

# A Preliminary Study to Spatial Data Mining

**Ho Tu Bao**

Japan Advanced  
Institute of Science  
and Technology

**Luong Chi Mai**

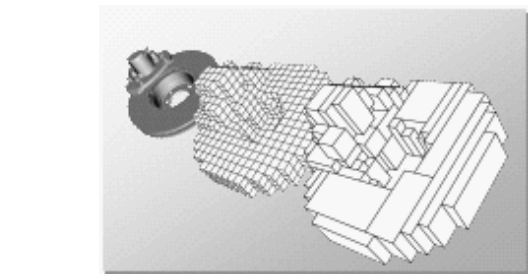
Institute of  
Information Technology  
Hanoi

# Outlines

---

- **What Is Spatial Data Mining?**
- Spatial Data Mining Tasks?
- Creating Primary Spatial Databases  
Case study: POPMAP
- Creating Secondary Spatial Features

\_\_\_\_\_



A photograph showing a rugged, rocky coastline. In the foreground, there are large, dark, jagged rock formations. The sea is visible in the middle ground, with several large, rounded rock formations protruding from the water. The sky is filled with soft, white clouds. The overall scene is a natural coastal landscape.

[illegible]

# Spatial Objects

- Spatial objects usually consist of both spatial and non-spatial data

## Spatial data

data related to spatial description of the objects such as coordinates, areas, latitudes, perimeters, spatial relations (distance, topology, direction), etc.

**Example:** earthquake points, town coordinates on map, etc.

## Non-spatial data

other data associated to spatial objects.

**Example:** earthquake degrees, population of a town, etc.

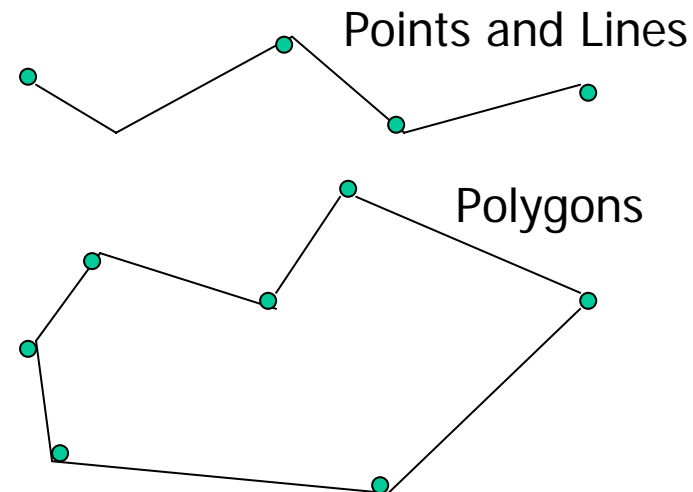
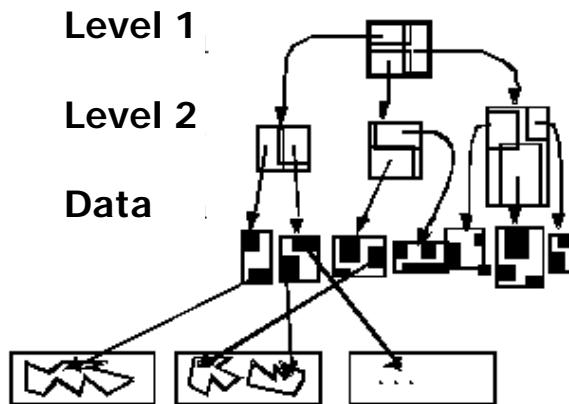
Spatial objects can be classified into different **object classes** (hospital class, airport class, etc.), several classes can be hierarchically organized in **layers** (countries, provinces, districts, etc.)

# Spatial Databases (GIS)

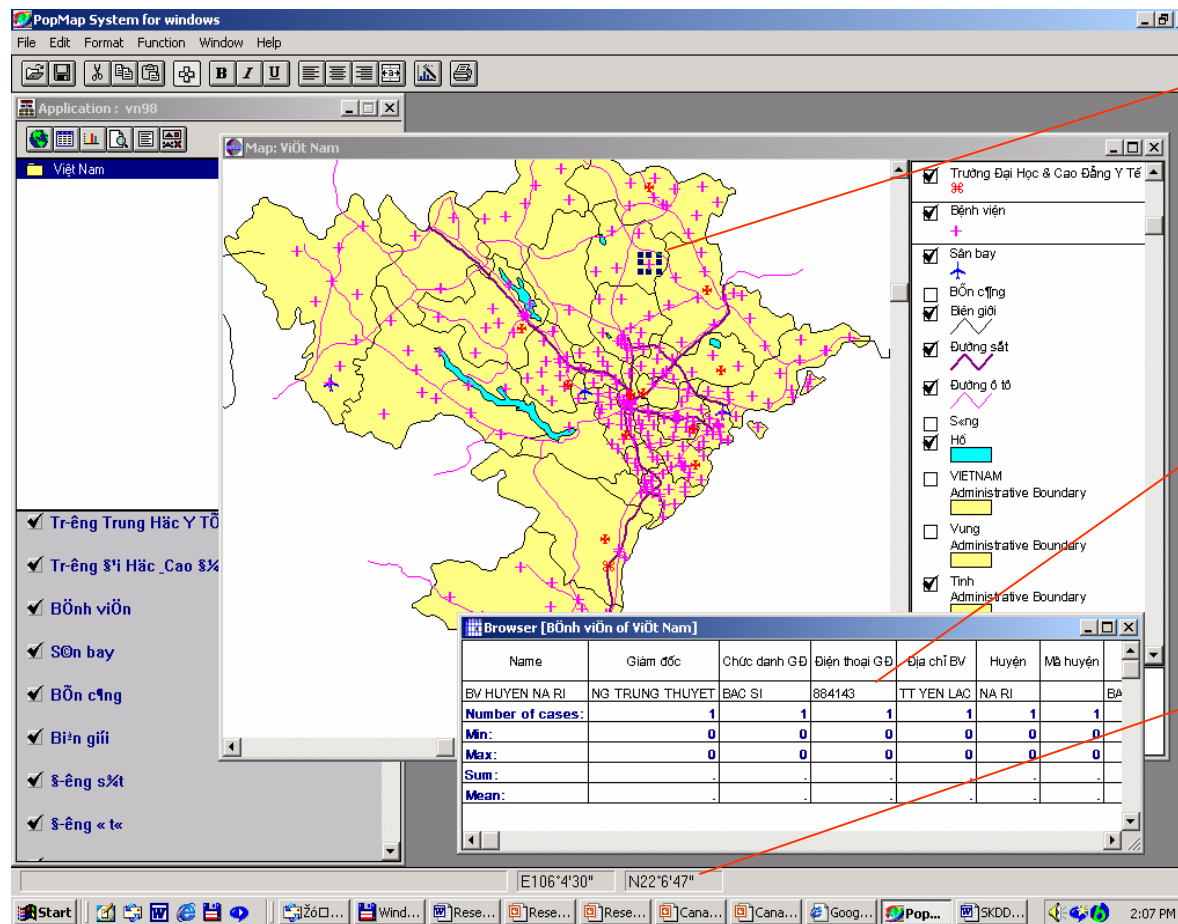
- Many **spatial database architectures** (models) proposed for managing both non-spatial and spatial components of spatial objects
- GIS: Geographical Information Systems
  - **Non-spatial** data can be stored in relational databases
  - **Spatial** data are typically described by
    - **Geometric properties**: **coordinates** (schools), **lines** (roads, rivers), **area** (countries, towns), etc.
    - **Relational properties**: **adjacency** (A is neighbor of B), **inclusion** (A is inside in B), etc.

