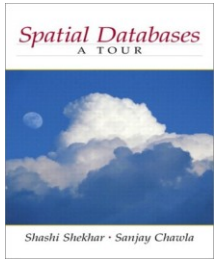
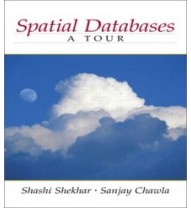


Introduction to Spatial Data Mining

7.1 Pattern Discovery
7.2 Motivation
7.3 Classification Techniques
7.4 Association Rule Discovery Techniques
7.5 Clustering
7.6 Outlier Detection

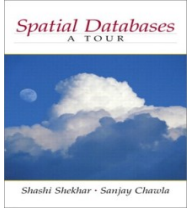


Learning Objectives

- Learning Objectives (LO)
  - LO1: Understand the concept of spatial data mining (SDM)
    - Describe the concepts of patterns and SDM
    - Describe the motivation for SDM
  - LO2 : Learn about patterns explored by SDM
  - LO3: Learn about techniques to find spatial patterns
- Focus on concepts not procedures!
- Mapping Sections to learning objectives
  - LO1 - 7.1
  - LO2 - 7.2.4
  - LO3 - 7.3 - 7.6

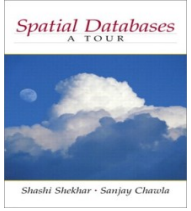
Examples of Spatial Patterns

- Historic Examples (section 7.1.5, pp. 186)
  - 1855 Asiatic Cholera in London : A water pump identified as the source
  - Fluoride and healthy gums near Colorado river
  - Theory of Gondwanaland - continents fit like pieces of a jigsaw puzlle
- Modern Examples
  - Cancer clusters to investigate environment health hazards
  - Crime hotspots for planning police patrol routes
  - Bald eagles nest on tall trees near open water
  - Nile virus spreading from north east USA to south and west
  - Unusual warming of Pacific ocean (El Nino) affects weather in USA



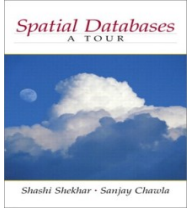
What is a Spatial Pattern ?

- What is not a pattern?
  - Random, haphazard, chance, stray, accidental, unexpected
  - Without definite direction, trend, rule, method, design, aim, purpose
  - Accidental - without design, outside regular course of things
  - Casual - absence of pre-arrangement, relatively unimportant
  - Fortuitous - What occurs without known cause
- What is a Pattern?
  - A frequent arrangement, configuration, composition, regularity
  - A rule, law, method, design, description
  - A major direction, trend, prediction
  - A significant surface irregularity or unevenness



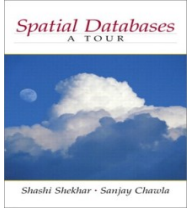
What is Spatial Data Mining?

- Metaphors
  - Mining nuggets of information embedded in large databases
    - Nuggets = interesting, useful, unexpected spatial patterns
    - Mining = looking for nuggets
  - Needle in a haystack
- Defining Spatial Data Mining
  - Search for spatial patterns
  - **Non-trivial search** - as "automated" as possible—reduce human effort
  - **Interesting, useful** and **unexpected** spatial pattern



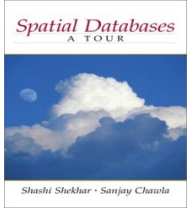
What is Spatial Data Mining? - 2

- Non-trivial search for **interesting** and **unexpected** spatial pattern
- Non-trivial Search
  - Large (e.g. exponential) search space of plausible hypothesis
  - Example - Figure 7.2, pp. 186
  - Ex. Asiatic cholera : causes: water, food, air, insects, …; water delivery mechanisms - numerous pumps, rivers, ponds, wells, pipes, ...
- Interesting
  - Useful in certain application domain
  - Ex. Shutting off identified Water pump => saved human life
- Unexpected
  - Pattern is not common knowledge
  - May provide a new understanding of world
  - Ex. Water pump - Cholera connection lead to the "germ" theory



What is 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 evenings
  - GPS product buyers are of 3 kinds:
    - outdoors enthusiasts, farmers, technology enthusiasts



Why Learn about Spatial Data Mining?

- Two basic reasons for new work
  - Consideration of use in certain application domains
  - Provide fundamental new understanding

- Application domains
  - Scale up secondary spatial (statistical) analysis to very large datasets
    - Describe/explain locations of human settlements in last 5000 years
    - Find cancer clusters to locate hazardous environments
    - Prepare land-use maps from satellite imagery
    - Predict habitat suitable for endangered species
  - Find new spatial patterns
    - Find groups of co-located geographic features

- Exercise. Name 2 application domains not listed above.

Why Learn about Spatial Data Mining? - 2

- New understanding of geographic processes for Critical questions
  - Ex. How is the health of planet Earth?
  - Ex. Characterize effects of human activity on environment and ecology
  - Ex. Predict effect of El Nino on weather, and economy
- Traditional approach: manually generate and test hypothesis
  - But, spatial data is growing too fast to analyze manually
    - Satellite imagery, GPS tracks, sensors on highways, …
  - Number of possible geographic hypothesis too large to explore manually
    - Large number of geographic features and locations
    - Number of interacting subsets of features grow exponentially
    - Ex. Find tele connections between weather events across ocean and land areas
- SDM may reduce the set of plausible hypothesis
  - Identify hypothesis supported by the data
  - For further exploration using traditional statistical methods

Spatial Data Mining: Actors

- Domain Expert -
  - Identifies SDM goals, spatial dataset,
  - Describe domain knowledge, e.g. well-known patterns, e.g. correlates
  - Validation of new patterns
- Data Mining Analyst
  - Helps identify pattern families, SDM techniques to be used
  - Explain the SDM outputs to Domain Expert
- Joint effort
  - Feature selection
  - Selection of patterns for further exploration

The Data Mining Process

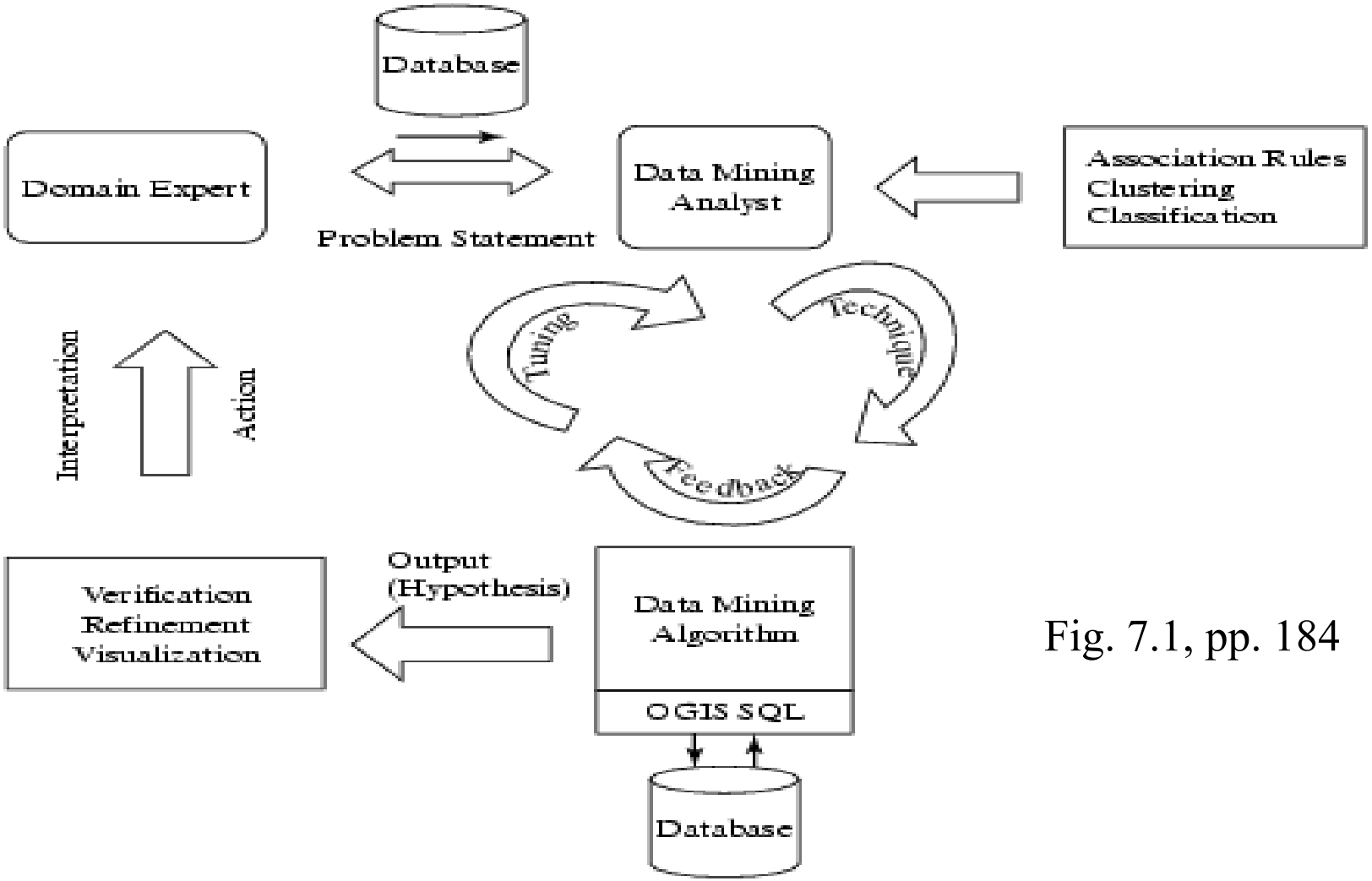
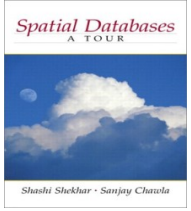


Fig. 7.1, pp. 184

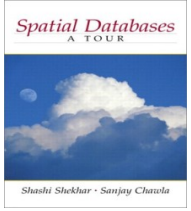
Choice of Methods

- 2 Approaches to mining Spatial Data
  - 1. Pick spatial features; use classical DM methods
  - 2. Use novel spatial data mining techniques
- Possible Approach:
  - Define the problem: capture special needs
  - Explore data using maps, other visualization
  - Try reusing classical DM methods
  - If classical DM perform poorly, try new methods
  - Evaluate chosen methods rigorously
  - Performance tuning as needed



Learning Objectives

- Learning Objectives (LO)
  - LO1: Understand the concept of spatial data mining (SDM)
  - LO2 : Learn about patterns explored by SDM
    - Recognize common spatial pattern families
    - Understand unique properties of spatial data and patterns
  - LO3: Learn about techniques to find spatial patterns
- Focus on concepts not procedures!
- Mapping Sections to learning objectives
  - LO1 - 7.1
  - LO2 - 7.2.4
  - LO3 - 7.3 - 7.6



7.2.4 Families of SDM Patterns

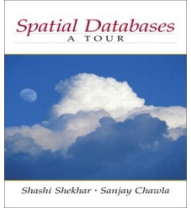
- Common families of spatial patterns
  - Location Prediction: Where will a phenomenon occur ?
  - Spatial Interaction: Which subsets of spatial phenomena interact?
  - Hot spots: Which locations are unusual ?
- Note:
  - Other families of spatial patterns may be defined
  - SDM is a growing field, which should accommodate new pattern families

7.2.4 Location Prediction

- Question addressed
  - Where will a phenomenon occur?
  - Which spatial events are predictable?
  - How can a spatial events be predicted from other spatial events?
    - Equations, rules, other methods,

- Examples:
  - Where will an endangered bird nest ?
  - Which areas are prone to fire given maps of vegetation, draught, etc.?
  - What should be recommended to a traveler in a given location?

