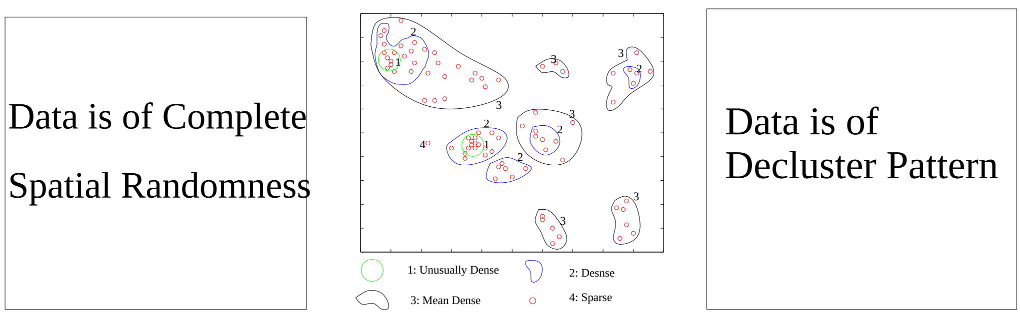


Figure 11: Spatial Clustering

Clustering

⋆ Similarity Measures

- Non-spatial: e.g., soundex
- Classical clustering: Euclidean, metric, graph-based
- Topological: **neighborhood EM**
  - Implicitly based on locations
- Interest measure:
  - spatial continuity
  - cartographic generalization
  - unusual density
  - keep nearest neighbors in common cluster

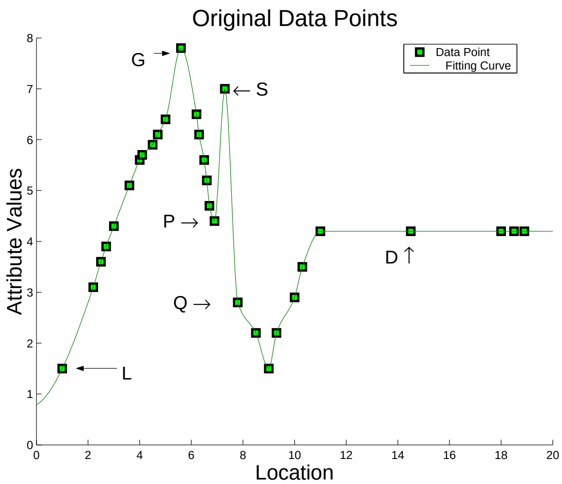
Outlier Detection

⋆ Spatial Outlier Detection

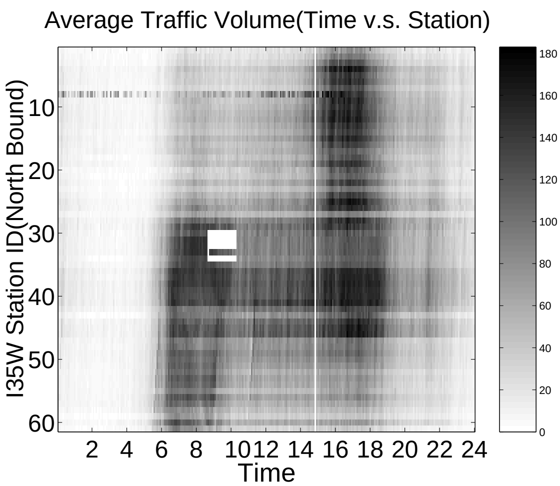
- Finding anomalous tuples
- Global and spatial outlier
- Detection Approaches
  - Graph-based outlier detection: variogram, moran scatter plot
  - Quantitative outlier detection: scatter plot, and z-score

⋆ Location-awareness

- All tuples/No tuple: classical
- Some tuple: locations for neighborhood and non-spatial attributes for difference test



(a) Outliers in Example Data



(b) Outliers in Traffic Data

Association

⋆ 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

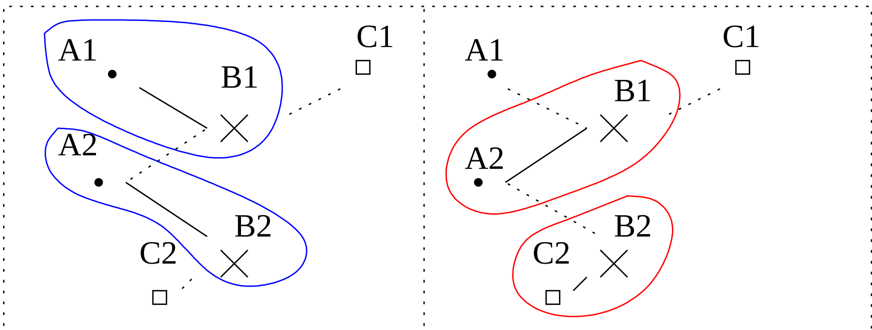


Figure 12: Different Transactionizing Schemes

⋆ Location-awareness

- All tuples: co-location mining
- No tuple: classical association rule mining
- Some tuple: future work

Output Patterns

⋆ Output Patterns vs. Input

SDM Techniques	Vector Data Point	Vector Data Line	Vector Data Polygon	Raster Data
classification	$\surd$	-	$\surd$	$\surd$
association	$\surd$	-		$\surd$
clustering	$\surd$	-		$\surd$
outlier detection	$\surd$			$\surd$

