Spatial Data Mining: Accomplishments and Research Needs

Shashi Shekhar

Department of Computer Science and Engineering

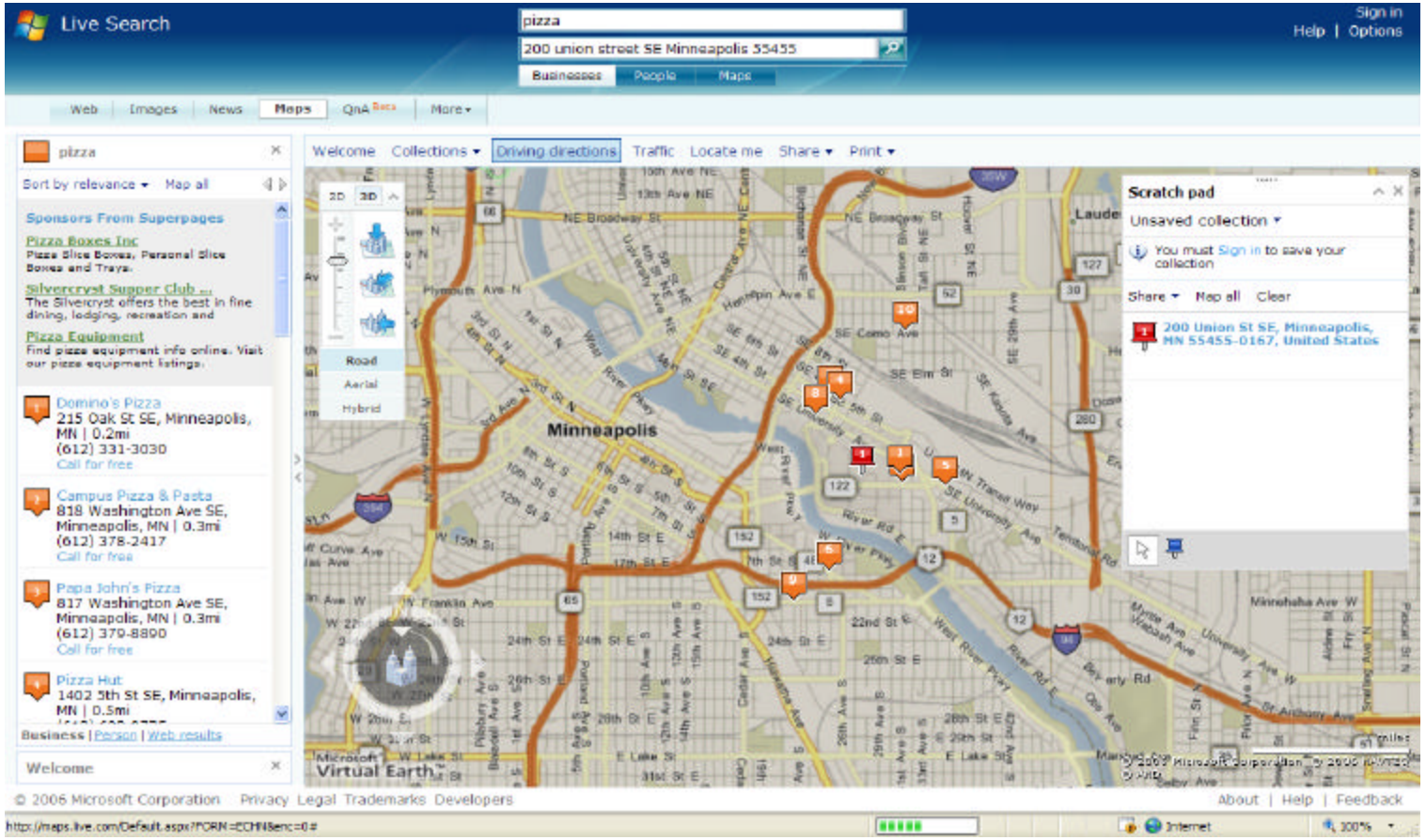
University of Minnesota

Why Data Mining?

- Holy Grail - Informed Decision Making
- Lots of Data are Being Collected
  - Business - Transactions, Web logs, GPS-track, …
  - Science - Remote sensing, Micro-array gene expression data, …
- Challenges:
  - Volume (data) >> number of human analysts
  - Some automation needed
- Data Mining may help!
  - Provide better and customized insights for business
  - Help scientists for hypothesis generation

Spatial Data

- Location-based Services
  - E.g.: MapPoint, MapQuest, Yahoo/Google Maps, …



Courtesy: Microsoft Live Search (http://maps.live.com)

Spatial Data

- In-car Navigation Device

Emerson In-Car Navigation System (Courtesy: 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 pattern families
  - Spatial outlier, discontinuities
  - Location prediction models
  - Spatial clusters
  - Co-location patterns
  - …

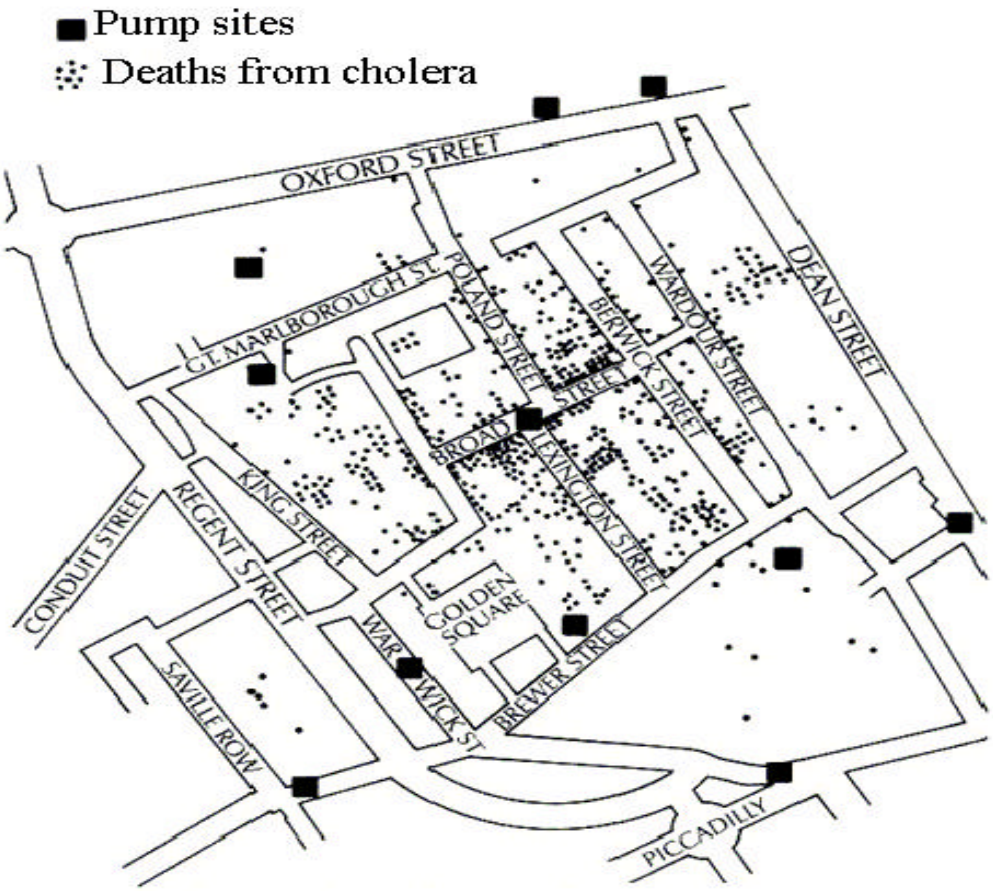
Spatial Data Mining and Science

- Understanding of a physical phenomenon
  - Though, final model may not involve location
    - Cause-effect e.g. Cholera caused by germs
  - Discovery of model may be aided by spatial patterns
    - Many phenomenon are embedded in space and time
    - Ex. 1854 London – Cholera deaths clustered around a water pump
    - Spatio-temporal process of disease spread => narrow down potential causes
    - Ex. Recent analysis of SARS

- Location helps bring rich contexts
  - Physical: e.g., rainfall, temperature, and wind
  - Demographical: e.g., age group, gender, and income type
  - Problem-specific, e.g. distance to highway or water

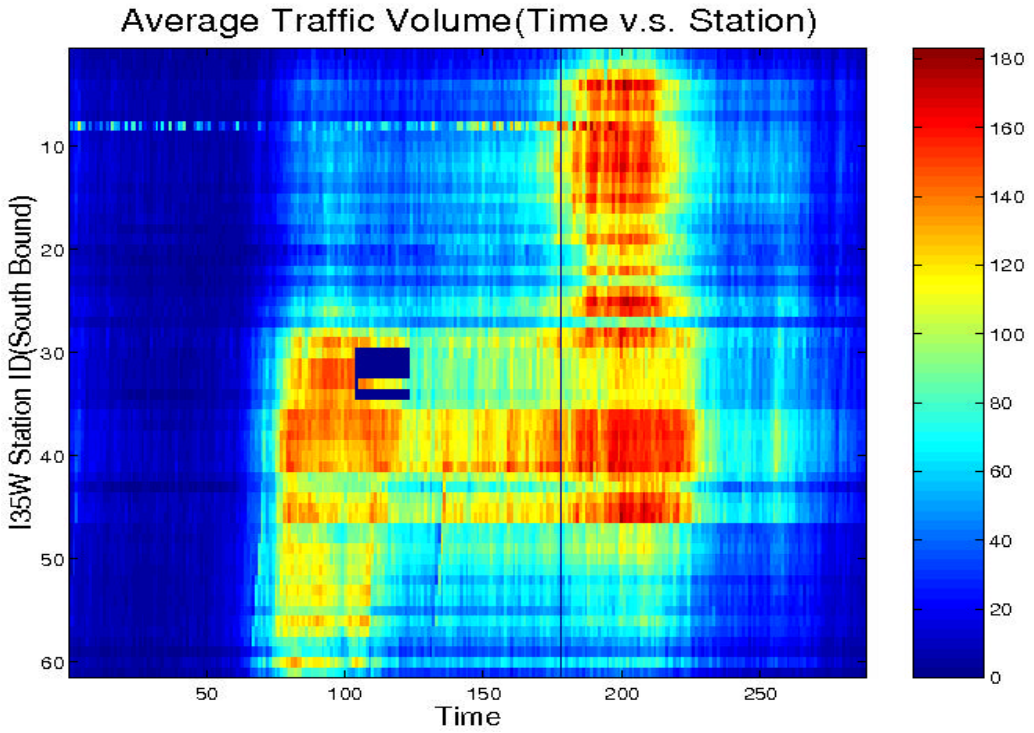
Example Pattern: Spatial Cluster

- The 1854 Asiatic Cholera in London



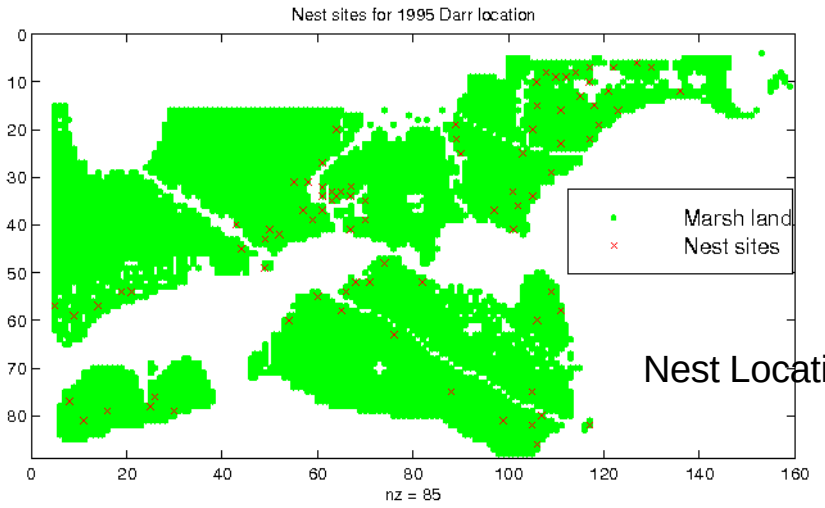
Example Pattern: Spatial Outliers

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

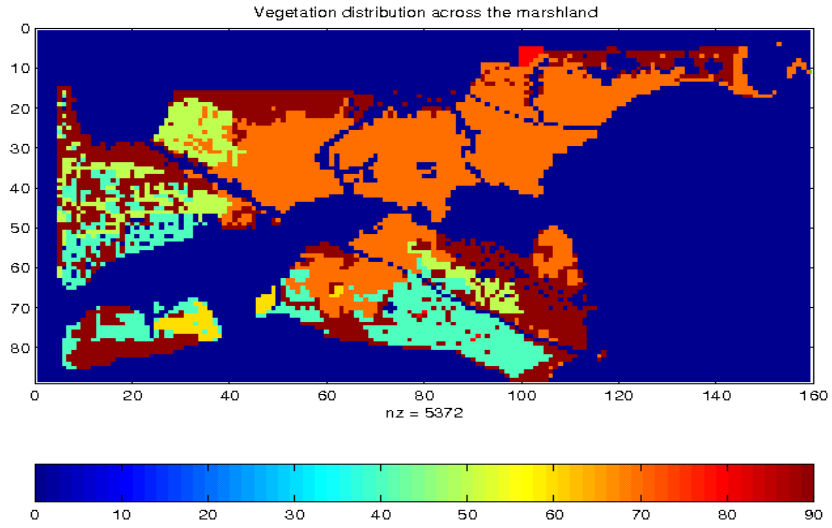


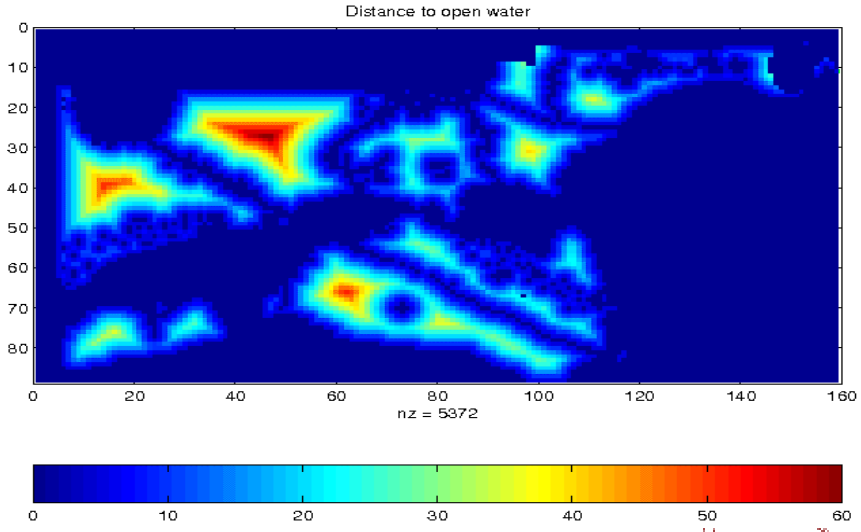
Example Pattern: Predictive Models

- Location Prediction:
  - Predict Bird Habitat Prediction
  - Using environmental variables



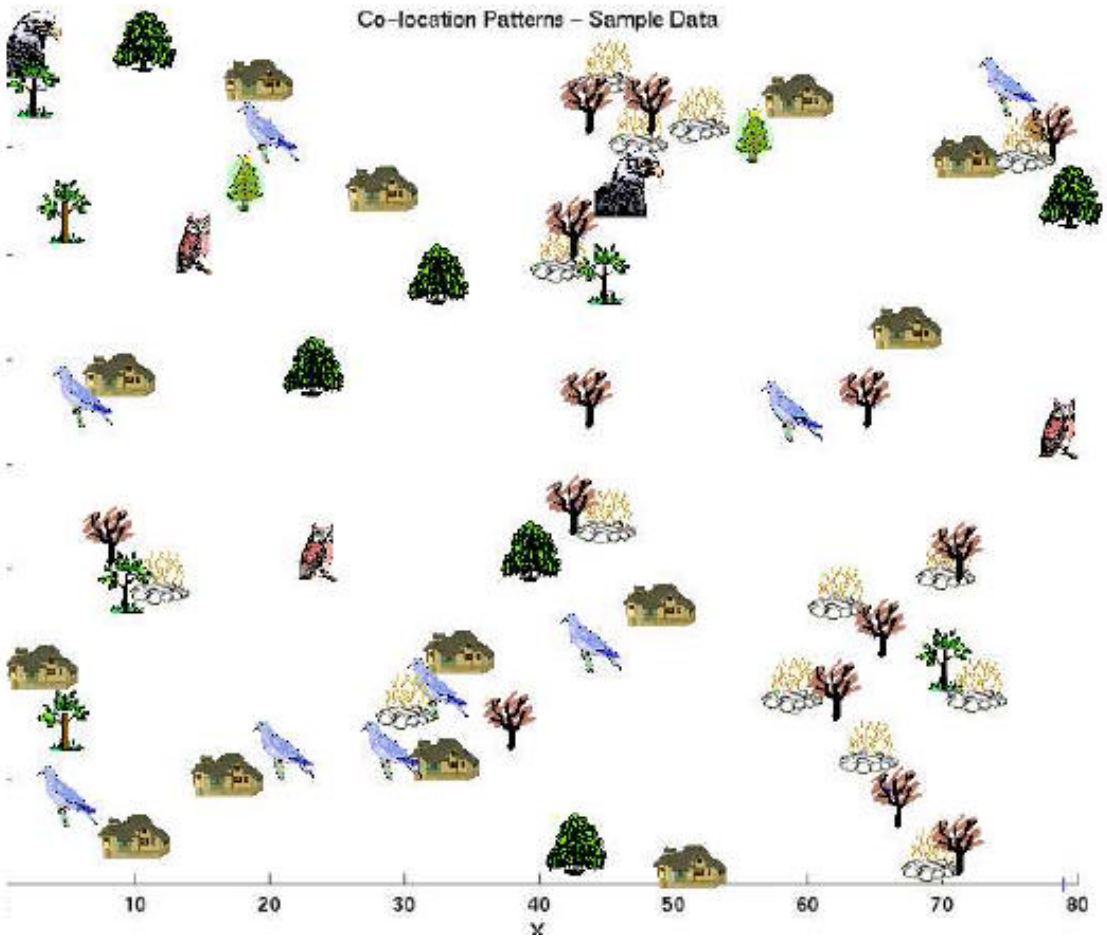
Nest Locations





Example Patterns: Co-locations

- Given: A collection of different types of spatial events
- Find: Co-located subsets of event types



