Analisi di Dati Spaziali

1. Oggetti spaziali e variabili spaziali
2. Analisi spaziali
3. Esempio

Oggetti Spaziali

Oggetti che hanno un riferimento spaziale (coordinate)

Punti
- epicentri di un terremoto
- eventi di un tipo di cancro
- posizioni di una specie di pianta

Linee
- fiumi, strade, linee elettriche

Zone
- gerarchia regionale: stati, regioni, province, …

Superfici
- zone tridimensionali

Variabili Spaziali

Misurano proprietà di oggetti spaziali

Variabili continue

virtualmente noti per tutti i punti nello spazio tridimensionale

pressione atmosferica, inquinamento, …

misurati per punti selezionati ed interpolati fra i punti

metodi per interpolare spazialmente (Kriging)

Variabili discrete

descrivono punti discreti, linee, zone

tipicamente aggregati: tassi di disoccupazione nelle regioni

Autocorrelazione e Cluster Spaziali

I valori di una variabile per oggetti vicini sono correlati
Distribuzione di una variabile spaziale:
"tutto è differente ma i valori di oggetti vicini sono simili"

Metodi per misurare autocorrelazione spaziale:

Moran´s I

-1 : i valori di oggetti vicini sono opposti

0 : i valori distribuiti casualmente in spazio

1 : i valori di oggetti vicini sono gli stessi

Clustering spaziale

gli oggetti con una certa proprietà sono più o meno uniformemente distibuiti nello spazio o esibiscono una regolarità spaziale?

Dati Multirelazionali

Parecchi strati spaziali

eventi di un malattia (p.e. SARS) punti
linee (fiumi, strade, …)
ospedali, aeroporti, … punti, zone

Join spaziali: con coordinate
intersezioni e distanze

Tipi di dati spaziali

vettori (poligono = serie di vettori di coordinate)
rappresentazione compatta, relazioni spaziali
raster

p.e griglia bidimensionale per rappresentare una variabile spaziale

Compiti Principali

Strutture di indexing spaziale

efficienti interrogazioni spaziali
trovare oggetti che intersecano una regione
con una distanza < 1000 m alla stazione
trovare coppie di oggetti rispettando un predicato spaziale (join)

Visualizzazione e amministrazione (GIS)

componenti cartograficche ed analitiche, amministrazione di dati

Clustering spaziale

per oggetti con una proprietà distinte (p.e. malattia, crimine)
- cerchi che contengono significativamente molti degli oggetti
- agglomerazione di zone elementari: cluster di zone vicine
- k-medoid (k-means), con distanze geometriche

Versioni spaziali di metodi di data mining

alberi di decisione, regole d´associazione, sottogruppi,…

GIS: Domande di natura Geografica

Che cosa c'è? **indicando su una mappa**

Dov'è …? **Specificando criteri**

Che cosa è cambiato da …? **Dati temporali**

Qual è il cammino migliore? **Criteri: distanza, tempo, costi**

Tecniche di visualizzazione

mappa choropleth **con zone colorate**
i valori di una variabile corrispondono a un colore

Spatial Subgroup Mining Integrated in an Object-Relational Database

1 Spatial Subgroup Mining
2 Database Integration
3 Visualization, Clustering & Causality

Willi Klösgen and Michael May
Fraunhofer AIS

1 Spatial Subgroup Mining

Goals of Spatial Data Mining

- **Identifying spatial patterns**
  Death rate is high in areas near power plant
- **Identifying information relevant for explaining the spatial pattern**
  attributive patterns: high rate of poor people in those areas
- **Presenting the information in a way that is intuitive to the analyst and supports further analysis (e.g. GIS, visualization)**

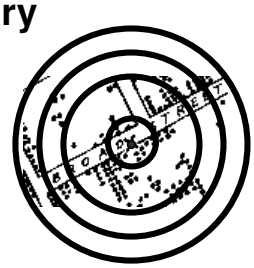
A classic example for spatial analysis

Dr. John Snow
Deaths of cholera epidemia
London, September 1854

Infected water pump?

