Spatial Data Mining

Vania Bogorny
Universidade Federal do Rio Grande do Sul
www.inf.ufrgs.br/~vbogorny
vbogorny@inf.ufrgs.br

Shashi Shekhar
University of Minnesota
www.cs.umn.edu/~shekhar
shekhar@cs.umn.edu

What is a Spatial Pattern ?

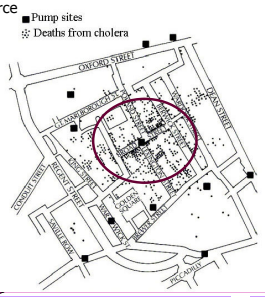
•What is not a pattern?
- Random
- Without definite direction, trend, rule, method
- Accidental - outside regular course of things
- Casual - relatively unimportant

•What is a Pattern?
- A frequent arrangement or regularity
- A rule or law
- A major direction, trend, prediction

Examples of Spatial Patterns

- Historic Example
  - 1855 Asiatic Cholera in London :

A water pump identified as the source

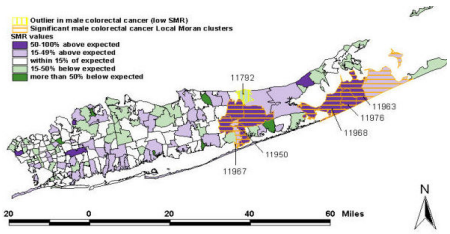


What is Spatial Data Mining?

- Search for **Interesting, useful** and **unexpected** spatial patterns
- **Non-trivial search**
  - Ex. Asiatic cholera : causes - water, food, air, insects, ...; water delivery mechanisms - numerous pumps, rivers, 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 the world
  - Ex. Connection between Water pump - Cholera

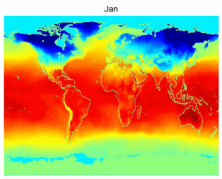
Example of Application Domains

- Questions from Epidemiology (Shekhar 2003)
  - What is the 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

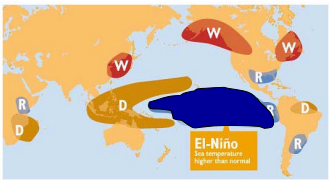


Modern Examples *(Shekhar 2003)*

Unusual warming of Pacific ocean (El Nino) affects weather



Average Monthly Temperature
(Courtesy: NASA, Prof. V. Kumar)



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

What is NOT Spatial Data Mining?

- Simple Querying of Spatial Data
  - Find neighbors of Florianopolis given names and boundaries of all cities
  - Find shortest path from SC to SP
- Uninteresting or obvious patterns in spatial data
  - Heavy rainfall in Florianopolis downtown is correlated with heavy rainfall in downtown São José, given that both cities are less than 20 Kilometers apart
  - Common knowledge: Nearby places have similar rainfall
- Mining of non-spatial data
  - Diaper and beer sales are correlated in evenings

Motivation for Spatial Data Mining

- Answer *Critical* questions:
  - Ex. How is the health of planet Earth?
  - Ex. Characterize or predict effects of human activity on the environment
  - Ex. Predict effect of El Nino on weather and economy
  - ....
- Spatial data is growing too fast to analyze manually
  - Satellite imagery, GPS tracks, sensors on highways, cell phones …

Families of Spatial Patterns

- Common families of spatial patterns
  - Co-location
  - Outliers
  - Classification / Location Prediction
  - Spatial Association Rules
  - Clustering
  - ..
- Other families of spatial patterns may be defined
  - SDM is a growing field, which should accommodate new pattern families

General Overview of Spatial Data Mining Literature

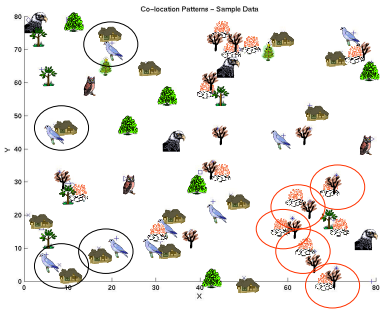
Transaction x Geometry DM

- Quantitative Spatial DM (Geometry-based)
  - Techniques*: Co-location, clustering*
  - Algorithms (SHEKHAR 2001, 2002) (HUANG 2004) (YOO 2005) (ZHANG2004)
    - Distance spatial relationships
    - Most use point spatial representation
    - Not implemented in toolkits
    - Single-granularity
- Qualitative Spatial DM (Transaction-based)
  - Techniques*: Spatial Association Rules, Classification, Clustering, Outlier detection*
  - Algorithms (APPICE 2003) (SHEKHAR, 2001a) (HAN, 2001) (BOGORNY 2006, 2008)
  - DMQL (LU, 1993) (KOPERSKI, 1995) (BIGOLIN 2003) (MALERBA, 2002) (BOGORNY 2008)
  - New operations to compute spatial relationships (ESTER 1997, 2000)
  - Semantic-based data mining (Bogorny 2006, 2007, 2008)
    - Any spatial relationship
    - Any spatial representation
    - Some tools
    - Multiple-Granularity

Co-Location

Co-location *(Shekhar 2003)*

Answers:

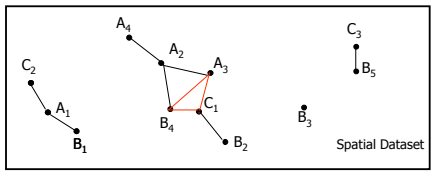


find patterns from the following sample dataset

Co-Location Patterns (Huang 2004, Yoo 2005)

- Input:
  - Spatial dataset
  - Distance threshold
  - Minimum participation index
- Method
  - Find neighbors
  - Find co-location candidates
  - Find frequent co-location sets
  - Extract co-location rules