Answers:   and  

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

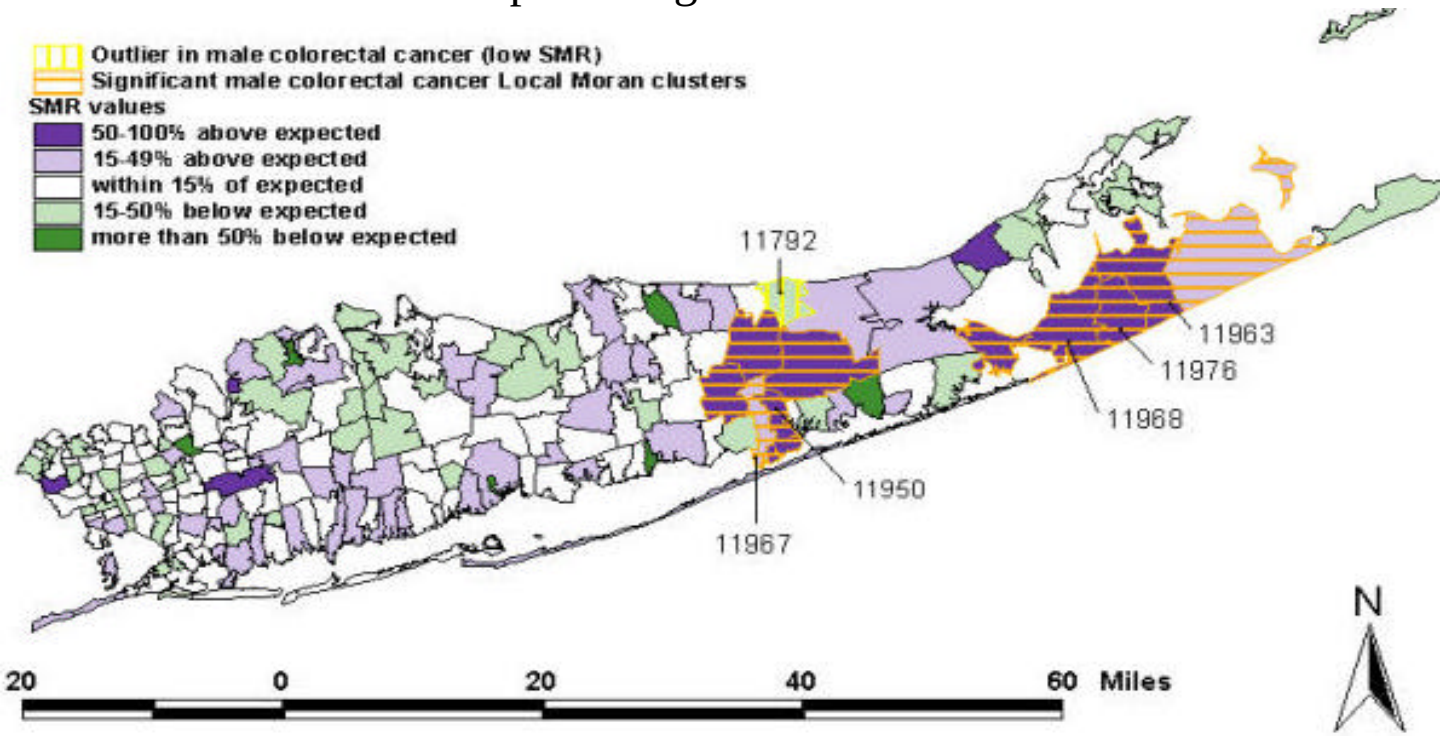
Application Domains

- 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 forcing 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



Geographic distribution of male colorectal cancer in Long Island, New York (Courtesy: 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
    - $\Pr[\text{person } p \text{ selects choice } c] \; a \; \text{perceived\_utility}(\forall c \in S, p)$
    - $\text{perceived\_utility}(\text{store } c, \text{person } p) = f(\text{square} - \text{footage}(c), \text{distance}(c, p), \text{parameters})$
  - 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

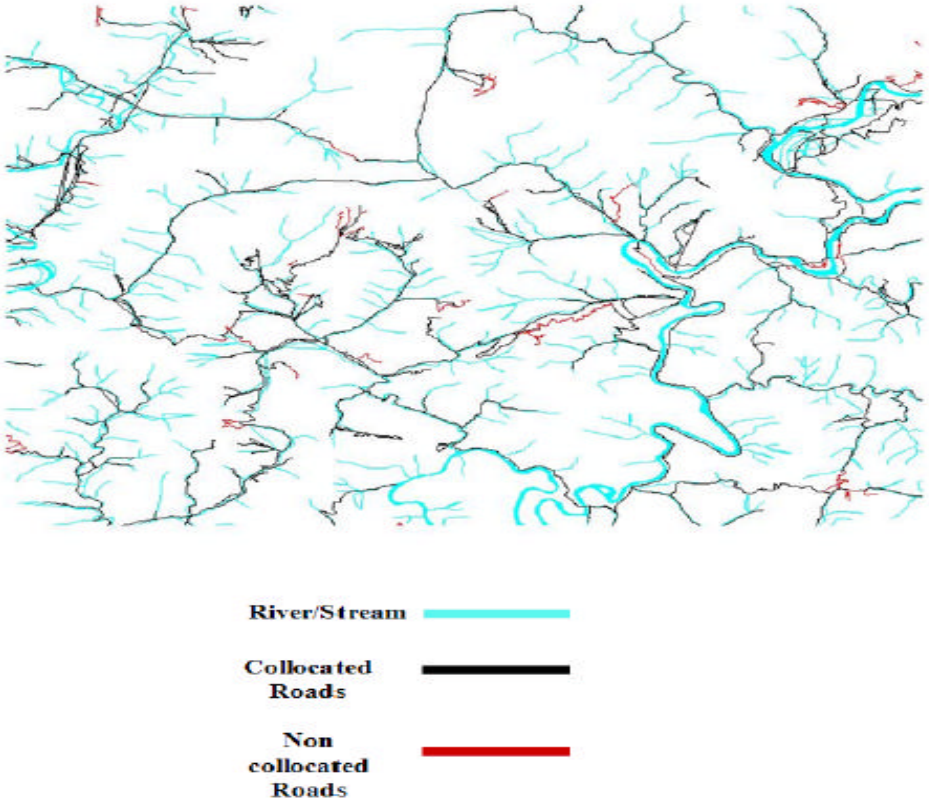
- Sample Questions
  - Which features are anomalous?
  - Which layers are related?
  - How can the gaps be filled?
- Korea Data
  - Latitude 37deg15min to 37deg30min
  - Longitude 128deg23min51sec to 128deg23min52sec
- Layers
  - Obstacles (Cut, embankment, depression)
  - Surface drainage (Canal, river/stream, 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

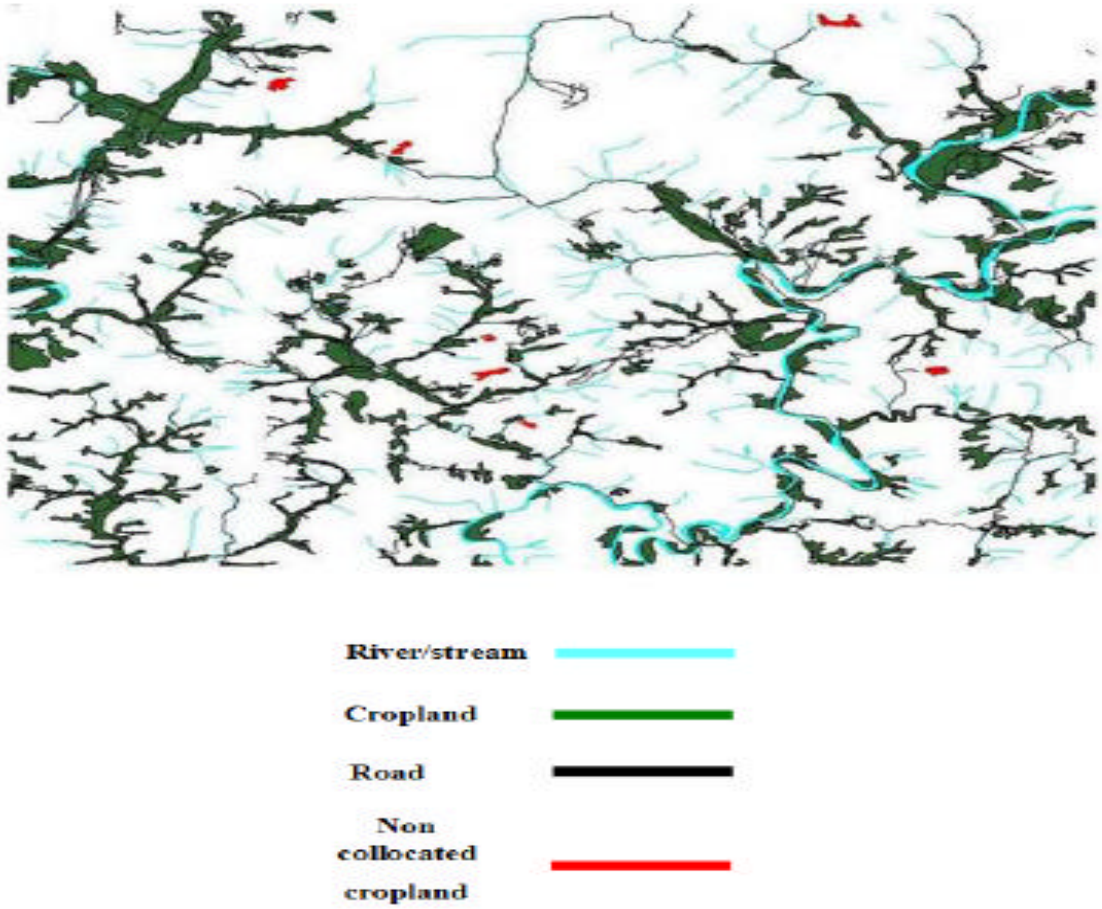
- Interestingness
  - Patterns to Non-Specialist vs. Exceptions to Specialist
- Road-River/Stream Colocation



Road-River Colocation Example
(Korea database, Courtesy: Architecture Technology Corporation)

A Complex Colocation Example

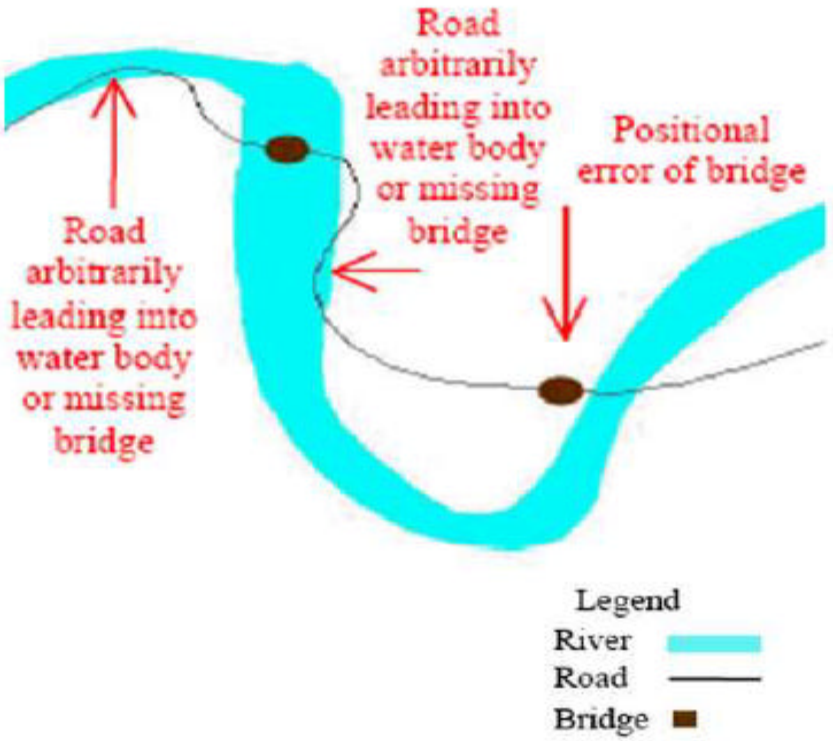
- Cropland colocated with river, stream or road



Complex Colocation Example
(Korea dataset, Courtesy: Architecture Technology Corporation)

Outliers in Example Data

- Outlier detection
  - Extra/erroneous features
  - Positional accuracy of features
  - Predict mislabeled/misclassified features
- Examples
  - Cropland not close to river and road
  - Overlapping road and river
    - without bridge



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
- Trends

Overview of Input

- Data
  - Table with many columns (attributes)

tid	f_1	f_2	…	f_n
0001	3.5	120	…	Yes
0002	4.0	121	…	No

Example of Input Data

    - 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

Raster Data for UMN Campus
Courtesy: UMN

Vector Data for UMN Campus
Courtesy: 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

OGC Model

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

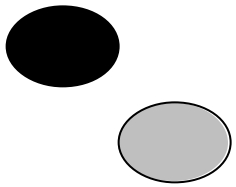
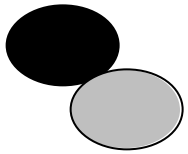
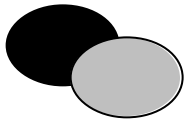
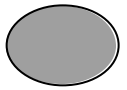
Operator Type	Operator Name
Basic Function	SpatialReference, Envelope, Boundary, Export, IsEmpty, IsSimple
Topological/Set Operations	Equal, Disjoint, Intersect, Touch, Cross, Within, Contains, Overlap
Spatial Analysis	Distance, Buffer, ConvexHull, Intersection, Union, Difference, SymmDiff

Examples of Operations in OGC Model

OGIS – Topological Operations

- Topology
  - 9-intersections using
    - Interior
    - boundary
    - exterior

$$\begin{pmatrix} (A^{o} \cap B^{o}) & (A^{o} \cap \partial B) & (A^{o} \cap B^{-}) \\ (\partial A \cap B^{o}) & (\partial A \cap \partial B) & (\partial A \cap B^{-}) \\ (A^{-} \cap B^{o}) & (A^{-} \cap \partial B) & (A^{-} \cap B^{-}) \end{pmatrix}$$

Topological Relationship	 disjoint	 meet	 overlap	 equal
9-intersection model	$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Mining Implicit Spatial Relationships

- Choices
  - Materialize spatial info + classical data mining
  - Customized spatial data mining techniques
- Example -
  - Distance:
    - Point: Euclidean, Extended objects: buffer-based, Graph: shortest path
  - Transactions: i.e., space partitions
    - Circles centered at reference features, Gridded cells, Voronoi diagram

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

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
  - Spatio-temporal
- Needs for New Research
  - Modeling semantically rich spatial properties
  - Moving objects
  - Spatio-temporal data models

Spatio-Temporal Issues

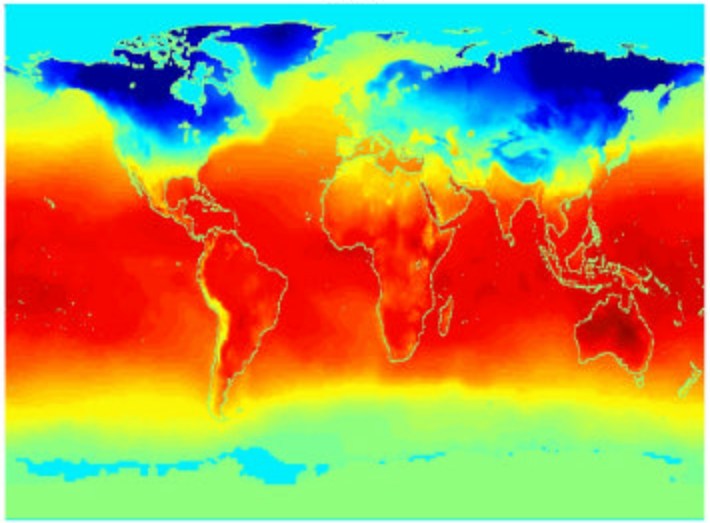
- Spatio-Temporal Data
  - Examples
- Spatio-Temporal Data Models
  - Emerging ideas

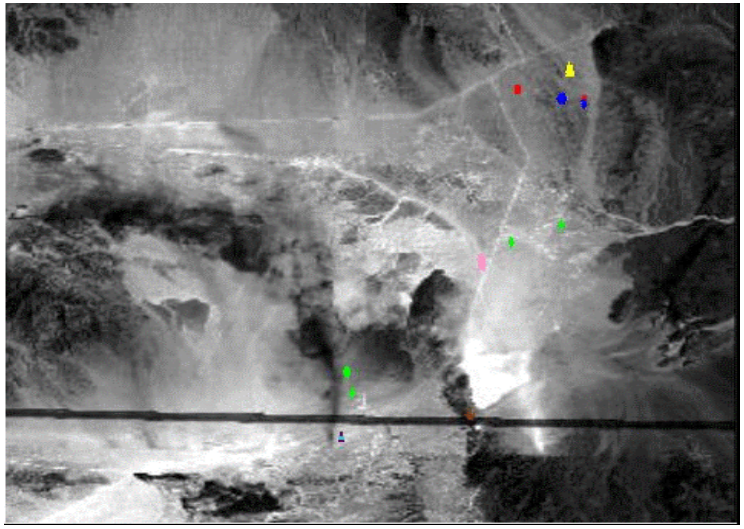
Spatio-Temporal Data

- Spatial Time Series Data
  - Space is fixed
  - Measurement value changes over a series of time
  - E.g. Global Climate Patterns, Army vehicle movement

Average Monthly Temperature

Jan





Army vehicle movement

Manpack stinger (1 Objects)

M2_IFV (3 Objects)

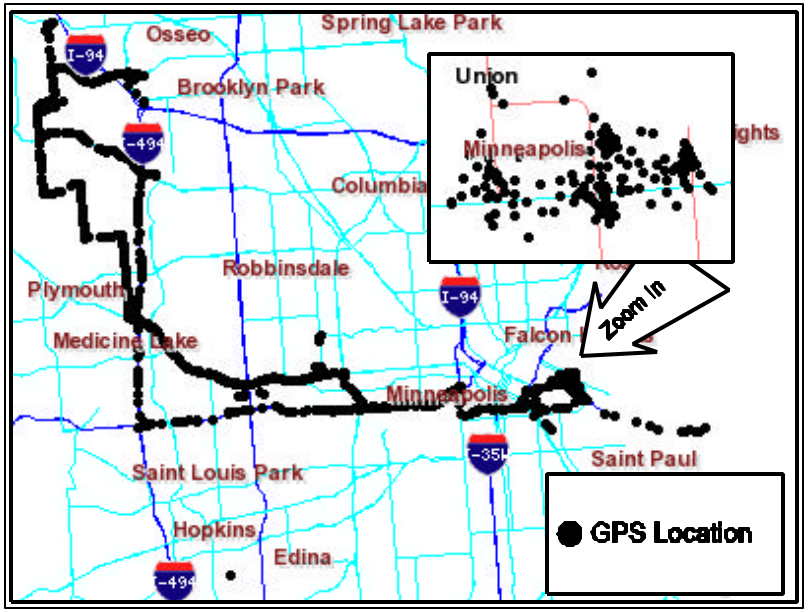
Field_Marker (6 Objects)

T80_tank (2 Objects)

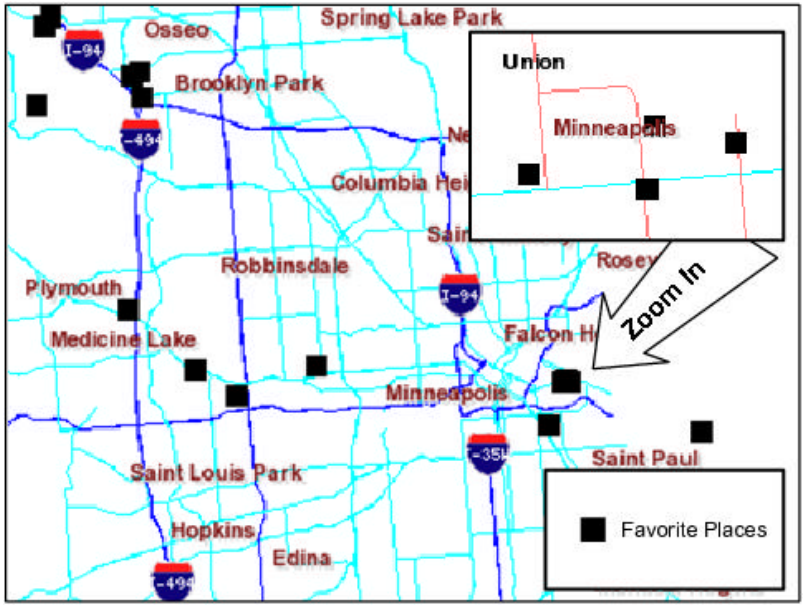
BRDM_AT5 (enemy) (1 Object)