# Spatial Data Structures (primary data)

- **Spatial data structure**: points, lines, polygons, etc. for representing geometric data
- Multidimensional trees used to **build indices** for spatial data structures: quad-tree, B-tree, R-tree, **R\*-tree**, etc.



# Example of Spatial Data in GIS



A spatial object

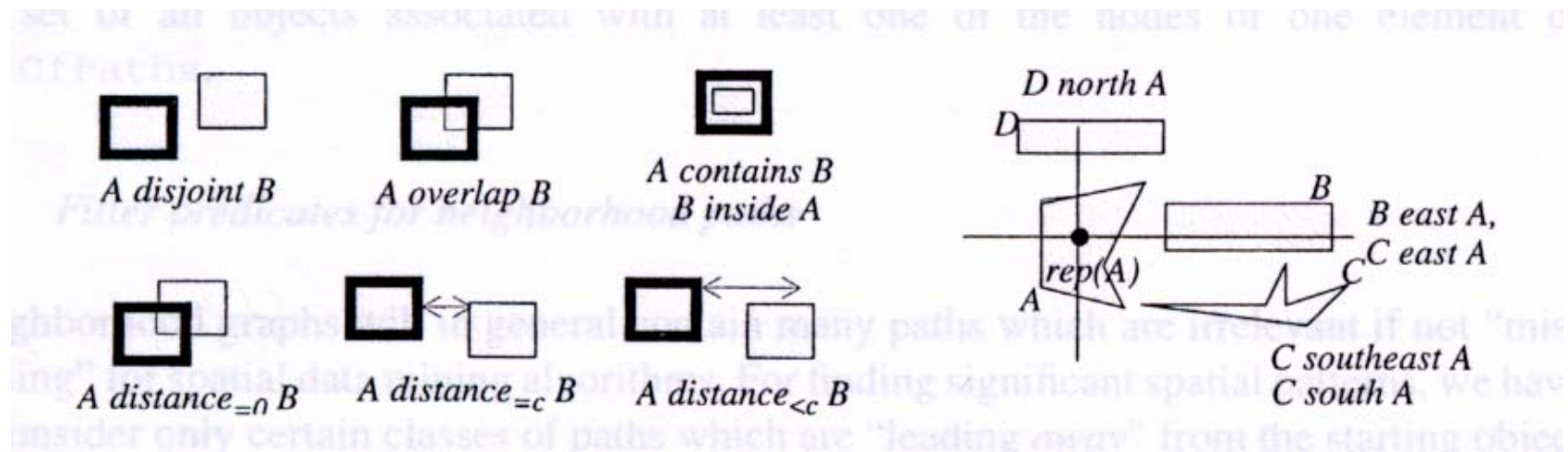
Traditional relational database for non-spatial data of this spatial object

Geometric spatial properties or features  
(primary spatial data)

# Spatial Computation (secondary data)

## Relations (functions/predicates)

- **Topological**: intersect, overlap, disjoint, etc.
- **Distance**: close\_to, far\_away, etc.
- **Direction**: north, south, east, west, northwest, etc.





# Examples of Spatial Databases

	$f_1$	$f_2$	$f_3$	$f_4$	
HouseID	Spatial F.	Yrs_old	Type	Num_pat	Class Label
00001	Close_to(railway)?	15yrs	I	3000	H
00002	Inside (city) ?	5yrs	T	2000	H
00003		10yrs	I	5000	H
00004		20yrs	T	1200	L
00005		12yrs	I	2100	L
...		3yrs	I	4000	...



Topological Features (Secondary Spatial Data)



Non-spatial Features

# What is Spatial Data Mining?

- **Data Mining (DM)** is the extraction of non-obvious, hidden knowledge (patterns/models) from large volumes of data

- **Spatial Data Mining (SDM)** is the extraction of interesting spatial patterns/models, and general relationships between spatial and non spatial data

The main difference is that the **neighbors of a spatial object** may have an influence on the object.