- Exercise:
  - List two prediction patterns.



7.2.4 Spatial Interactions

- Question addressed
  - Which spatial events are related to each other?
  - Which spatial phenomena depend on other phenomenon?
- Examples:

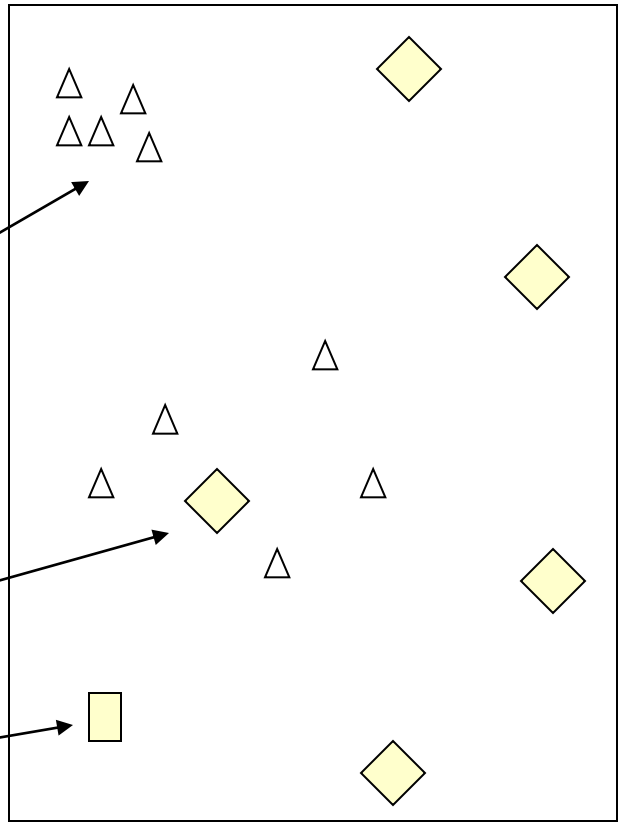
Table 1: Examples of Co-location Patterns

Domains	Example Features	Example Co-location Patterns
Ecology	Species	(Nile crocodile,Egyptian plover)
Earth science	climate and disturbance events	(wild fire, hot, dry, lightning)
Economics	industry types	(suppliers, producers, consultants)
Epidemiology	disease types and environmental events	(West Nile disease, stagnant water sources, dead birds, mosquitoes)
Location-based service	service type requests	(tow, police, ambulance)
Weather	fronts, precipitation	(cold front, warm front, snow fall)
Transportation	delivery service tracks	(US Postal Service, UPS, newspaper delivery)

- Exercise: List two interaction patterns.

7.2.4 Hot spots

- Question addressed
  - Is a phenomenon spatially clustered?
  - Which spatial entities or clusters are unusual?
  - Which spatial entities share common characteristics?
- Examples:
  - Cancer clusters [CDC] to launch investigations
  - Crime hot spots to plan police patrols
- Defining unusual
  - Comparison group:
    - neighborhood
    - entire population
  - Significance: probability of being unusual is high



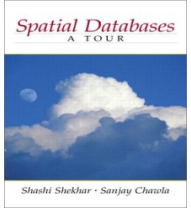
7.2.4 Categorizing Families of SDM Patterns

- Recall spatial data model concepts from Chapter 2
  - Entities - Categories of distinct, identifiable, relevant things
  - Attribute: Properties, features, or characteristics of entities
  - Instance of an entity - individual occurrence of entities
  - Relationship: interactions or connection among entities, e.g. neighbor
    - Degree - number of participating entities
    - Cardinality - number of instance of an entity in an instance of relationship
    - Self-referencing - interaction among instance of a single entity
  - Instance of a relationship - individual occurrence of relationships

- Pattern families (PF) in entity relationship models
  - Relationships among entities, e.g. neighbor
  - Value-based interactions among attributes,
    - e.g. Value of Student.age is determined by Student.date-of-birth

7.2.4 Families of SDM Patterns

- Common families of spatial patterns
  - Location Prediction:
    - Determination of value of a special attribute of an entity is by values of other attributes of the same entity
  - Spatial Interaction:
    - N-ry interaction among subsets of entities
    - N-ry interactions among categorical attributes of an entity
  - Hot spots: self-referencing interaction among instances of an entity
  - ...
- Note:
  - Other families of spatial patterns may be defined
  - SDM is a growing field, which should accommodate new pattern families

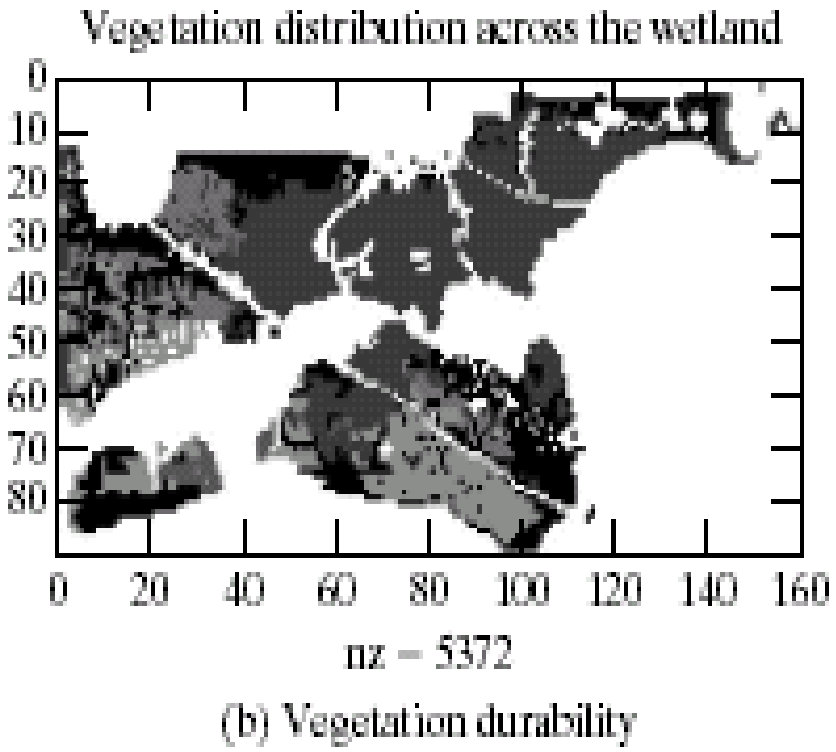


Unique Properties of Spatial Patterns

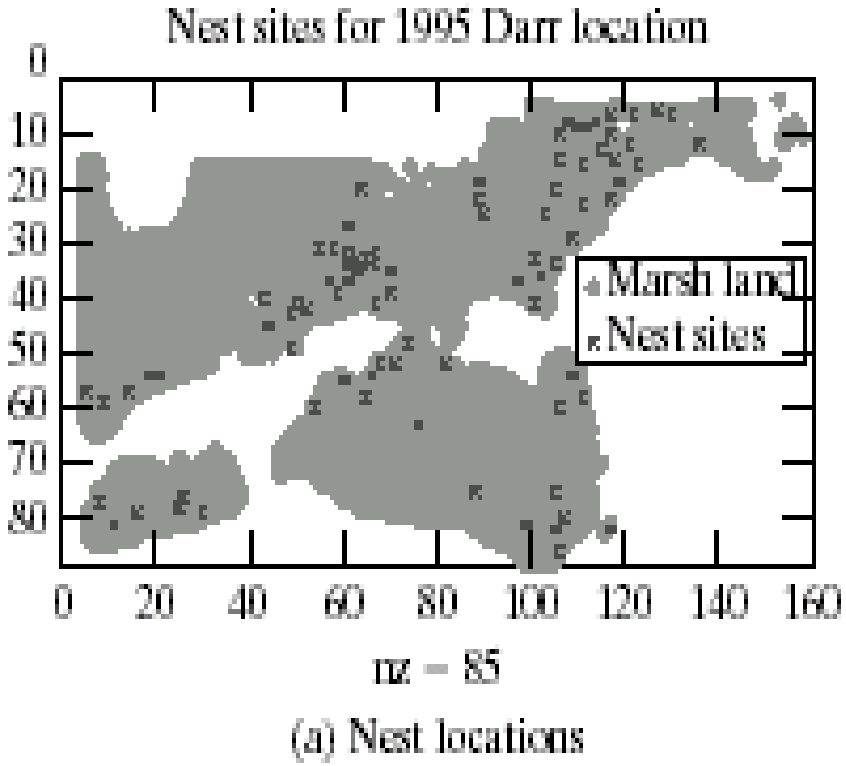
- Items in a traditional data are independent of each other,
  - whereas properties of locations in a map are often "**auto-correlated**".
- Traditional data deals with simple domains, e.g. numbers and symbols,
  - whereas spatial data types are complex
- Items in traditional data describe discrete objects
  - whereas spatial data is continuous
- First law of geography [Tobler]:
  - Everything is related to everything, but nearby things are more related than distant things.
  - People with similar backgrounds tend to live in the same area
  - Economies of nearby regions tend to be similar
  - Changes in temperature occur gradually over space(and time)

Example: Clusterng and Auto-correlation

- Note clustering of nest sites and smooth variation of spatial attributes
- (Figure 7.3, pp. 188 includes maps of two other attributes)
- Also see Fig. 7.4 (pp. 189) for distributions with no autocorrelation



(b) Vegetation durability



(a) Nest locations

Moran's I: A measure of spatial autocorrelation

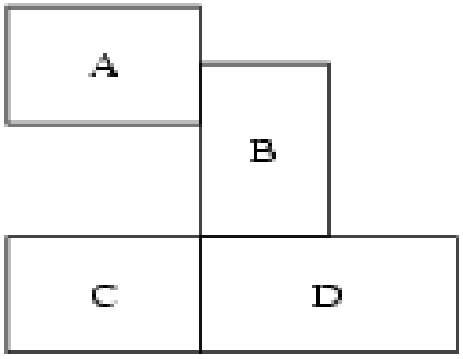
- Given $x = \{x_1, \ldots x_n\}$ sampled over n locations. Moran I is defined as

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

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

and W is a normalized contiguity matrix.

Fig. 7.5, pp. 190



(a) Map

	A	B	C	D
A	0	1	0	0
B	1	0	1	1
C	0	1	0	1
D	0	1	1	0

(b) Boolean W

	A	B	C	D
A	0	1	0	0
B	0.3	0	0.3	0.3
C	0	0.5	0	0.5
D	0	0.5	0.5	0

(c) Row-normalized W

Moran I - example

Figure 7.5, pp. 190

3	g	h	i
2	d	e	f
1	a	b	c
	1	2	3

(a)

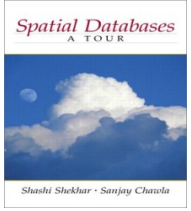
3	5	7	11
2	6	10	13
1	8	14	16
	1	2	3

(b)

3	5	7	13
2	8	16	14
1	10	11	6
	1	2	3

(c)

- Pixel value set in (b) and (c) are same Moran I is different.
- Q? Which dataset between (b) and (c) has higher spatial autocorrelation?