T=Death rate S=blocks near pump B
spatial objects: housing blocks, streets, pumps, ...

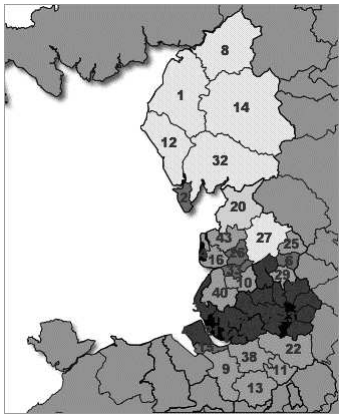
To solve the problem, a good representation is necessary



Represent spatial objects of several types
Represent spatial relation of objects to *other* objects

It is not only important where a cluster is but also, what else is there (e.g. a water-pump)!

North-West England



North West England
with 43 local authorities
and 1011 wards

Census data

Neigborhood Statistics

Census Tables

Table S09 Economic position and ethnic group: Residents aged 16 and over

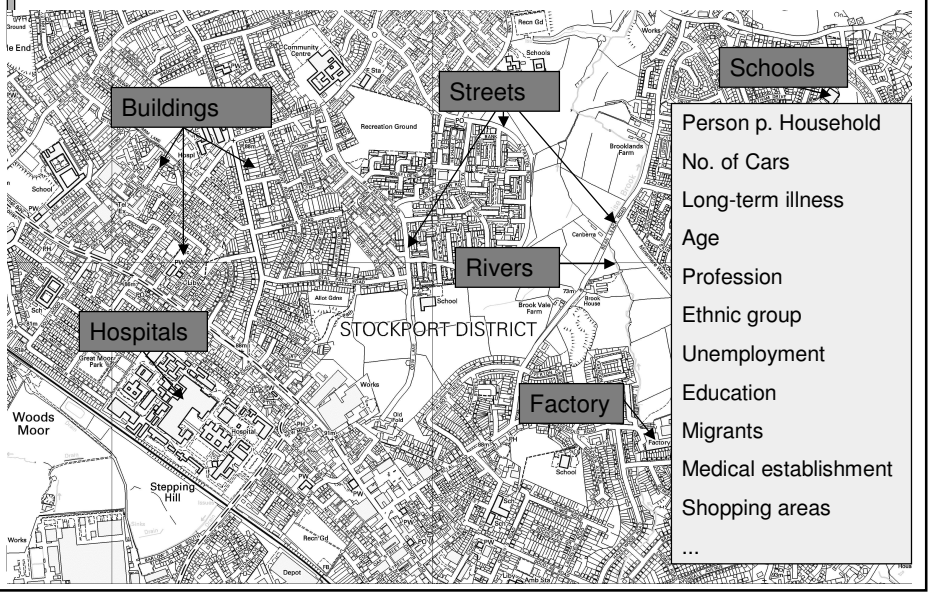
Economic position	TOTAL PERSONS	White	Black groups	Indian Pakistani and Bangladeshi	Chinese and other groups	Persons born in Ireland
TOTAL PERSONS	1	2	3	4	5	6
Males 16 and over	7	8	9	10	11	12
Economically active	13	14	15	16	17	18
Unemployed	19	20	21	22	23	24
Economically inactive	25	26	27	28	29	30
Females 16 and over	31	32	33	34	35	36
Economically active	37	38	39	40	41	42
Unemployed	43	44	45	46	47	48
Economically inactive	49	50	51	52	53	54