Co-location Mining

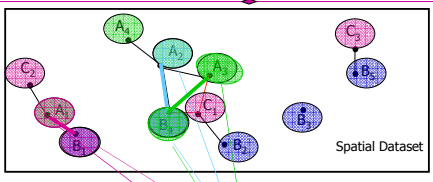


A, B, C: Spatial Feature Types

A1, A2... Spatial Feature Instances

Edges: neighbor

Co-location Mining



Set of Spatial Feature Types {A, B, C}

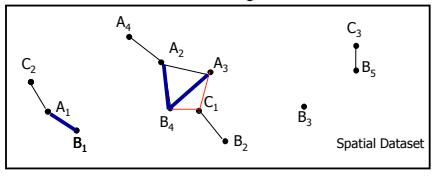
Candidates of size k=1

Candidates of size k=2

A B | A C | B C → Co-location

instances

Co-location Mining



Candidates of size k=1

Candidates of size k=2

A B | A C | B C → Co-location

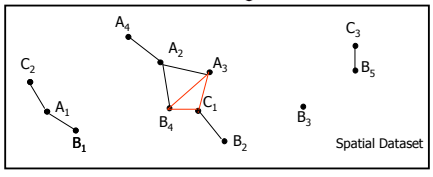
instances

2/4 2/3

3/4 2/5

3/5 2/3 → Participation ratio

Co-location Mining



Candidates of size k=2

A B | A C | B C → Co-location

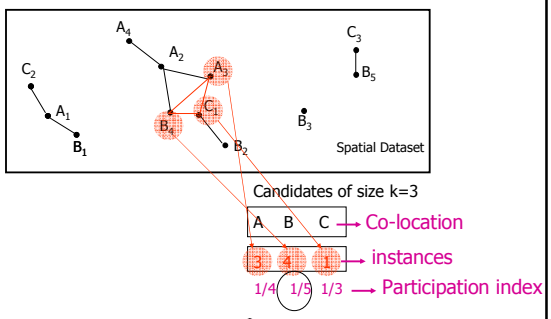
instances

2/4

2/5 3/5

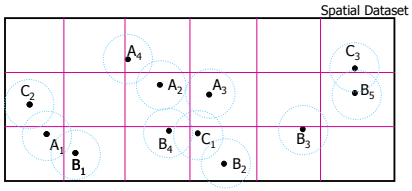
Participation Index (Lowest index) (If participIndex>minPartIndex) → frequent set

Co-location Mining



Co-Location Mining (Zhang 2004)

- Divide the space in cells (with size at least 2d)
- Buffer on each object,
  - object belongs to all cells that the buffer intersects (most 4 cells)
- All objects in a cell should fit in memory (are stored in a bucket)
- For each cell, objects are co-located if they are close



Co-location Example (Shekhar 2003)

Cropland with Roads
Roads with Bridges

Cropland
Roads
Bridges

Outliers?

Outliers

Outliers

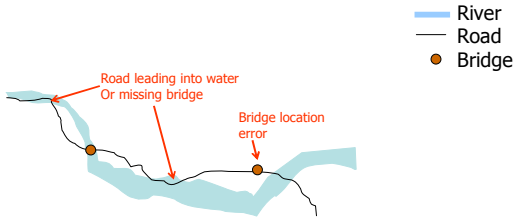
- What is an outlier?
  - Observations inconsistent with the rest of the dataset
- What is a spatial outlier?
  - Observations inconsistent with their neighborhoods
  - A local instability or discontinuity

Outliers (Shekhar 2001, 2003)

- *Global* outliers are observations of data inconsistent with the rest of the data in the database
  - has a number of practical applications in areas such as *credit card fraud, athlete performance analysis, voting irregularity, and severe weather prediction*
- A *spatial* outlier is a spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood.
  - For example, a new house in an old neighborhood is a spatial outlier based on the non-spatial attribute house age
- Tests to detect spatial outliers separate the spatial attributes from the non-spatial attributes.
  - Spatial attributes are used to characterize location, neighborhood, and distance.
  - Non-spatial attributes are used to compare a spatial referenced object to its neighbors.

Outliers – Examples *(Shekhar 2003)*

- Map Production
  - Error identification
  - E.g., spatial object violation



Spatial Association Rules

Association Rules (Agrawal 1993)

- Association rule is an implication of form

X → Y (support) (confidence)

Support : #(X∪Y) / #, where # is the number of rows in the dataset
Confidence : suport(X∪Y) / suport(X)

1 Generate frequent itemsets

Tid	Itemset
1	A, C, D, T, W
2	C, D, W
3	A, D, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

Set *k*	Frequent itemsets with minsup 50%
k=1	{A}, {C}, {D}, {T}, {W}
k=2	{A,C}, {A,D}, {A,T}, {A,W}, {C,D}, {C,T}, {C,W}, {D,T}, {D,W}, {T,W}
k=3	{A,C,D}, {A,C,W}, {A,D,T}, {A,D,W}, {A,T,W}, {C,D,T}, {C,D,W}, {D,T,W}
k=4	{A,C,D,W}, {A,D,T,W}

Support {AC} = 3/6 (50%)

2 Extract association rules

Confidence A→C = 3/4 (75%)

Association Rules

- Main Problem: Generate hundreds or thousands of rules
- **Frequent Itemsets:** generate all possible frequent itemsets
  - Apriori-like (generate candidates) (Agrawal, 1994)
  - Pattern-growth (without candidate generation) (Han, 2000)
- **Closed frequent itemsets:** generate non-redundant frequent itemsets
  - Apriori-like (generate candidates) (Pasquier, 1999) (Zaki, 2000)
  - Pattern-growth (without candidate generation) (Han, 2001) (Zaki 2002).....