- The **explicit** location data (primary data)
  - The **implicit** relations of spatial neighborhood (secondary data)
- } used by spatial data mining algorithms

# Outlines

---

- What Is Spatial Data Mining?
- **Spatial Data Mining Tasks?**
- Creating A Primary Spatial Database.  
Case study: POPMAP
- Creating Secondary Spatial Features

# Spatial Data Mining Tasks

- Geo-Spatial Warehousing and OLAP
- Spatial data classification/predictive modeling
- Spatial clustering/segmentation
- Spatial association and correlation analysis
- Spatial regression analysis
- Time-related spatial pattern analysis: trends, sequential patterns, partial periodicity analysis
- Many more to be explored

# Example of Spatial Classification

MINE CLASSIFICATION RULES

ANALYZE crimes100000R

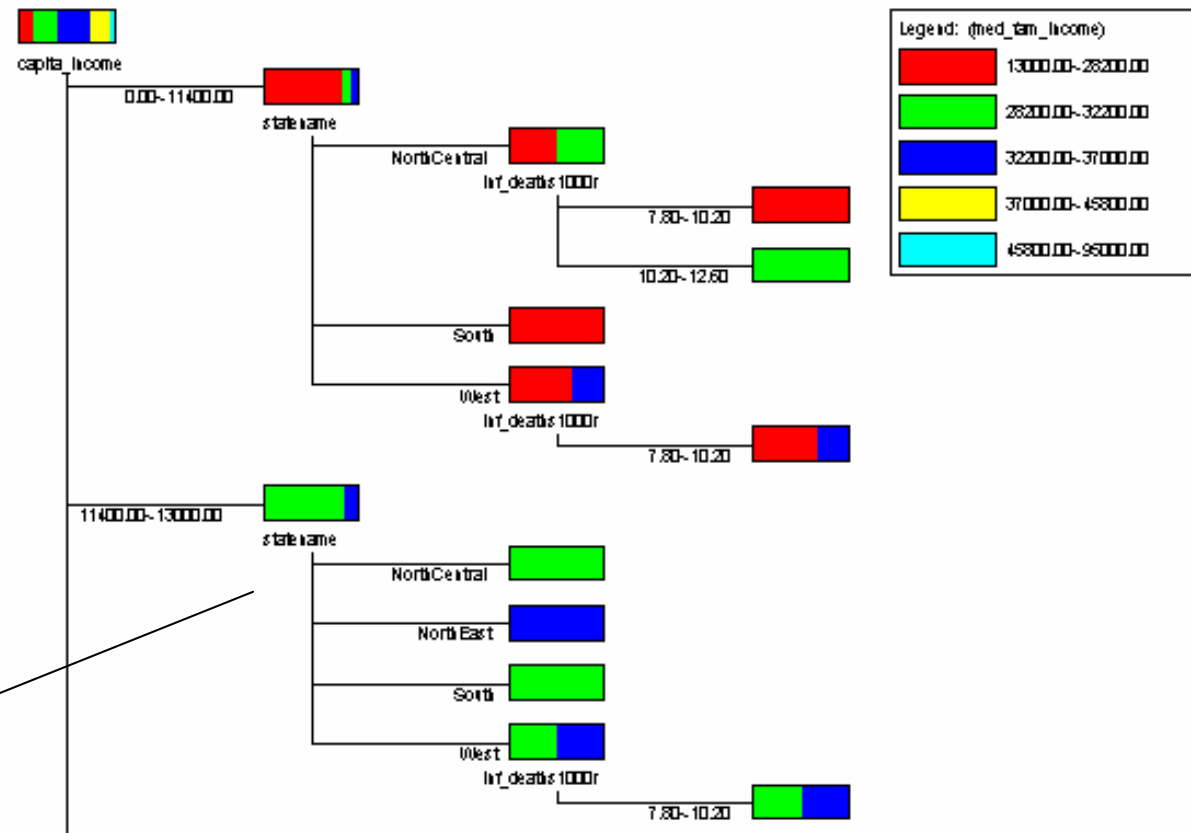
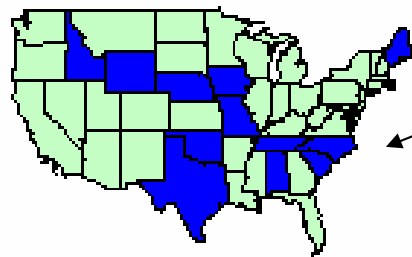
WITH RESPECT TO

states\_census.geo, statename,

capita\_income,

with\_bachelor\_degp

FROM states\_census



# Example of Spatial Clustering

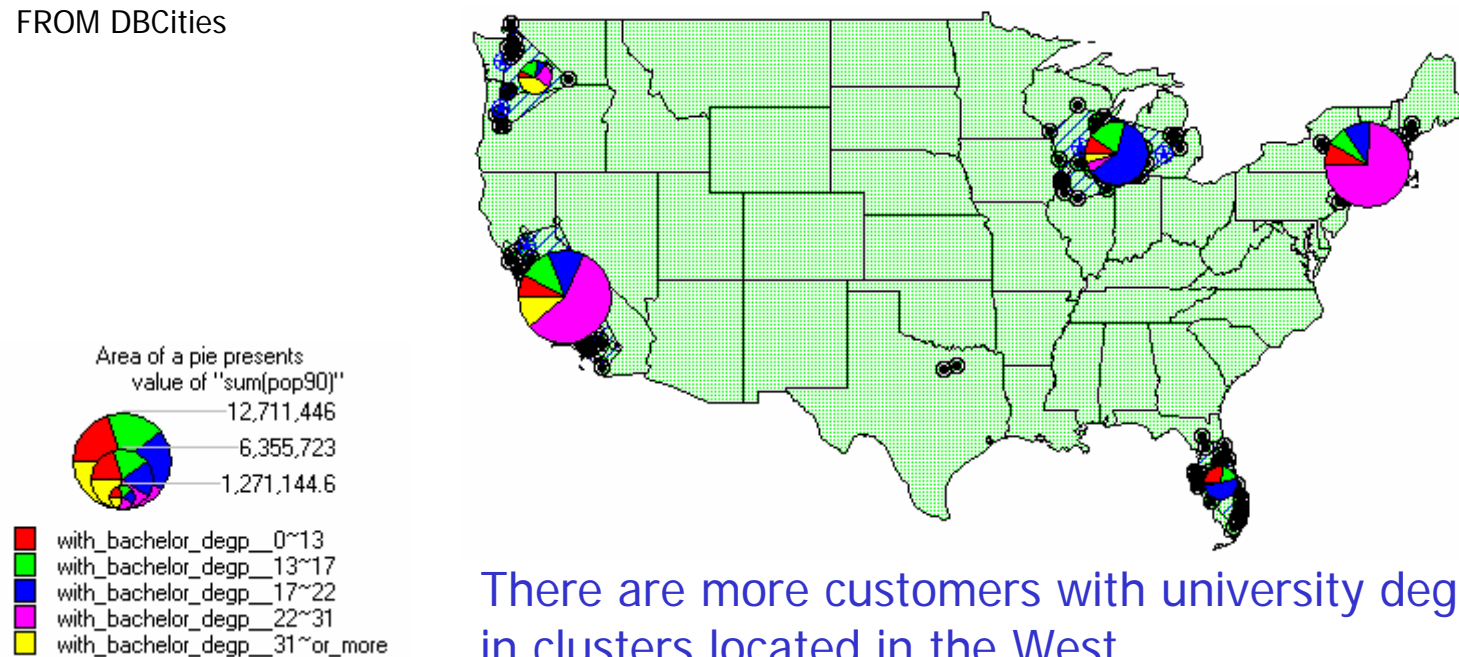
MINE CLUSTERS AS ``DBCities''