Table 7: Output Patterns vs. Input

⋆ Output Patterns vs. Interest Measures

	Traditional Non-spatial	Spatial	Mixture
Predictive Model	Classification accuracy	Spatial accuracy, e.g., avg dist(actual site, predicted site)	Future Work
Cluster	Low coupling and high cohesion in feature space	Spatial continuity, unusual density, cartographic generalization	Future Work
Outlier	Different from population or neighbors in feature space	Geographically distant from neighbors	Significant attribute discontinuity in geographic space
Association	Subset prevalence, $Pr[B \in T \mid A \in T]$, Correlation: e.g.,	Clique prevalence $Pr[B \in N(L) \mid A\, at\, L]$ Cross K-Function	Future Work

Table 8: Output Patterns vs. Interest Measures

Output Patterns vs. Location Awareness

⋆ Output Patterns vs. Location Awareness

- No awareness: no location info
- Total awareness: location info available for all tuples
- Partial awareness: location info missing for some tuples

	No Awareness	Total Awareness	Partial Awareness
Prediction	Decision tree, nearest neighbor, Bayesian classifier, neural network, regression	kriging, MRF Bayesian classifier, self-organizing map, spatial autoregressive model	future work
Clustering	EM in feature space, k-means, density-based, graph-based	Neighborhood EM	future work
Outliers	Neighbor def: feature domain Difference test def: feature domain	Neighbor def: geographic domain Difference test def: feature domain	future work
Association	Association rules	Co-location	future work

Table 9: Output vs. Location Awareness

Overview

- √ Input
- √ Statistical Foundation
- √ Output
- ⇒ 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 10: 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 13: Global Influence of El Nino during the Northern Hemisphere Winter(D: Dry; W:Warm; R:Rainfall)

- ⋆ Challenge: high dim(e.g., 600) feature space
- ⋆ Computational Efficiency Idea
  - Observation: Spatial autocorrelation
  - Spatial indexing to organize locations
    - – Top-down tree traversal is a strong filter
    - – Spatial join query: filter-and-refine
      - ∗ 50 year long monthly data on 67k land locations and 100k ocean locations
      - ∗ save 40% to 98% computational cost at $\theta = 0.3$ to 0.9

Summary

⋆ What's Special About Spatial Data Mining?

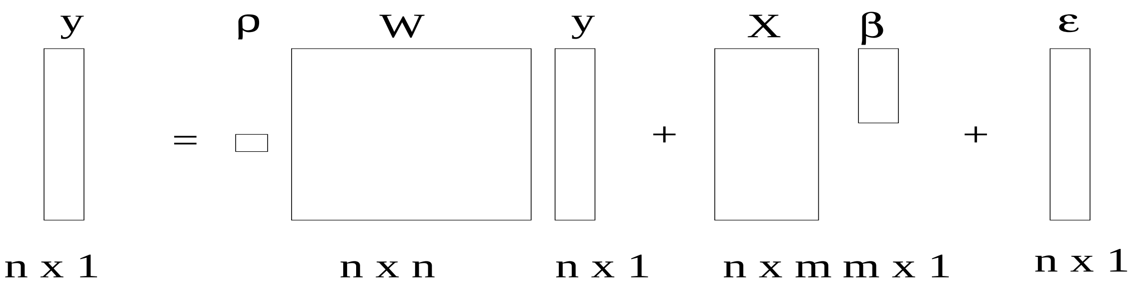
- Input Data
- Statistical Foundation
- Output Patterns
- Computational Process

	Classical DM	Spatial DM
Input	All explicit, simple types	often Implicit relationships, complex types
Stat Foundation	Independence of samples	spatial autocorrelation
Output	Interest Measures: set-based	Location-awareness
Computational Process	Combinatorial optimization	Computational efficiency opportunity
		Spatial autocorrelation, plane-sweeping
	Numerical alg.	New complexity: SAR, co-location mining
		Estimation of anisotropic W is nontrivial

Table 11: Summary of Spatial Data Mining

⋆ A Hard Problem:

- Estimate W besides ρ and β for $y = \rho W y + X\beta + \epsilon$



Research Needs

⋆ Research Issues:

- Classical DM techniques vs. SDM techniques
- Statistical interpretation models for spatial patterns
  - e.g., co-location and Ripley's K-function
- Spatial interest measures: e.g., spatial accuracy
- Modeling semantically rich spatial properties
- Visualization
- Improving computational efficiency
- Preprocessing

Conclusions

⋆ Applications of Spatial Data Mining

- Businesses, e.g. logistics, marketing, ...
- Government - almost all branches e.g. defense, public safety, ...

⋆ Rationale for spatial data mining

- Simpson's paradox and 2nd law of Geography
- Space as a surrogate variable
  - Ex. co-location(water, cholera) led to Germ theory
- Unique properties of spatial data, e.g. auto-correlation

⋆ Approaches to mine spatial data

- A. Traditional DM methods + spatial feature selection
  + Easy to start with
  − But results are weak due to spatial-autocorrelation etc.
- B. Novel spatial DM methods
  + Better models unique properties of spatial data
  + Often improves results
  + Sometime reduces computation costs

References

- ⋆ Email: shekhar@cs.umn.edu
- ⋆ More – http://www.cs.umn.edu/research/shashi-group
- ⋆ References

- [Cressie, 1991], N. Cressie, ***Statistics for Spatial Data***, John Wiley and Sons, 1991
- [Miller, Han, 2001], H. Miller and J. Han(eds), ***Geographic Data Mining and Knowledge Discovery***, Taylor and Francis, 2001
- [Kazar at al., 2003], B. Kazar, S. Shekhar, and D. Lilja, ***Parallel Formulation of Spatial Auto-Regression***, Army High-Performance Computing Research Center (AHPCRC) Technical Report no. 2003-125, August 2003
- [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
- [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

References

⋆ References

- [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
- [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 at al, 2003], P. Zhang, Y. Huang, S. Shekhar, and V. Kumar, ***Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries***, SSTD, 2003

Spatial Databases: A Tour