Table 1: An aggregated cross tabulation available e.g. for all wards

Layer name	Description	Type	Objects
Motorway	Motorway	Line	494
	Motorway (over), Motorway tunnel		
PrimRoad	Primary route, dual carriageway	Line	3945
	Primary route, dual carriageway (over)		
	Primary route, single carriageway		
	Primary route, single carriageway (over)		
	Primary route, narrow		
	Primary route, narrow (over)		
	Primary route tunnel		
A_Road	A road, dual carriageway	Line	3882
	Other subtypes: see PrimRoad		
B_Road	B road, dual carriageway	Line	4368
	Other subtypes: see PrimRoad		
Mnr_Rd4o	Minor road over 4 meters wide	Line	9705
	Minor road over 4 meters wide (over)		
	Minor road over 4 meters wide tunnel		
Mnr_Rd4u	Minor road under 4 meters wide / over / tunnel	Line	8756
Railway	Railway, standard gauge	Line	4231
	Railway, standard gauge (over)		
	Railway, narrow gauge / narrow gauge (over)		
	Railway tunnel / Railway station		
UrbAreaL	Large Urban Area (outer limit)	Line	384
	Large Urban Area (inner limit)		
UrbAreaS	Small Urban Area (outer limit) / (inner limit)	Line	2235
Water	Inland water (inner limit)	Line	438
	Inland water (outer limit)		
River	River (primary), source / middle / lower	Line	12103
	River (secondary), source / middle / lower		
	River (other and drains)		
Canal	Canal	Line	968
	Canal tunnel / Canal (over)		
Wood	Wood/Forest (inner limit)	Line	859
	Wood/Forest (outer limit)		
Foreshor	Foreshore (sand, inner limit)	Line	209
	Foreshore (other) and offshore rocks (il)		
	Foreshore (sand, outer limit)		
	Foreshore (other) and offshore rocks (ol)		
National	National boundary	Line	12
County	County boundary	Line	88
District	District boundary	Line	61
Park	National park/forest park	Line	11
CampCara	Camping and caravanning combined sites	Point	212
...			

Table 2: Geographic Layers (spatial objects of type line / point)

UK, Greater Manchester, Stockport



Multirelational Description Language
Spatial Query Language for Subgroups

Domain:

{O_1, ..., O_n} set of object classes, e.g.: **EnumerationDistricts**, Rivers, Streets, ...

attribute schema for each object class

R = {R_1, ..., R_k} graph of relations $R_i(O_{i1}, O_{i2})$: intersects(ED,Riv), intersects(ED,Str)

Description Language:

Multirelational subgroups are represented by a concept set $C = \{C_i\}$, where each C_i consists of a set of attribute value-pairs $\{A_1=v_1,...,A_n=v_n\}$ for O_i

```
C={{ED.unemployment=high,ED.age35_60=high},
        {Street.name=Manchester Road}, {River.type=primary}}
```

*"**Enumeration districts** with high rate of unemplyoment and of 35-60 year old persons and crossed by Manchester Road and crossed by (**at least one**)primary river"*

1) cross product $O = O_1 x ... x O_n$

2) R defines a subset of O: $R1(O) \subseteq O$

3) C_i represents a subset of O_i and C a subset of $R1(O)$: $R2(O) \subseteq R1(O) \subseteq O$

4) Projection of $R2(O)$ onto O_1

at least one: existential quantifier for 1 to m relations (count > 0)

aggregation functions (count, sum, avg, min, max, ...) *sum (river.pollution)=high*

Representation of spatial data in GIS

A set of spatial object classes $O_1,...,O_n$ each O_i has a geometry attribute G_i

then O_i can be linked (joined) to O_k over G_i and G_k

- Geometry attributes G_i consist of ordered sets of x,y-pairs defining points, lines, or polygons
- Different types of spatial objects are organized in different tables O_i (geographic layers), e.g. streets, rivers, enumeration districts, buildings, and
- each layer can have its own set of attributes $A_1,..., A_n$ and at most one geometry attribute G