Redundant Rules

A Redundant rule has same support and confidence of another rule generated from the same set of transactions

Frequent Itemsets

Tid	Itemset
1	A, C, D, T, W
2	C, D, W
3	A, D, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

TidSet	Frequent itemsets with *minsup* 50%
123456	{D}
12456	{C} {C,D}
12345	{W} {D,W}
1245	{C,W} {C,D,W}
1345	{A} {A,D}, {A,W} {A,D,W}
1356	{T}, {D,T}
145	{A,C}, {A,C,W}, {A,C,D}, {A,C,D,W}
135	{A,T}, {T,W}, {A,D,T}, {A,T,W}, {D,T,W} {A,D,T,W}
156	{C,T},{C,D,T}

A→W (suport= 4/6) (confidence = 4/4)

A→ DW (suport = 4/6) (confidence = 4/4)

25 frequent itemsets / 9 closed frequent itemsets

Spatial *Association Rules*

- Spatial association rule is an implication of the form

X → Y (support)(confidence)

- at least one element in *X* or *Y* is a spatial predicate
  - is_a(island) → **within**(river)
  - ***closeTo(slum)*** → *criminalityRate=High*

Different Spatial Objects are Stored in Different Relations

Street

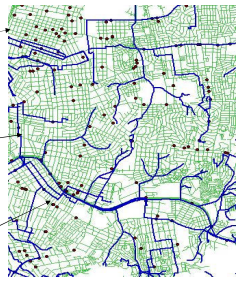
Gid	Name	Shape
1	Ijui	Multiline [(x1,y1),(x2,y2),..]
2	Lavras	Multiline [(x1,y1),(x2,y2),..]

WaterResource

Gid	Name	Shape
1	Jacui	Multiline [(x1,y1),(x2,y2),..]
2	Guaiba	Multiline [(x1,y1),(x2,y2),..]
3	Uruguai	Multiline [(x1,y1),(x2,y2),..]

GasStation

Gid	Name	VolDiesel	VolGas	Shape
1	BR	20000	85000	Point[(x1,y1)]
2	IPF	30000	95000	Point[(x1,y1)]
3	Esso	25000	120000	Point[(x1,y1)]



Most Spatial Association Rule Mining algorithms have a single table/file INPUT format

Different Relations (tables) need to be Spatially Joined

Preprocessed Geographic Data for Transaction-Based Data Mining

Target feature

Tuple (city)	Spatial Predicates							
1	contains(Port)	contains(Hospital)	contains(Street),	contains(Factory)	crosses(Water Body)
2		contains(Hospital)	contains(Street),		crosses(Water Body)
3	contains(Port)		contains(Street),	contains(Factory),	crosses(Water Body)
4	contains(Port)	contains(Hospital)	contains(Street),		crosses(Water Body)
5	contains(Port)	contains(Hospital)	contains(Street),	contains(Factory)	crosses(Water Body)
6		contains(Hospital)	contains(Street),	contains(Factory)	

Relevant features

Spatial Association Rules

- Are computed in 3 main steps:
  - Data preprocessing: compute spatial relationships (spatial joins). Most expensive step
  - Compute frequent itemsets
  - Generate association rules

Transaction Dataset X Preprocessed Spatial Dataset

Transactional Dataset

Transaction	Items
1	milk, bread, butter, cereal
2	milk, bread
3	beer, bread, chocolate
4	cereal, meet, milk
5	milk, beer, nuts, orange, cereal

- rows are transactions
- attributes are items, supposed to be independent

Spatial Dataset

Tuple (city)	Spatial Predicates
1	contains(Port), contains(Hospital), contains(TreatedWaterNet), contains(Factory), crosses(WaterBody)
2	contains(Hospital), contains(TreatedWaterNet), crosses(WaterBody)
3	contains(Port), contains(TreatedWaterNet), contains(Factory), crosses(WaterBody)
4	contains(Hospital), contains(TreatedWaterNet), crosses(
5	contains(Port), contains(Hospital), contains(TreatedWaterNet), contains(Factory), crosses(
6	contains(Hospital), contains(TreatedWaterNet), contains(Factory)

- rows are instances of the target feature type
- attributes are predicates
- spatial predicates are spatial relationships between the target feature type and relevant feature types

Spatial Predicate Computation (Preprocessing)

Given: D, //geographic database

e.g. [river, bridge, city, district, waterBody, island, road, cellularAntenna, gasStation, hospital, school, treatedWaterNetwork, port, industry]

$T = \{t_1, t_2, ..., t_n\}$, // target feature type

e.g. [city]

$S = \{O_1, O_2, ..., O_m\}$, // set of relevant feature types

e.g. [river, road, waterBody, hospital, school, gasStation, industry, port]

R //spatial relationships

e.g. [topological]

Find: a spatial dataset Ψ for mining SAR;

$\boldsymbol{T = \{t_1, t_2, ..., t_n\}}$

$\boldsymbol{S = \{O_1, O_2, ..., O_m\}}$, **where** $O_i = \{o_1, o_2, ..., o_k\}$,

Spatial Join - bottleneck

Some Spatial Association Rule Mining Algorithms

- Koperski 1995
- Spada (Appice 2003)
- Clementini (2003)
- Apriori-KC (Bogorny 2006)
- Max-FGP (Bogorny 2006a)
- ...

- Preprocess geographic data and apply classical DM algorithms

(Koperski 1995)

- In a first step spatial approximations are calculated (distance),
- In a second step, more precise spatial relationships are computed to the result of the first step (touches, contains, crosses, etc)
- Minimum support is used to extract only *frequent spatial relationships*.
- Multiple-granularity approach

Spada (Appice 2003)

- Inductive Logic Programming (ILP) approach
  - compute all spatial relationships in preprocessing steps
    - transform the result into a deductive relational database (set of predicates)
  - Compute frequent itemsets (as in Apriori)
  - Generate association rules
  - Filter association rules with declarative bias *a posteriori*
    - Pattern_constraint (*AtomList, Min_occur*),
    - Example: pattern_constraint (crossesRiver, 5)
- A large amount of background knowledge is required from the data mining user, which has to define all possible frequent patterns to be eliminated.