ANALYZE sum(pop90)

WITH RESPECT TO DBCities.geo, pop90,  
med\_fam\_income, with\_bachelor\_degp

FROM DBCities

- How can we cluster points?
- What are the distinct features of the clusters?



There are more customers with university degrees  
in clusters located in the West.

Thus, we can use different marketing strategies!

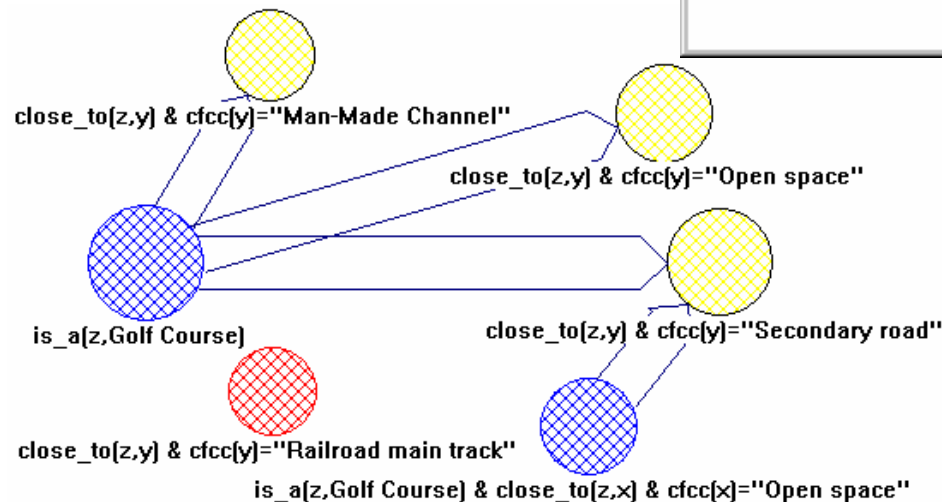
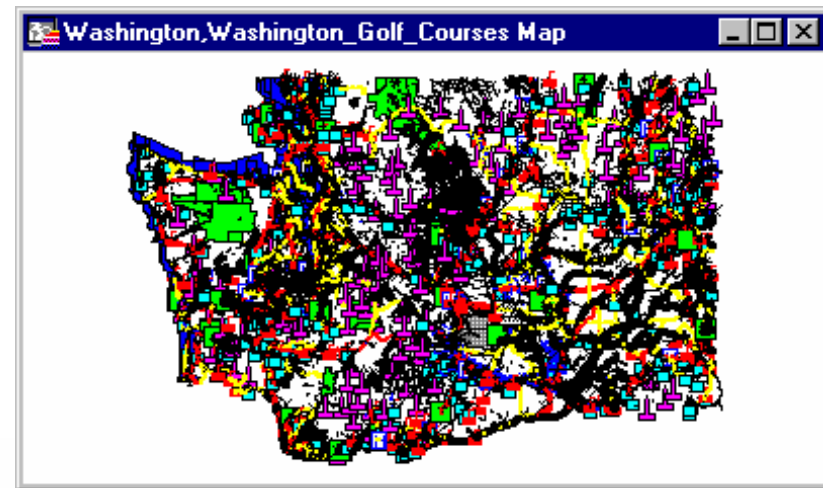
# Spatial Association Mining

- What kinds of spatial objects are close to each other in B.C.?"
  - Kinds of objects: cities, water, forests, usa\_boundary, mines, etc.
- Rules mined:
  - $\text{is\_a}(x, \text{large\_town}) \wedge \text{intersect}(x, \text{highway}) \rightarrow \text{adjacent\_to}(x, \text{water})$  [7%, 85%]
  - $\text{is\_a}(x, \text{large\_town}) \wedge \text{adjacent\_to}(x, \text{georgia\_strait}) \rightarrow \text{close\_to}(x, \text{u.s.a.})$  [1%, 78%]
- Mining method: Apriori + multi-level association + geo-spatial algorithms (from rough to high precision)

# Example of Spatial Association Mining

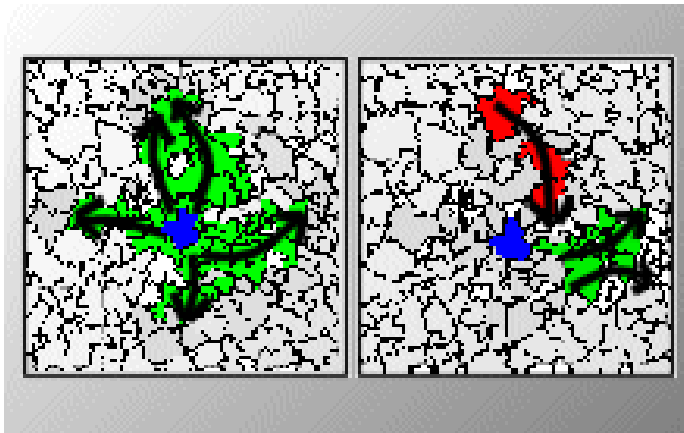
```

FIND SPATIAL ASSOCIATION RULE DESCRIBING
"Golf Course"
FROM Washington_Golf_courses, Washington
WHERE CLOSE_TO(Washington_Golf_courses. Obj,
    Washington. Obj, "3 km")
    AND Washington.CFCC <> "D81"
IN RELEVANCE TO Washington_Golf_courses. Obj,
    Washington. Obj, CFCC
SET SUPPORT THRESHOLD 0.5
SET CONFIDENCE THRESHOLD 0.5
    
```