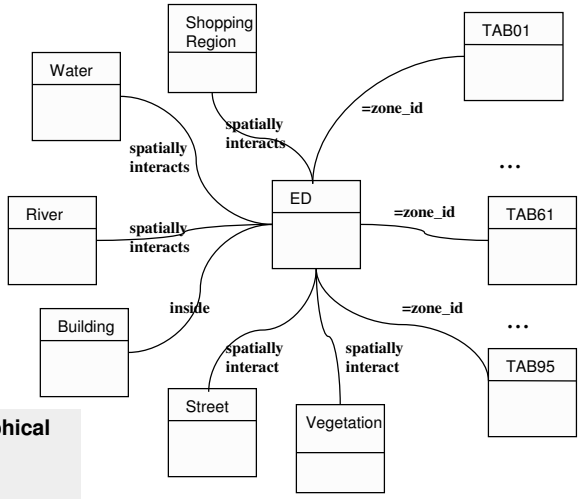
Spatial Predicates in SDBS

Topological relation (Egenhofer 1991)

A disjoint B, B disjoint A	
A meets B, B meets A	
A overlaps B, B overlaps A	
A equals B, B equals A	
A covers B, B covered by A	
A covered-by B, B covers A	
A contains B, B inside A	
A inside B, B contains A	

Distance relation: Minimum distance between 2 points

Stockport Database Schema



Attribute data

95 tables with census data,

~8000 attributes

Spatial Hierarchy

- County
- District
- Wards
- Enumeration district

Geographical Layers

85 tables

Possible approaches

Pre-processing part of the spatial data deriving aggregated attributes for target object class

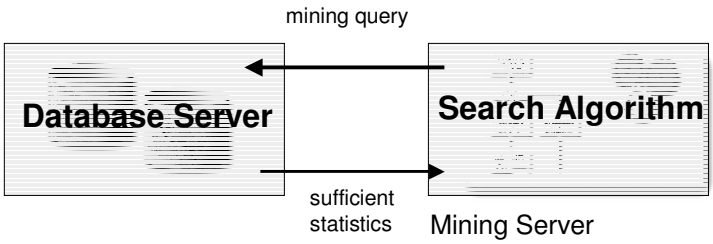
Embedding data mining in a spatial database (GIS) dynamically join object classes

- Advantage:
  does not restrict hypothesis space

Our approach

2 Database Integration

Division of labour between RDBMS and Search Manager



- Database integration: efficiently organize mining queries
- Mining query delivers statistics (aggregations) sufficient for evaluating **many** hypotheses
- all subgroups of the next expansion level

- search in hypothesis space
- generation and evaluation of hypotheses (subgroup patterns)

Multirelational queries require joins

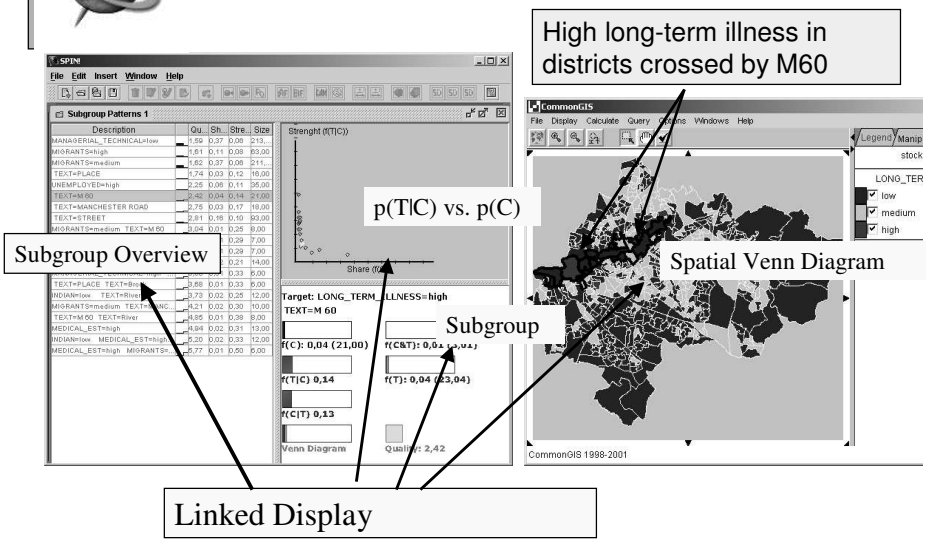
Expansion of a parent subgroup -
adding an additional selector for a concept -
adding a new concept (join another table)

```
select ED.T,ROAD.A from
    (select distinct ED.KEY,ED.T,ROAD.A from ED,ROAD
        where MDSYS.SDO_FILTER(ED.GEOM,ROAD.GEOM,´mask=anyinteract´...))
    group by ED.T, ROAD.A
```