Spatio-Temporal Data

- Moving objects Data
  - Area of interest changes with the moving object
  - E.g. GPS track of a vehicle, Personal Gazetteers



GPS Tracks of a User



Personal Gazetteer

(a personal gazetteer records places meaningful for a specific person)

Spatio-Temporal Data: Modeling

	Spatial	Spatio-Temporal	
		Differentiation	Aggregation
Topology	9-Intersection Matrix, OGIS	*d/dt*(9-Intersection Matrix)	Open Time series of 9-Intersection Matrix
Vector Space	Location OGIS – direction, distance, area, perimeter	Speed, Velocity, *d/dt*(area)	Time series of points, lines, polygons (tracks) Visualized as helixes (linear/angular motion)
Spatial properties of objects		Motion – Translation, Rotation, Deformation *d/dt*(position, orientation, shape)	Open e.g. Helix Track = (t_i, x_i, y_i) – moving object databases
Aspatial properties of objects		*d/dt*(mass)	Time-series of velocities

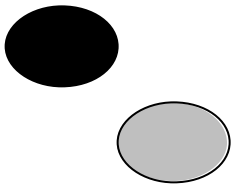
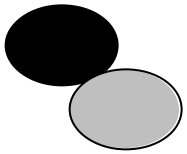
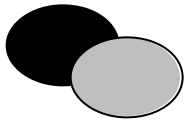
Spatio-Temporal Data: Modeling

- Topology
  - Differentiation

A, *B* - objects

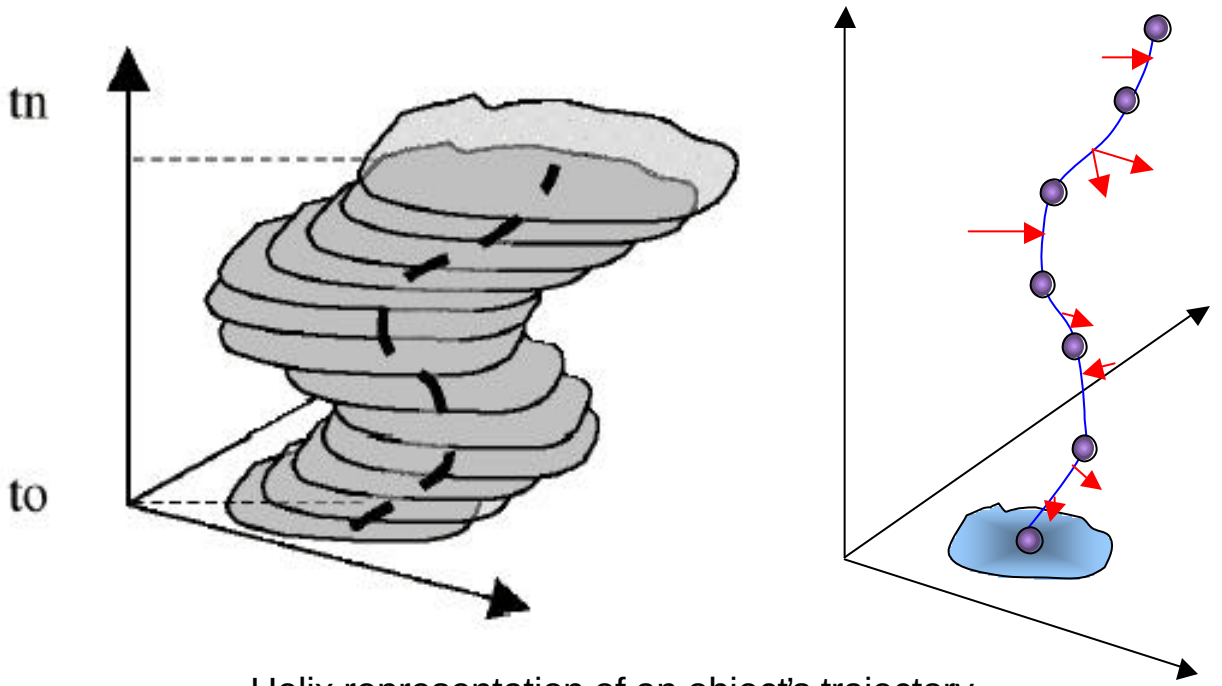
$$\begin{pmatrix} \frac{d}{dt}(A^o \cap B^o) & \frac{d}{dt}(A^o \cap \partial B) & \frac{d}{dt}(A^o \cap B^-) \\ \frac{d}{dt}(\partial A \cap B^o) & \frac{d}{dt}(\partial A \cap \partial B) & \frac{d}{dt}(\partial A \cap B^-) \\ \frac{d}{dt}(A^- \cap B^o) & \frac{d}{dt}(A^- \cap \partial B) & \frac{d}{dt}(A^- \cap B^-) \end{pmatrix}$$

  - Aggregation

Time	1	2	3
Relation	 disjoint	 meet	 overlap
9-intersection model	$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

Spatio-Temporal Data: Modeling

- Open Problems
  - Aggregation Modeling – Helix
- Helix
  - Representation of trajectory and boundary changes in an object over time



Helix representation of an object's trajectory and change in shape over time

Spine – represents trajectory of the object

Prongs – represents deformation of the object

Courtesy: University of Maine

Overview

- ✓ Input
- ➢ Statistical Foundation
- ■ Output
- ■ Computational Process
- ■ Trends

Statistics in Spatial Data Mining

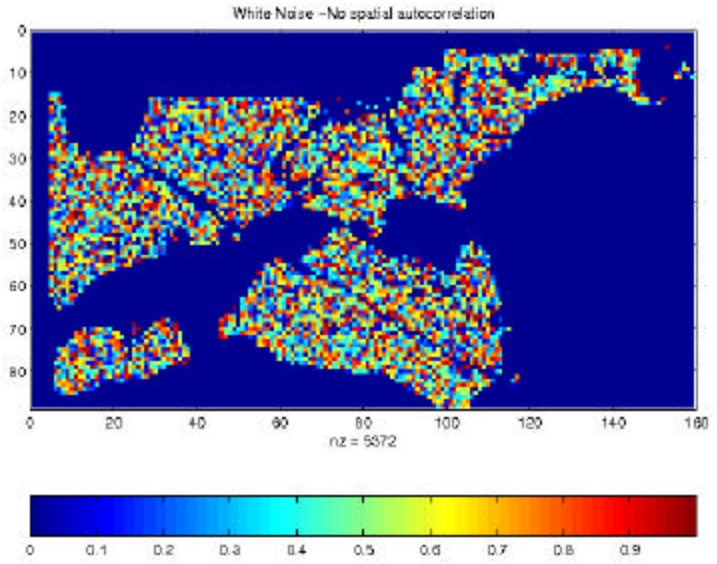
- Classical Data Mining
  - Learning samples are independently distributed
  - Cross-correlation measures, e.g., Chi-square, Pearson
- Spatial Data Mining
  - Learning sample are **not independent**
  - Spatial Autocorrelation
    - Measures:
      - distance-based (e.g., K-function)
      - neighbor-based (e.g., Moran's I)
- Spatial Cross-Correlation
  - Measures: distance-based, e.g., cross K-function
- Spatial Heterogeneity

Overview of Statistical Foundation

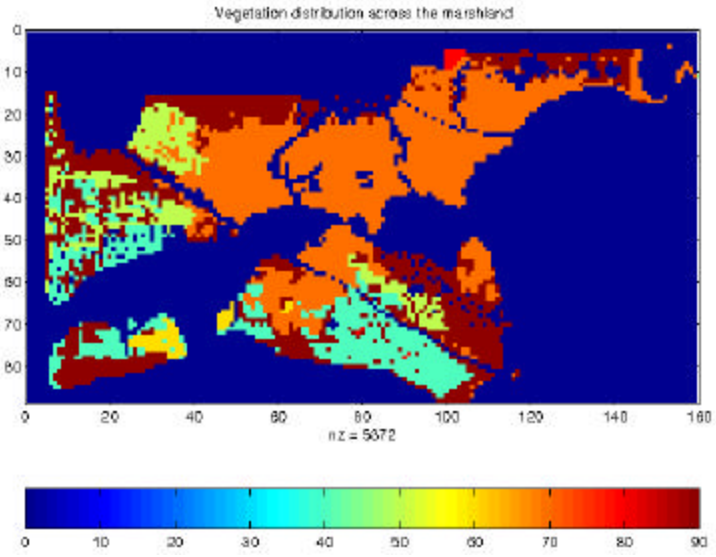
- Spatial Statistics [Cressie, 1991][Hanning, 2003]
  - Geostatistics
    - Continuous
    - Variogram: measure how similarity decreases with distance
    - Spatial prediction: spatial autocorrelation
  - Lattice-based statistics
    - Discrete location, neighbor relationship graph
    - Spatial Gaussian models
      - Conditionally specified, 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]"



Pixel property with independent identical distribution



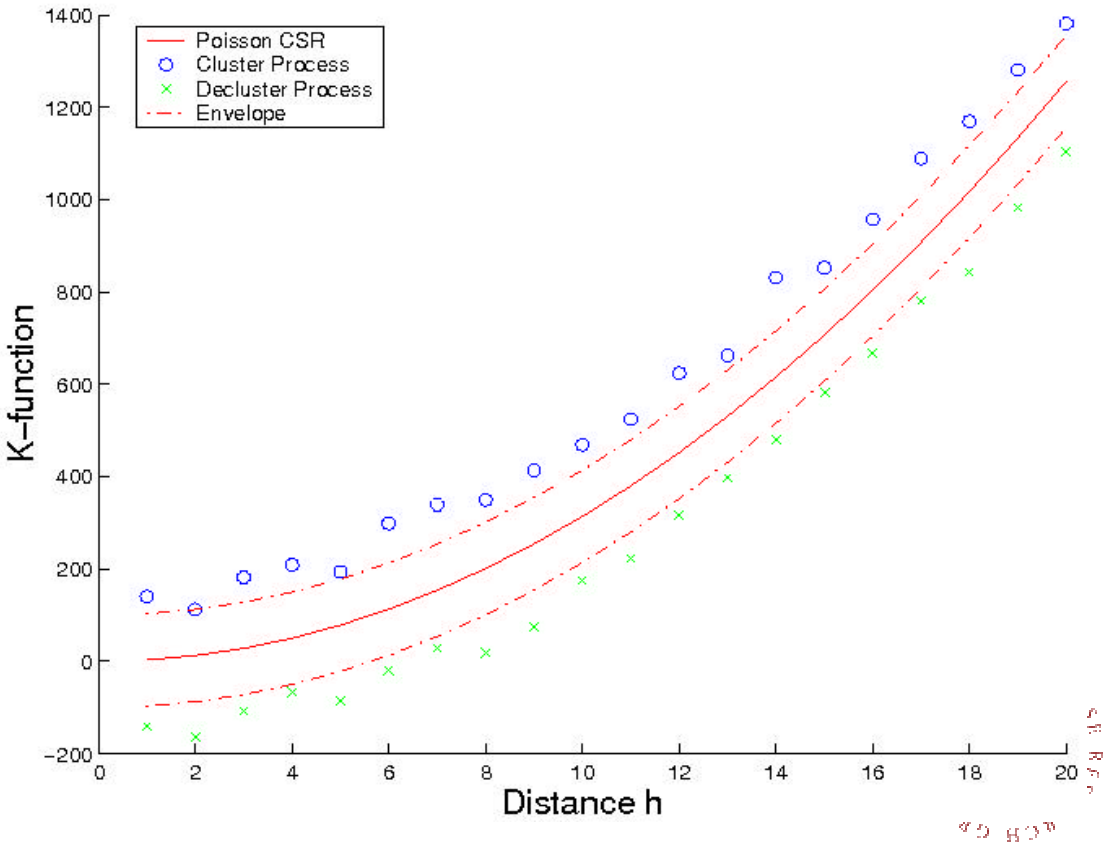
Vegetation Durability with SA

- 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, …

Spatial Autocorrelation: Distance-based measure

- *K*-function Definition
  - Test against randomness for point pattern
  - $K(h) = l^{-1} E$[number of events within distance h of an arbitrary event]
    - ? is intensity of event
  - For Poisson complete spatial randomness (CSR): $K(h) = ph^2$
    - >: cluster
    - <: decluster/regularity

K-Function based
Spatial Autocorrelation



Spatial Autocorrelation: Topological Measure

- Moran's I Measure Definition
  - ***W***: the contiguity matrix
- Ranges between -1 and +1
  - higher positive value
    - => high SA, Cluster, Attract
  - lower negative value
    - => interspersed, de-clustered, repel
  - Example
    - spatial randomness => MI = 0
    - checker board => MI = -1
    - distribution of vegetation durability => MI = 0.7

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

$$z = \{x_1 - \bar{x}, \ldots, x_n - \bar{x}\}$$

x_i : data values

$\bar{x}$: mean of x

n : number of data

Cross-Correlation

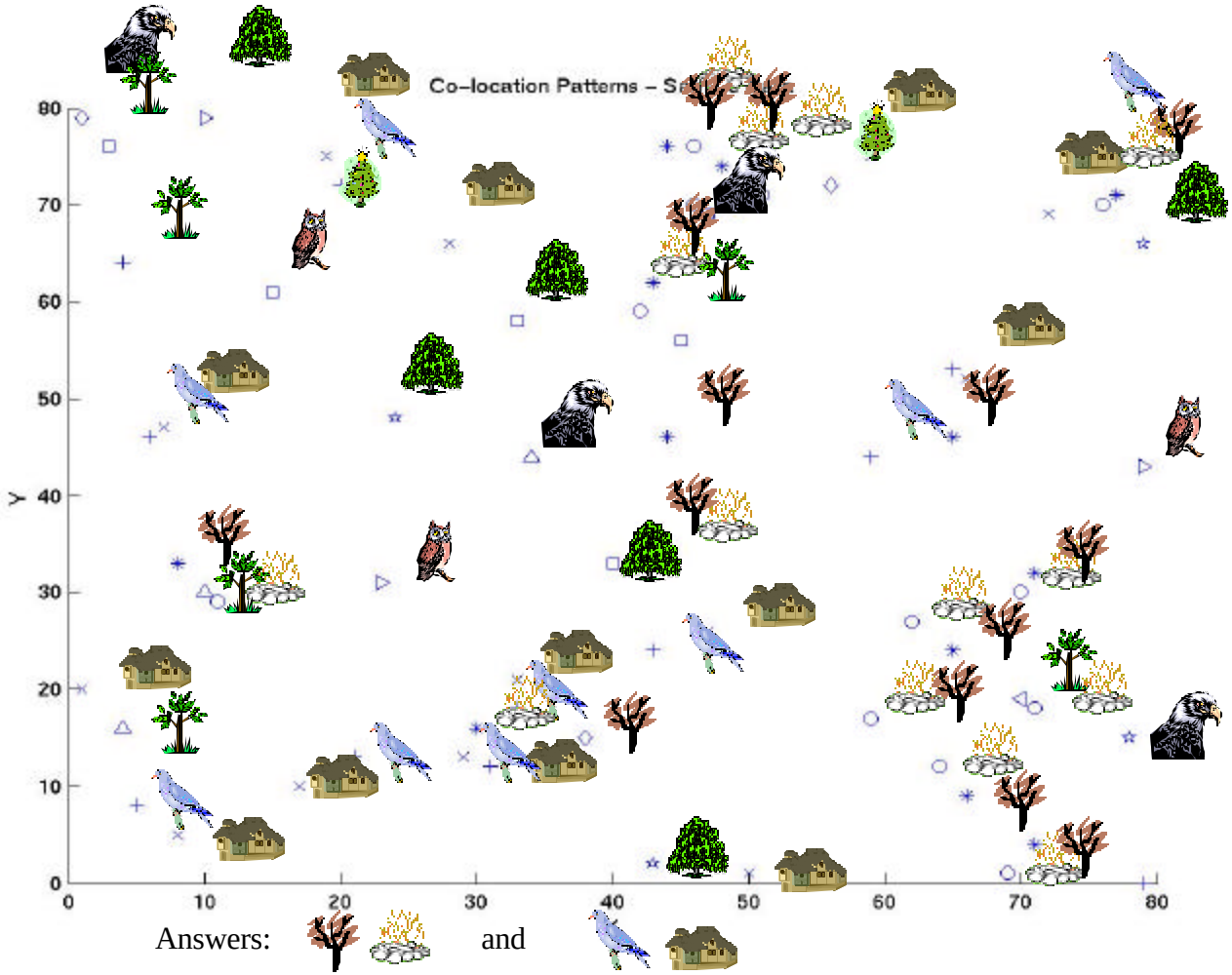
- Cross *K*-Function Definition

$$K_{ij}(h) = l_j^{-1} E[\text{number of type } \boldsymbol{j} \text{ event within distance } h \text{ of a randomly chosen type } \boldsymbol{i} \text{ event}]$$

  - Cross *K*-function of some pair of spatial feature types
  - Example
    - Which pairs are frequently co-located
    - Statistical significance

