Data Mining:
Principles and Algorithms

— Chapter 10.1 —
— Spatiotemporal Data Mining —

©Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
www.cs.uiuc.edu/~hanj

Generalizing Spatial and Multimedia Data

- Spatial data:
  - Generalize detailed geographic points into clustered regions, such as business, residential, industrial, or agricultural areas, according to land usage
  - Require the merge of a set of geographic areas by spatial operations
- Image data:
  - Extracted by aggregation and/or approximation
  - Size, color, shape, texture, orientation, and relative positions and structures of the contained objects or regions in the image
- Music data:
  - Summarize its melody: based on the approximate patterns that repeatedly occur in the segment
  - Summarized its style: based on its tone, tempo, or the major musical instruments played

What Is a Spatial Database System?

- Geometric, geographic or spatial data: space-related data
  - Example: Geographic space (2-D abstraction of earth surface), VLSI design, model of human brain, 3-D space representing the arrangement of chains of protein molecule.
- Spatial database system vs. image database systems.
  - Image database system: handling digital raster image (e.g., satellite sensing, computer tomography), may also contain techniques for object analysis and extraction from images and some spatial database functionality.
  - Spatial (geometric, geographic) database system: handling objects in space that have identity and well-defined extents, locations, and relationships.

GIS (Geographic Information System)

- GIS (Geographic Information System)
  - Analysis and visualization of geographic data
- Common analysis functions of GIS
  - Search (thematic search, search by region)
  - Location analysis (buffer, corridor, overlay)
  - Terrain analysis (slope/aspect, drainage network)
  - Flow analysis (connectivity, shortest path)
  - Distribution (nearest neighbor, proximity, change detection)
  - Spatial analysis/statistics (pattern, centrality, similarity, topology)
  - Measurements (distance, perimeter, shape, adjacency, direction)

Spatial DBMS (SDBMS)

- SDBMS is a software system that
  - supports spatial data models, spatial ADTs, and a query language supporting them
  - supports spatial indexing, spatial operations efficiently, and query optimization
  - can work with an underlying DBMS
- Examples
  - Oracle Spatial Data Catridge
  - ESRI Spatial Data Engine

Modeling Spatial Objects

- What needs to be represented?
- Two important alternative views
  - Single objects: distinct entities arranged in space each of which has its own geometric description
    - modeling cities, forests, rivers
  - Spatially related collection of objects: describe space itself (about every point in space)
    - modeling land use, partition of a country into districts

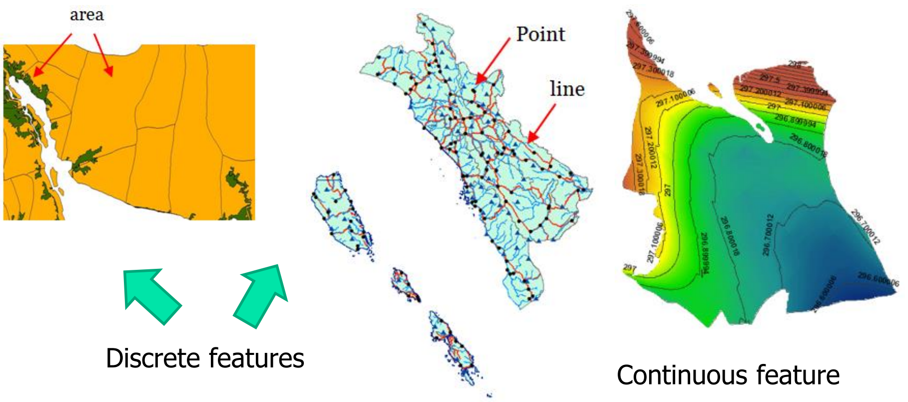
Modeling Single Objects: Point, Line and Region

- Point: location only but not extent
- Line (or a curve usually represented by a polyline, a sequence of line segment):
  - moving through space, or connections in space (roads, rivers, cables, etc.)
- Region:
  - Something having extent in 2D-space (country, lake, park). It may have a hole or consist of several disjoint pieces.