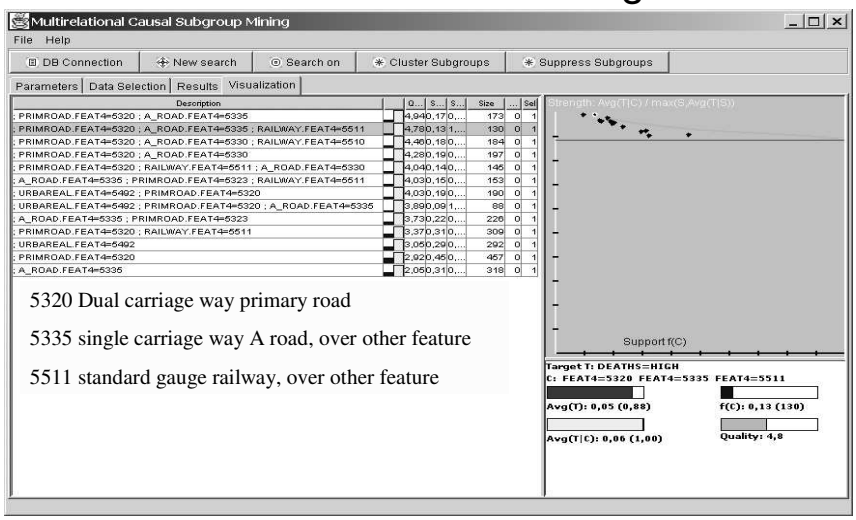
1011 EDs, 3882 A_ROADs : 1 sec
1011 EDs, 3882 A_ROADs, 4368 B_ROADs: 12 sec