Cross-Correlation

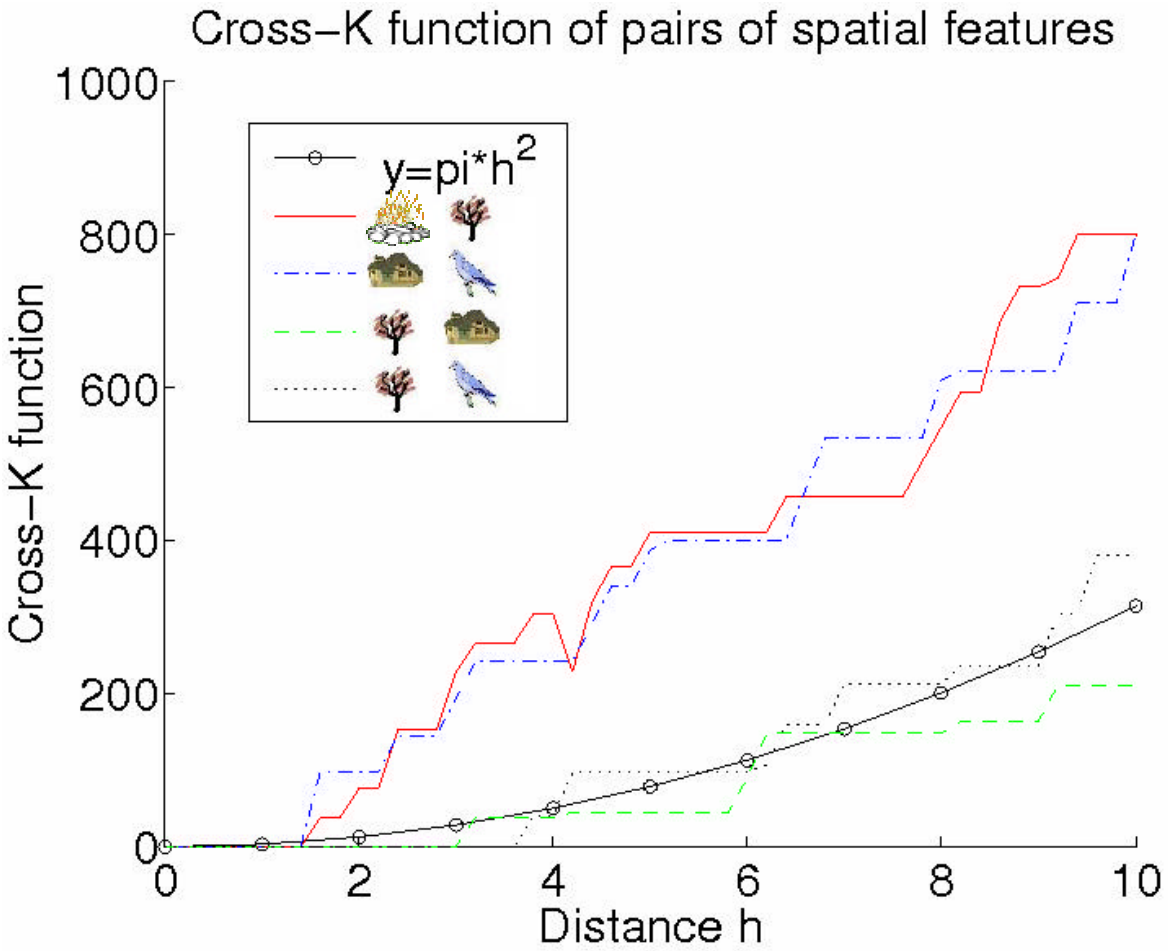
Find Patterns in the following data:



Answers: and

Illustration of Cross-Correlation

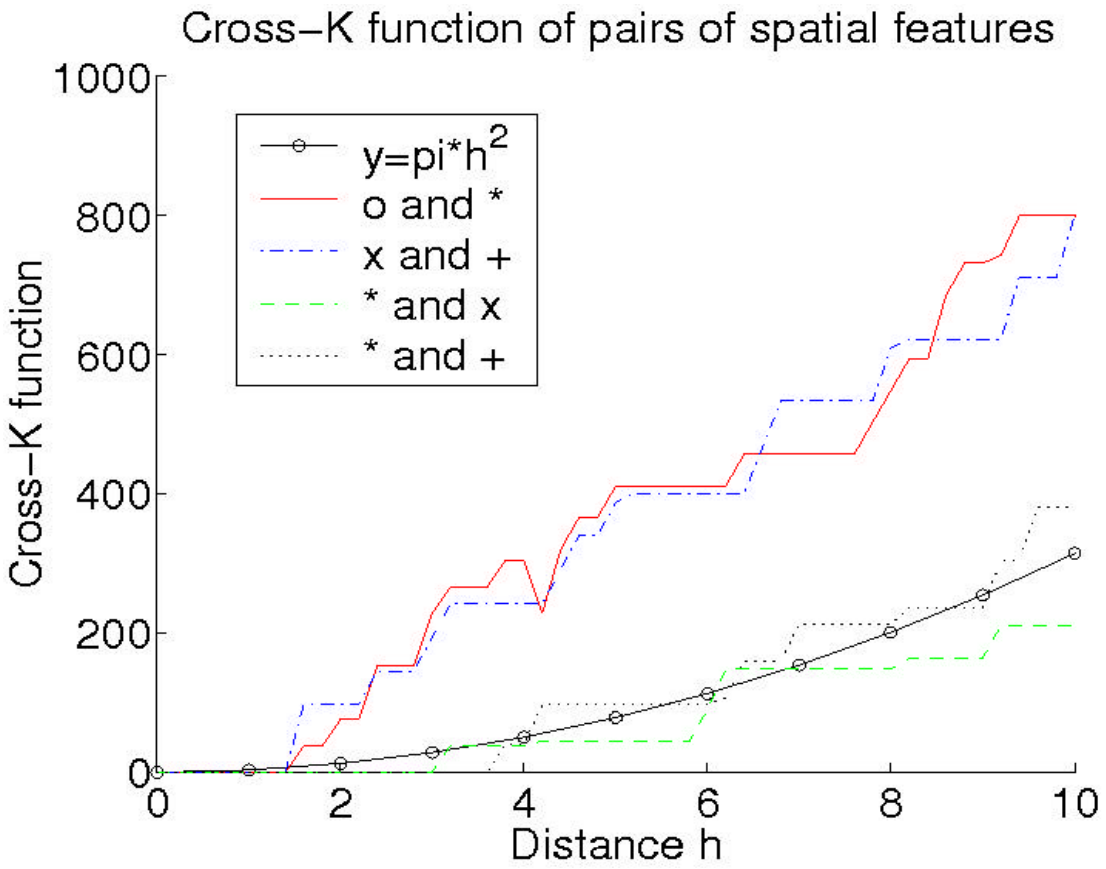
- Illustration of Cross *K*-function for Example Data



Cross-K Function for Example Data

Illustration of Cross-Correlation

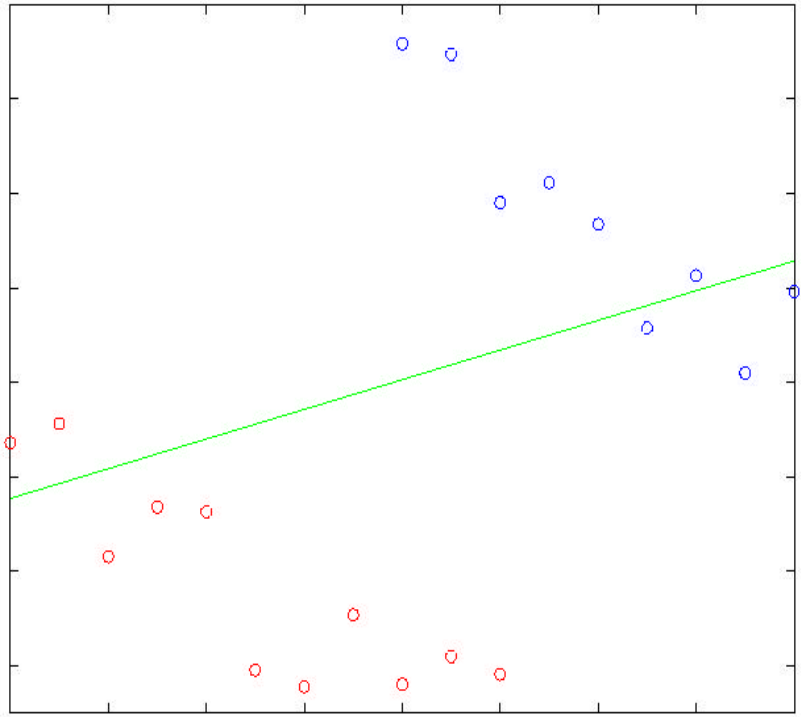
- Illustration of Cross *K*-function for Example Data

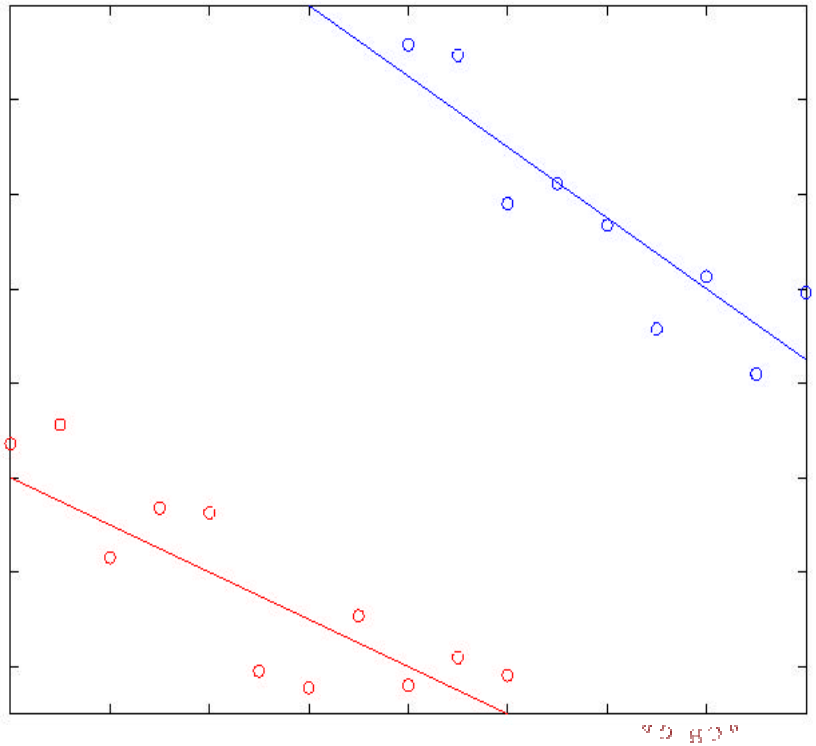


Cross-K Function for Example Data

Spatial Slicing

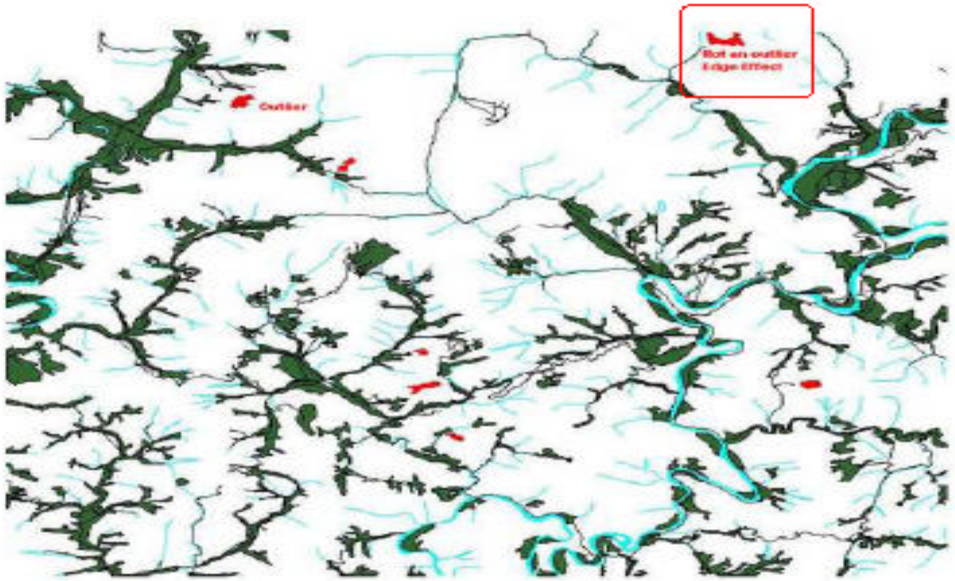
- Spatial heterogeneity
  - "Second law of geography" [M. Goodchild, UCGIS 2003]
  - Global model might be inconsistent with regional models
    - spatial Simpson's Paradox
- Spatial Slicing
  - May improve the effectiveness of SDM, show support regions of a pattern

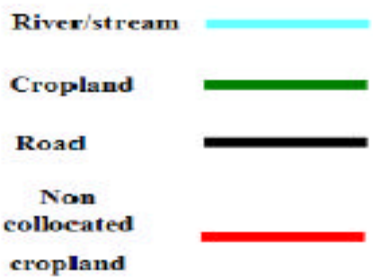




Edge Effect

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





Korea Dataset, Courtesy: Architecture Technology Corporation

Research Challenges of Spatial Statistics

- State-of-the-art of Spatial Statistics

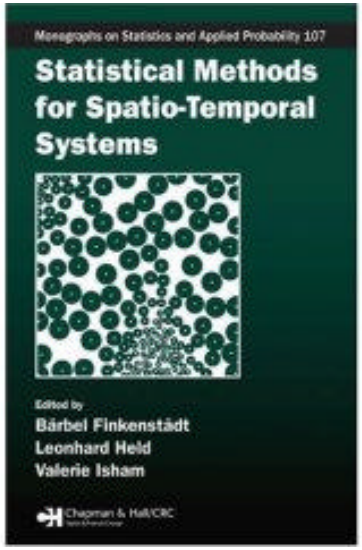
		Point Process	Lattice	Geostatistics
raster			v	v
Vector	Point	v	v	v
	Line			v
	Polygon		v	v
graph				

Data Types and Statistical Models

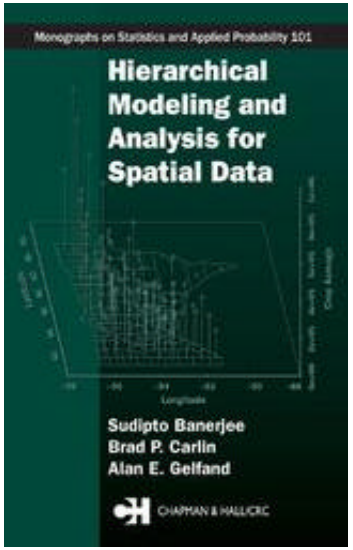
- Research Needs
  - Correlating extended features, road, rivers, cropland
  - Edge effect
  - Relationship to classical statistics
    - Ex. SVM with spatial basis function vs. SAR
  - Spatio-temporal statistics

Spatio-Temporal Statistics

- Emerging topic



"First" statistics book on Spatio-temporal models, 1st edition, 2007



Chapter on Bayesian-based Spatio-Temporal modeling, 2004

UNIVERSITY of ARKANSAS 1871

32nd Spring Lecture Series, 2007

Principal Lecturer: Noel Cressie

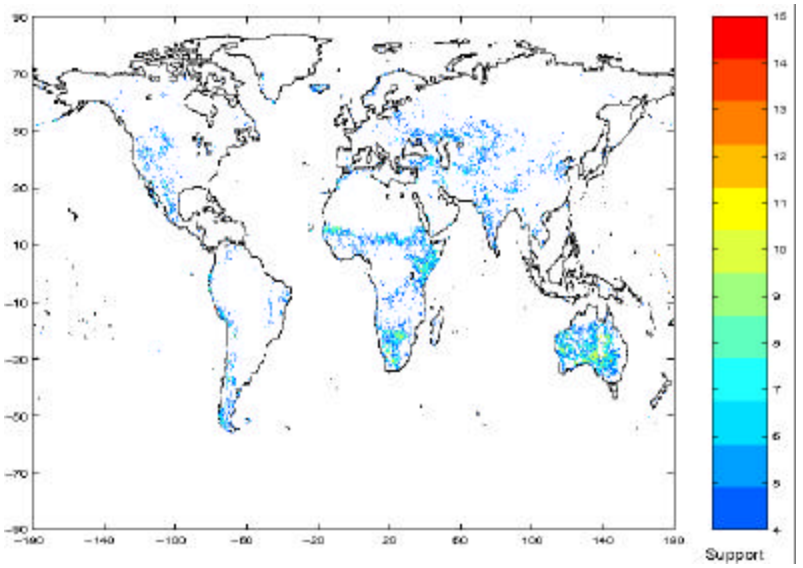
SPATIAL DATABASE RESEARCH GROUP

Overview

- ✓ Input
- ✓ Statistical Foundation
- ➢ Output
- ■ Computational Process
- ■ Trends

Three General Approaches in SDM

- A. 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
- B. Spatial slicing, use classical DM
  - Ex. association rule with support map [P. Tan et al]
  - commercial tools: e.g., Matlab, SAS, R, Splus
- C. 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



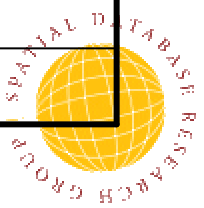
Association rule with support map
(FPAR-high -> NPP-high)

Overview of Data Mining Output

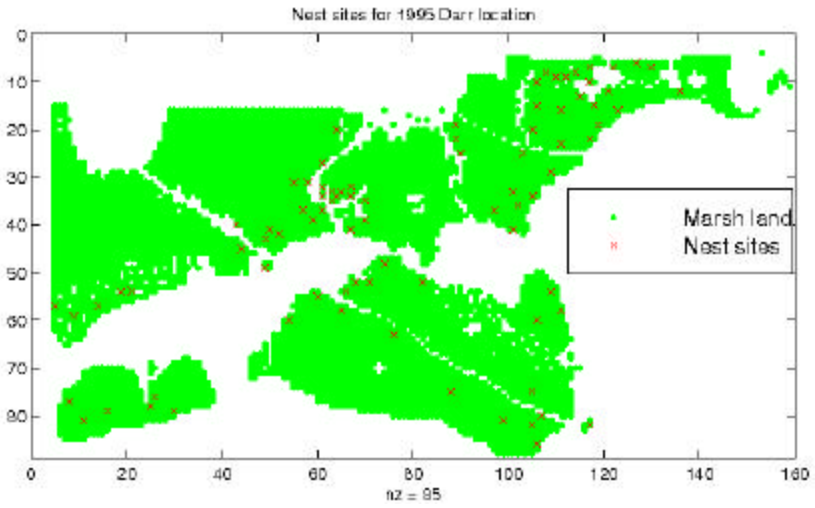
- Supervised Learning: Prediction
  - Classification
  - Trend
- Unsupervised Learning:
  - Clustering
  - Outlier Detection
  - Association
- Output Patterns vs. Statistical Models

Patterns	Point Process	Lattice	Geostatistics
Prediction	v	v	
Trend			v
Clustering	v	v	
Outliers	v	v	v
Associations	v	v	

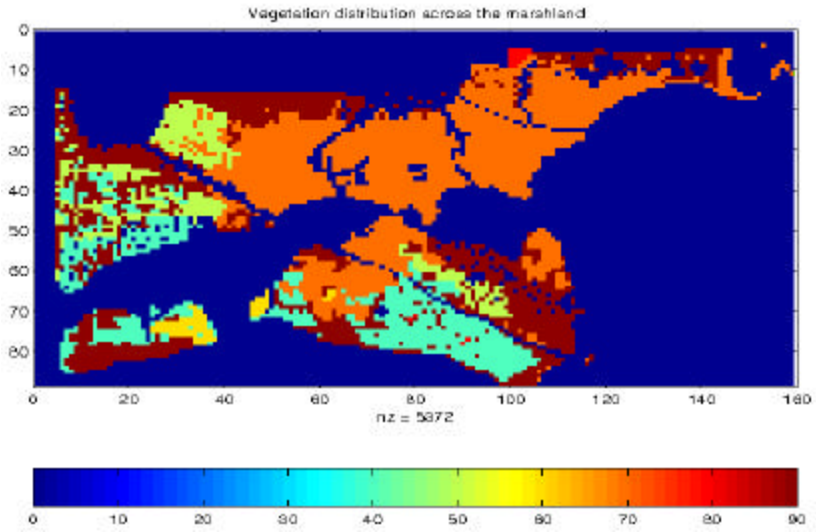
Output Patterns vs. Statistical Models



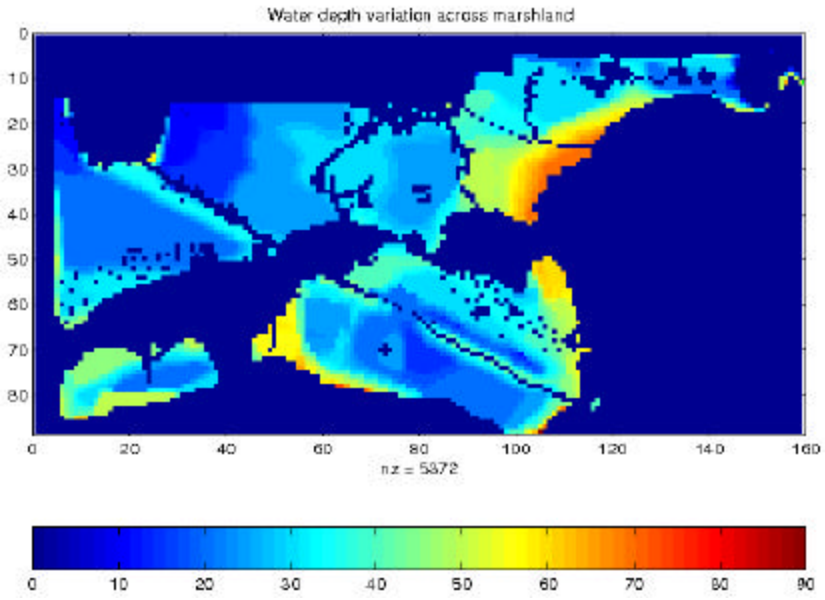
Illustrative Application to Location Prediction