Representations of spatial features (Chang, 2008)

- Discrete features are individually distinguishable features that do not exist between observations
  - include points (e.g schools), lines (e.g roads and rivers), and polygons (e.g land use).
- Continuous features are features that exist spatially between observations
  - E.g. elevation and precipitation

Representations of spatial features (Chang, 2008)

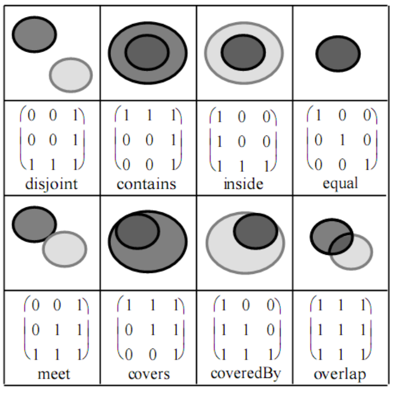


Spatial Dataset and Spatial Relationship

- In spatial data mining, a spatial relationship is essential because it makes spatial data mining algorithms differ from non-spatial data mining algorithms.
- Spatial relationships are usually stored implicitly in spatial databases.
- Three groups of relations between a spatial object and its neighborhood (Ester et al., 1997):
  - Topological-relations, e.g. meet, overlap, covered-by, contains, inside, equal
  - Metric-relation, e.g. distance < d
  - Direction-relation, e.g. north, south, west, east.

Spatial Dataset and Spatial Relationship

- The eight topological relations between two spatial regions and their corresponding 9-intersection matrices (Clementini et al., 1994)

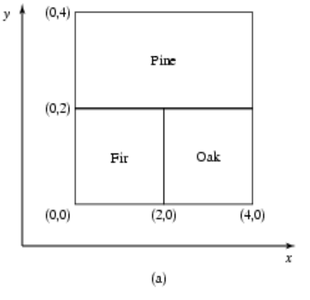


$$\Im_9(A,B)=\begin{pmatrix} A^0\cap B^0 & A^0\cap\partial B & A^0\cap B^- \\ \partial A\cap B^0 & \partial A\cap\partial B & \partial A\cap B^- \\ A^-\cap B^0 & A^-\cap\partial B & A^-\cap B^- \end{pmatrix}$$

A's boundary (∂A), A's interior (A^o), A's exterior (A$^-$), B's boundary (∂B), B's interior (B^o) and B's exterior (B$^-$)

Spatial Data Types and Models

- Field-based model: raster data
  - framework: partitioning of space
- Object-based model: vector model
  - point, line, polygon, Objects, Attributes



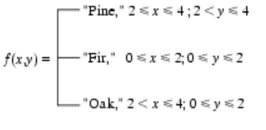
(a)

Object Viewpoint of Forest Stands

Area-ID	Dominant Tree Species	Area/Boundary
FS1	Pine	[(0,2),(4,2),(4,4),(0,4)]
FS2	Fir	[(0,0),(2,0),(2,2),(0,2)]
FS3	Oak	[(2,0),(4,0),(4,2),(2,2)]

(b)

Field Viewpoint of Forest Stands



(c)

Spatial Query Language

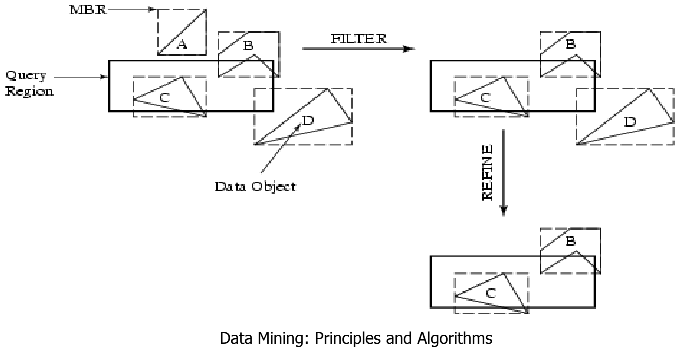
- Spatial query language
  - Spatial data types, e.g. point, line segment, polygon, …
  - Spatial operations, e.g. overlap, distance, nearest neighbor, …
  - Callable from a query language (e.g. SQL3) of underlying DBMS

```
SELECT S.name
FROM  Senator S
WHERE S.district.Area() > 300
```

- Standards
  - SQL3 (a.k.a. SQL 1999) is a standard for query languages
  - OGIS is a standard for spatial data types and operators
  - Both standards enjoy wide support in industry

Query Processing

- Efficient algorithms to answer spatial queries
- Common Strategy: filter and refine
  - Filter: Query Region overlaps with MBRs (minimum bounding rectangles) of B, C, D
  - Refine: Query Region overlaps with B, C



