Knowledge Discovery Applications to the Public Sector

Slides borrowed from Donato Malerba - University of Bari

Outline of the presentation

- Public Sector Information in the Information Society
- Mining official data
- eGovernance/eGovernment
- Environment
- Health Care

Information Society and Knowledge

- Profound transformation in our society

  industrial society ➔ **information society *IS***

- the majority of workers will soon be producing, handling and distributing information or codified **knowledge**
- the generation and utilization of knowledge is recognized as the driving force of productivity and economic growth of *IS*

Information Society and Knowledge

- The only sustainable source of competitive advantage for Europe is in the
  - development,
  - accumulation,
  - sharing,
  - conversion and
  - application

These activities are the basis of innovation

of knowledge in all production and decision-making processes.

Innovation: Who are the actors?

- A principle role is not only played by academic and research institutions,
- but also by private companies, professional associations,
- and above all, by **public and governmental agencies**.

- The public sector, because of its size and scope of activities, has been identified as …

Innovation: Who are the actors?

"*the biggest single information content resource for the creation of value-added information content and services. Studies have shown that the bulk of commercial information services in the EU information market, consists of services in areas where the public sector holds very important resources.*"

Green Paper on Public Sector Information in the Information Society [COM (1998) 555] EC, 1998

Services for Whom?

Although progress has been made by the EU countries in providing sophisticated online public services, they **almost exclusively** concern online services for **commercial activities**.

Services for Whom?

In almost every country, public services for businesses score significantly higher than those for **citizens**, and this gap is growing.

"*European Governments now need to focus more on other services, particularly those most relevant to citizens" – and, overall, ensure that they have the right approach to implement successfully*"

Stanislas Cozon Vice President, Cap Gemini Ernst & Young, commenting the report prepared for the European Commission. Press release, February 6th 2003

KD for Knowledge-Based Services

- Developing **knowledge-based services** for
  - Commercial activities
  - Citizens

  is crucial for Information Societies.
- Public sector plays an important role
- **Knowledge discovery technologies** are fundamental for the development of such knowledge based services.
- They are essential to the transformation of large amount of public data knowledge

KD for Knowledge-Based Services

- There is a great potential for KD or DM applications in the **public sector**, since all European governmental departments and local authorities collect huge amounts of data that are practically unexploited in the decision-making process.
- KD can help to manage the rapidly increasing demand for services from the communities being served and the necessity for the logical rationing of the services.

Examples of problems …

- … encountered by public administration officials:
  - the expanding elderly population, leading to great pressure on the available health services,
  - increasing crime, leading to serious difficulties in the organization of penal institutions,
  - the widespread use of automobiles, which cause severe problems regarding transport and environmental issues.
- Complex decision making

Examples of problems …

- National and local authorities need to use the valuable information hidden in the data collected by governmental agencies and other public institutions,
  - schools, universities and colleges,
  - hospitals and health care institutions,
  - the military and social service institutions.

Outline of the presentation

- Public Sector Information in the Information Society
- Mining official data
- eGovernance/eGovernment
- Environment
- Health Care

Mining Official Data

- In statistics, the term "*official data*" denotes data collected in censuses and statistical surveys by National Statistics Institutes (NSIs), as well as administrative and registration records collected by government departments and local authorities.
- They are used to produce "*official statistics*"
  - E.g. *inflation rate*

Mining Official Data

- Why not mining official data to extract knowledge?
- Some examples:
  - Mining census data for spatial effects on mortality ➔ improve health services
  - Mining georeferenced census data for urban accessibility ➔ improve publ. transp. services
  - Calculating economic household indexes ➔ improve CRM

Mining census data for spatial effects on mortality

- Problem: explore possible factors for high mortality rate
- Factors:
  - Geographical
  - Deprivation
  - Transportation
- Data:
  - Census data (UK census 1991)
  - Deprivation data (Office of National Statistics, 1998/9)
  - Health data (Office of National Statistics, 1998/9)
  - Spatial data (provided by Ordnance Survey)

W. Kloesgen, M. May, & J. Petch (2003). Mining census data for spatial effects on mortality, *Intelligent Data Analysis*, 7(6): 521,540.