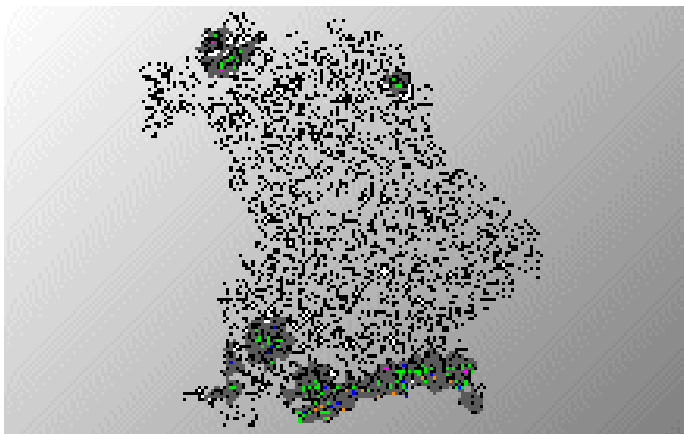




# Spatial Trend Detection & Characterization



**Spatial trends** describe a regular change of non-spatial attributes when moving away from certain start objects. Global and local trends can be distinguished



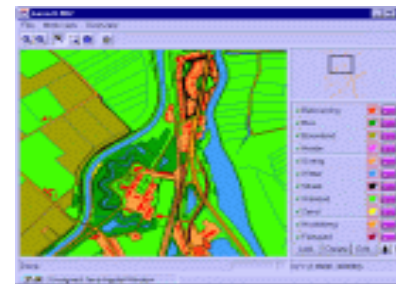
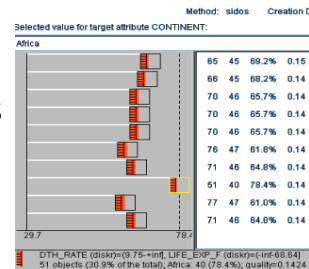
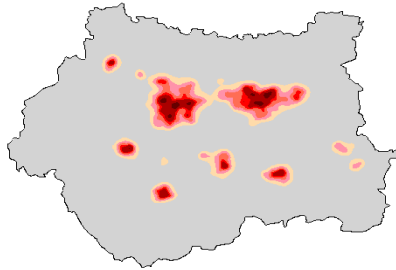
**Spatial (region) characterization** does not only consider the attributes of the target regions but also neighboring regions and their properties

# Spatial Trend Detection & Characterization

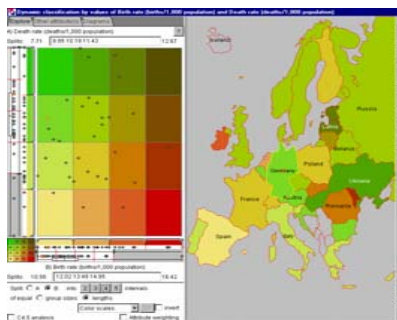
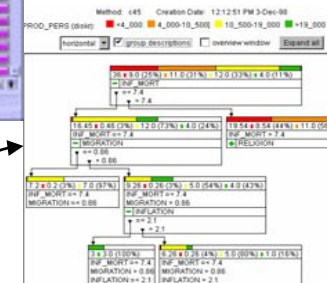
- **Spatial trend predictive modeling**
  - Discover centers: local maximal of some non-spatial attribute
  - Determine the (theoretical) trend of some non-spatial attribute, when moving away from the centers
  - Discover deviations (from the theoretical trend)
  - Explain the deviations
- **Example**
  - Trend of unemployment rate change according to the distance to Osaka
  - Trend of temperature with the altitude, degree of pollution in relevance to the regions of population density, etc.

# Spatial Data Mining: SPIN Elements

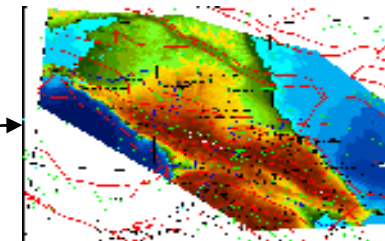
clusters of disease incidences



manipulations of a GIS system with layers (object classes) such as blocks, rivers, mountains.



two dimensional interactive classification of a GIS tool



subgroup mining and decision trees

# Outlines

---

- What Is Spatial Data Mining?
- Spatial Data Mining Tasks?
- **Creating Primary Spatial Databases.  
Case study: POPMAP**
- Creating Secondary Spatial Features

# Creating A Primary Spatial Database

## Case study: PopMap

- PopMap (UN Software and Support for Population Activities Project, 1992-1999 in IOIT)
  - Integrated Software Package for Geographical Information, Map and Graphics Database
  - Population Activities
- PopMap Features
  - Tools for **creating, editing** and **maintaining** databases of spatial and non-spatial data
  - Capabilities for retrieving and and processing data in worksheet, statistical graphs, etc.