Spatial Data Warehousing

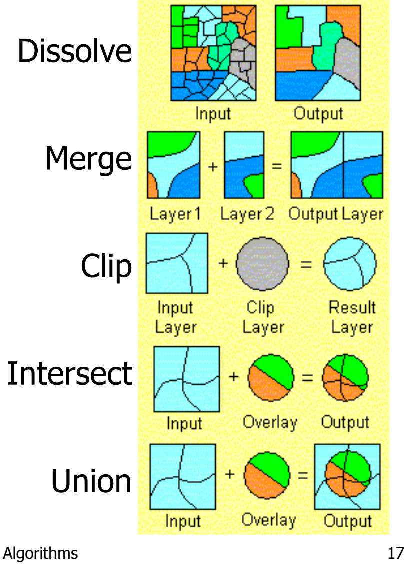
- Spatial data warehouse: Integrated, subject-oriented, time-variant, and nonvolatile spatial data repository
- Spatial data integration: a big issue
  - Structure-specific formats (raster- vs. vector-based, OO vs. relational models, different storage and indexing, etc.)
  - Vendor-specific formats (ESRI, MapInfo, Integraph, IDRISI, etc.)
  - Geo-specific formats (geographic vs. equal area projection, etc.)
- Spatial data cube: multidimensional spatial database
  - Both dimensions and measures may contain spatial components

Dimensions and Measures in Spatial Data Warehouse

- Dimensions
  - non-spatial
    - e.g. *"25-30 degrees"* generalizes to *"hot"* (both are strings)
  - spatial-to-nonspatial
    - e.g. *Seattle* generalizes to description *"Pacific Northwest"* (as a string)
  - spatial-to-spatial
    - e.g. *Seattle* generalizes to *Pacific Northwest* (as a spatial region)
- Measures
  - numerical (e.g. monthly revenue of a region)
    - distributive (e.g. count, sum)
    - algebraic (e.g. average)
    - holistic (e.g. median, rank)
  - spatial
    - collection of spatial pointers (e.g. pointers to all regions with temperature of 25-30 degrees in July)

Spatial-to-Spatial Generalization

- Generalize detailed geographic points into clustered regions, such as businesses, residential, industrial, or agricultural areas, according to land usage
- Requires the merging of a set of geographic areas by spatial operations

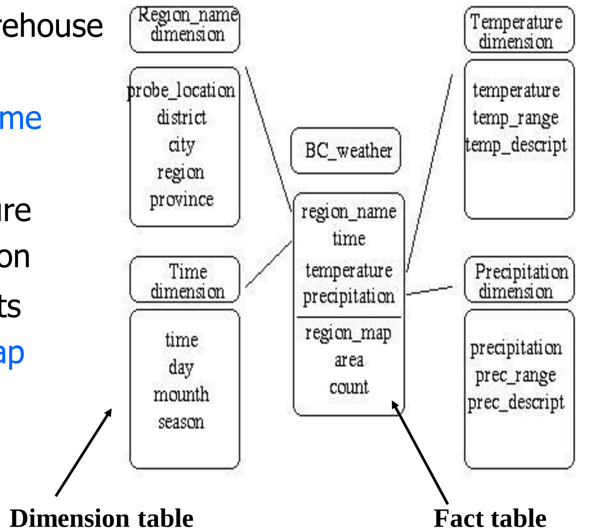


Example: British Columbia Weather Pattern Analysis

- Input
  - A map with about 3,000 weather probes scattered in B.C.
  - Daily data for temperature, precipitation, wind velocity, etc.
  - Data warehouse using star schema
- Output
  - A map that reveals patterns: merged (similar) regions
- Goals
  - Interactive analysis (drill-down, slice, dice, pivot, roll-up)
  - Fast response time
  - Minimizing storage space used
- Challenge
  - A merged region may contain hundreds of "primitive" regions (polygons)

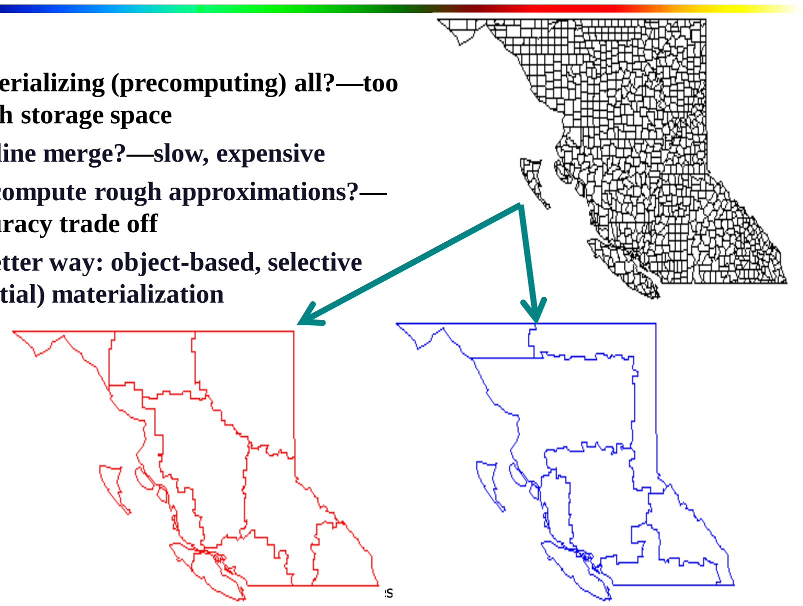
Star Schema of the BC Weather Warehouse

- Spatial data warehouse
  - Dimensions
    - region_name
    - time
    - temperature
    - precipitation
  - Measurements
    - region_map
    - area
    - count