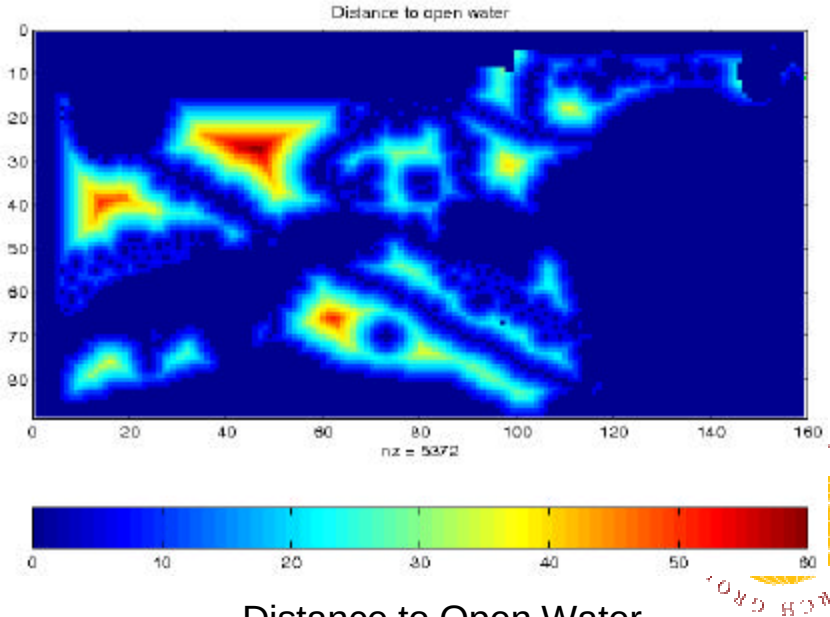
Nest Locations



Vegetation



Water Depth



Distance to Open Water

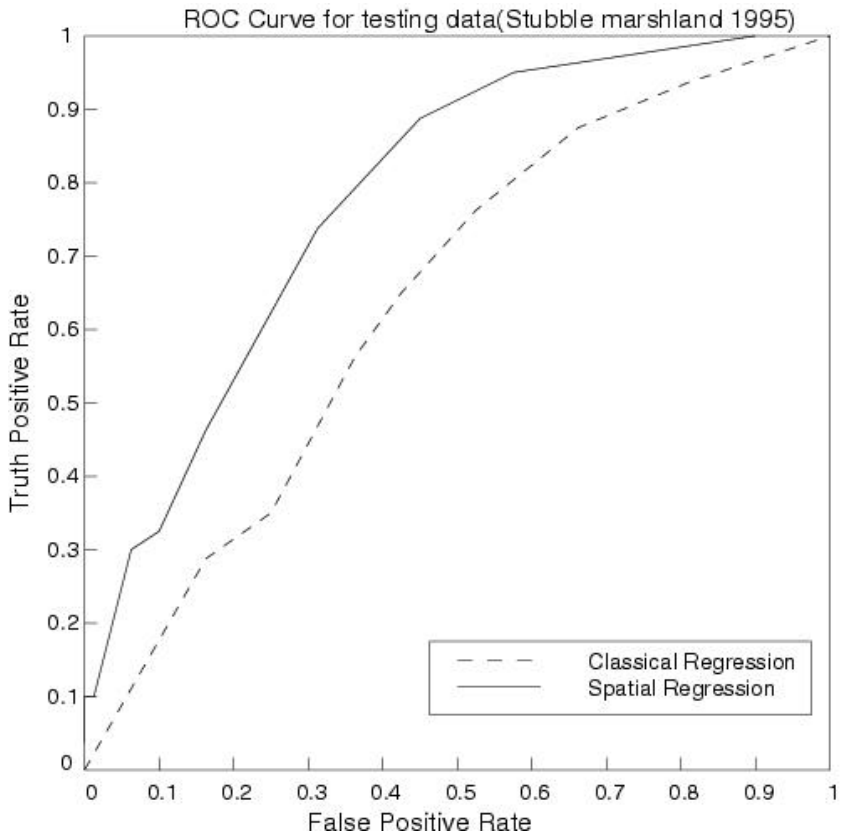
Prediction and Trend

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

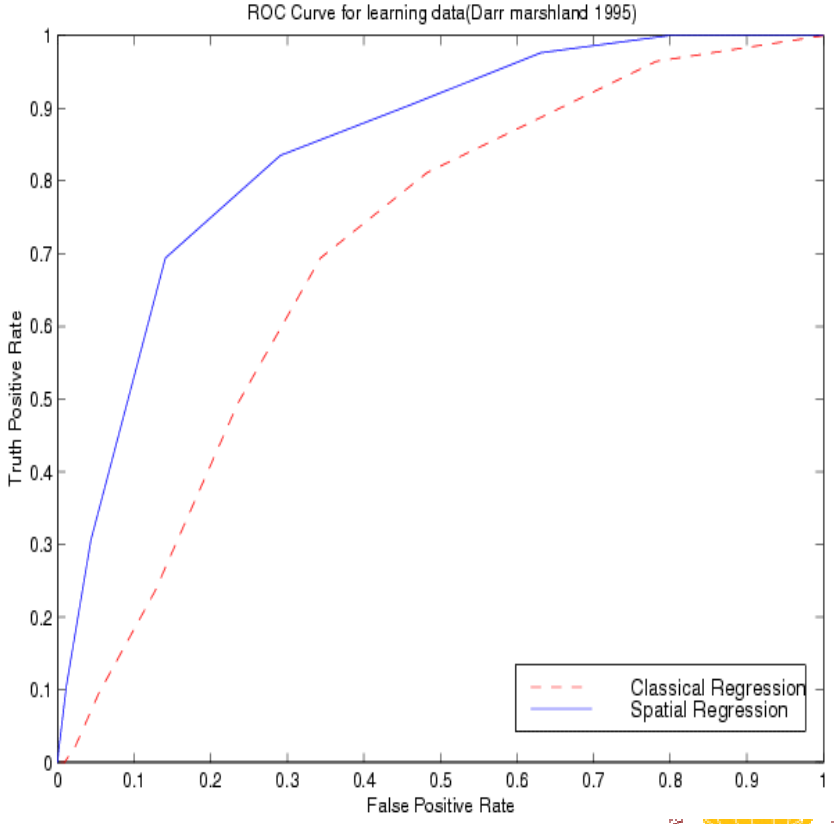
Classical	Spatial
$y = Xb + e$ $\Pr(C_i \mid X) = \frac{\Pr(X \mid C_i)\Pr(C_i)}{\Pr(X)}$	$y = rWy + Xb + e$ $\Pr(c_i \mid X, C_N) = \frac{\Pr(C_i)\Pr(X, C_N \mid c_i)}{\Pr(X, C_N)}$

Prediction and Trend

- Linear Regression $y = Xb + e$
- Spatial Regression $y = rWy + Xb + e$
- Spatial model is better



ROC Curve for learning



ROC Curve for testing

Spatial Contextual Model: SAR

- Spatial Autoregressive Model (SAR)

$$y = r\,\mathrm{W}y + Xb + e$$

  - Assume that dependent values y_i are related to each other
    - $y_i = f(y_i)\ i \neq j$
  - 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: $y = Xb' + e'$
  - where b' and e' are location dependent

Spatial Contextual Model: MRF

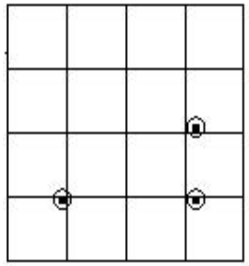
- Markov Random Fields Gaussian Mixture Model (MRF-GMM)
  - Undirected graph to represent the interdependency relationship of random variables
  - A variable depends only on neighbors
  - Independent of all other variables
  - $f_C(S_i)$ independent of $f_C(S_i)$, if $\boldsymbol{W}(s_i, s_j) = 0$
  - Predict $f_C(S_i)$, given feature value $\boldsymbol{X}$ and neighborhood class label C_N

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

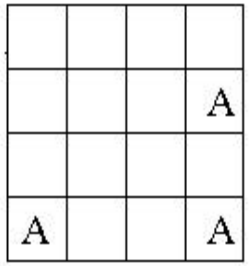
  - Assume: $Pr(c_i)$; $Pr(\boldsymbol{X}, C_N|c_i)$; and $Pr(\boldsymbol{X}, C_N)$ are mixture of Gaussian distributions.

Research Needs for Spatial Classification

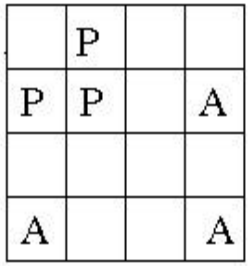
- Open Problems
  - Estimate W for SAR and MRF-BC
  - Scaling issue in SAR
    - Scale difference: $r\mathrm{W}y$ vs. $\mathrm{X}b$
  - Spatial interest measure: e.g., avg, dist(actual, predicted)



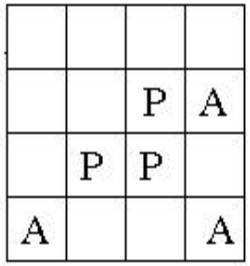
(a)
Actual Sites



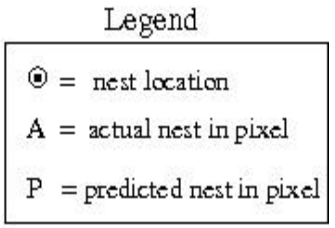
(b)
Pixels with actual sites



(c)
Prediction 1

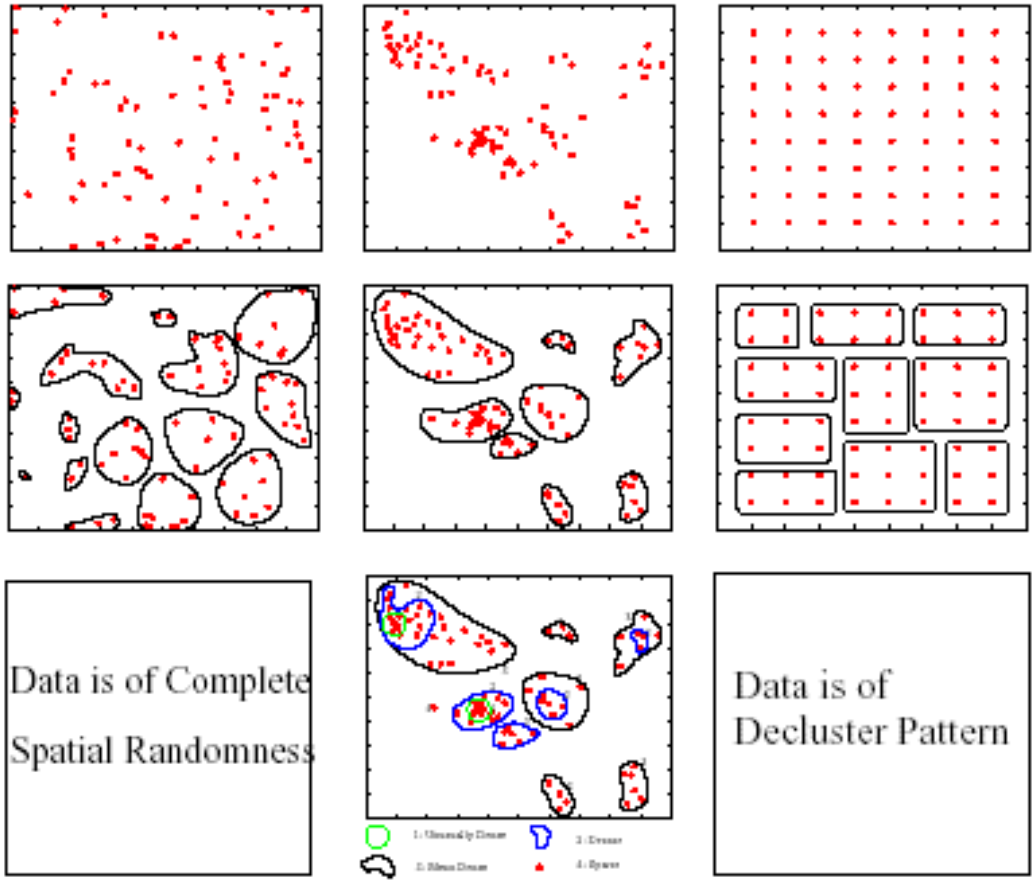


(d)
Prediction 2.
Spatially more accurate than Prediction 1



Clustering

- Clustering: Find groups of tuples
- Statistical Significance
  - Complete spatial randomness, cluster, and decluster



Inputs:
Complete Spatial Random (CSR),
Cluster,
Decluster

Classical Clustering

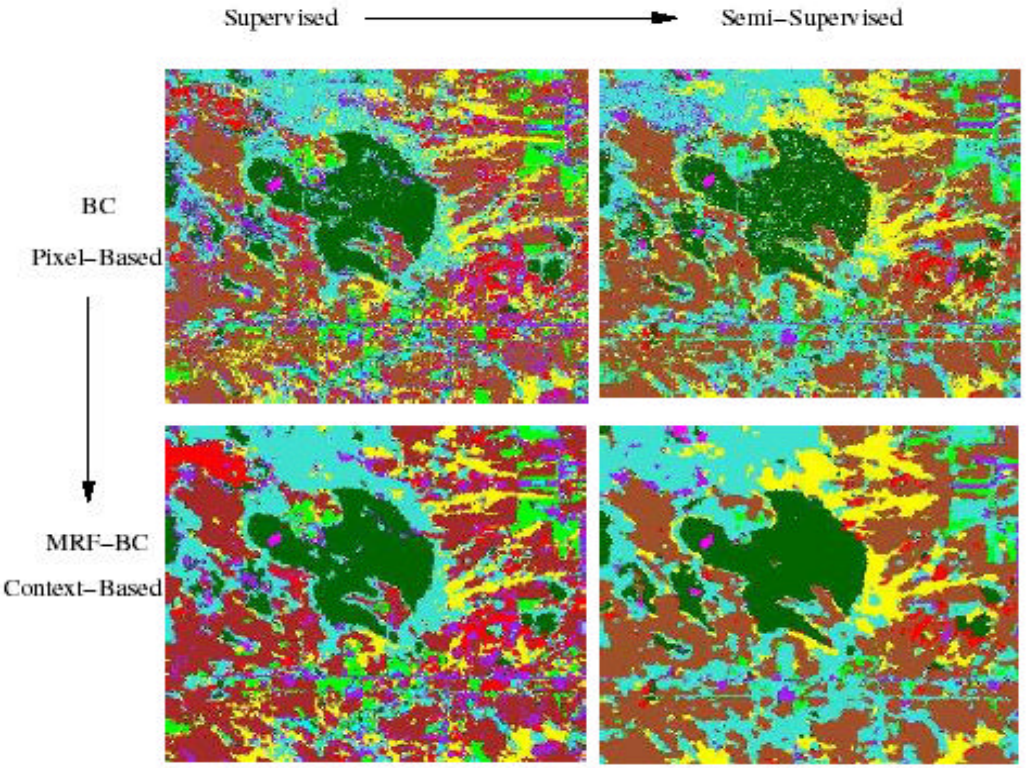
Spatial Clustering

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
  - Ex. Rivers, mountain ranges, etc

Semi-Supervised Bayesian Classification

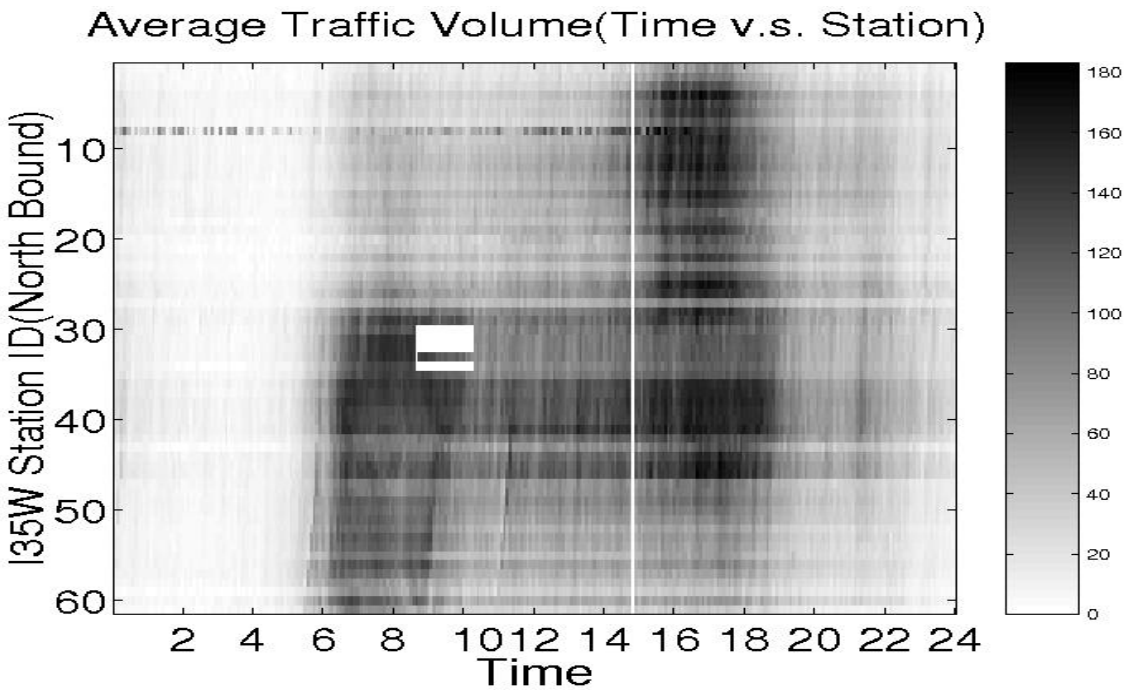
- Motivation: high cost of collecting labeled samples
- Semi-supervised MRF
  - Idea: use unlabeled samples to improve classification
    - Ex. reduce salt-N-pepper noise
  - Effects on land-use data - smoothing