# Creating A Primary Spatial Database

## Case study: PopMap

---

- PopMap databases
  - Can manage up to 999 different object classes and hierarchical substructures up to 7 layers each is an object class
  - Spatial data: points, lines, polygons
  - Non-spatial data: e.g., population, number of households... stored in tabular dBASE compatible files
- Tools to build spatial databases
  - Using digitizers
  - Automatic Raster-to-Vector conversion
  - Tools for describing non-spatial data

# Outlines

---

- What Is Spatial Data Mining?
- Spatial Data Mining Tasks?
- Creating A Primary Spatial Database.  
Case study: POPMAP
- **Creating Secondary Spatial Features**

# Why Create Secondary Spatial Features ?

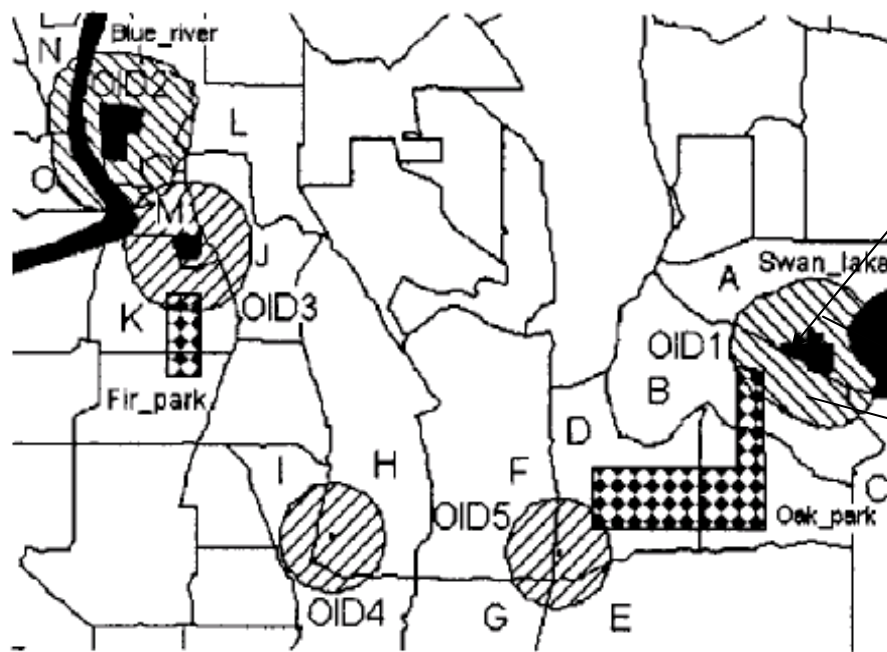
- There are some spatial operations (buffering, overlay) in GIS to discover some kinds of relations in spatial databases
- For spatial data mining tasks, a set of spatial operations to compute **implicit** relations on spatial neighborhood becomes necessary



# Requirements for Spatial Operations

- Spatial operations are **I/O-intensive**
  - **I/O-spatial operation time** is time to access to required primitive data in a database
  - Due to large amounts of spatial objects, I/O access is time consuming
- Spatial operations are **CPU-intensive**
  - **CPU-spatial operation time** is time to process calculations between objects
  - Due to complex spatial operations, their execution is time intensive

# Example of Intersect Operation

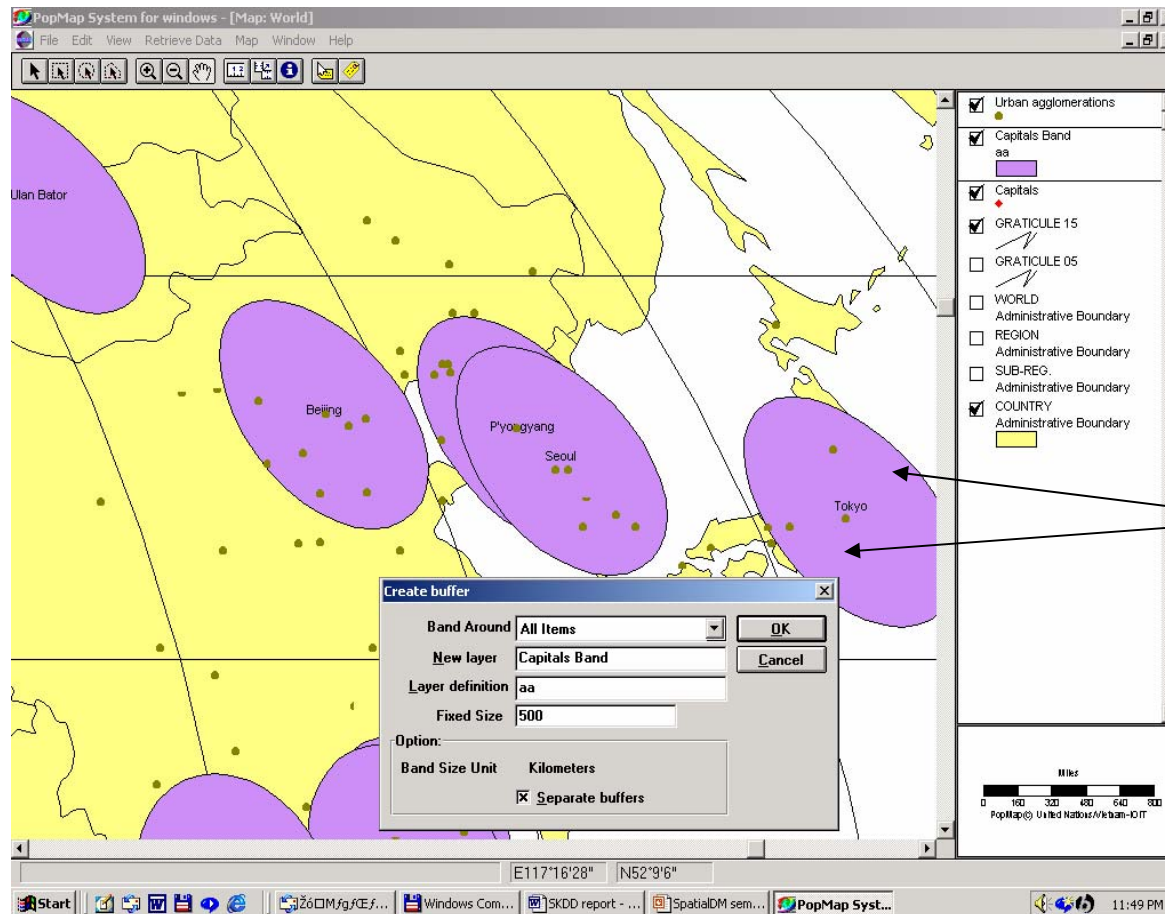


OID	high_profit	others
1	Y	close_to(x,Oak_park), close_to(x,Swan_lake)
2	Y	close_to(x,Blue_river)
3	N	close_to(x,Fir_park), close_to(x,Blue_river)
4	N	
5	N	close_to(x,Oak_park)

BlockID	population	avg_income	crimes
▲ A	5000	\$25,500	50
→ B	4500	\$27,500	40
C	5500	\$28,300	45
H	6000	\$31,500	50
I	5000	\$25,500	30
...	...	...	...