Dynamic Merging of Spatial Objects

- **Materializing (precomputing) all?—too much storage space**
- **On-line merge?—slow, expensive**
- **Precompute rough approximations?—accuracy trade off**
- **A better way: object-based, selective (partial) materialization**



Methods for Computing Spatial Data Cubes

- On-line aggregation: collect and store pointers to spatial objects in a spatial data cube
  - expensive and slow, need efficient aggregation techniques
- Precompute and store all the possible combinations
  - huge space overhead
- Precompute and store rough approximations in a spatial data cube
  - accuracy trade-off
- Selective computation: only materialize those which will be accessed frequently
  - a reasonable choice

Spatial Association Analysis

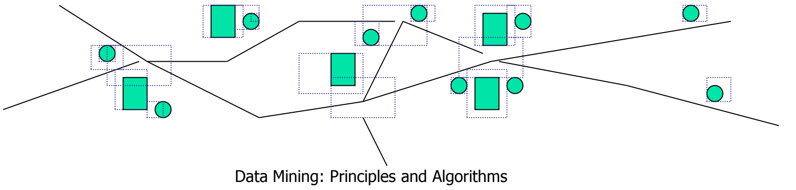
- Spatial association rule: $A \Rightarrow B\,[s\%, c\%]$
  - A and B are sets of spatial or non-spatial predicates
    - Topological relations: *intersects, overlaps, disjoint,* etc.
    - Spatial orientations: *left_of, west_of, under,* etc.
    - Distance information: *close_to, within_distance,* etc.
  - *s%* is the support and *c%* is the confidence of the rule
- Examples

1) *is_a(x, large_town) ^ intersect(x, highway) → adjacent_to(x, water)* *[7%, 85%]*

2) What kinds of objects are typically located close to golf courses?

Progressive Refinement Mining of Spatial Association Rules

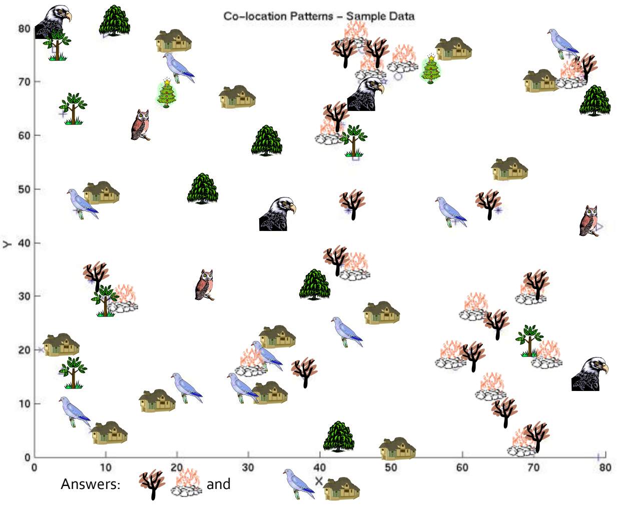
- Hierarchy of spatial relationship:
  - *g_close_to*: *near_by*, *touch*, *intersect*, *contain*, etc.
  - First search for rough relationship and then refine it
- Two-step mining of spatial association:
  - Step 1: Rough spatial computation (as a filter)
    - Using MBR or R-tree for rough estimation
  - Step2: Detailed spatial algorithm (as refinement)
    - Apply only to those objects which have passed the rough spatial association test (no less than *min_support*)



Mining Spatial Co-location Rules

- Co-location rule is similar to association rule but explore more relying spatial auto-correlation
- It leads to efficient processing
- It can be integrated with progressive refinement to further improve its performance
- Spatial co-location mining idea can be applied to clustering, classification, outlier analysis and other potential mining tasks