Semantic-based Spatial Association Rule Mining

Semantic-based SAR Mining - Motivation

- Existing approaches for spatial data mining, in general, do not make use of background knowledge
  - Use *syntactic constraints* for frequent set and rule prunning
  - Only the data is considered, not the schema
- Result
  - Same associations explicitly represented in the schema (database designer) are extracted by SAR mining algoritms
- Bogorny (2006) and Bogorny (2007, 2008) introduced the idea of using ***background knowledge***
  - in data preprocessing, to reduce spatial joins
  - in spatial association rule mining, to eliminate well known patterns

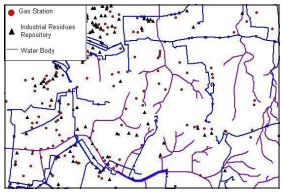
Spatial Relationships

- Mandatory (Spatial constraints) Dependences:
  <island> <inside> <1><1> <Water Body>
- Prohibited:
  <River> <*contains*> <0><0> <Road>
- **Possible**: Normally undefined
  Road *crosses* River

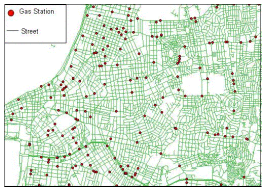
For data mining and knowledge discovery,
only POSSIBLE/PROHIBITED RELATIONSHIPS are interesting!!!!
Mandatory relationships are well known.

Well Known Geographic Dependences

Non-obvious spatial relationships

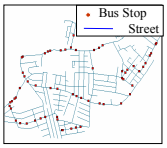


Well known dependences



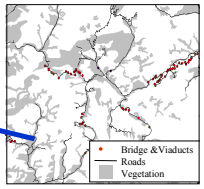
Is_a(gasStation) → intersects(street) (100%)

Well Known relationships X Association Rules

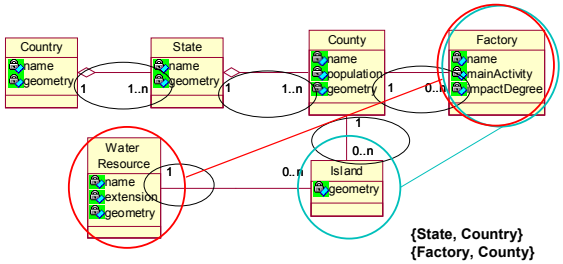


intersects(busStop) → intersects(Street) (100%)

Contains(viaduct) → contains(road) (100%)

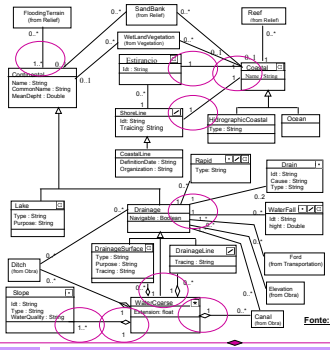


Well Known Associations – Conceptual Schemas



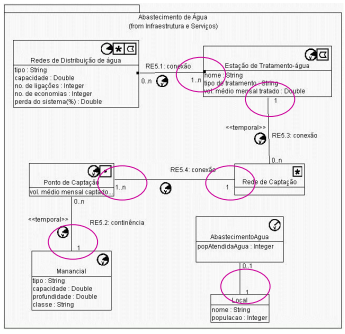
{State, Country}
{Factory, County}
{Island, WaterBody}
....

Well Known Associations – Conceptual Schemas

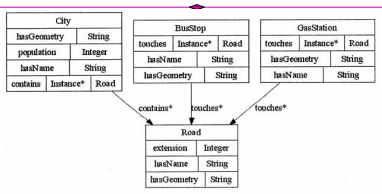


Fonte: 1ª Divisão de Levantamento do Exército Brasileiro

Well Known Associations – Conceptual Schemas



Well Known Associations – Geo-Ontologies




```
<owl:Class rdf:ID="GasStation">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:minCardinality rdf:datatype="&int;" >1</owl:minCardinality>
      <owl:onProperty>
        <owl:ObjectProperty rdf:about="#touches"/>
      </owl:onProperty>
      <owl:valuesFrom rdf:resource="#Road"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
</owl:Class>
```


Well known dependences X Spatial Association Rules (SAR)

- Well knonw dependences affect the 3 main steps in the process of mining SAR:
  - Spatial predicate computation: compute unnecessary relationships
  - Frequent set generation: generate frequent itemsets with well known patterns
  - Association rule extraction: produce a high number of rules with well known dependences

How do Well Known Dependences appear in SAR Mining?

Problem I - Geographic Dependences between the Target Feature and Relevant Features

Dependence = City and Street → contains(Hospital) → contains(Street)

Tuple (city)	Spatial Predicates		
1	contains(Port), contains(Hospital),	contains(Street),	contains(Factory), crosses(Water Body)
2	contains(Hospital),	contains(Street),	crosses(Water Body)
3	contains(Port),	contains(Street),	contains(Factory), crosses(Water Body)
4	contains(Port), contains(Hospital),	contains(Street),	crosses(Water Body)
5	contains(Port), contains(Hospital),	contains(Street),	contains(Factory), crosses(Water Body)
6	contains(Hospital),	contains(Street),	contains(Factory)

Minconf=70%

100% de support

Min Sup %	Total FrequentSets/ Rules	Rules with Dependence / Rules without Dependence	FrequentSets with dependence / FrequentSets without dependence
20	31 / 180	130 / 50	16 / 15
50	25 / 96	72 / 24	13 / 12

Problem II - Dependences among Relevant Feature Types

Dependence = {Port, WaterBody}

Tuple (city)	Spatial Predicates		
1	contains(Port),	contains(Hospital), contains(Street), contains(Factory)	crosses(Water Body)
2	contains(Hospital),	contains(Street),	crosses(Water Body)
3	contains(Port),	contains(Street), contains(Factory)	crosses(Water Body)
4	contains(Port),	contains(Hospital), contains(Street),	crosses(Water Body)
5	contains(Port),	contains(Hospital), contains(Street), contains(Factory)	crosses(Water Body)
6	contains(Hospital), contains(Street), contains(Factory)		

Minsup=50%

25 frequent sets (6 contain the dependence)

9 closed frequent sets (3 have the dependence)

contains(Port) → crosses(WaterBody)

Problem III - Redundant Frequent Itemsets

- For MinSup 50% we have:
  - 25 frequent sets
  - 9 closed frequent sets (3 contain the dependence)
  - 16 redundant frequent sets

Dependence {A,W}

Dataset

Tid (city)	Predicate Set
1	A, C, D,T, W
2	C, D, W
3	A, D, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

c) predicates

A = contains(Port)
C = contains(Hospital)
W = crosses(WaterBody)
D = contains(Street)
T = contains(Factory)