To build decision trees for classifying objects OID1, ..., OID5 (to Y or N high\_profit) based on (1) characterized objects by areas closed to objects by building buffers around objects, (2) census blocks groups **intersecting buffers** and other polygons.

# Example of Inside Operation



Want to evaluate  
the density  
population of all  
cities within 500  
km from Capital



Inside Operation  
finds all cities (in  
the buffers) that  
match this  
condition

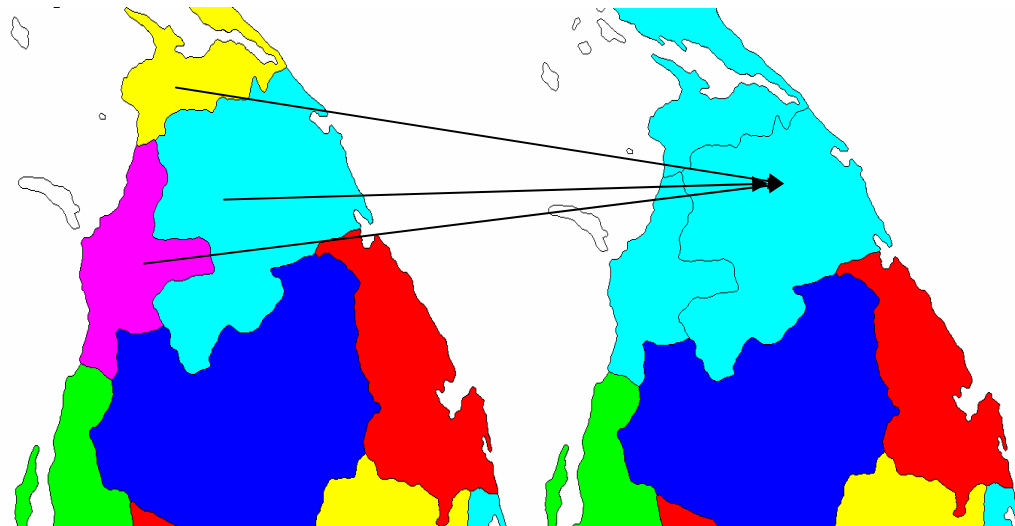
# Example of Merging Neighboring Regions

## Input

- a map with about 3,000 weather probes scattered in some countries.
- daily data for temperature, precipitation, wind velocity, etc.
- attributes are organized in hierarchies

## Output

- a map that reveals patterns : merged (similar) regions!

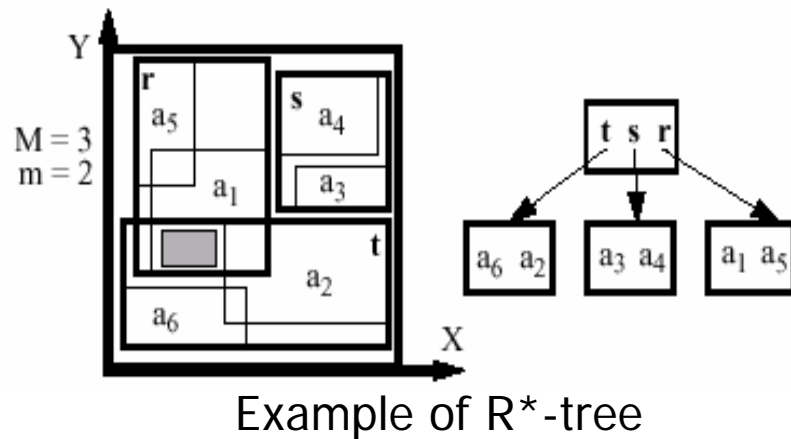


Merging 3 regions (yellow, magenta, blue) to one region (colored by blue)

# Some Methods for Spatial Joins

- **Spatial join**: one of the most important operations for combining spatial objects of several relations
- Multi-step processing of **spatial joins using R\*-tree**
- Efficient polygon **amalgamation methods**: compute the boundary of the union of a set of polygons
- **Indexing** support for spatial joins
- **Schedule** of join operations
- etc.