Bayesian Classifiers

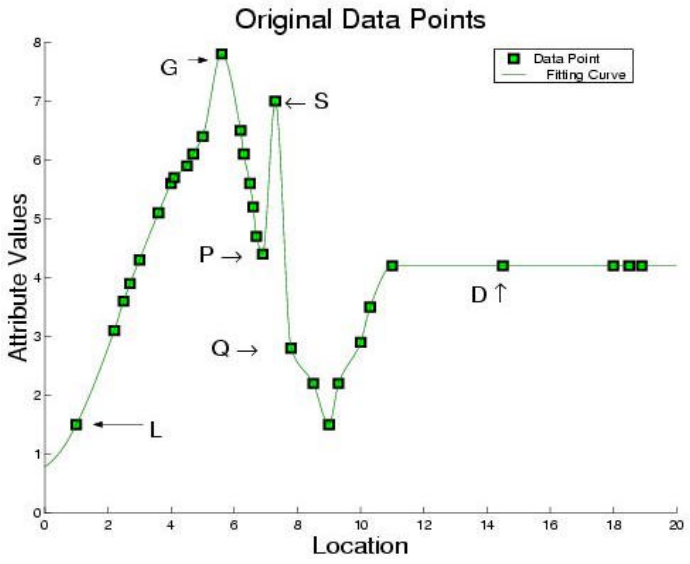
Outlier Detection

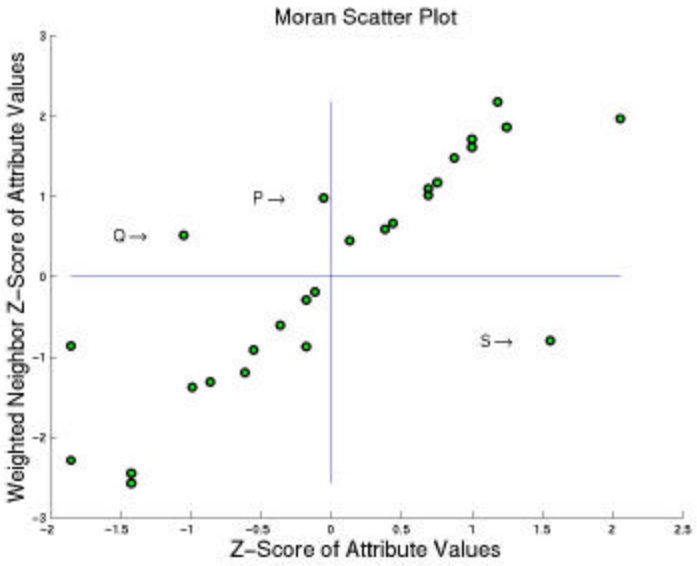
- Spatial Outlier Detection
  - Finding anomalous tuples
  - Global and spatial outlier
  - Detection Approaches
    - Graph-based outlier detection: variogram, Moran scatter plot
    - Quantitative outlier detection: scatter plot, and z-score

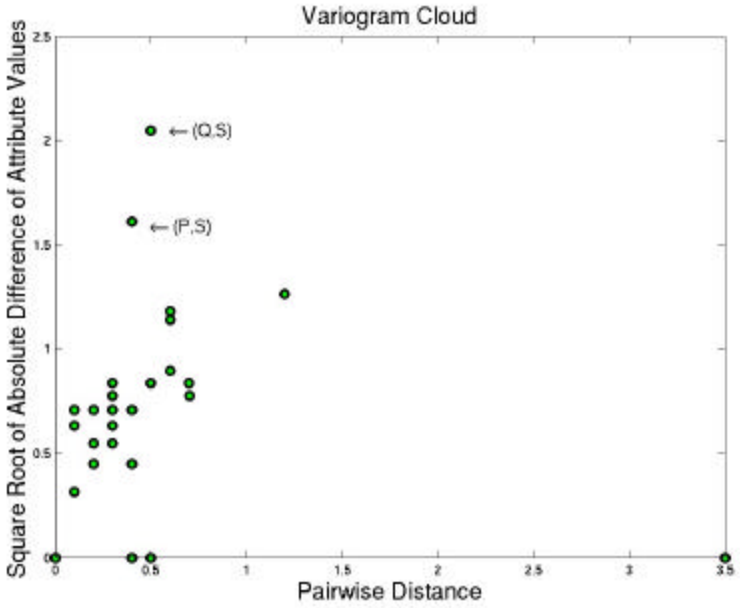


Outlier Detection

- Graphical Tests
  - Moran Scatter Plot
  - Variogram Cloud

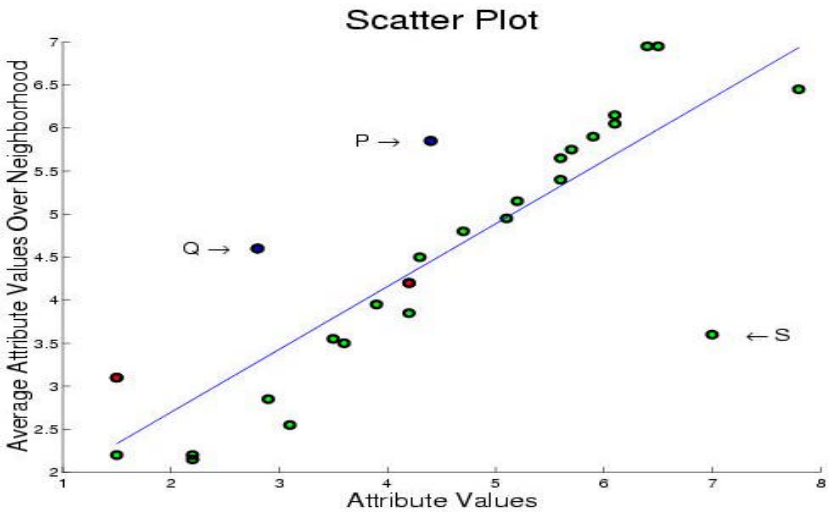


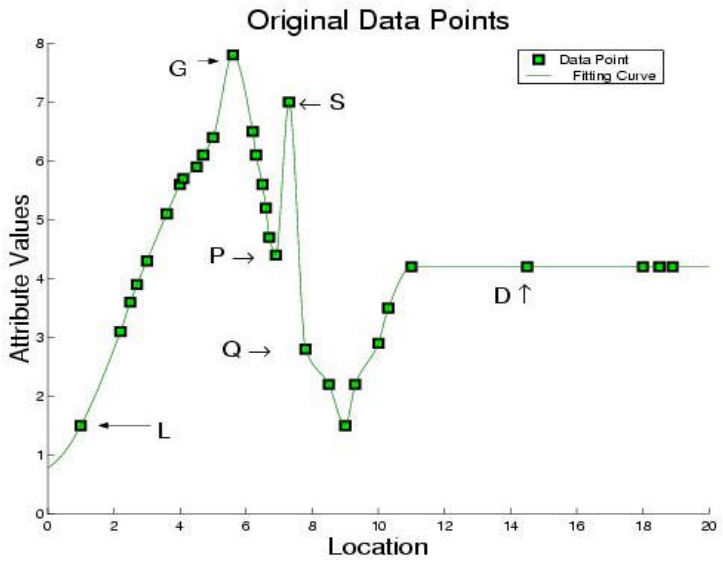


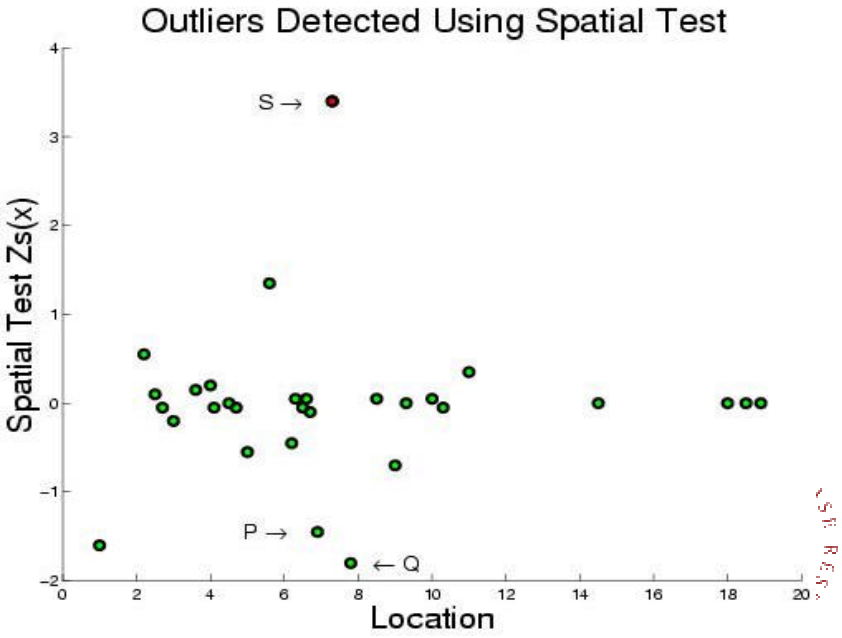


Outlier Detection – Quantitative Tests

- Quantitative Tests:
  - Scatter Plot
  - Spatial Z-test
  - Algorithmic Structure
    - Spatial Join on neighbor relation







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)

Transaction	Items Bought
1	{socks,  , milk,  , beef, egg, …}
2	{pillow,  , toothbrush, ice-cream, muffin, …}
3	{  ,  , pacifier, formula, blanket, …}
…	…
n	{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 [Agarwal, Srikant, VLDB94]
  - Support based pruning using monotonicity
- Note: **Transaction is a core concept!**

Spatial Colocation

- Comparison with Association rules

	Association rules	Colocation rules
underlying space	discrete sets	continuous space
item-types	item-types	events /Boolean spatial features
collections	transactions	neighborhoods
prevalence measure	support	participation index
conditional probability measure	Pr.[A in T | B in T]	Pr.[A in N(L) | B at L]

Participation index

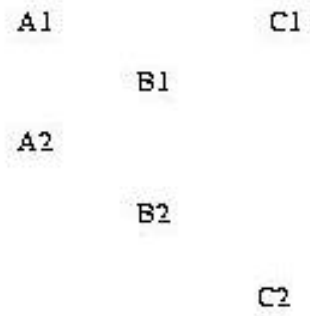
Participation ratio pr(f_i, c) of feature f_i in colocation c = {f_1, f_2, …, f_k}: fraction of instances of f_i with feature {f_1, …, f_{i-1}, f_{i+1}, …, f_k} nearby. Participation index = min{pr(f_i, c)}

Algorithm

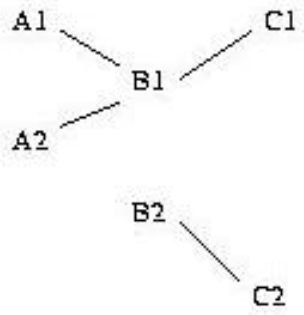
Hybrid Colocation Miner

Spatial Colocation: Approaches

Input Dataset

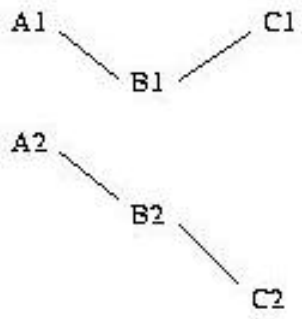


Spatial feature A,B, C, and their instances



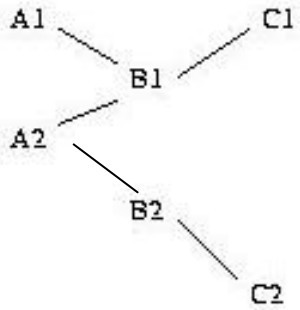
Neighbor relationship

- Partition approach



Support A,B =2 B,C=2

- Colocation

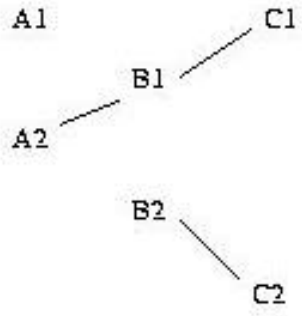


Support(A,B)=min(2/2,3/3)=1

Support(B,C)=min(2/2,2/2)=1

- Reference feature approach

C as reference feature
Transactions: (B1) (B2)
Support (A,B) = ?



Support A,B=1 B,C=2

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” trasactionazation
  - Statistical interpretation
  - Computational complexity
    - Large number of joins
    - Join predicate is a conjunction of
      - Neighbor
      - Distinct item types

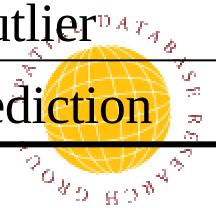
Spatio-Temporal Patterns

- Outlier Detection
  - Emerging Hot-spots
  - ST Discontinuity
- Prediction
  - Location Prediction for moving objects
  - Temporal generalization
- Clusters
  - Cluster of moving objects, e.g. Flock
  - Grid-based: Moving Clusters
- Associations, Co-locations
  - Grid-based: ST Association Rules
  - Join-based: Mixed-Drove

Summary

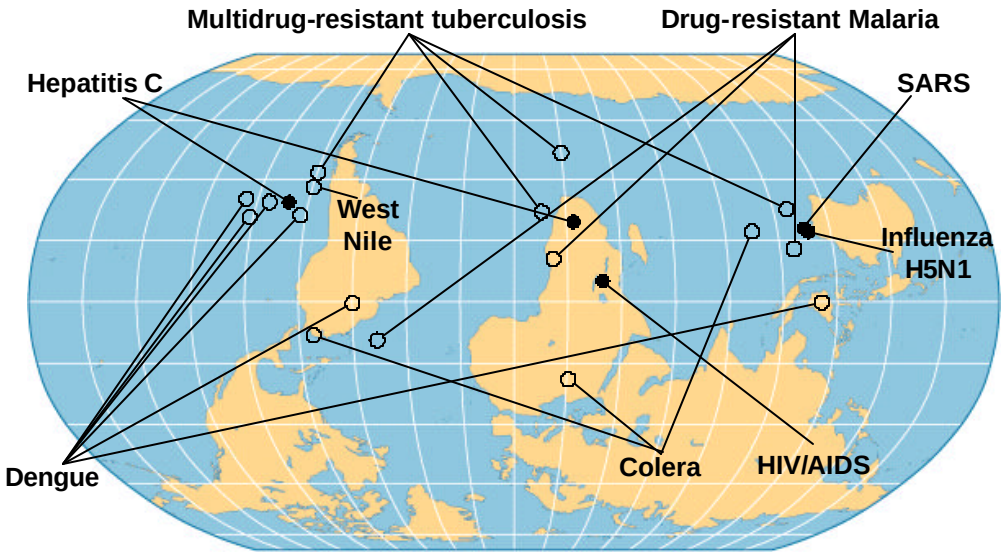
What's Special About Spatio-Temporal Data Mining ?

		Spatial DM	Spatio-Temporal DM
Input Data		Often implicit relationships, complex types	Another dimension – Time. Implicit relationships changing over time
Statistical Foundation		Spatial autocorrelation	Spatial autocorrelation and Temporal correlation
Output	Association	Colocation	Spatio-Temporal association Mixed-Drove pattern Sustained Emerging pattern
	Clusters	Hot-spots	Flock pattern Moving Clusters
	Outlier	Spatial outlier	Spatio-Temporal outlier
	Prediction	Location prediction	Future Location prediction



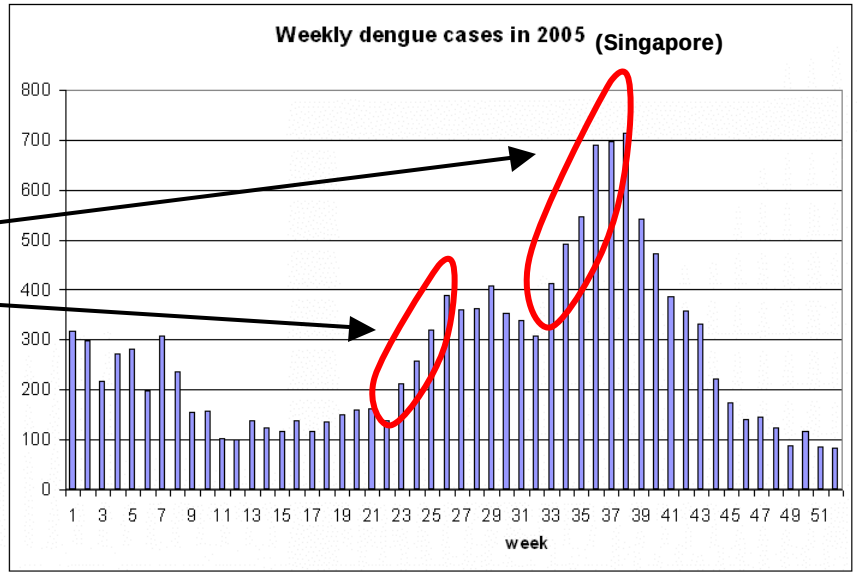
ST Patterns: Sustained Emerging Hotspots

- Sustained Emerging
  - Public health (Infectious emerging diseases - dengue fever)
  - homeland defense (looking for growing “events”, bio-defense)



• Newly emerging diseases o Re-emerging diseases

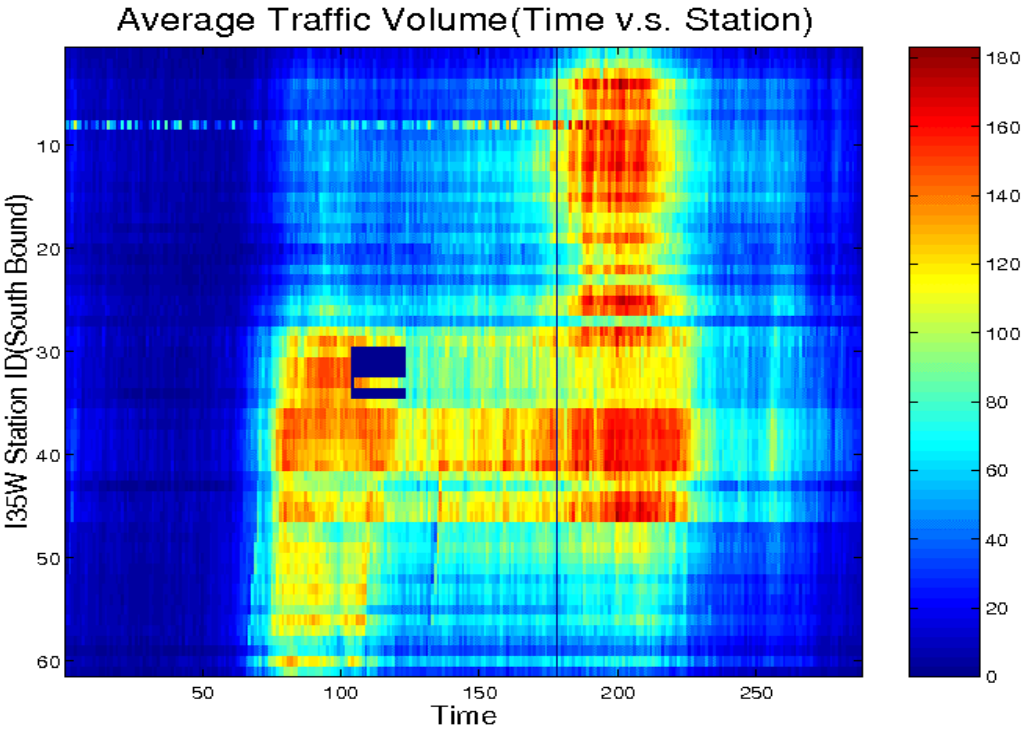
Instances of sustained emerging patterns