3 Visualization, Clustering, Causality of Subgroups

Visualization of spatial sugroups



Mining health care data: Deathrates in North West England

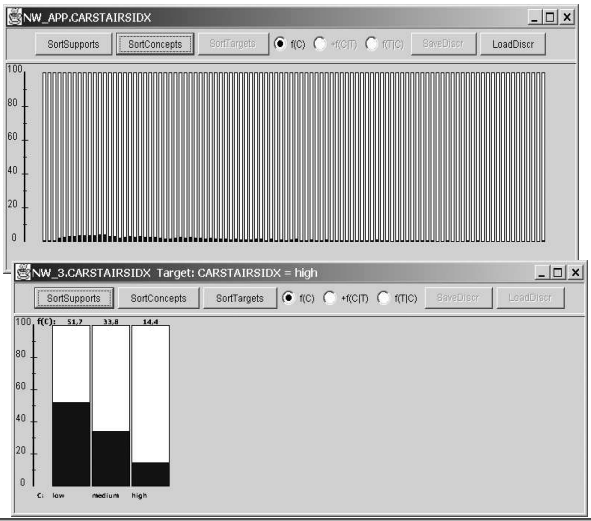


5320 Dual carriage way primary road

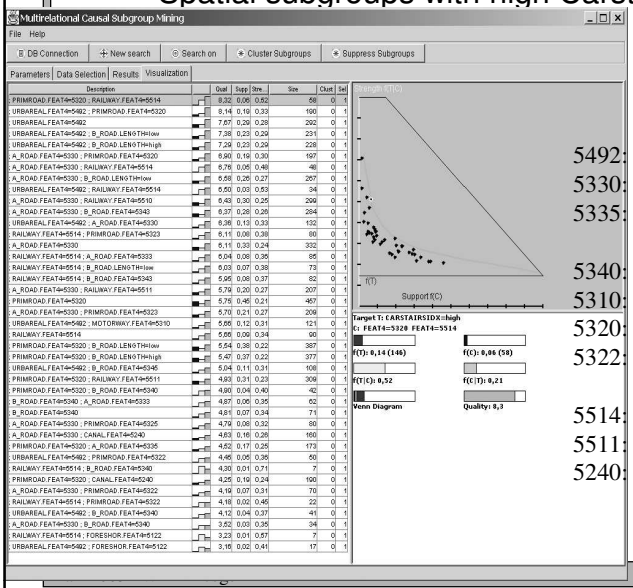
5335 single carriage way A road, over other feature

5511 standard gauge railway, over other feature

Carstairs Index

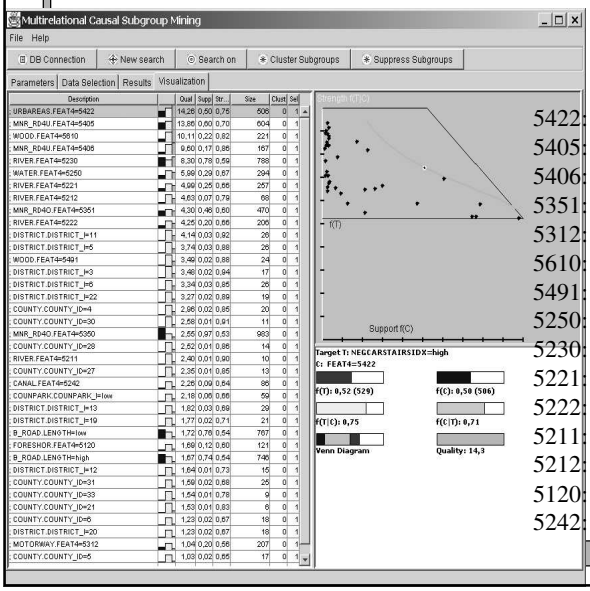


Spatial subgroups with high Carstairs index



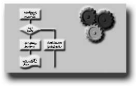
5492: inner limit of large urban areas
5330: dual carriageway ARoad
5335: single carriageway Aroad over other feature
5340: dual carriageway Broad
5310: motorway, normal
5320: dual carriageway PrimRoad
5322: dual carriageway PrimRoad, over other feature
5514: railway tunnel
5511: standard gauge, over other
5240: canal, normal

Spatial subgroups with low Carstairs index

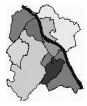


5422: outer limit of small urban areas
5405: minor roads under 4 m wide, normal
5406: minor roads under 4 m wide, over o.
5351: minor roads over 4 m wide, over o.
5312: motorway, over other
5610: outer margin of woods
5491: inner limits of woods
5250: outer limits of waters
5230: other rivers and drains
5221: source of secondary rivers
5222: middle of secondary rivers
5211: source of primary rivers
5212: middle of primary rivers
5120: foreshore, sand, outer limit
5242: canal, over other

Approach to Spatial Knowledge Discovery

Data Mining  $\sqrt{\frac{n}{p_0 \cdot (1-p_0)}}(p - p_0)$

+

Geographic Information Systems 

=

SPIN! 

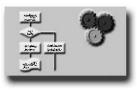
Approach: Translation of Spatial Subgroup Mining to SQL

- Representing subgroups in object-relational SQL, i.e. multi-relational representation
- Using representation for spatial geometry based on Spatial Database
- Division of work between RDBMS and Search Manager
- Combining visualization in abstract and physical space

Conclusion

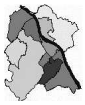
SPIN! combines spatial data mining methods and Geographical Information Systems to enhance the analysis of spatial data

Combination with RDBMS allows application to many kinds of real-world applications, e.g. geomarketing, site selection, urban planning, environmental planning



$$\sqrt{\frac{n}{p_0 \cdot (1-p_0)}}(p-p_0)$$

+



=