TidSet	Frequent sets L (MinSup=50%)
123456	{D}
12456	{C}, {C,D}
12345	{W}, {D,W}
1245	{C,W}, {C,D,W}
1345	{A}, {A,D}, {A,W}, {A,D,W}
1356	{T}, {D,T}
145	{A,C}, {A,C,W}, {A,C,D}, {A,C,D,W}
135	{A,T}, {T,W}, {A,D,T}, {A,T,W}, {D,T,W}, {A,D,T,W}
156	{C,T}, {C,D,T}

9 closed frequent itemsets

Pruning Methods using Background Knowledge

Frequent Set Pruning (Apriori-KC) (Bogorny, 2006a)

{Island,water}

Given: Φ, // set of knowledge constraints
Ψ, // *dataset generated with spatial_predicate_extraction*

minsup, // minimum support
L_1 = {large 1-predicate sets};
For (k = 2; L_{k-1} != ∅; k++) do begin
C_k = apriori_gen(L_{k-1}); // Generates new candidates
If (k=2)
// remove pairs with dependences
(step 1) Delete from C_2 all pairs with a dependence in Φ;

Forall rows w ∈ Ψ do begin
C_w = subset(C_k, w); // Candidates contained in w
Forall candidates c ∈ C_w do
c.count++;
End;
L_k = {c ∈ C_k | c.count ≥ minsup};
End;
Answer = $\cup_k L_k$

Max-FGP (Bogorny 2006c)

Given: *L; // frequent sets without dependences (Apriori-KC)*
Ψ; // dataset generated with spatial_predicate_extraction

Find: Maximal M

```
// find maximal generalized predicate sets
M = L;
For ( k = 2; Mk != ∅; k++ ) do  begin
    For ( j = k+1;  Mj!=0;  j++ ) do  begin
        If  (tidSet (Mk) = tidSet (Mj))
                If (Mk ⊂ Mj)   // Mj is more general than Mk
                        Delete Mk from M;
   End;
End;
Answer = M;
```


Summary

- Well known dependences exist in several non-spatial application domains
  - Biology/Bioinformatics
    - Pregnant → Female (confidence=100%)
    - Breast_cancer → Female (confidence 100%)
    - ...
- Almost no data mining approaches consider background knowledge or domain knowledge

Spatial Classification

Classification

- Given a set of instances, the role of classification is to discover the classes of the instances
- Spatial objects may be characterized (classified) by different types of information (Koperski 1998):
  - non-spatial attributes (e.g. population);
  - spatially related attributes with non-spatial values (e.g. *total population living within 100 meters from cellular antennas);*
  - spatial predicates (e.g. *closeTo_beach);*

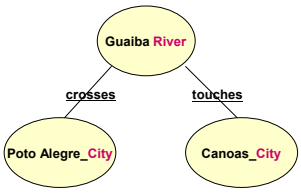
Ester (1997, 2001)

- Proposed a graph-based approach for spatial neighbourhood computation
- Idea is to integrate data mining into database systems, with new database primitives for the computation of spatial relationships
- and explicitly represent spatial relationships that are normally implicit in spatial databases

Ester (1997, 2001)

- A neighborhood graph for a relation "neighbor" in a geographic database is a graph G(N,E), where
  - *N* are nodes
  - *E* are edges
- Each node *N* is an object in the database connected via some *edge* to another node if the neighbor holds.
- Two objects are neighbors if any spatial relationship (topological, distance or order) holds

Ester (1997, 2001)

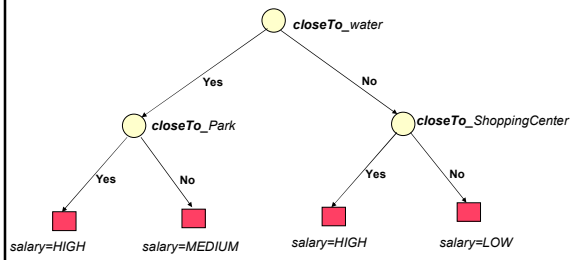


Ester (1997, 2001)

- Operations
  - ***Get_nGraph(db, rel)***: computes all relationships
  - ***Get_neighborhood(graph, o, pred)***: retrieves all objects *o* directly connected via some edge in the *graph* satisfying a condition in *pred*
  - ***Create_nPaths (objects, graph, pred, i):*** *creates a set of all paths from one object following the edges of the neighborhood graph with length<i*
    - the influence of neighboring objects and their attributes decreases with increasing distance
    - the length of the relevant neighborhood paths are limited by an input parameter *max-length*.

Ester (1997, 2001)

Class is a non-spatial attribute = *salary*
Class values: high, medium, low



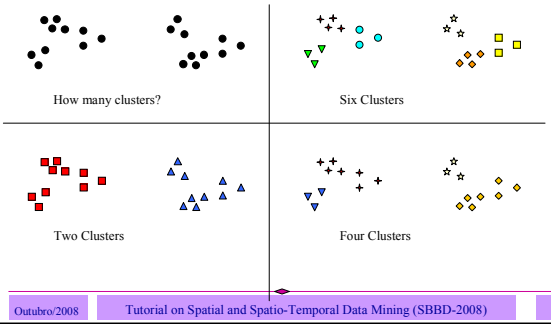
Clustering

Clustering (cluster analysis)

- Clustering is a process of partitioning a set of data into a set of groups called *clusters*
- A cluster is a set of data (objects) with
  - similar characteristics
  - that can be collectively treated as one group
- Clustering is an unsupervised method
  - no predefined classes

Clustering Analysis (Kumar 2005)

Different ways of clustering the same set of points



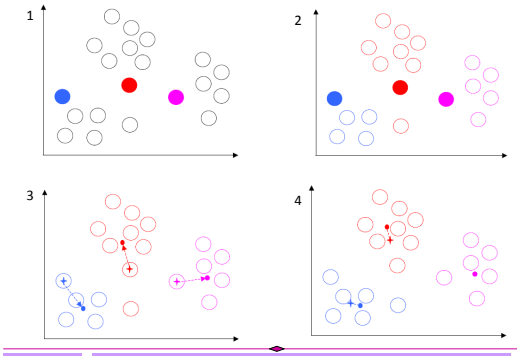
Main Clustering Approaches

- Partitioning
  - A division of data objects into non-overlapping subsets (clusters) such that *each object* is in exactly one subset
- Hierarchical
  - A set of nested clusters organized as a hierarchical tree
- Density-based
  - Find clusters based on density of regions
- Grid-based
  - Find clusters based on the number of points in each cell

K-means

- Partitional clustering approach
- Each cluster is associated with a centroid
- Each point is assigned to the cluster with the closest centroid
- A drawback of the k-means is that the number of clusters *k* is an input parameter.