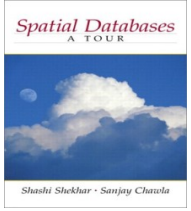
Basic of Probability Calculus

- Given a set of events Ω , the probability P is a function from into [0,1] which satisfies the following two axioms
  - and $P(\Omega) = 1$
  - If A and B are mutually exclusive events then P(AB) = P(A)P(B)

- Conditional Probability:
  - Given that an event B has occurred the conditional probability that event A will occur is P(A|B). A basic rule is
  - P(AB) = P(A|B)P(B) = P(B|A)P(A)

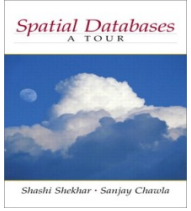
- Baye's rule: allows inversions of probabilities $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$

- Well known regression equation $Y = X\beta + \varepsilon$
  - allows derivation of linear models



Learning Objectives

- Learning Objectives (LO)
  - LO1: Understand the concept of spatial data mining (SDM)
  - LO2 : Learn about patterns explored by SDM
  - LO3: Learn about techniques to find spatial patterns
    - Mapping SDM pattern families to techniques
    - classification techniques
    - Association Rule techniques
    - Clustering techniques
    - Outlier Detection techniques
- Focus on concepts not procedures!
- Mapping Sections to learning objectives
  - LO1 - 7.1
  - LO2 - 7.2.4
  - LO3 - 7.3 - 7.6



Mapping Techniques to Spatial Pattern Families

- Overview
  - There are many techniques to find a spatial pattern familiy
  - Choice of technique depends on feature selection, spatial data, etc.
- Spatial pattern families vs. Techniques
  - Location Prediction: Classification, function determination
  - Interaction : Correlation, Association, Colocations
  - Hot spots: Clustering, Outlier Detection
- We discuss these techniques now
  - With emphasis on spatial problems
  - Even though these techniques apply to non-spatial datasets too

Location Prediction as a classification problem

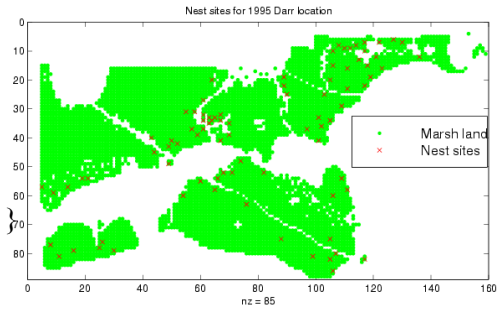
Given:

1. Spatial Framework $S = \{s_1, \ldots s_n\}$

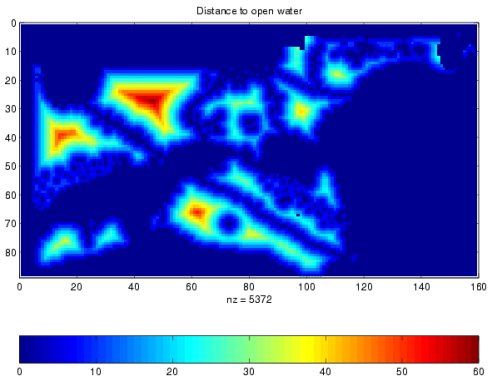
2. Explanatory functions: $f_{X_k} : S \rightarrow \Re$

3. A dependent class: $f_C : S \rightarrow \Im = \{c_1, \ldots c_M\}$

4. A family $\Im$ of function mappings: $R \times \ldots \times \Re \rightarrow \Im$



Nest locations



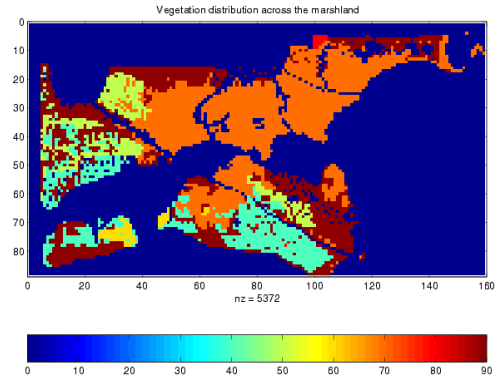
Distance to open water

Find: Classification model: $\hat{f}_c \in \Im$

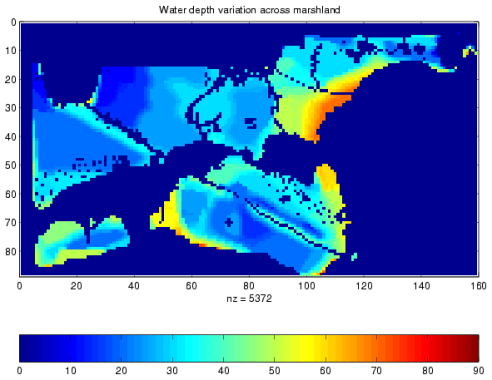
Objective: maximize classification_accuracy $(\hat{f}_c, f_c)$

Constraints:

Spatial Autocorrelation exists



Vegetation durability

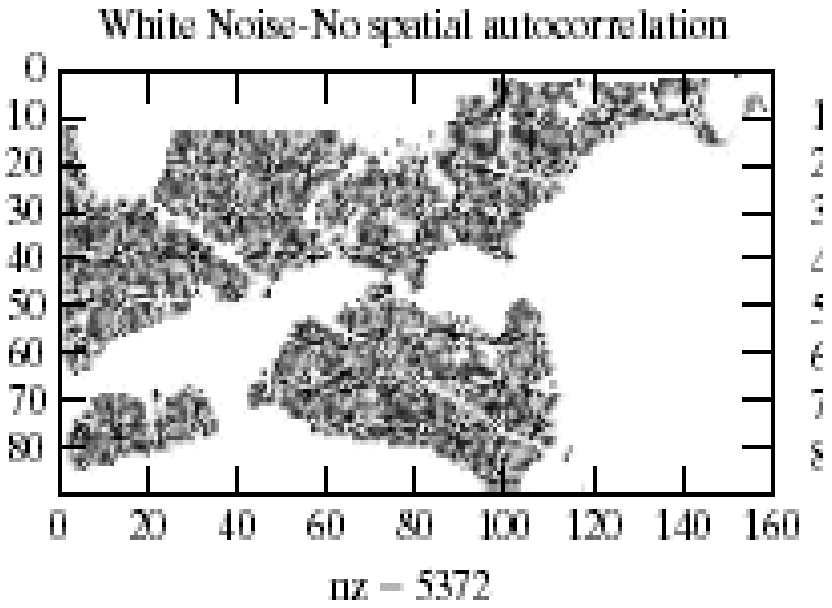


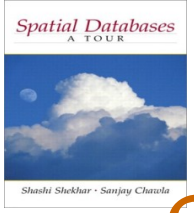
Water depth

Color version of Fig. 7.3, pp. 188

Techniques for Location Prediction

- Classical method:
  - logistic regression, decision trees, bayesian classifier
  - assumes learning samples are independent of each other
  - Spatial auto-correlation violates this assumption!
  - Q? What will a map look like where the properties of a pixel was independent of the properties of other pixels? (see below - Fig. 7.4, pp. 189)
- New spatial methods
  - Spatial auto-regression (SAR),
  - Markov random field
    - bayesian classifier



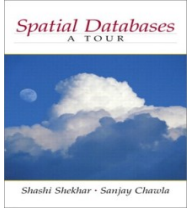


Spatial AutoRegression (SAR)

- Spatial Autoregression Model (SAR)
  - $y = \rho Wy + X\beta + \varepsilon$
    - W models neighborhood relationships
    - ρ models strength of spatial dependencies
    - ε error vector
  - Solutions
    - ρ and β - can be estimated using ML or Bayesian stat.
    - e.g., spatial econometrics package uses Bayesian approach using sampling-based Markov Chain Monte Carlo (MCMC) method.
    - Likelihood-based estimation requires $O(n^3)$ ops.
    - Other alternatives – divide and conquer, sparse matrix, LU decomposition, etc.

Model Evaluation

- Confusion matrix M for 2 class problems
  - 2 Rows: actual nest (True), actual non-nest (False)
  - 2 Columns: predicted nests (Positive), predicted non-nest (Negative)
  - 4 cells listing number of pixels in following groups
    - Figure 7.7 (pp. 196)
    - Nest is correctly predicted—True Positive(TP)
    - Model can predict nest where there was none—False Positive(FP)
    - No-nest is correctly classified--(True Negative)(TN)
    - No-nest is predicted at a nest--(False Negative)(FN)



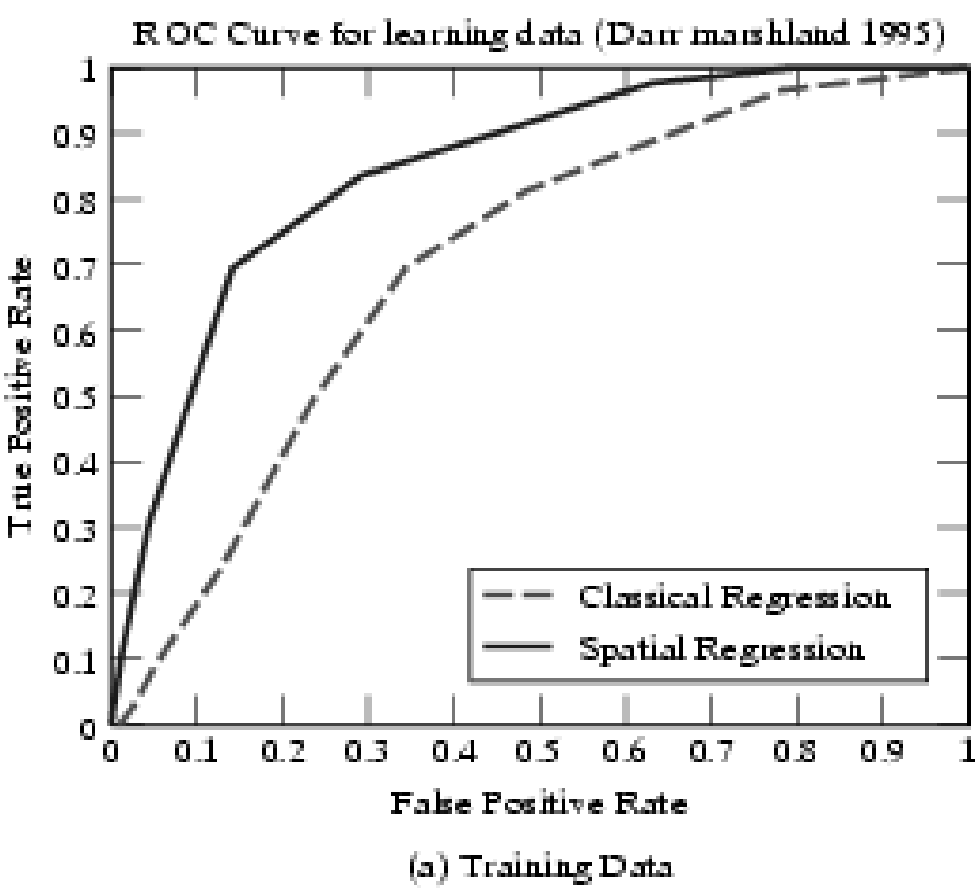
Model evaluation…cont

- Outcomes of classification algorithms are typically probabilities
- Probabilities are converted to class-labels by choosing a threshold level b.
- For example probability > b is "nest" and probability < b is "no-nest"
- TPR is the True Positive Rate, FPR is the False Positive Rate