Associations, Spatial associations, Co-location



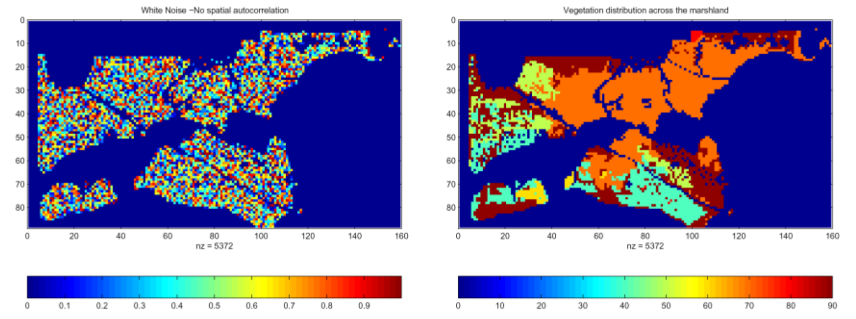
find patterns from the sample dataset?

Spatial Autocorrelation

- Spatial data tends to be highly self-correlated
  - Example: Neighborhood, Temperature
  - Items in a traditional data are independent of each other, whereas properties of locations in a map are often "**auto-correlated**".
- First law of geography:

  "*Everything is related to everything, but nearby things are more related than distant things*."

Spatial Autocorrelation (cont'd)



(a) Pixel property with independent identical distribution

(b) Vegetation Durability with SA

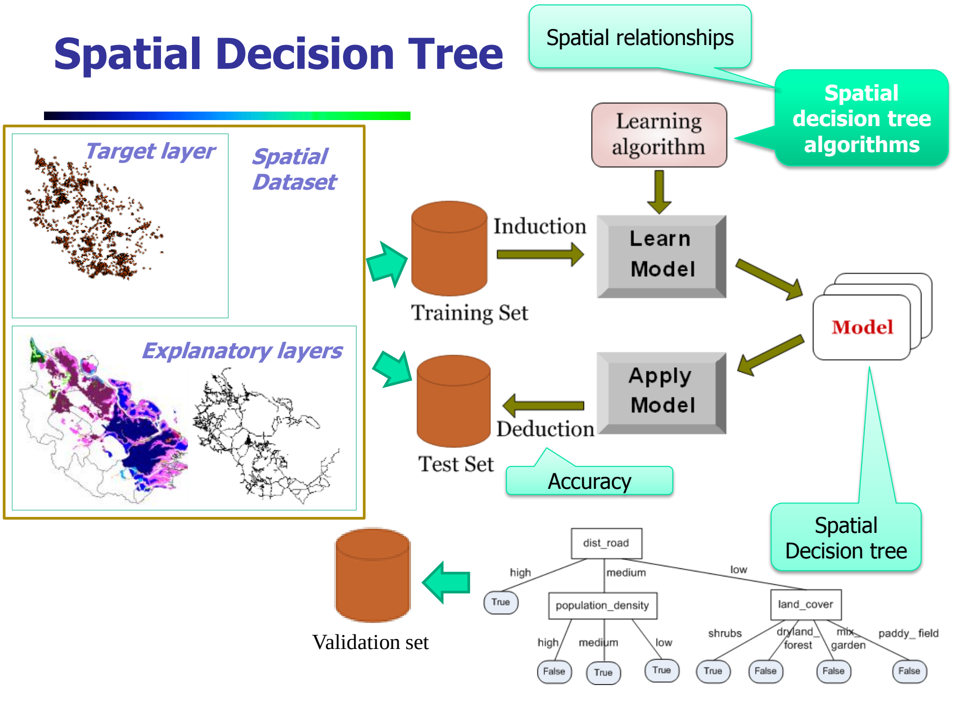
Spatial Classification

- Methods in classification
  - Decision-tree classification, Naïve-Bayesian classifier + boosting, neural network, logistic regression, etc.
  - Association-based multi-dimensional classification - Example: classifying house value based on proximity to lakes, highways, mountains, etc.
- Assuming learning samples are independent of each other
  - Spatial auto-correlation violates this assumption!
- Popular spatial classification methods
  - Spatial auto-regression (SAR)
  - Markov random field (MRF)

Spatial Decision Tree

- The works that apply the decision tree algorithms on spatial data can be divided into two groups
  - Applications of conventional decision tree algorithms (non-spatial algorithms) on spatial data.
    - Spatial data preprocessing steps need to be performed to prepare task relevant data
  - Applications of spatial decision tree algorithms on spatial data
    - Improvement of non-spatial algorithms is required involving spatial relationships