K-means



K-medoid methods

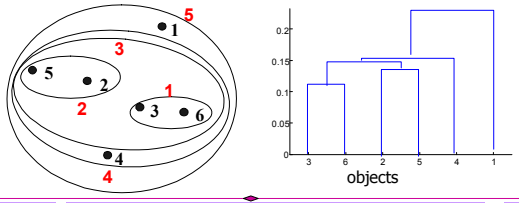
Instead of ***means***, use representative objects called ***medoids***

- *PAM* (Partitioning Around Medoids, 1987) - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
- *CLARA* (Clustering Large Applications, 1990) – It draws *multiple samples* of the data set, applies *PAM* on each sample, and gives the best clustering as the output
- *CLARANS* (NG 2002) - is an improved *k*-medoid method

Hierarchical Clustering

Two main types: Agglomerative and Divisive

- Agglomerative
  - Start with all objects as individual clusters
  - At each step, merge the two most similar clusters
  - Until rests one cluster (or k clusters)



Hierarchical Clustering

- Divisive
  - Start with one cluster (with all objects)
  - At each step, split a cluster in two
  - Until each cluster contains only one object (or k clusters)

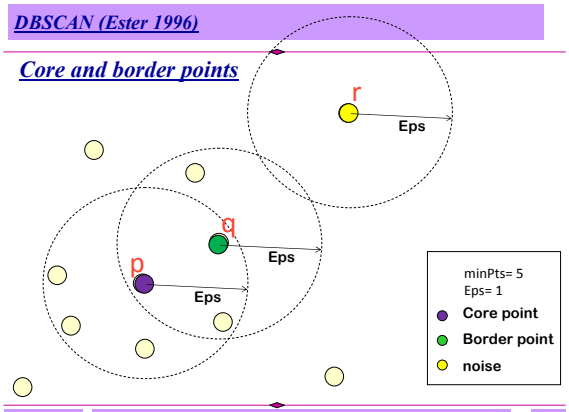
Similarity can be euclidean distance or any other measure

DBSCAN (Ester 1996)

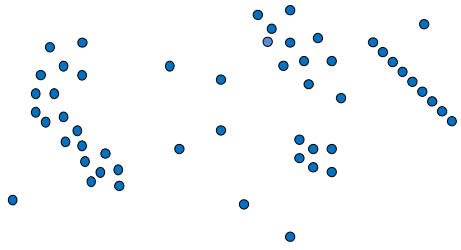
- DBSCAN is a density-based algorithm
  - Density = number of points within a specified radius (*Eps*)
  - A point is a core point if it has more than a specified number of points (*MinPts*) within *Eps*
  - A border point has less than *MinPts* within *Eps*, but it is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point.

DBSCAN (Ester 1996)

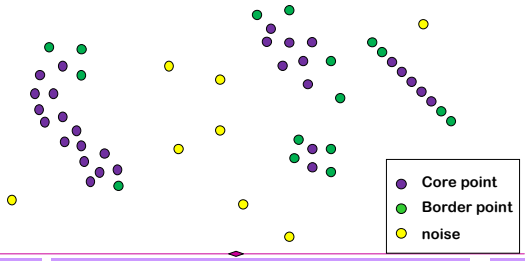
Core and border points



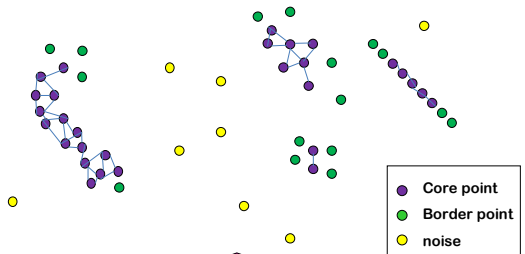
DBSCAN example



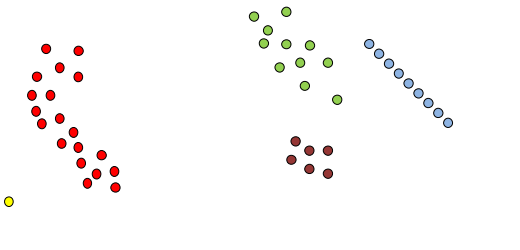
Identifying core, border and noise points



Computing distance



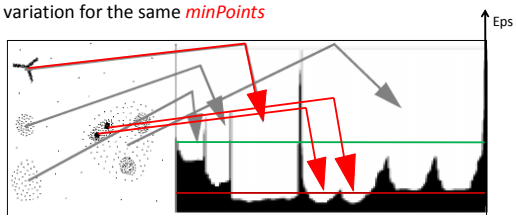
Final Clusters



OPTICS - Ordering Points to Identify the Clustering Structure (Ankerst 1999)

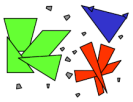
Produces a special order of the database with respect to its density-based clustering structure

EPS variation for the same *minPoints*



GDBSCAN (Generalized DBSCAN) (Sander 1998)

- Generalized version of DBSCAN
- Clusters are formed based on spatial or non-spatial attributes
- Any spatial relationshiop is used to compute neighbors, and spatial objects may have any representation
  - *NPred* : "neighbor",
  - *wCard*: cardinality>=MinCard, (generalizes the condition *NEps(o)>= MinPts)*
  - *MinWeight(N)*: *aggr* (non-spatial values) >= *threshold OR MinPts*
  - *ExampleI:* NPred: distance <= *Eps*, *wCard*: sum of areas, *MinWeight* : >= *MinPts*
  - *ExampleII:* *NPred*: "*intersects*" or "*touches*", *MinWeight(N)*: sum of areas >= *MinArea,*