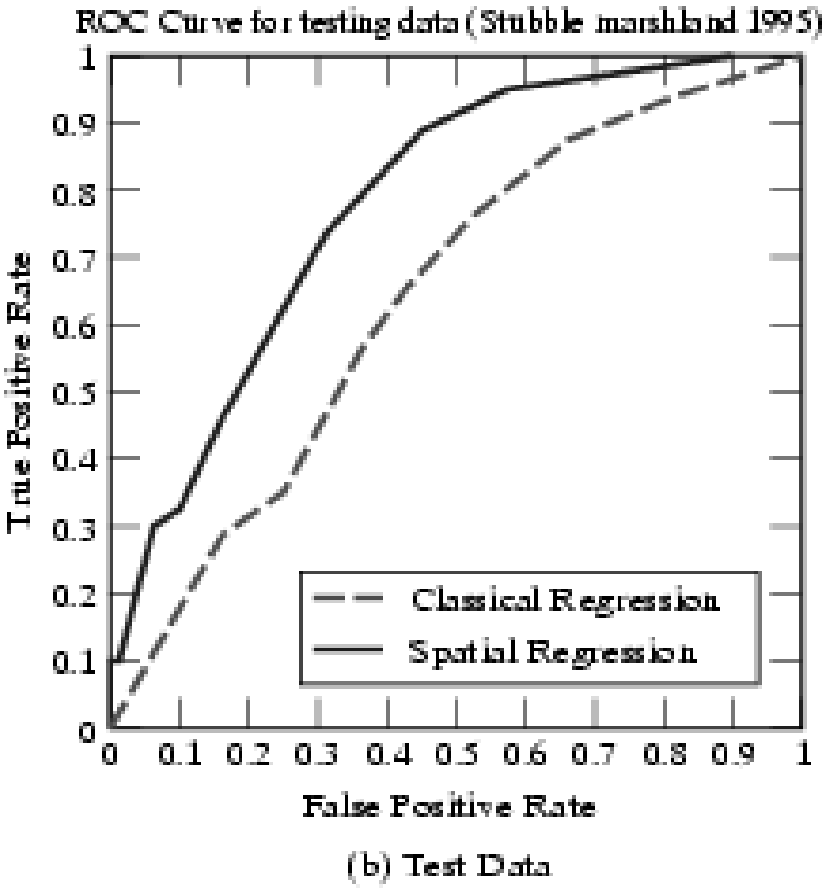
$$TPR\ (b) = \frac{TP\ (b)}{TP\ (b) + FN\ (b)}$$

$$FPR\ (b) = \frac{FP\ (b)}{FP\ (b) + TN\ (b)}$$

Comparing Linear and Spatial Regression



(a) Training Data

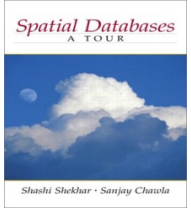


(b) Test Data

- The further the curve away from the the line TPR=FPR the better
- SAR provides better predictions than regression model. (Fig. 7.8, pp. 197)

MRF Bayesian Classifier

- Markov Random Field based Bayesian Classifiers
  - $Pr(l_i \mid X, L_i) = Pr(X|l_i, L_i)\ Pr(l_i \mid L_i) / Pr(X)$
    - $Pr(l_i \mid L_i)$ can be estimated from training data
    - L_i denotes set of labels in the neighborhood of si excluding labels at si
    - $Pr(X|l_i, L_i)$ can be estimated using kernel functions
  - Solutions
    - stochastic relaxation [Geman]
    - Iterated conditional modes [Besag]
    - Graph cut [Boykov]

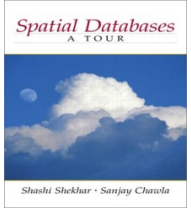


Comparison (MRF-BC vs. SAR)

- SAR can be rewritten as $y = (QX)\,\beta + Q\varepsilon$
  - where $Q = (I - \rho W)^{-1}$, a spatial transform.
  - SAR assumes linear separability of classes in transformed feature space
- MRF model may yields better classification accuracies than SAR,
  - if classes are not linearly separable in transformed space.
- The relationship between SAR and MRF are analogous to the relationship between logistic regression and Bayesian classifiers.

MRF vs. SAR (Summary)

Criteria	Spatial Autoregression model	Markov Random Field Bayesian Classifier
Input	$f_{x1}, \ldots, f_{xk}, f_l$	$f_{x1}, \ldots, f_{xk}, f_l$
Intermediate Results	β, ρ	$Pr(l_i \mid L_i)$, $Pr(X \mid l_i, L_i)$
Output	$Pr(l_i \mid X, L_i)$ based on β, ρ	$Pr(l_i \mid X, L_i)$ based on
Decision	Select most likely class For a given feature value	Select most likely class For a given feature value
Assumptions - $Pr(X \mid l_i)$ - Class bndry - Autocorrelation	 Exponential family Linearly separable Yes	 - - Yes



Learning Objectives

- Learning Objectives (LO)
  - LO1: Understand the concept of spatial data mining (SDM)
  - LO2 : Learn about patterns explored by SDM
  - LO3: Learn about techniques to find spatial patterns
    - Mapping SDM pattern families to techniques
    - classification techniques
    - Association Rule techniques
    - Clustering techniques
    - Outlier Detection techniques
- Focus on concepts not procedures!
- Mapping Sections to learning objectives
  - LO1 - 7.1
  - LO2 - 7.2.4
  - LO3 - 7.3 - 7.6

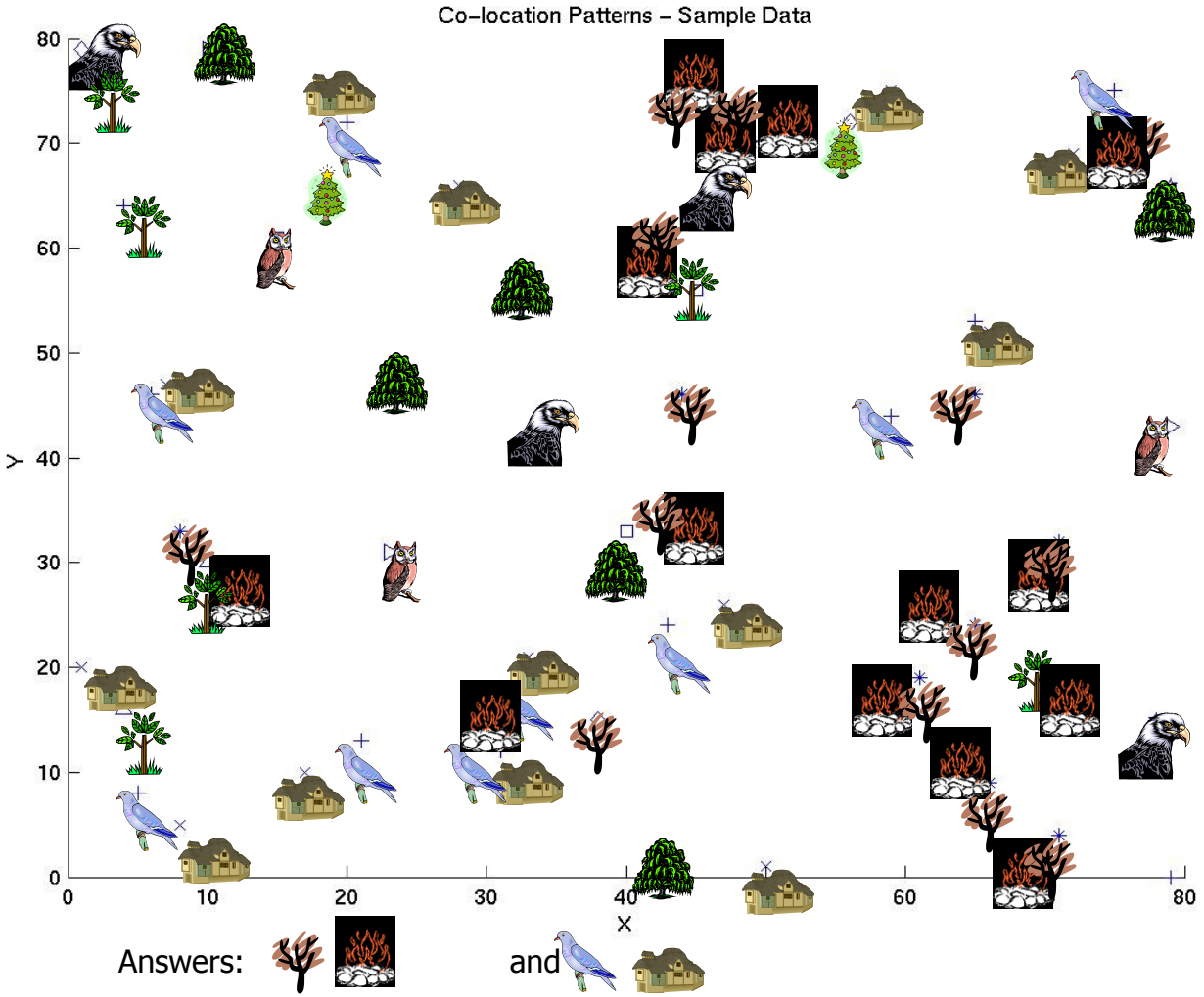
Techniques for Association Mining

- Classical method:
  - Association rule given item-types and transactions
  - assumes spatial data can be decomposed into transactions
  - However, such decomposition may alter spatial patterns
- New spatial methods
  - Spatial association rules
  - Spatial co-locations

- Note: Association rule or co-location rules are fast filters to reduce the number of pairs for rigorous statistical analysis, e.g correlation analysis, cross-K-function for spatial interaction etc.