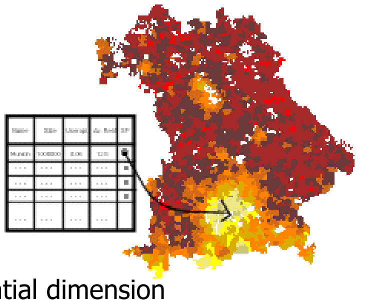
Spatial Decision Tree



Spatial Decision Tree

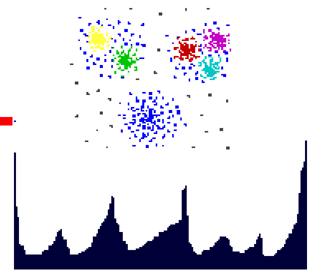
Algorithm	Type of spatial features	Spatial relationship
Spatial Decision Tree Algorithm based on the ID3 algorithm (Ester et al., 1997)	Point, line, and polygon	Distance
Spatial binary tree algorithm (Koperski et al. 1998)	Point, line, and polygon	Distance using buffer
SCART (Spatial Classification and Regression Trees) as the extension of the CART method (Chelghoum et al., 2002)	Point, line, and polygon	Topological and distance relationships applied in data pre-processing
Spatial Decision Tree based on the ID3 algorithm (Rinzivillo and Franco, 2004)	Polygon	Intersection
Spatial entropy-based decision tree method (Li and Claramunt, 2006)	Point, polygon	Distance
Extended Spatial ID3 algorithm (Sitanggang et al., 2011; Sitanggang et al., 2013)	Point, line, and polygon	Topological and distance relationships

Spatial Trend Analysis

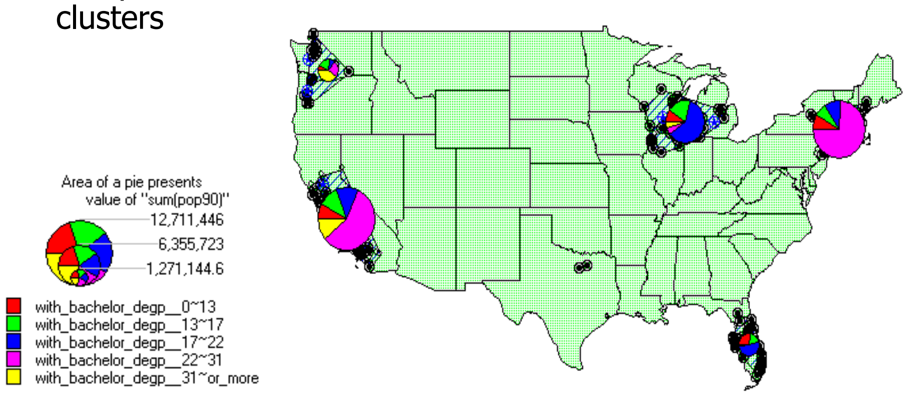


- Function
  - Detect changes and trends along a spatial dimension
  - Study the trend of non-spatial or spatial data changing with space
- Application examples
  - Observe the trend of changes of the climate or vegetation with increasing distance from an ocean
  - Crime rate or unemployment rate change with regard to city geo-distribution

Spatial Cluster Analysis



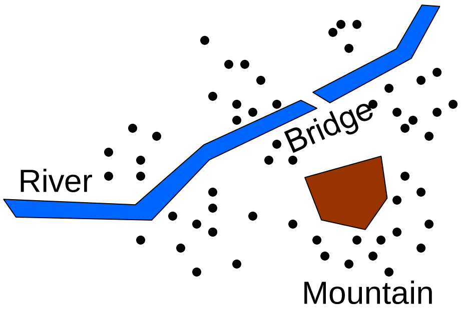
- Mining clusters—k-means, k-medoids, hierarchical, density-based, etc.
- Analysis of distinct features of the clusters



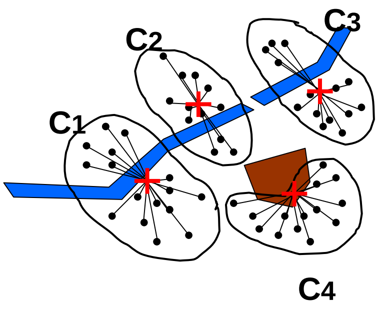
Constraints-Based Clustering

- Constraints on individual objects
  - Simple selection of relevant objects before clustering
- Clustering parameters as constraints
  - K-means, density-based: radius, min-# of points
- Constraints specified on clusters using SQL aggregates
  - Sum of the profits in each cluster > $1 million
- Constraints imposed by physical obstacles
  - Clustering with obstructed distance