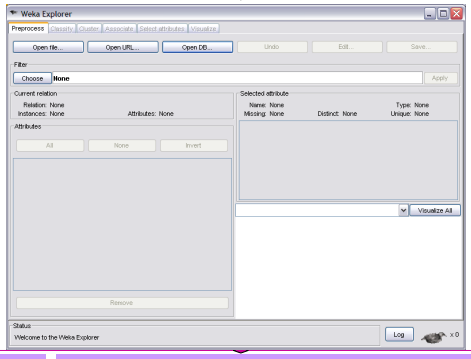
GDBSCAN (Generalized DBSCAN) (Sander 1998)

- Example I: Two cities can be close if they have similar non-spatial attributes
- Using non-spatial attributes as a weight for objects one can "induce" different densities, even if the objects are equally distributed in the space of the spatial attributes.

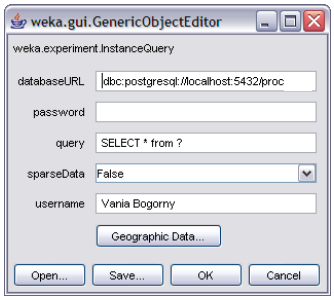
Tools

- GeoMiner (Han 1997)
- INGENS (Malerba 2001)
- Ares (Appice 2005)
- Weka-GDPM (Bogorny 2006d)

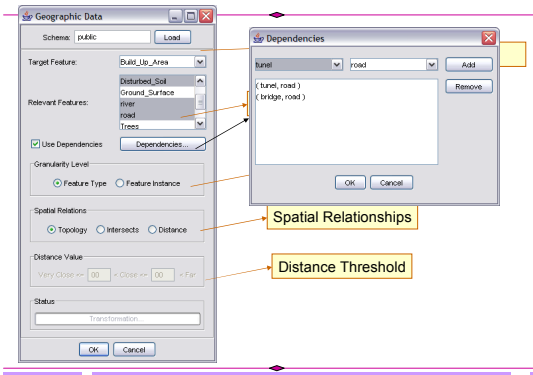
Weka GUI Explorer (Frank 2005)



Weka-GDPM



Weka-GDPM



References – association rules

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, SIGMOD, 1993, Washington, D.C. Proceedings... New York: ACM Press, 1993. p. 207-216.

HAN, J.; PEI, J.; YIN, Y. Mining Frequent Patterns without Candidate Generation. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, SIGMOD, 2000, Dallas. Proceedings...[S.l.]: ACM Press, 2000. p. 1-12.

BOGORNY, V.; CAMARGO, S.; ENGEL, P. M.; ALVARES, L.O. Towards elimination of well known geographic domain patterns in spatial association rule mining. In: IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS, IEEE-IS, 3., 2006, London. Proceedings... IEEE Computer Society, 2006a. p. 532-537.

BOGORNY, V.; CAMARGO, S.; ENGEL, P.; ALVARES, L. O. Mining Frequent Geographic Patterns with Knowledge Constraints. In: ACM INTERNATIONAL SYMPOSIUM ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, ACM-GIS, 14., 2006, Arlington. Proceedings... 2006b.

BOGORNY, V.; VALIATI, J.; CAMARGO, S.; ENGEL, P.; KUIJPERS, B.; ALVARES, L.O.: *Mining Maximal Generalized Frequent Geographic Patterns with Knowledge Constraints*. In: Proc. of the 6rd IEEE International Conference On Data Mining - (IEEE-ICDM'06),Hong Kong, pp.813-817 (2006c).

BOGORNY, V.; ENGEL, P.; ALVARES, L.O.: *Enhancing the Process of Knowledge Discovery in Geographic Databases using Geo-Ontologies (Chapter IX)*. In: NIGRO, H. O., CISARO, S.G., XODO, D. (Ed.) Data Mining with Ontologies: Implementations, Findings, and Frameworks. Idea Group Inc. (2007). pp. 160-181.

Bogorny, V., Kuijpers, B. & Alvares, L. O. (2008). *Reducing uninteresting spatial association rules in geographic databases using background knowledge: a summary of results*. International Journal of Geographical Information Science. Taylor and Francis, VOL 22, pp. 361-386.

MALERBA, D.; LISI, F. A. Discovering Associations between Spatial Objects: An ILP Application. In: INTERNATIONAL WORKSHOP ON INDUCTIVE LOGIC PROGRAMMING, 2001. Proceedings... Berlin: Springer, 2001. p.156-163.

References - Tools

HAN, J.; KOPERSKI, K.; STEFANVIC, N. GeoMiner: a system prototype for spatial data mining. In: ACM-SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, SIGMOD, 1997, Tucson. Proceedings... [S.l.]: ACM

KOPERSKI, K.; HAN, J. Discovery of Spatial Association Rules In Geographic Information Databases. In: INTERNATIONAL SYMPOSIUM ON LARGE GEOGRAPHICAL DATABASES, SSD, 4., 1995, Portland. Proceedings... [S.l.]: Springer, 1995. p.47-66.

APPICE, A. et al. Mining and Filtering Multi-level Spatial Association Rules with ARES. In: INTERNATIONAL SYMPOSIUM ON METHODOLOGIES FOR INTELLIGENT SYSTEMS, ISMIS, 15., 2005, New York. Proceedings... [S.l.]: Springer, 2005. p.342-353 (Lecture Notes in Computer Science, 3488).

BOGORNY, V.; PALMA, A.; ENGEL, P.; ALVARES, L.O.: *Weka-GDPM: Integrating Classical Data Mining Toolkit to Geographic Information Systems*. In: SBBD Workshop on Data Mining Algorithms and Aplications(WAAMD'06), Florianopolis, Brazil, October 16-20, (2006d). pp.9-16.