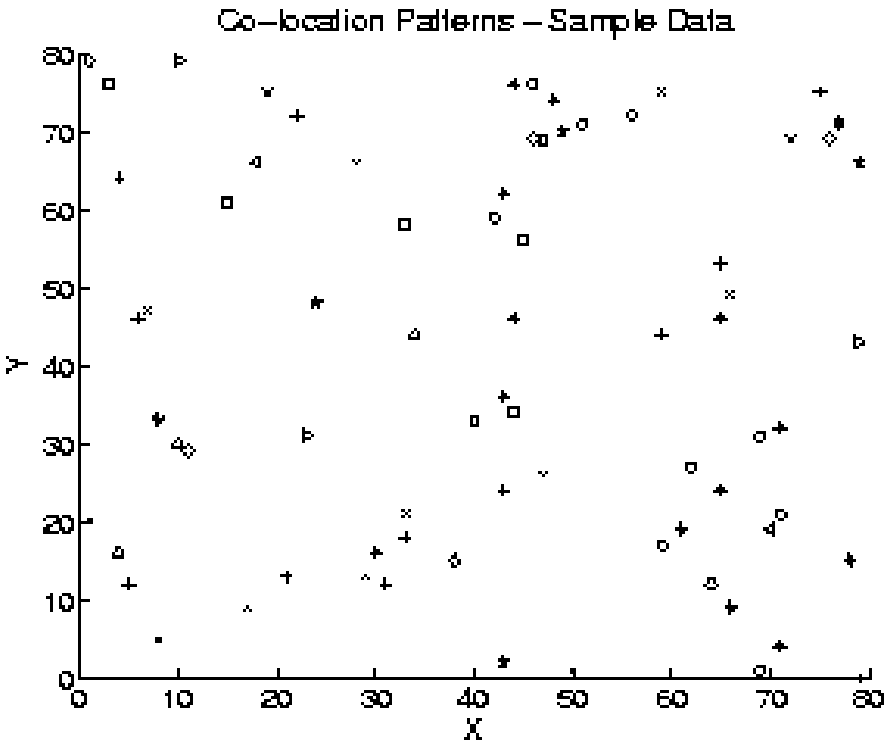
- Motivating example - next slide

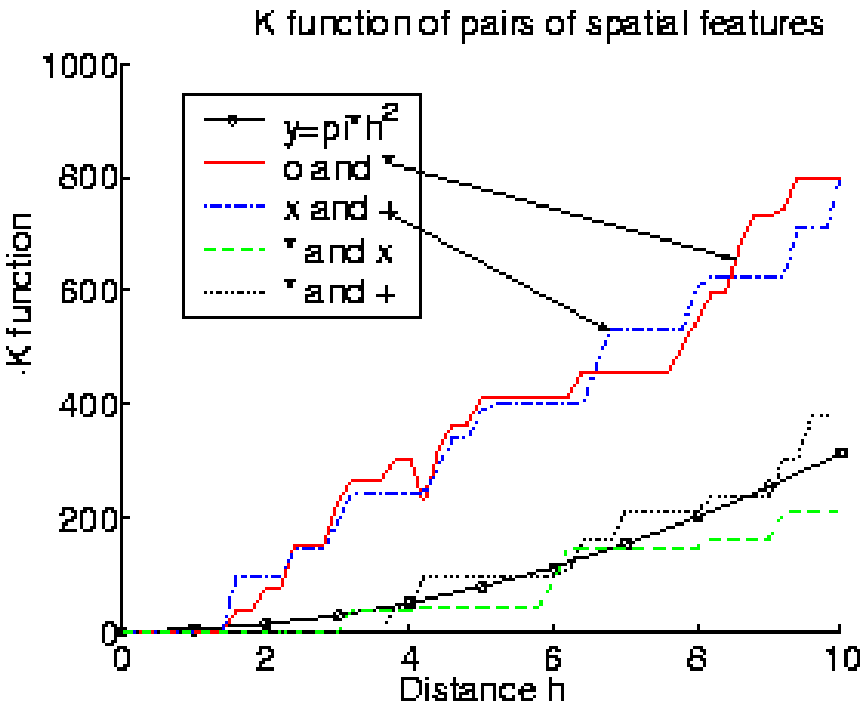
Associations, Spatial associations, Co-location



find patterns from the following sample dataset?

Colocation Rules – Spatial Interest Measures





Association Rules Discovery

- Association rules has three parts
  - rule: X→Y or antecedent (X) implies consequent (Y)
  - Support = the number of time a rule shows up in a database
  - Confidence = Conditional probability of Y given X
- Examples
  - Generic - Diaper-beer sell together weekday evenings [Walmart]
  - Spatial:
    - (bedrock type = limestone), (soil depth < 50 feet) => (sink hole risk = high)
    - support = 20 percent, confidence = 0.8
    - Interpretation: Locations with limestone bedrock and low soil depth have high risk of sink hole formation.

Association Rules: Formal Definitions

- Consider a set of items, $I = \{i_1, \ldots, i_k\}$
- Consider a set of transactions $T = \{t_1, \ldots, t_n\}$
  - where each t_i is a subset of I.
- Support of C $\sigma(C) = \{t \mid t \in T, C \subset t\}$
- Then $i_1 \rightarrow i_2$ iff
  - Support: occurs in at least s percent of the transactions: $\frac{\sigma(i_1 \wedge i_2)}{|T|}$
  - Confidence: Atleast c% $\frac{\sigma(i_1 \wedge i_2)}{\sigma(i_1)}$
- Example: Table 7.4 (pp. 202) using data in Section 7.4

Apriori Algorithm to mine association rules

- Key challenge
  - Very large search space
  - N item-types => power(2, N) possible associations
- Key assumption
  - Few associations are support above given threshold
  - Associations with low support are not intresting
- Key Insight - Monotonicity
  - If an association item set has high support, ten so do all its subsets
- Details
  - Psuedo code on pp. 203
  - Execution trace example - Fig. 7.11 (pp. 203) on next slide

Association Rules:Example

ITEMS

Car CD Player	D
Car Alarm	A
TV	T
VCR	V
Computer	C

DATABASE

1	D A V C
2	A T C
3	D A V C
4	D A T C
5	D A T V C
6	A T V

FREQUENT ITEMSETS

SUPPORT	ITEMSETS
100% (6)	A
83%(5)	C, AC
67%(4)	C, T, V, DA, DC, AT, AV, DAC
50%(3)	DV, TC, VC, DAV, DVC, ATC, AVC, DAVC

ASSOCIATION RULES WITH CONFIDENCE = 100 %

D → A (4/4)	D → A (4/4)	VC → A (3/3)
D → C (4/4)	D → A (3/3)	DV → A (3/3)
D → AC (4/4)	D → A (3/3)	VC → A (3/3)
T → C (4/4)	D → A (4/4)	DAV → A (3/3)
V → A (4/4)	D → A (3/3)	DVC → A (3/3)
C → A (5/5)	D → A (3/3)	AVC → A (3/3)

ASSOCIATION RULES WITH CONFIDENCE >=80%

C → D (4/5)	A → C (5/6)	C → DA (4/5)

Spatial Association Rules

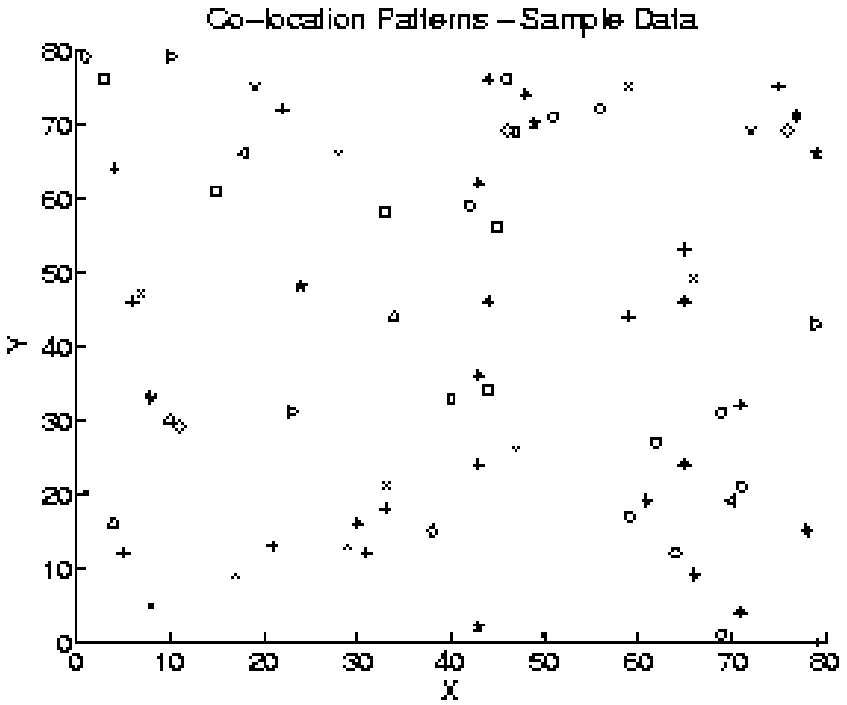
- Spatial Association Rules
  - A special reference spatial feature
  - Transactions are defined around instance of special spatial feature
  - Item-types = spatial predicates
  - Example: Table 7.5 (pp. 204)

Spatial Association Rule	Sup.	Conf.
$Stem\_height(x, high) \wedge Distance\_to\_edge(x, far) \rightarrow Vegetation\_Durability(x, moderate)$	0.1	0.94
$Vegetation\_Durability(x, moderate) \wedge Distance\_to\_water(x, close) \rightarrow Stem\_Height(x, high)$	0.05	0.95
$Distance\_towater(x, far) \wedge Water\_Depth(x, shallow) \rightarrow Stem\_Height(x, high)$	0.05	0.94

Colocation Rules

- Motivation
  - Association rules need transactions (subsets of instance of item-types)
  - Spatial data is continuous
  - Decomposing spatial data into transactions may alter patterns
- Co-location Rules
  - For point data in space
  - Does not need transaction, works directly with continuous space
  - Use neighborhood definition and spatial joins
  - "Natural approach"

Colocation Rules



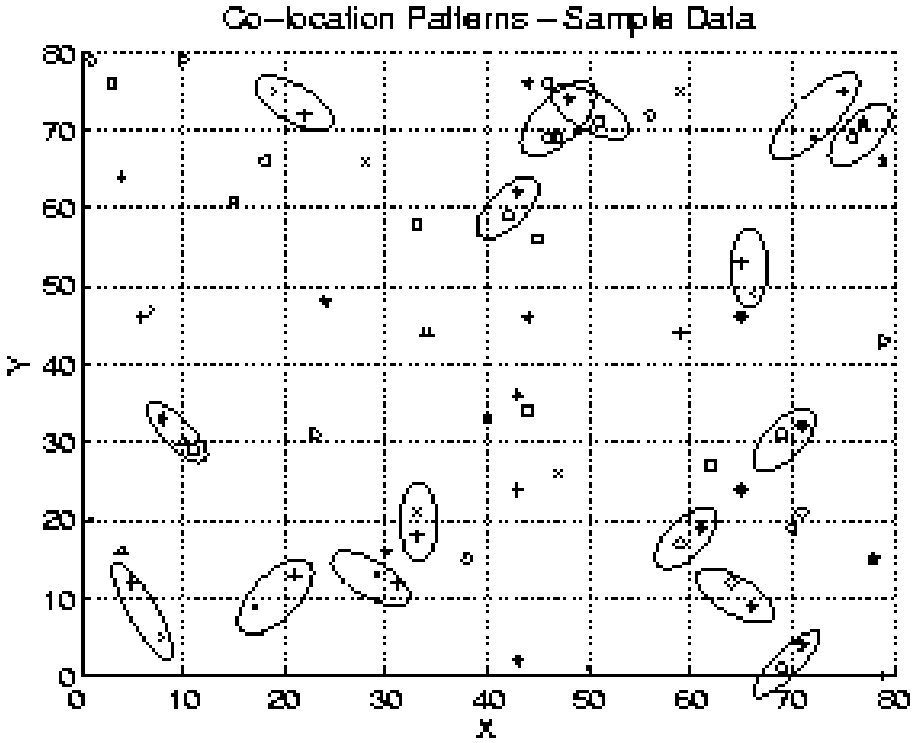
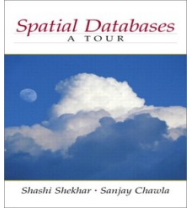


Figure 2: Transactions split circled instances of co-location patterns



Co-location rules vs. association rules

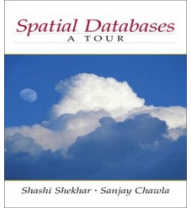
	Association rules	Co-location rules
Underlying space	discrete sets	continuous space
item-types	item-types	events /Boolean spatial features
collection	Transaction (T)	Neighborhood (N)
prevalence measure	support	participation index
conditional probability metric	Pr.[A in T | B in T]	Pr.[A in N(L) | B at location L]