Constrained Clustering: Planning ATM Locations



Spatial data with obstacles



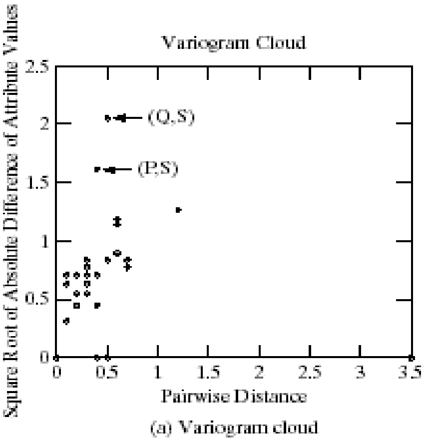
Clustering *without* taking obstacles into consideration

Spatial Outlier Detection

- Outlier
  - Global outliers: Observations which is inconsistent with the rest of the data
  - Spatial outliers: A local instability of non-spatial attributes
- Spatial outlier detection
  - Graphical tests
    - Variogram clouds
    - Moran scatterplots
  - Quantitative tests
    - Scatterplots
    - Spatial Statistic Z(S(x))
  - Quantitative tests are more accurate than Graphical tests

Spatial Outlier Detection—Variogram Clouds

- Graphical method
  - For each pair of locations, the square-root of the absolute difference between attribute values at the locations versus the Euclidean distance between the locations are plotted
  - Nearby locations with large attribute difference indicate a spatial outlier
- Quantitative method
  - Compute spatial statistic Z(S(x))



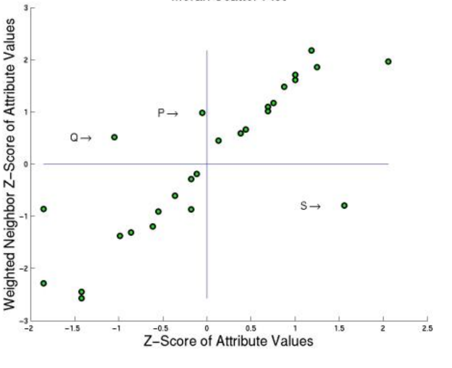
(a) Variogram cloud

Spatial Outlier Detection—Moran Scatterplots

- Graphical tests
  - A plot of normalized attribute value *Z* against the neighborhood average of normalized attribute values (*W•Z*)

$$Z[f(i)] = \frac{f(i) - u_f}{\sigma_f}$$

  - The upper left and lower right quadrants indicate a spatial outlier
- Computation method
  - Fit a linear regression line
  - Select points (e.g. P, Q, S) which are from the regression line greater than specified residual error θ



Mining Spatiotemporal Data

- Spatiotemporal data
  - Data has spatial extensions and changes with time
  - Ex: Forest fire, moving objects, hurricane & earthquakes
- Automatic anomaly detection in massive moving objects
  - Moving objects are ubiquitous: GPS, radar, etc.
  - Ex: Maritime vessel surveillance
  - Problem: Automatic anomaly detection

Tools for conducting spatial data mining

- Quantum GIS for spatial data analysis and visualization (http://www.qgis.org/)
- PostgreSQL as the database management system (http://www.postgresql.org/)
- PostGIS for spatial data analysis (http://postgis.net/)
- Python for programming (http://www.python.org/)
- R for statistical computation (http://www.r-project.org/)
- Ilwis for satellite image processing (http://www.ilwis.org/)
- The data mining tool kit Weka (http://www.cs.waikato.ac.nz/ml/weka/)

Summary

- Mining object data needs feature/attribute-based generalization methods
- Spatial, spatiotemporal and multimedia data mining is one of important research frontiers in data mining with broad applications
- Spatial data warehousing, OLAP and mining facilitates multidimensional spatial analysis and finding spatial associations, classifications and trends
- Multimedia data mining needs content-based retrieval and similarity search integrated with mining methods

References on Spatial Data Mining

- H. Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2001.
- Ester M., Frommelt A., Kriegel H.-P., Sander J.: Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support, Data Mining and Knowledge Discovery, *4: 193-216, 2000.*
- J. Han, M. Kamber, and A. K. H. Tung, "Spatial Clustering Methods in Data Mining: A Survey", in H. Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2000.
- Y. Bedard, T. Merrett, and J. Han, "Fundamentals of Geospatial Data Warehousing for Geographic Knowledge Discovery", in H. Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2000
- K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. SSD'95.
- Shashi Shekhar and Sanjay Chawla, Spatial Databases: A Tour , Prentice Hall, 2003 (ISBN 013-017480-7). Chapter 7.: Introduction to Spatial Data Mining
- X. Li, J. Han, and S. Kim, Motion-Alert: Automatic Anomaly Detection in Massive Moving Objects", IEEE Int. Conf. on Intelligence and Security Informatics (ISI'06).