Courtesy: Wikipedia

ST Patterns: Outliers

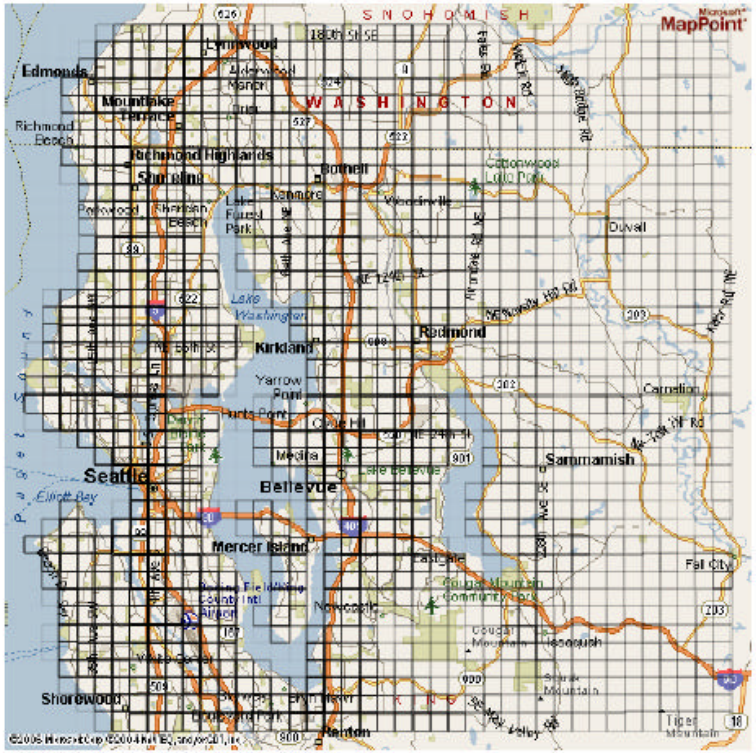
- Spatio-Temporal Outliers
  - Example Application: Sensor Networks - Traffic Data in Twin Cities
  - Abnormal Sensor Detections
  - Example: Sensor 9 (spatial) at time 0-60 (temporal)



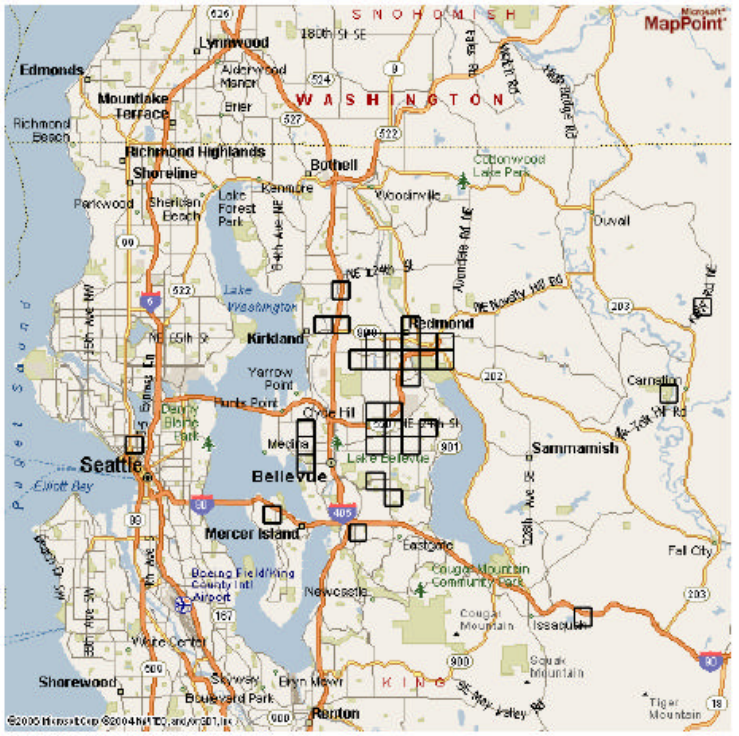
ST Patterns: Prediction

- Predict driver's destinations
  - From driver's gps track, destination history and behavior

Destination cells for a driver



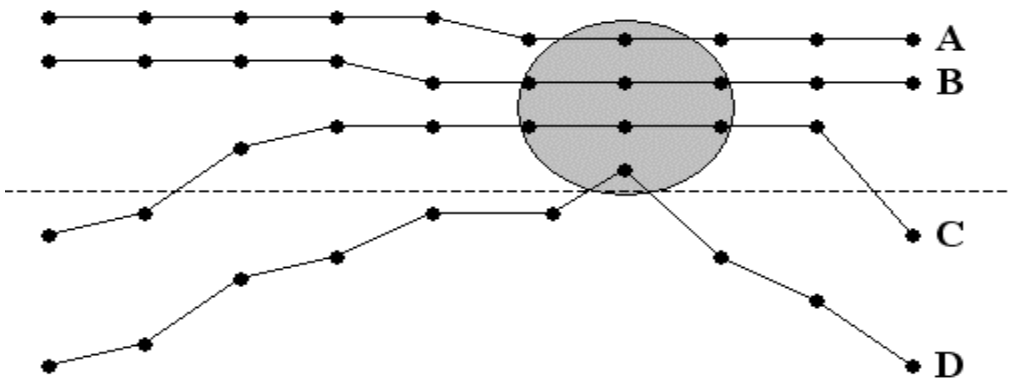
Probabilistic destinations, darker outlines are cells with higher probability



Courtesy: John Krumm, Eric Horovitz,Microsoft Research

ST Patterns: Mixed Drove

- Flock Pattern Mining



- Flock Pattern [Gudmundsson05]
  - Each time step treated separately

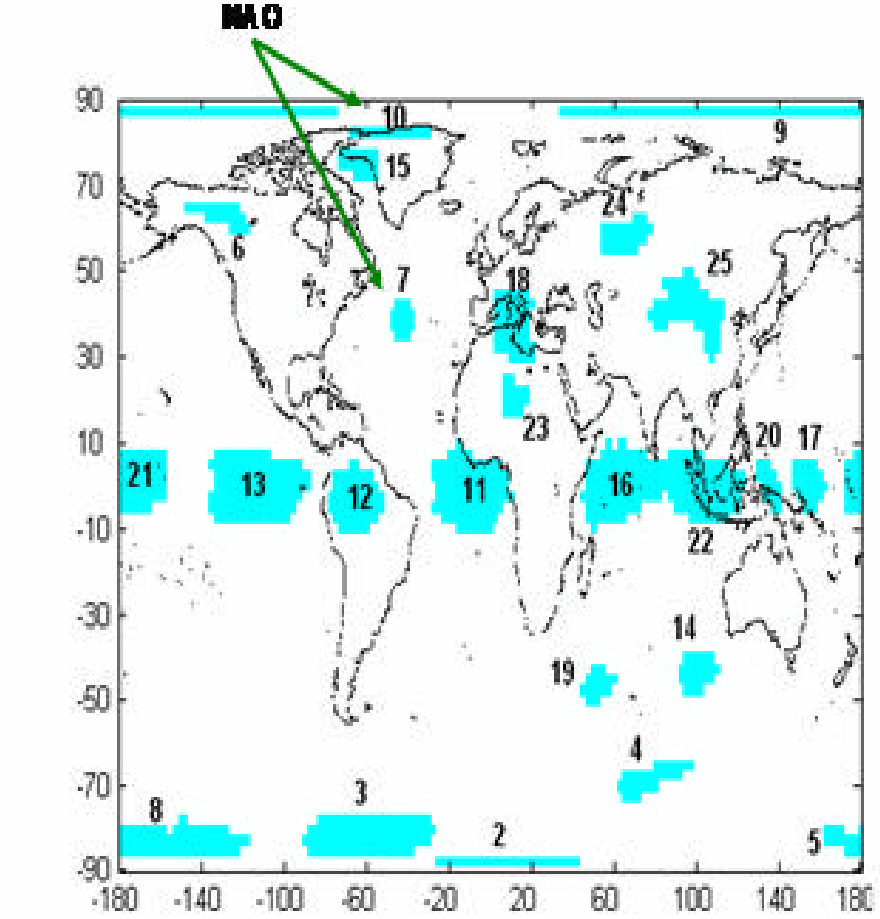
Time	Patterns	Time	Patterns
1-10	A B	7	A D
3-9	A C	7	B D
3-9	B C	7	C D
3-9	A B C	7	A B C B

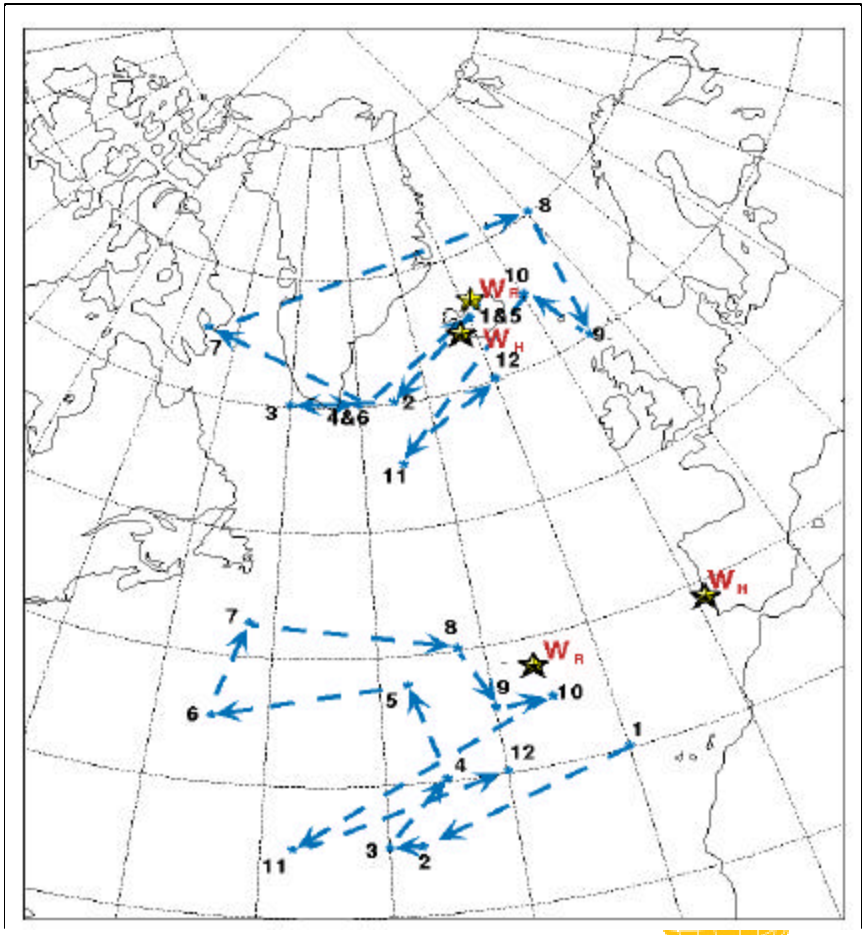
- Significant Flock Pattern

Patterns	Interest Measure (threshold 0.5)
(A B)	1
(A C)	0.7
(B C)	0.7
(A B C)	0.7
others	below threshold

ST Patterns: Moving Clusters

- Moving Clusters
  - North Atlantic Oscillation

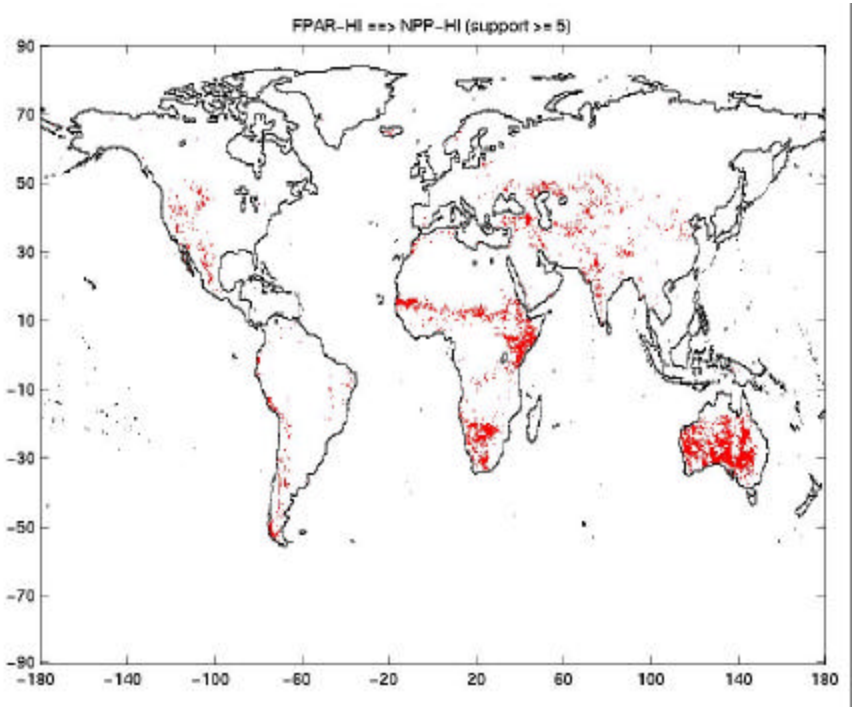




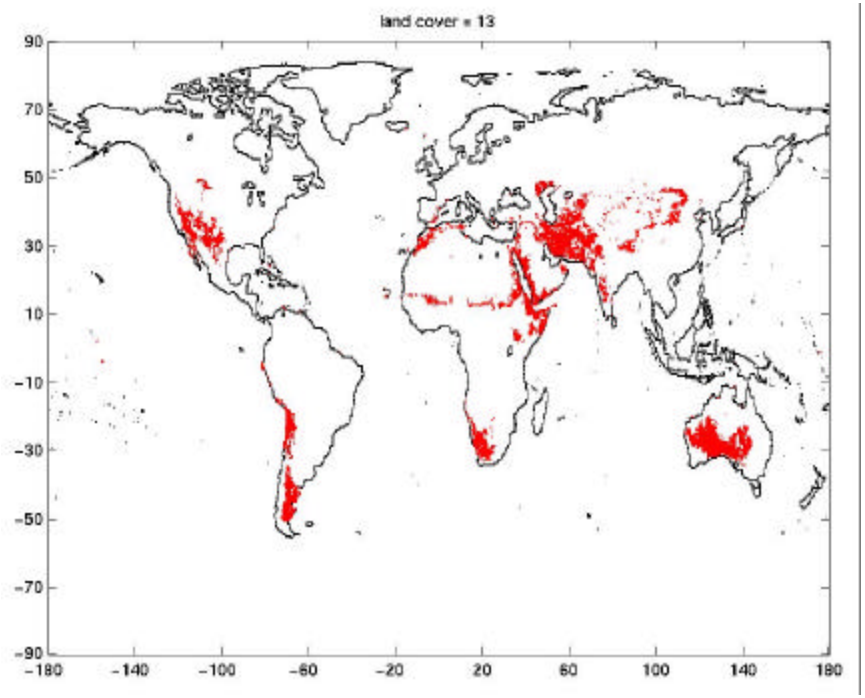
Source: Portis et al, Seasonality of the NAO, AGU Chapman Conference, 2000.

ST Patterns: Association

- Spatio-temporal Associations in Climate Data
  - ST Grid (latitude degree, longitude degree, month) defines transactions



FPAR-Hi ==> NPP-Hi (sup=5.9%, conf=55.7%)



Grassland/Shrubland areas

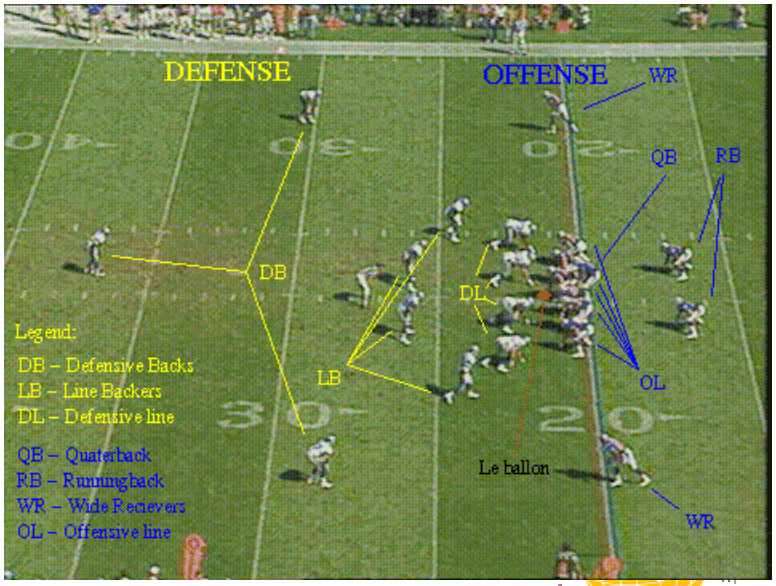
Association rule is interesting because it appears mainly in regions with grassland/shrubland vegetation type

Courtesy: Tan et al 2001

ST Patterns: Mixed Drove

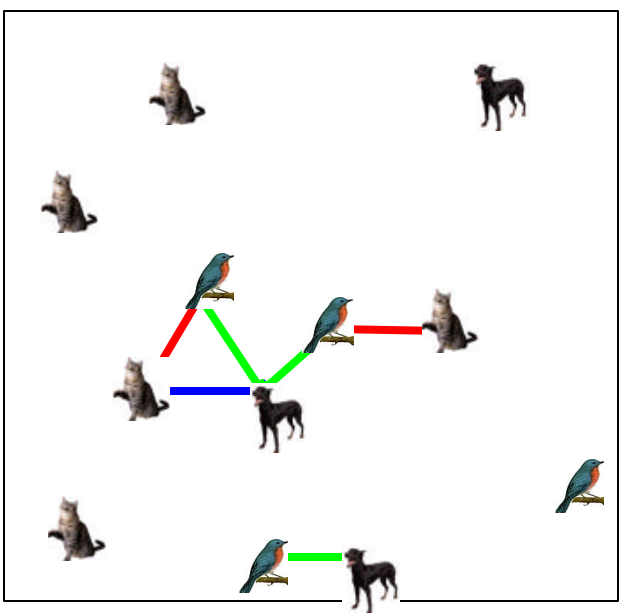
- Ecology
  - Animal movements (migration, predator-prey, encounter)
  - Species relocation and extinction (wolf – deer)

- Games
  - Game tactics of opponent team (soccer, American football, …)
  - Co-occurring role patterns

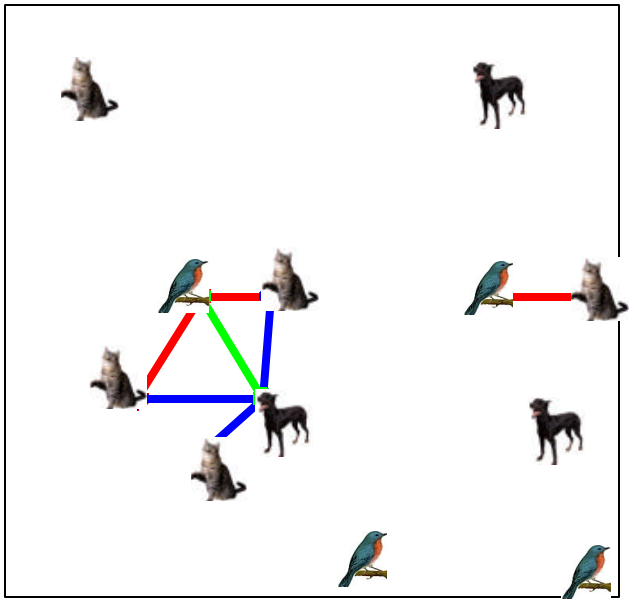


ST Patterns: Sustained Emerging

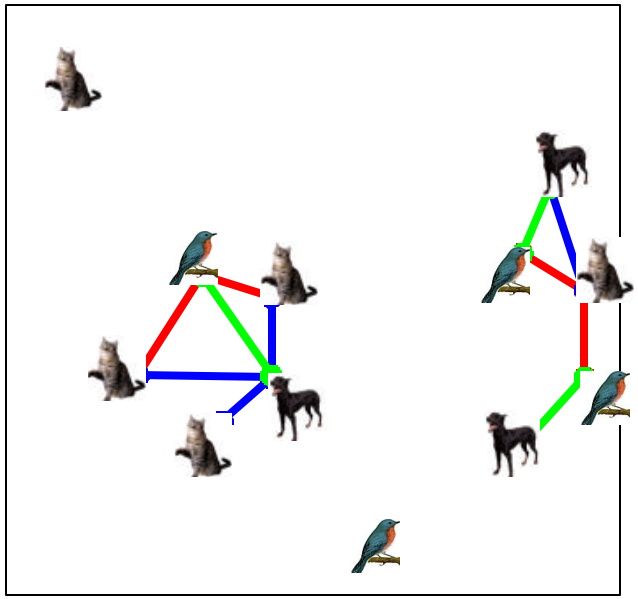
- Sustained Emerging



time slot t=0

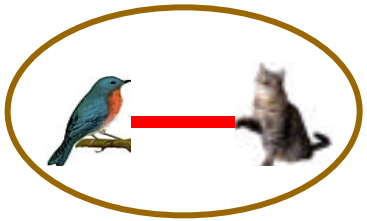


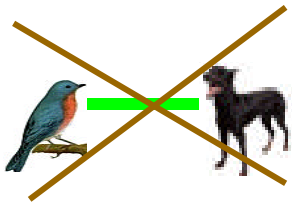
time slot t=1

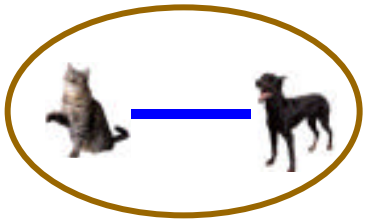


time slot t=2

Which pairs are sustained emerging patterns?