Participation index = $\min\{pr(f_i, c)\}$

Where $pr(f_i, c)$ of feature f_i in co-location $c = \{f_1, f_2, \ldots, f_k\}$:

= fraction of instances of f_i with feature $\{f_1, \ldots, f_{i-1}, f_{i+1}, \ldots, f_k\}$ nearby

N(L) = neighborhood of location L

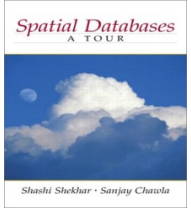


Learning Objectives

- Learning Objectives (LO)
  - LO1: Understand the concept of spatial data mining (SDM)
  - LO2 : Learn about patterns explored by SDM
  - LO3: Learn about techniques to find spatial patterns
    - Mapping SDM pattern families to techniques
    - classification techniques
    - Association Rule techniques
    - Clustering techniques
    - Outlier Detection techniques
- Focus on concepts not procedures!
- Mapping Sections to learning objectives
  - LO1 - 7.1
  - LO2 - 7.2.4
  - LO3 - 7.3 - 7.6

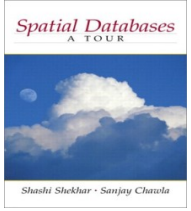
Idea of Clustering

- Clustering
  - process of discovering groups in large databases.
  - Spatial view: rows in a database = points in a multi-dimensional space
  - Visualization may reveal interesting groups
- A diverse family of techniques based on available group descriptions
- Example: census 2001
  - Attribute based groups
    - Homogeneous groups, e.g. urban core, suburbs, rural
    - Central places or major population centers
    - Hierarchical groups: NE corridor, Metropolitan area, major cities, neighborhoods
    - Areas with unusually high population growth/decline
  - Purpose based groups, e.g. segment population by consumer behaviour
    - Data driven grouping with little a priori description of groups
    - Many different ways of grouping using age, income, spending, ethnicity, ...



Spatial Clustering Example

- Example data: population density
  - Fig. 7.13 (pp. 207) on next slide
- Grouping Goal - central places
  - identify locations that dominate surroundings,
  - groups are S1 and S2
- Grouping goal - homogeneous areas
  - groups are A1 and A2
- Note: Clustering literature may not identify the grouping goals explicitly.
  - Such clustering methods may be used for purpose based group finding



Spatial Clustering Example

- Example data: population density
  - Fig. 7.13 (pp. 207)
- Grouping Goal - central places
  - identify locations that dominate surroundings,
  - groups are S1 and S2
- Grouping goal - homogeneous areas
  - groups are A1 and A2

Spatial Clustering Example

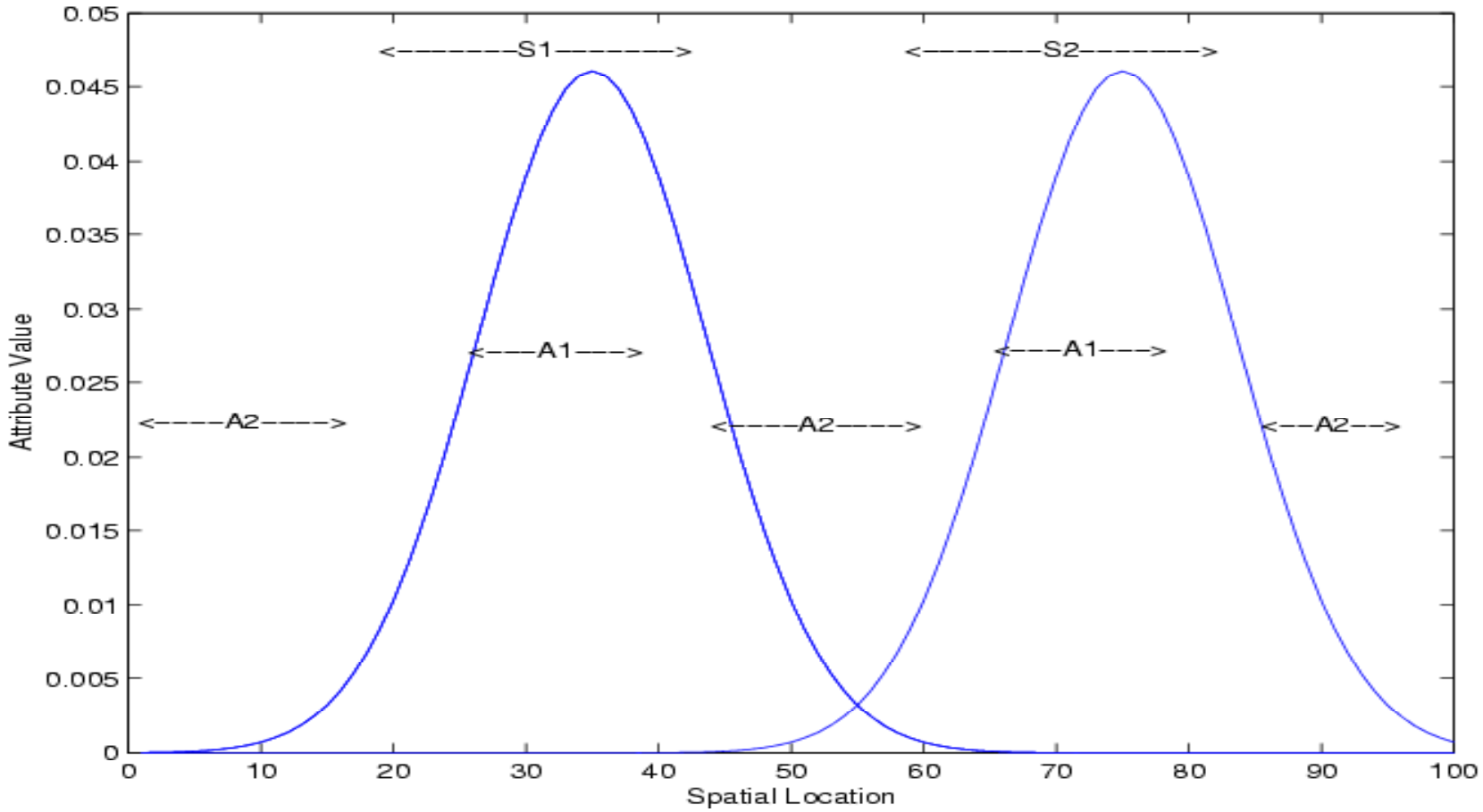


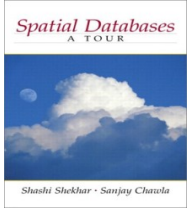
Figure 7.13 (pp. 206)

Techniques for Clustering

- Categorizing classical methods:
  - Hierarchical methods
  - Partitioning methods, e.g. K-mean, K-medoid
  - Density based methods
  - Grid based methods
- New spatial methods
  - Comparison with complete spatial random processes
  - Neighborhood EM
- Our focus:
  - Section 7.5: Partitioning methods and new spatial methods
  - Section 7.6 on outlier detection has methods similar to density based methods

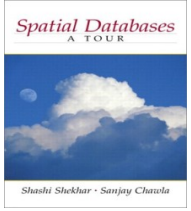
Algorithmic Ideas in Clustering

- Hierarchical—
  - All points in one clusters
  - then splits and merges till a stopping criterion is reached
- Partitional—
  - Start with random central points
  - assign points to nearest central point
  - update the central points
  - Approach with statistical rigor
- Density
  - Find clusters based on density of regions
- Grid-based—
  - Quantize the clustering space into finite number of cells
  - use thresholding to pick high density cells
  - merge neighboring cells to form clusters



Learning Objectives

- Learning Objectives (LO)
  - LO1: Understand the concept of spatial data mining (SDM)
  - LO2 : Learn about patterns explored by SDM
  - LO3: Learn about techniques to find spatial patterns
    - Mapping SDM pattern families to techniques
    - classification techniques
    - Association Rule techniques
    - Clustering techniques
    - Outlier Detection techniques
- Focus on concepts not procedures!
- Mapping Sections to learning objectives
  - LO1 - 7.1
  - LO2 - 7.2.4
  - LO3 - 7.3 - 7.6



Idea of Outliers

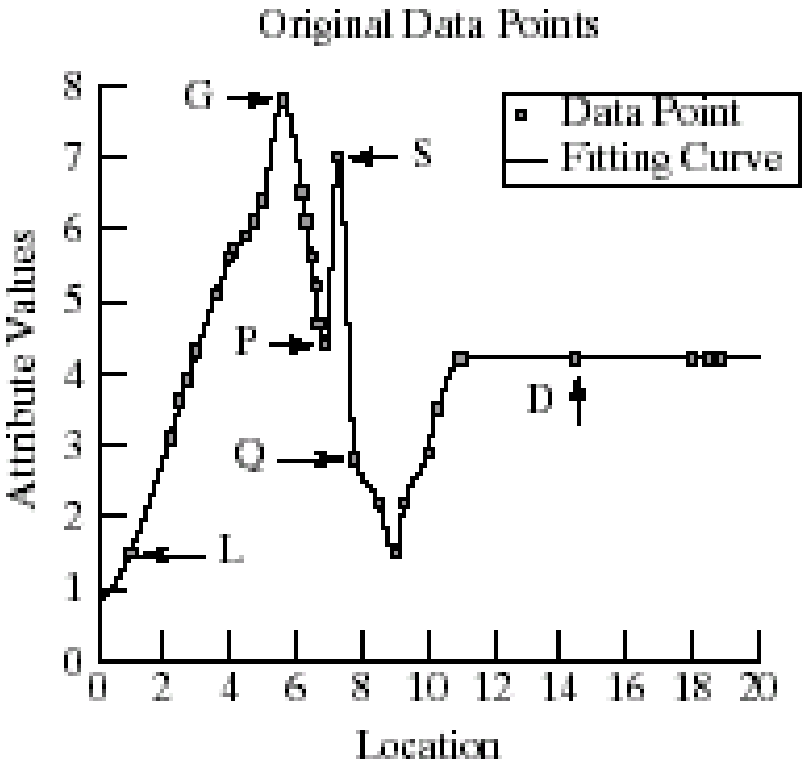
- What is an outlier?
  - Observations inconsistent with rest of the dataset
  - Ex. Point D, L or G in Fig. 7.16(a), pp. 216
  - Techniques for global outliers
    - Statistical tests based on membership in a distribution
      - Pr.[item in population] is low
    - Non-statistical tests based on distance, nearest neighbors, convex hull, etc.