References

- (Chang, 2008) Chang, K. T. 2008. Introduction to Geographic Information Systems . 4th edn. McGraw-Hill.
- (Ester et al., 1997) Ester, M., Kriegel, H.-P. and Sander, J. 1997. Spatial Data Mining: A Database Approach. Symposium on Large Spatial Databases 1262: 47-66.
- (Clementini et al., 1994) Clementini, E., Sharma, J. and Egenhofer, M. J. 1994. Modelling topological spatial relations: Strategies for query processing. Computers and Graphics 18 (6): 815-822.
- (Chung, 2012) Chung, C., Chiang, Y. and Chang, F. 2012. A spatial neural fuzzy network for estimating pan evaporation at ungauged sites. Hydrology and Earth System Sciences 16 (1): 255-266.
- (Kuo et al, 2010) Kuo, B. C., Huang, C. S., Hung, C. C., Liu, Y. L. and Chen, I. L. 2010. Spatial information based support vector machine for hyperspectral image classification. In Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS), 832 -835.

References

- (Koperski et al., 1998) Koperski, K., Han, J. and Stefanovic, N. 1998. An Eficient Two-Step Method for Classification of Spatial Data. In Proceedings of International Symposium on Spatial Data Handling.
- (Chelghoum et al., 2002) Chelghoum, N., Zeitouni, K. and Boulmakoul, A. 2002. A decision tree for multi-layered spatial data. In Proceedings of the Joint International Symposium on Geospatial Theory, Processing and Applications , 8-12.
- (Rinzivillo and Franco, 2004) Rinzivillo, S. and Franco, T. 2004. Classification in Geographical Information Systems , 374-385. Berlin Heidelberg: Springer-Verlag.
- (Han and Kamber, 2006) Han, J. and Kamber, M. 2006. Data Mining: Concepts and Techniques . 2nd edn. Morgan Kaufmann.
- (Koperski and Han, 1995) Koperski, K. and Han, J. 1995. Discovery of Spatial Association Rules in Geographic Information Databases. In Proceedings of the 4th International Symposium on Advances in Spatial Databases , 47{66. London, UK, UK:Springer-Verlag.

References

- (Appice et al., 2003) Appice, A., Ceci, M., Lanza, A., Lisi, F. A. and Malerba, D. 2003. Discovery of spatial association rules in geo-referenced census data: A relational mining approach. Intell. Data Anal. 7 (6): 541-566.
- (Wang et al., 2005) Wang, L., Xie, K., Chen, T. and Ma, X. 2005. Eficient discovery of multilevel spatial association rules using partitions. Journal of Information and Software Technology 47 (13): 829-840
- (Mennis and Liu, 2005) Mennis, J. and Liu, J. W. 2005. Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change. Transactions in GIS 9 (1): 5{17.
- (Han et al. 2001) Han, J., Kamber, M. and Tung, A. K. H. 2001. Spatial Clustering Methods in Data Mining: A Survey . Taylor and Francis.
- (Lu-He et al., 2005) Lu-He, W., Yi-Jun, L., Wan-Yu, L. and Zhang, D.-Y. 2005. Application and Study of Spatial Cluster and Customer Partitioning. In Proceedings of the Fourth International Conference on Machine Learning and Cybernetics .

References

- (Sander et al., 1998) Sander, J., Ester, M., Kriegel, H. P. and Xu, X. 1998. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications , Data Mining and Knowledge Discovery , vol. 2, 169-194.
- (Zeitouni and Chelghoum, 2001) Zeitouni, K. and Chelghoum, N. 2001. Spatial Decision Tree-Application to Traffic Risk Analysis. In Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications , 203{207. Washington, DC, USA: IEEE Computer Society.
- (Li and Claramunt, 2006) Li, X. and Claramunt, C. 2006. A Spatial Entropy-Based Decision Tree for Classification of Geographical Information. Transactions in GIS 10 (3): 451-467.
- (Sitanggang et al., 2011) Sitanggang, I. S**.**, R. Yaakob, N. Mustapha, A. A. Nuruddin, *an Extended ID3 Decision Tree Algorithm for Spatial Data*, A paper published in the Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services, June 29-July 1, 2011, Fuzhou, China.
- (I.S. Sitanggang, 2013) I.S. Sitanggang, R. Yaakob, N. Mustapha and A.N. Ainuddin, 2013. *Predictive Models for Hotspots Occurrence using Decision Tree Algorithms and Logistic Regression*. Journal of Applied Sciences, 13(2).