# Multi-step Spatial Joins using R\*-tree



Set of spatial objects

$A = \{a_1, \dots, a_n\}$

$B = \{b_1, \dots, b_m\}$

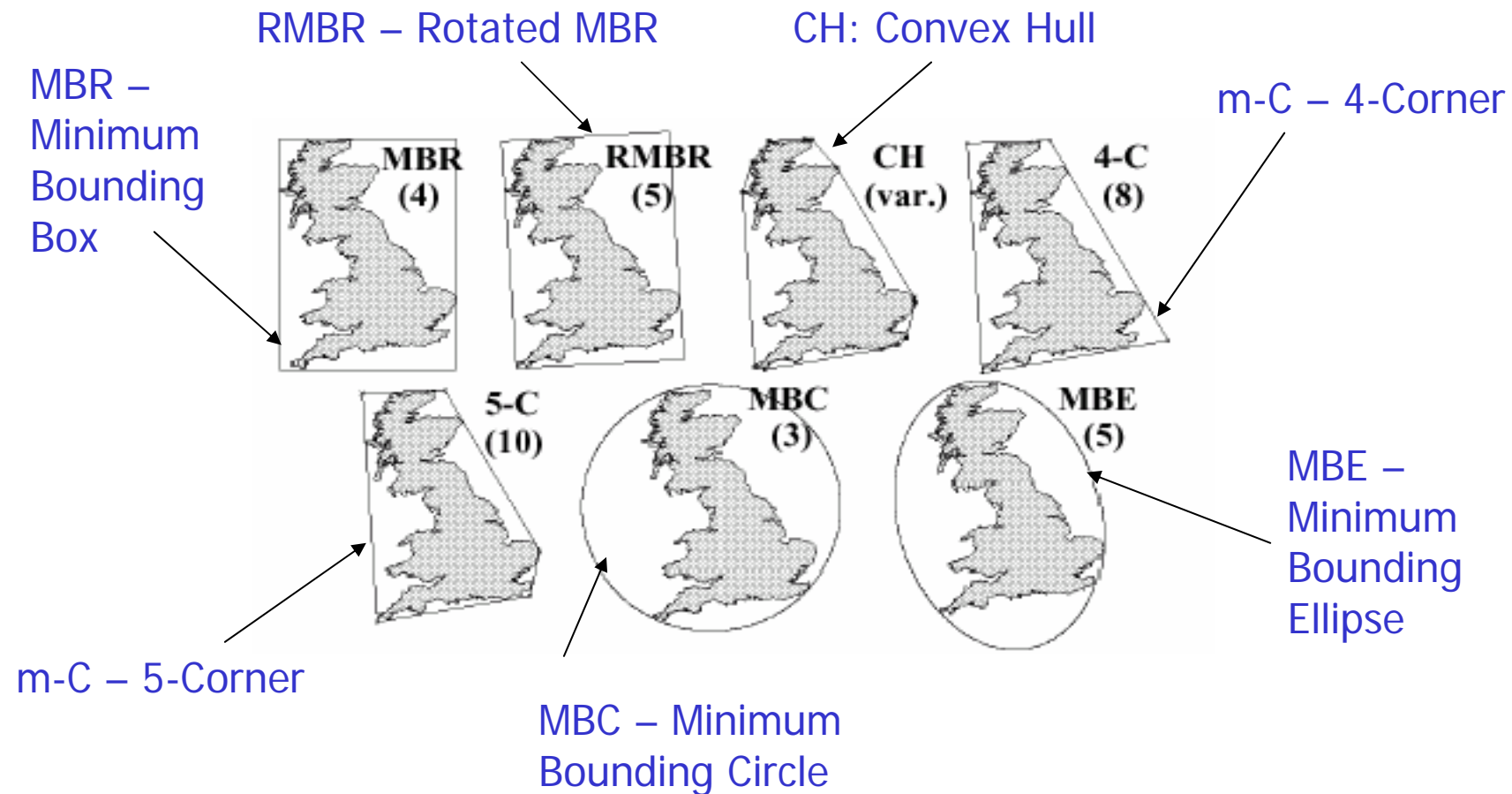
Id – function to assign an unique identifier to each object

Mbr- function to compute Maximum Bounding Box

## Spatial joins

1. **MBR-spatial-join**: Compute all pairs  $(Id(a_i), Id(b_j))$  with  $Mbr(a_i) \cap Mbr(b_j) \neq \emptyset$
2. **Id-spatial-join**: Compute all pairs  $(Id(a_i), Id(b_j))$  with  $a_i \cap b_j \neq \emptyset$
3. **Object-spatial-join**: Compute  $a_i \cap b_j$  with  $a_i \cap b_j \neq \emptyset$

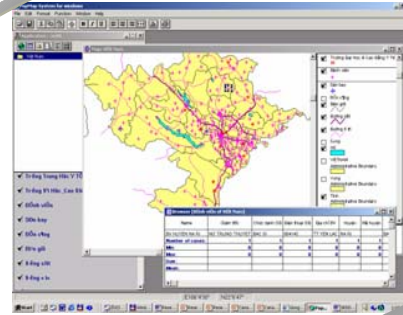
# Approximations in Multi-step Spatial Joins



# Project under Consideration

POPMAP lab (IOIT):  
Methods and tools to  
create primary  
spatial databases

OCR lab (IOIT+HUS):  
Methods and tools to  
create secondary  
spatial databases



CSA lab (JAIST):  
Spatial data mining  
methods and  
applications

KDC lab (JAIST):  
Methods and tools  
for spatial  
data mining

To create high quality methods and tools for spatial data mining



# Pattern Recognition Lab

## Institute of Information Technology, Vietnam NCST

- VnOCR: Software for recognizing Vietnamese characters
- Integrated in products of HP and Fujitsu, and used in many organizations
- Awards: 1<sup>st</sup> prize for scientific research 2000 (MOSTE), 2<sup>nd</sup> prize of VN PC World, etc.
- Experience on image and spatial objects processing algorithms



Cảng Marseille khỏi lửa. Thông tin ban đầu cho biết cảnh sát đã can thiệp bằng ...

# Current work

- Convert POPMAP data structure into MIF format of MapInfo
- Development of Multi-Layer DBSCAN and Visual Support k-Means. Algorithms for finding common edges of polygons, etc.
- Trials with hospitals in Vietnam (593 districts, 551 hospitals and medical care centers in Vietnam): density, impacts of geographical factors, population, etc.



**Seminar in Hanoi, 3.2001**

# non-spatial attribute

- 1 bac sy phuc vu duoc khoang bao nhieu dan
- So bac sy cho 10000 dan
- Su phan bo cua mot so benh theo vung
- Thong tin chi tiet ve cac bac sy, y ta... cho tung huyen, xa cua cac tinh trong ca nuoc.

# Geometrical computation

- Given a point (hospital) within a district and  $\delta$ , find polygons closed to the point within  $\delta \rightarrow$  find polygons adj
- Cho 1 diem (benh vien) nam trong 1 huyen nao do va mot delta (ban kinh trong DBSCAN), phai tim nhung polygon nao cach diem cho truoc voi ban kin delta. Van de nay dan den phai xac dinh cac polygon ke voi polygon co diem nam trong (truoc tien phai xac dinh diem thuoc polygon nao). Cung co nhieu thuat toan cho van de nay nhung cai nao hieu qua cho du lieu lon? Binh va Minh da tim hieu duoc cau truc du lieu nay va da xac dinh thuat toan tim cac canh chung cua cac polygon.
- Co 1 diem cho truoc --> tim polygon ma no nam trong --> tim so cac diem thuoc polygon do.