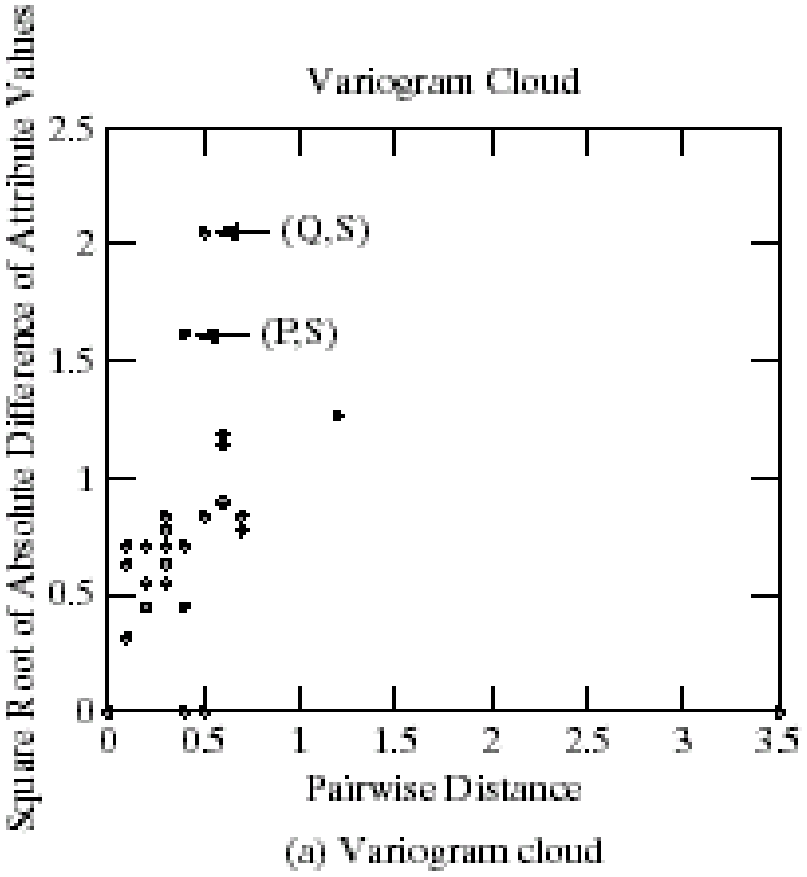
- What is a special outliers?
  - Observations inconsistent with their neighborhoods
  - A local instability or discontinuity
  - Ex. Point S in Fig. 7.16(a), pp. 216

- New techniques for spatial outliers
  - Graphical - Variogram cloud, Moran scatterplot
  - Algebraic - Scatterplot, Z(S(x))

Graphical Test 1- Variogram Cloud

- Create a variogram by plotting (attribute difference, distance) for each pair of points
- Select points (e.g. S) common to many outlying pairs, e.g. (P,S), (Q,S)

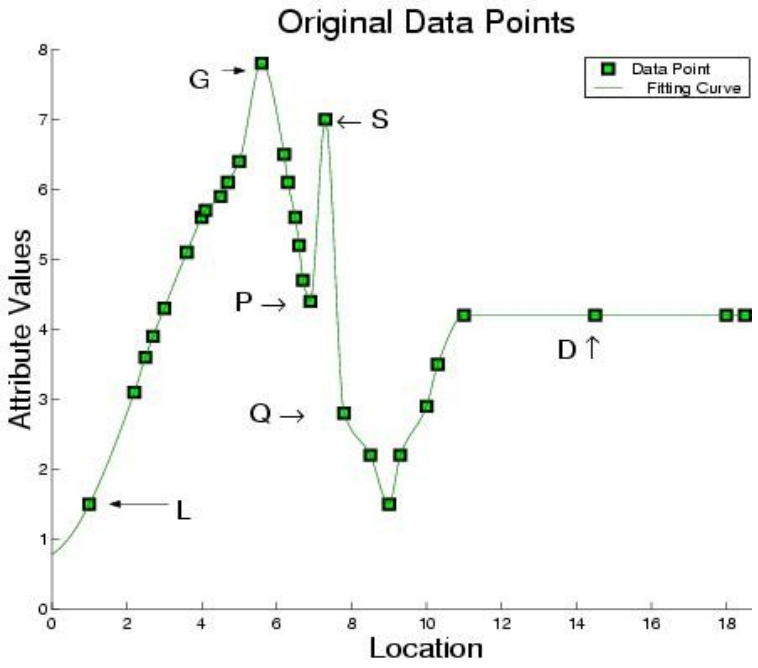




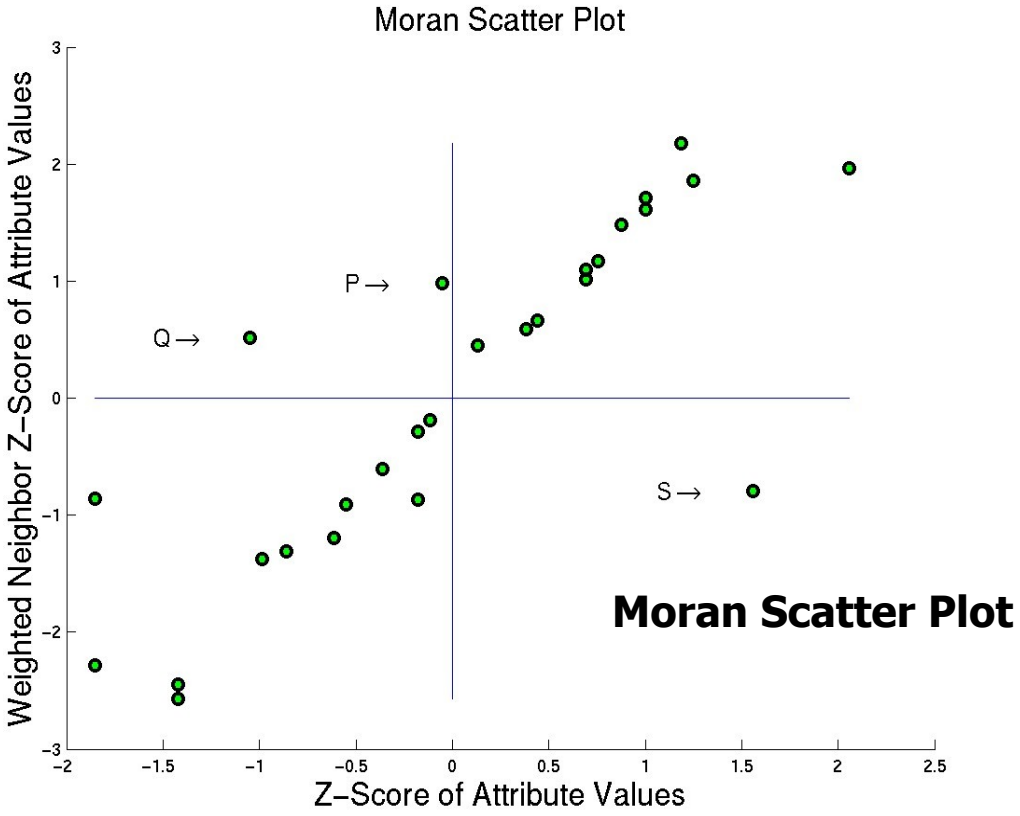
(a) Variogram cloud

Graphical Test 2- Moran Scatter Plot

- Plot (normalized attribute value, weighted average in the neighborhood) for each location
- Select points (e.g. P, Q, S) in upper left and lower right quadrant



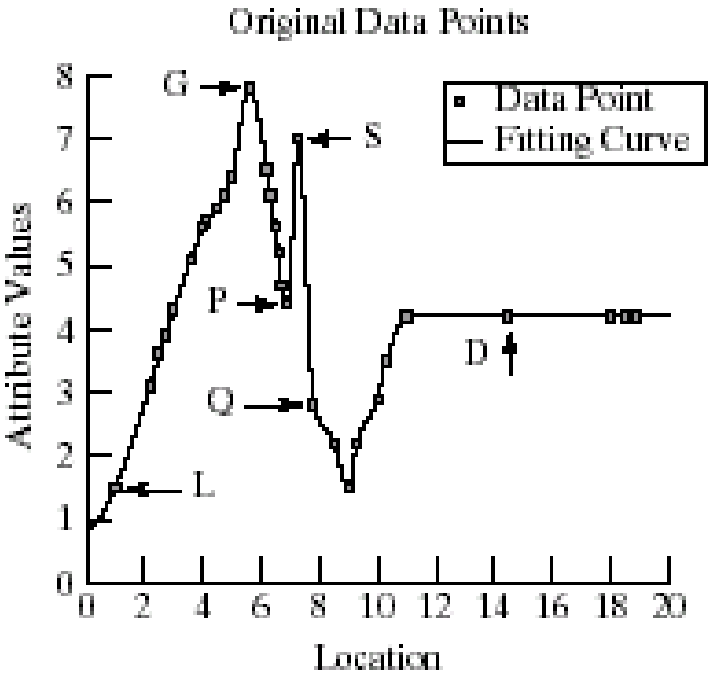
Original Data

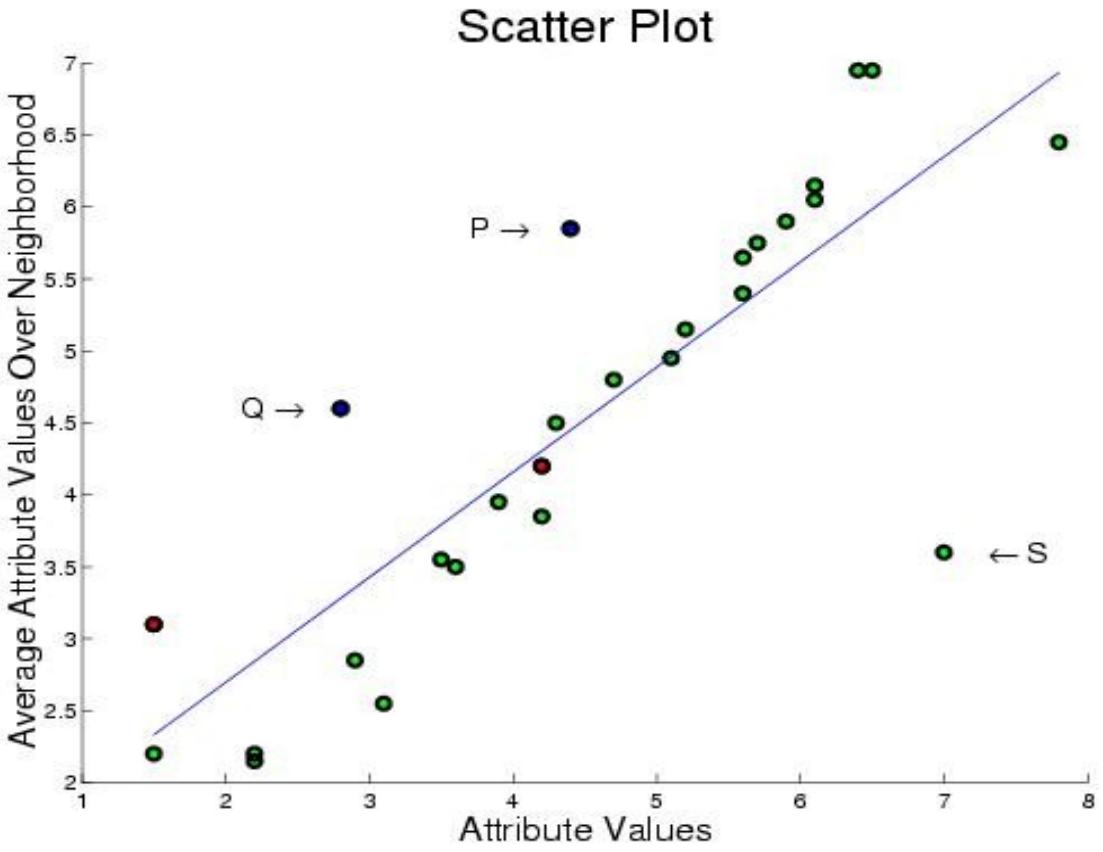


Moran Scatter Plot

Quantitative Test 1 : Scatterplot

- Plot (normalized attribute value, weighted average in the neighborhood) for each location
- Fit a linear regression line
- Select points (e.g. P, Q, S) which are unusually far from the regression line