Overview

- ✓ Input
- ✓ Statistical Foundation
- ✓ Output
- ➢ Computational Process
- ■ Trends

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 loss of information
Filter-and-Refine	Minimum-Bounding Rectangle (MBR), Predicate Approximation	
Ordering	Plane Sweeping, Space Filling Curve	
Hierarchical Structures	Spatial Index, Tree Matching	
Parameter Estimation	Parameter estimation with spatial autocorrelation	

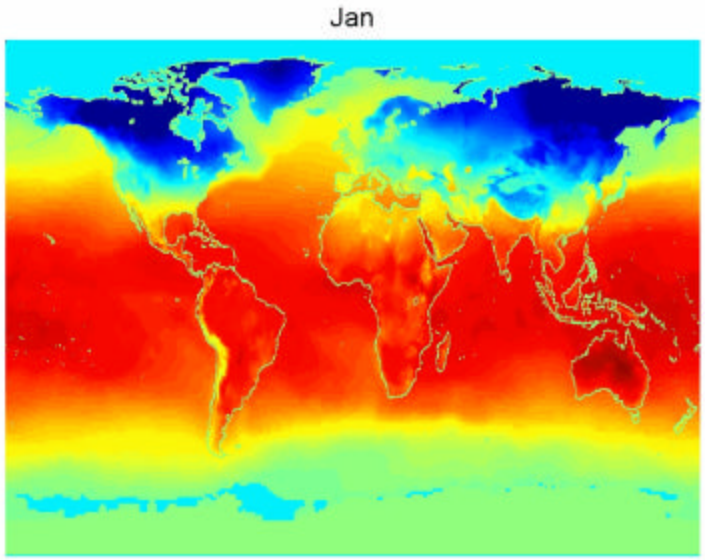
Algorithmic Strategies in Spatial Data Mining

Computational Process

- 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

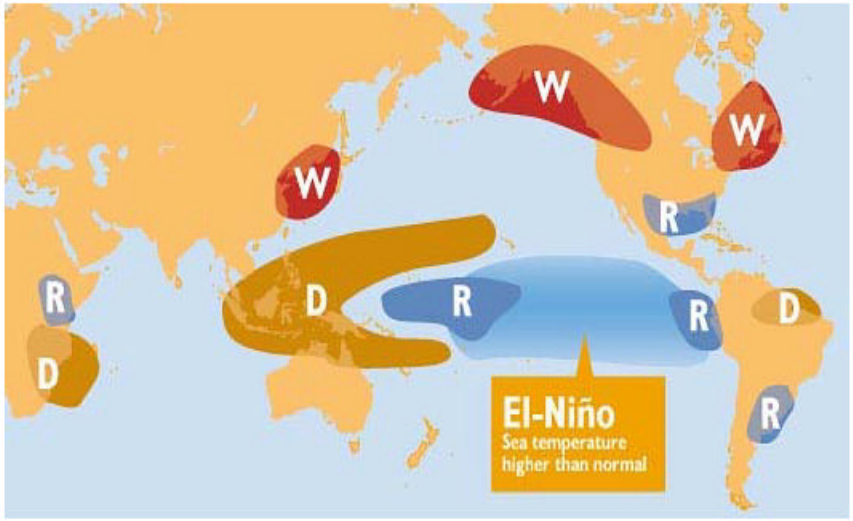
Example of Computational Process

- Teleconnection
  - Find (land location, ocean location) pairs with correlated climate changes
    - Ex. El Nino affects climate at many land locations



Average Monthly Temperature

(Courtesy: NASA, Prof. V. Kumar)



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

Example: Teleconnection (Cont')

- Challenge
  - high dimensional (e.g., 600) feature space
  - 67k land locations and 100k ocean locations (degree by degree grid)
  - 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 ? = 0.3 to 0.9

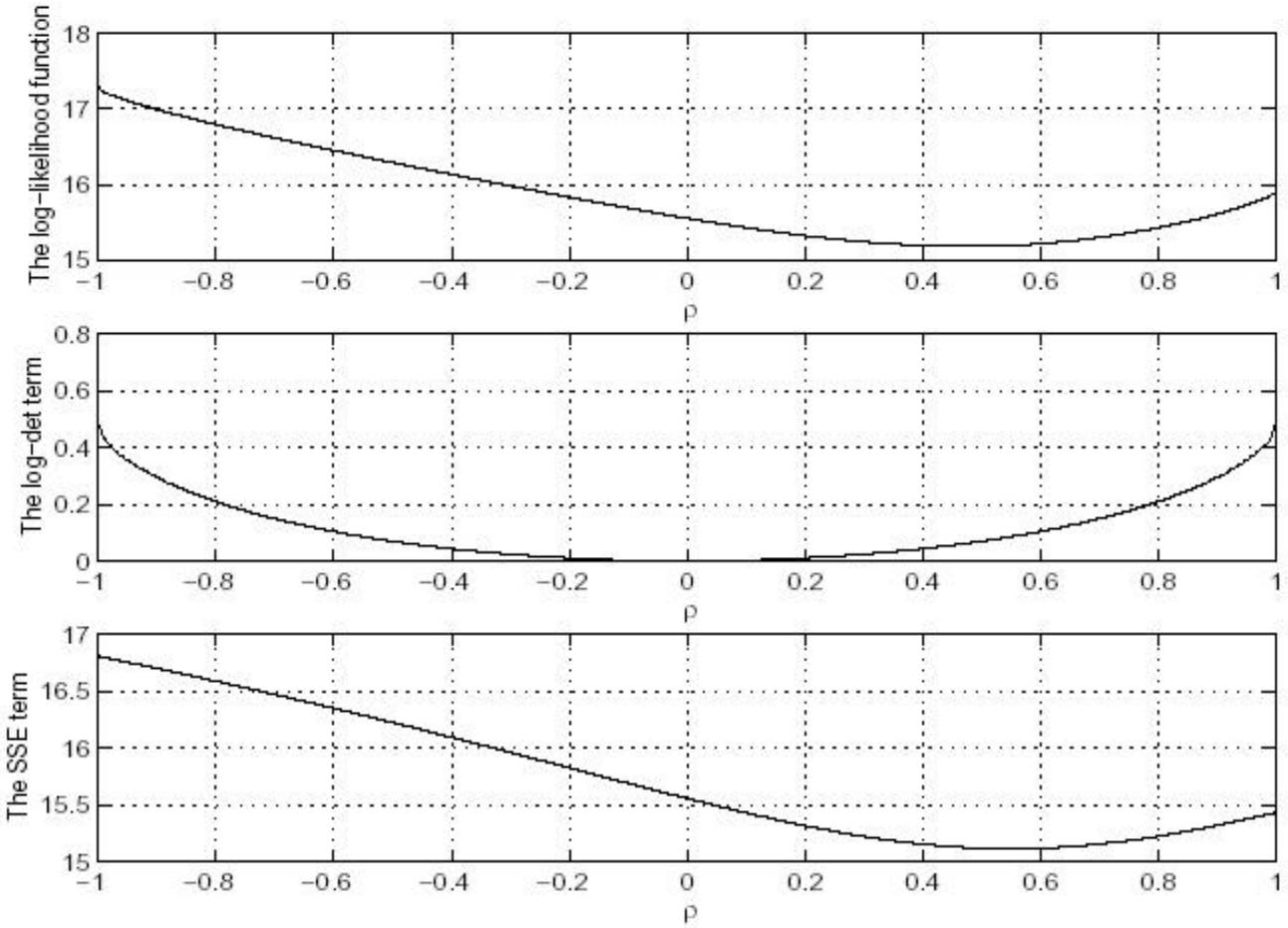
Parameter estimation of SAR

- Spatial Auto-Regression Model
  - Estimate *?* and *ß* for $y = rWy + Xb + e$
  - The estimation uses maximum-likelihood (ML) theory

- Log-likelihood function LLF = log-det + SSE + const
  - log-det = ln|**I-** *?**W***|
  - SSE = $\frac{1}{2s^2}\{y^T(I - r\mathrm{W})^T M^T (I - r\mathrm{W})y\}$

Parameter estimation of SAR

- Computational Insight:
  - *LLF* is uni-model [Kazar et al., 2005]: breakthrough result
  - Optimal *?* found by Golden Section Search or Binary Search

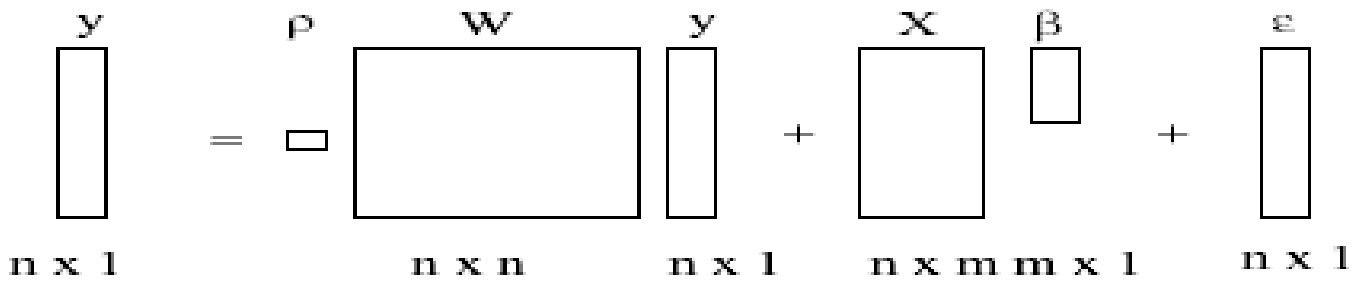


Reducing Computational Cost

- Exact Solution
  - Bottleneck = evaluation of log-det
  - Reduce cost by getting a seed for *?* 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 >> Taylor Series [Kazar et.al., 2004]

Reducing Computational Cost

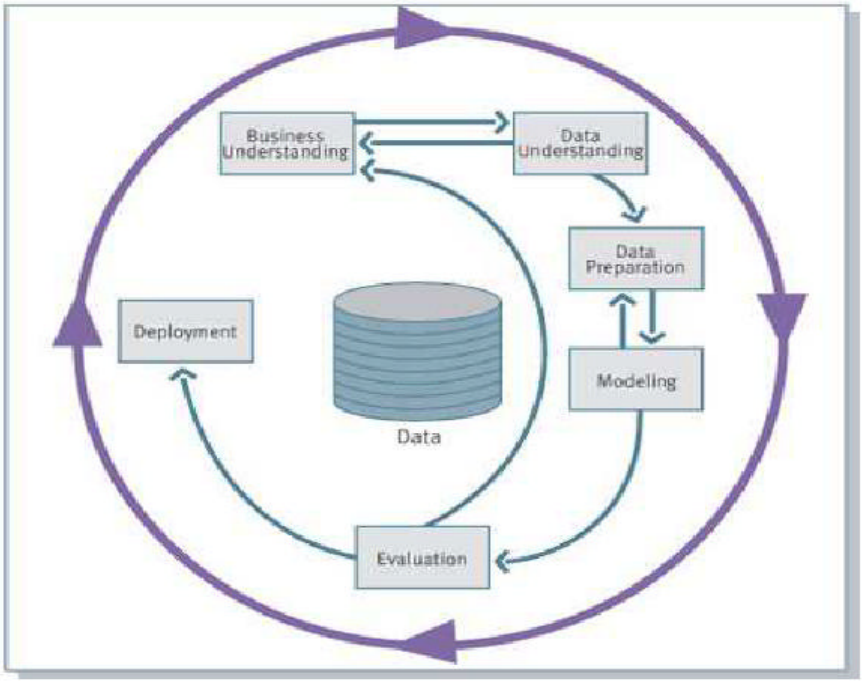
- Parallel Solution



- Computational Challenges
  - Eigenvalue + Least square + ML
  - 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



Phases of CRISP-DM

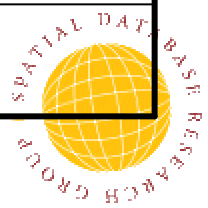
Is CRISP-DM adequate for Spatial Data Mining?

[1] CRISP-DM URL: http://www.crisp-dm.org

Summary

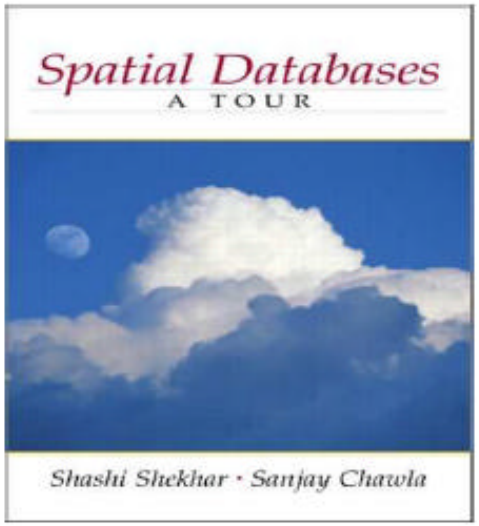
- What's Special About Spatial Data Mining

	Classical DM	Spatial DM
Input Data	All explicit, simple types	Often implicit relationships, complex types
Statistical Foundation	Independence of samples	Spatial autocorrelation
Output	Interest Measures: set-based	Location-awareness
Computational Process	Combinatorial optimization, Numerical Algorithms	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



Book

http://www.spatial.cs.umn.edu



References

- N. Cressie, Statistics for Spatial Data, John Wiley and Sons, 1991
- M. Degroot and M. Schervish, *Probability and Statistics (Third Ed.)*, Addison Wesley, 2002
- A. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression : The Analysis of Spatially Varying Relationships*, John Wiley, 2002
- M. Goodchild, *Spatial Analysis and GIS*, 2001 ESRI User Conference Pre-Conference Seminar
- R. Hanning, *Spatial Data Analysis : Theory and Practice*, Cambridge University Press, 2003
- Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, 2001
- D. Huff, *A Probabilistic Analysis of Shopping Center Trade Areas*, Lan Economics, 1963
- 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

References

- K. Kopperski and J. Han, Discovery of Spatial Association Rules in Geographic Information Database, SSTD, 1995
- K. Kopperski, J. Adhikary, and J. Han, Spatial Data Mining: Progress and Challenges, DMKD, 1996
- 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.
- H. Miller and J. Han(eds), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2001
- J. Roddick, K. Hornsby and M. Spiliopoulou, Yet Another Bibliography of Temporal, Spatial Spatio-temporal Data Mining Research, KDD Workshop, 2001
- S. Shekhar, C. T. Lu, and P. Zhang, A Unified Approach to Detecting Spatial Outliers, GeoInformatica, 7(2), Kluwer Academic Publishers, 2003

References

- S. Shekhar and S. Chawla, Spatial Databases: A Tour, Prentice Hall, 2003
- 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
- S. Shekhar and Y. Huang, Discovering Spatial Co-location Patterns: A Summary of Results, SSTD, 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
- W. Tobler, A Computer Movie Simulating Urban Growth of Detroit Region, Economic Geography, 46:236-240, 1970
- P. Zhang, Y. Huang, S. Shekhar, and V. Kumar, Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries, SSTD, 2003
- 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

References

- K. Eickhorst, A. Croitoru, P. Agouris & A. Stefanidis (2004): Spatiotemporal Helixes for Environmental Data Modeling, *IEEE CompSAC*, Hong Kong, Vol. 2, pp. 138-141.

- H. Cao, N. Mamoulis, and D. W. Cheung, "Discovery of Periodic Patterns in Spatiotemporal Sequences," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, to appear.

- Marios Hadjieletheriou, George Kollios, Petko Bakalov, and Vassilis Tsotras. Complex Spatio-Temporal Pattern Queries. Proc. of the 31st *International Conference on Very Large Data Bases (VLDB)*, Trondheim, Norway, August 2005.

- Nikos Mamoulis, Huping Cao, George Kollios, Marios Hadjieleftheriou, Yufei Tao, and David Cheung. Mining, Indexing, and Querying Historical Spatio-Temporal Data. Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Seattle, WA, August 2004.

- Sanjay Chawla, Florian Verhein. Mining Spatio-Temporal Association Rules, Sources, Sinks, Stationary Regions and ThouroughFares in Object Mobility Databases" *Proc. of 11th International Conference on Database Systems for Advanced Applications (DASFAA'06)*

- B. Arunasalam, S. Chawla and P. Sun, *Striking Two Birds With One Stone: Simultaneous Mining of Positive and Negative Spatial Patterns*, In Proceedings of the Fifth SIAM International Conference on Data Mining, Newport Beach, CA, 2005.