Mining census data for spatial effects on mortality

- Census data can be aggregated to different levels of spatial unit

County

⇑

District

⇑

Ward

⇑

Enumeration District (ED)

Aggregation is necessary to preserve confidentiality.

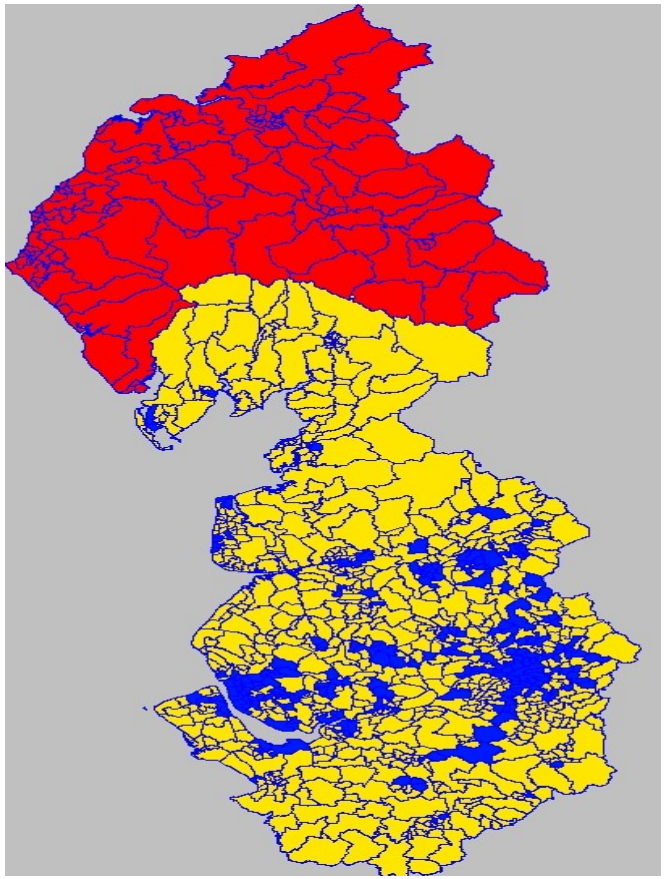
NSI's are not allowed to disclose micro-data

Mining census data for spatial effects on mortality

- Issue: which is the appropriate level of analysis?
  - Lower levels ensure higher homogeneity of aggregated variables → higher potential to identify and evaluate hypotheses about individuals (persons)
  - Lower levels require scalable methods ← the number of the objects in an analysis can get very large for wide regions
  - The appropriate level also depends on the available secondary data (e.g. on deprivation and health)

In this study: 1011 wards situate in the 43 local authorities of North West England