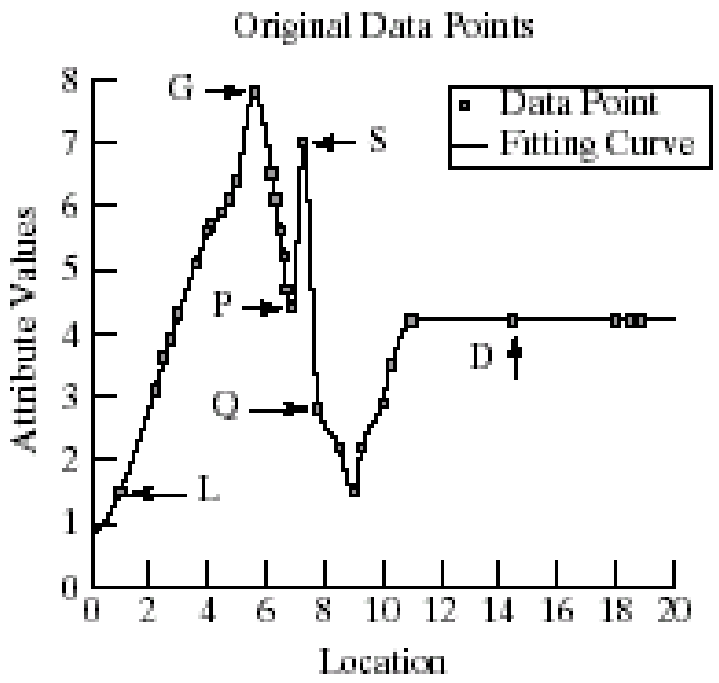


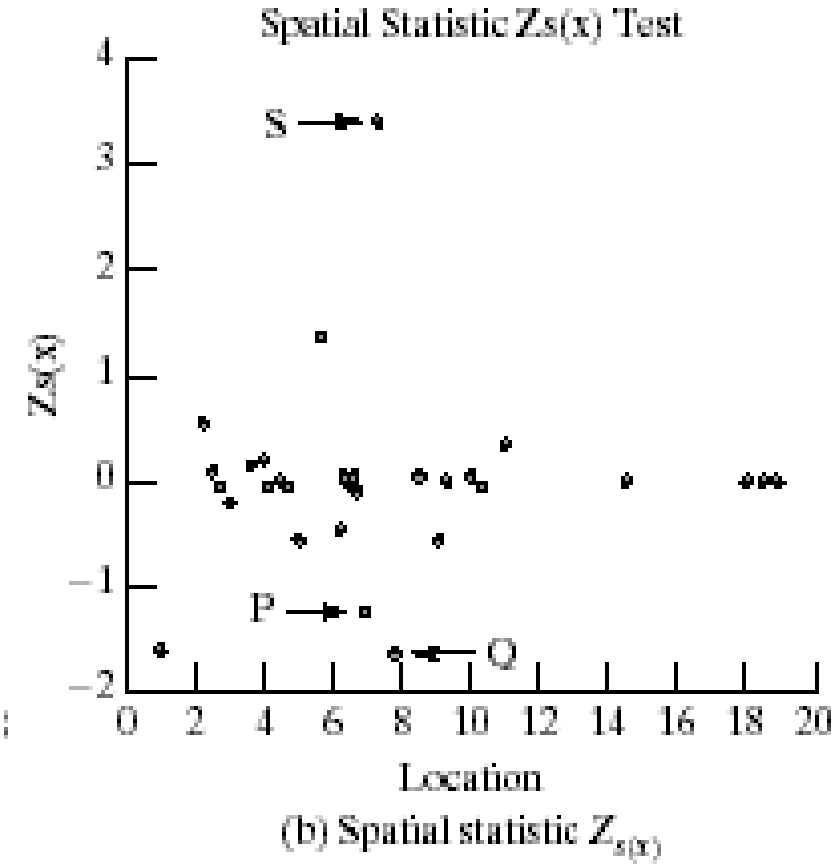


Quantitative Test 2 : Z(S(x)) Method

- Compute $Z_{S(x)} = \frac{| S(x) - \iota_s |}{\sigma(S)}$ where $S(x) = [f(x) - E_{y \in \cdot(x)}(f(y))]$

- Select points (e.g. S with Z(S(x)) above 3





(b) Spatial statistic $Z_{s(x)}$

Spatial Outlier Detection: Example

Given

A spatial graph G={V,E}

A neighbor relationship (K neighbors)

An attribute function f : V -> R

Find

O = {v_i | $v_i \in$V, v_i is a spatial outlier}

Spatial Outlier Detection Test

1. Choice of Spatial Statistic

$S(x) = [f(x) - E_{y \in N(x)}(f(y))]$

2. Test for Outlier Detection

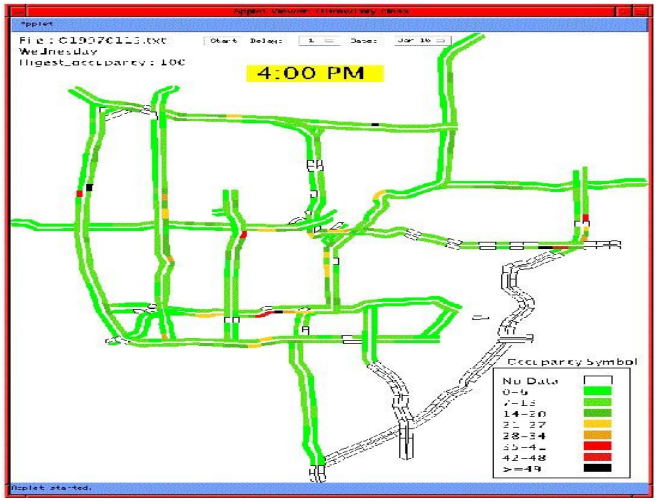
$| (S(x) - \mu_s) / \sigma_s | > \theta$

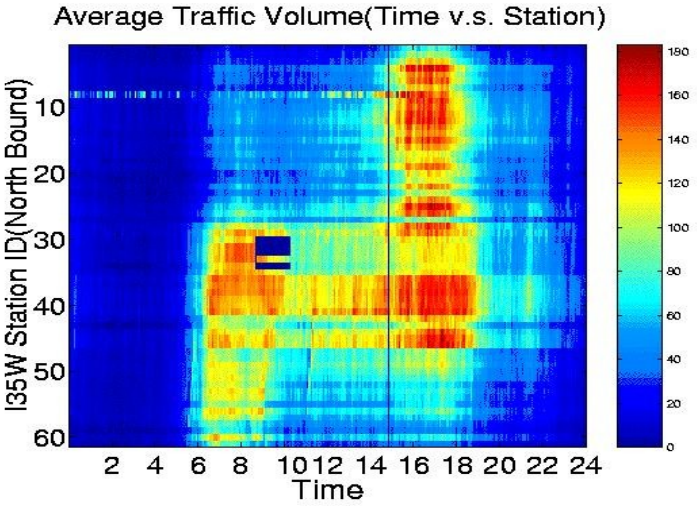
Rationale:

Theorem: S(x) is normally distributed

if f(x) is normally distributed

Color version of Fig. 7.19 pp. 219

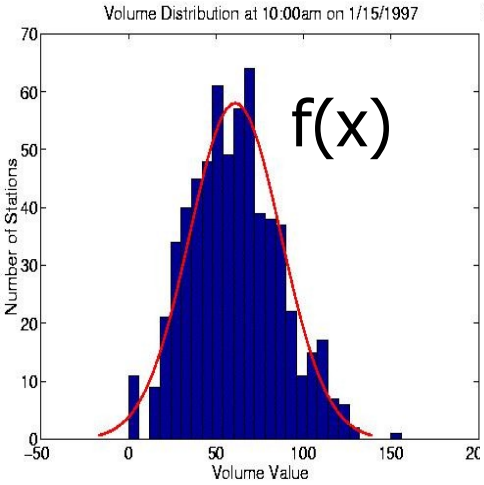


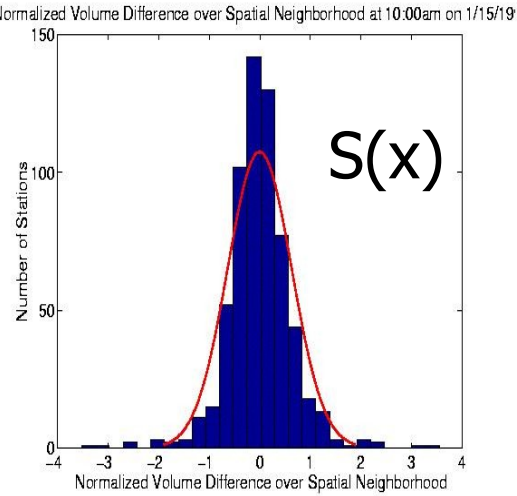


Color version of Fig. 7.21(a) pp. 220

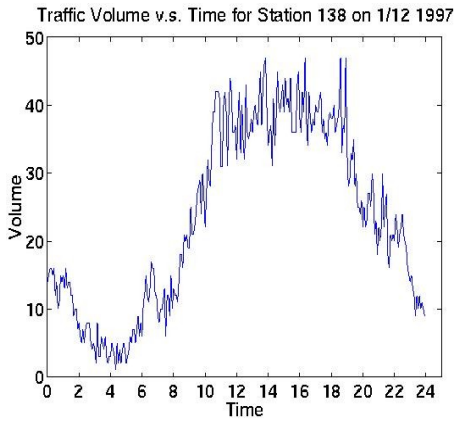
Spatial Outlier Detection- Case Study

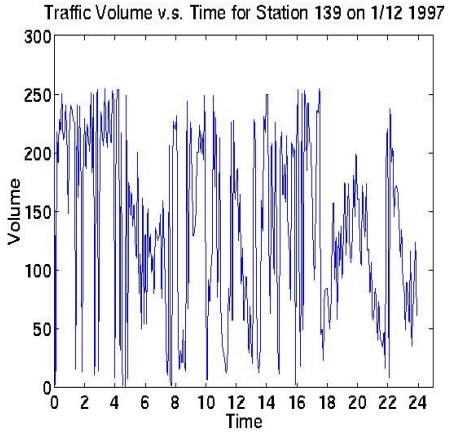
Verifying normal distribution of f(x) and S(x)

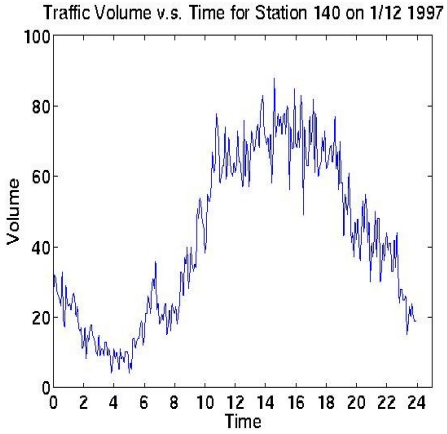


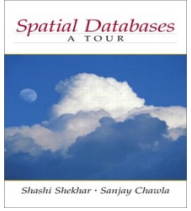


Comparing behaviour of spatial outlier (e.g. bad sensor) detexted by a test with two neighbors









Conclusions

- Patterns are opposite of random
- Common spatial patterns: location prediction, feature interaction, hot spots,
- SDM = search for unexpected interesting patterns in large spatial databases
- Spatial patterns may be discovered using
  - Techniques like classification, associations, clustering and outlier detection
  - New techniques are needed for SDM due to
    - Spatial Auto-correlation
    - Continuity of space