References – closed frequent itemsets

PASQUIER, N. et al. Discovering frequent closed itemsets for association rules. In: INTERNATIONAL CONFERENCE ON DATABASE THEORY, ICDT, 7., 1999, Jerusalem. Proceedings... [S.l.]: Springer, 1999a. p.398-416.

PASQUIER, N. et al. Efficient Mining of Association Rules using Closed Itemset Lattices. Information Systems, [S.l.], v.24, n.1, p.25-46, Mar. 1999b.

ZAKI. M. Generating Non-redundant Association Rules. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, KDD, 6., 2000, Boston. Proceedings... [S.l.]: ACM, 2000. p.34-43.

ZAKI., M.; HSIAO, C. CHARM: An Efficient Algorithm for Closed Itemset Mining. In: INTERNATIONAL CONFERENCE ON DATA MINING, SIAM, 2., 2002, Arlington. Proceedings... [S.l.]:SIAM, 2002.

References - clustering

[NG e HAN 94] NG, R. T.; HAN, J. Efficient and Effecctive Clustering Methods for Spatial Data Mining. In: Twentieth International Conference on Very Large Data Base, Santiago, 1994.

ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X., 1996, "A density-based algorithm for discovering clusters in large spatial databases.", *In Proc. 1996 Int. Conf. Knowledge Discovery and Data Mining (KDD'96)*, pp. 226–231

SHEIKHOLESLAMI, G., CHATTERJEE, S., ZHANG, A., 1998, "WaveCluster: A multi-resolution clustering approach for very large spatial databases.", *In: Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pp. 428–439, New York, NY, Aug.

SANDER J., ESTER M., KRIEGEL H.-P., Xu X.: "Densisty-Based Clustering in Spatial Darabases: The Algorithm GDBSCAN and its Applications. In Data Mining and Knowledge Discovery, Kluwer Academic Publishers, Vol. 2, 1998.

TUNG, A. K. H., HOU, J., HAN. J., 2001, "Spatial clustering in the presence of obstacles", *In: Proc. 17th International Conference on Data Engineering*, pp. 359- 367, Heidelberg, Germany.

NG, R., HAN, J., 2002, "CLARANS: A Method for Clustering Objects for Spatial Data Mining", *IEEE Trans. Knowledge & Data Engineering* , v.14, n. 5 (Set), pp. 1003-1016.

ANKERST, M., BREUNIG, M., KRIEGEL, H.-P., SANDER, J., 1999, "OPTICS: Ordering points to identify the clustering structure.", *In: Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99)*, pp. 49–60, Philadelphia, June.

References - classification

KOPERSKI, K., HAN, J., STEFANOVIC, N.: An Efficient Two-Step Method For Classification Of Spatial Data. In: International Symposium On Spatial Data Handling, Vancouver, BC, Canada (1998).

LEE, C-H.; GREINER, R.; SCHMIDT, M.: Support Vector Random Fields for Spatial Classification. In Proceedings of the (*PKDD'2005)* 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Springer-Verlag, Porto, Portugal, 2005, 121-132.

MALERBA, D.; APPICE, A.; VACCA N.: SDMOQL: an OQL-based data mining query language for map interpretation tasks. In: WORKSHOP ON DATABASE TECHNOLOGIES FOR DATA MINING, DTDM, Prague, 2002. Proceedings... [S.l.]: Springer. 2002.

ESTER, M. et al. Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support. Journal of Data Mining and Knowledge Discovery, [S.l.], v.4, n.2-3, p.193-216, July 2000.

References – Co-Location and Outliers

Shekhar, S.; Huang, Y.Discovering Spatial Co-location Patterns: A Summary of Results , Proc. of 7th International Symposium on Spatial and Temporal Databases(SSTD01), L.A., CA, July 2001

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.

HUANG, Y.; SHEKHAR, S.; XIONG, H. Discovering Co-location Patterns from Spatial Datasets: A General Approach. IEEE Transactions on Knowledge and Data Engineering, v.16, n.12, Dec. 2004.

SHEKHAR, S.; LU, C.-T.; ZHANG, P. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, KDD, 7., 2001, San Francisco. Proceedings… ACM, 2001. p.371-376.

SHEKHAR, S., CHAWLA, S. Spatial databases: a tour. Upper Saddle River, NJ: Prentice Hall, 2003.

YOO, J.S.; SHEKHAR, S; CELIK, M. A Join-less Approach for Co-location Pattern Mining: A Summary of Results. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING, ICDM, 5., 2005, Houston. Proceedings… IEEE Computer Society, 2005. p.813-816.

Zhang, N. Mamoulis, D. W. L. Cheung, and Y. Shou, "Fast Mining of Spatial Collocations," Proceedings of the *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 384-393, Seattle, WA, August 2004.

References - Data Mining

Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining. Addison-Wesley, 2005. ISBN : 0321321367.

References - Data Mining Query Languages

Han, J., Fu, Y., Wang, W., Koperski, K. and Zaiane, O., 1996, Dmql: A data mining query language for relational databases. In Proceedings of the SIGMOD'96 Workshop on Research Issues in Data Mining and Knowledge Discovery, Montreal, Canada, pp. 27-33.

Meo, R., Psaila, G. and Ceri, S., 1996, A New SQL-like Operator for Mining Association Rules. In Proceedings of the VLDB, T.M. Vijayaraman, A.P. Buchmann, C. Mohan and N.L. Sarda (Eds) (Morgan Kaufmann), pp. 122-133.

Wang, H. and Zaniolo, C., 2003, ATLaS: A Native Extension of SQL for Data Mining.. In Proceedings of the SDM, D. Barbara and C. Kamath (Eds) (SIAM).

Malerba, D., Appice, A. and Ceci, M., 2004, A Data Mining Query Language for Knowledge Discovery in a Geographical Information System.. In Proceedings of the Database Support for Data Mining Applications, pp. 95-116.

Boulicaut, J.F. and Masson, C., 2005, Data Mining Query Languages.. In The Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach (Eds) (Springer), pp. 715{727.

Chen, X. and Petrounias, I., 1998, Language Support for Temporal Data Mining. In Proceedings of the Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery (London, UK: Springer-Verlag), pp. 282-290.