Mining census data for spatial effects on mortality



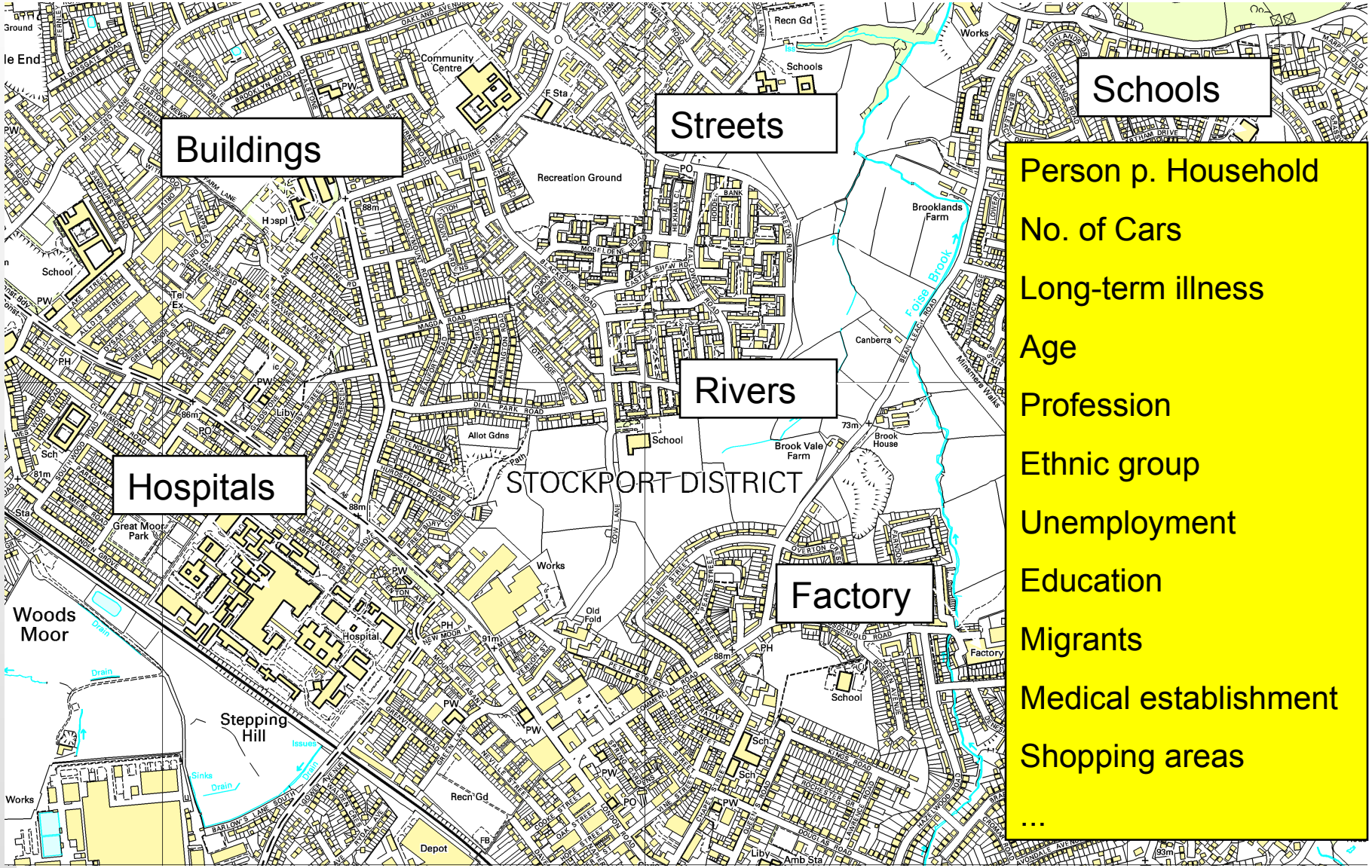
Wards level analysis of mortality rate. Red → missing data, yellow → not-high, blue → high mortality rate

Mining census data for spatial effects on mortality

- Several detailed geographic layers are available in the MeridianTM product of the Ordnance Survey
  - Roads, railway lines, rivers, buildings, …

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		
…			

UK, Greater Manchester, Stockport



Mining census data for spatial effects on mortality

- Issue: how to represent spatial relations between spatial objects?
  - 9-intersection model formalized by Egenhofer

A disjoint B, B disjoint A

A meets B, B meets A

A overlaps B, B overlaps A

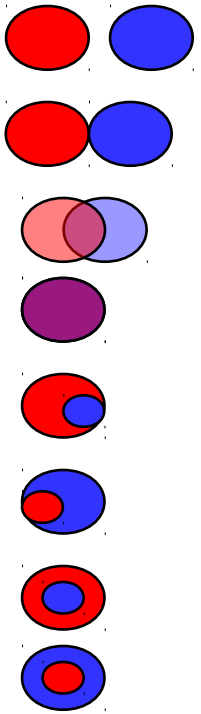
A equals B, B equals A

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

Mining census data for spatial effects on mortality

- Spatial relations can be pre-computed before the data mining step or computed on-the-fly by means of appropriate queries during the data mining step.

In this work, spatial and non-spatial joins are executed dynamically during the data mining step.

Mining census data for spatial effects on mortality

- Deprivation indices are provided by the UK Office of National Statistics at ward level for each year
- They are used as additional explanatory variables
- A number of different such indices have been developed for different applications
  - Jarman
  - Townsend
  - Carstairs
  - DoE

Mining census data for spatial effects on mortality

- Issue: Deprivation indices are ordinal. How to use them?
  - Not all data mining systems handle ordinal variables
  - Discretization may be necessary

In this work, deprivation measures are used as ranks.

Mining census data for spatial effects on mortality

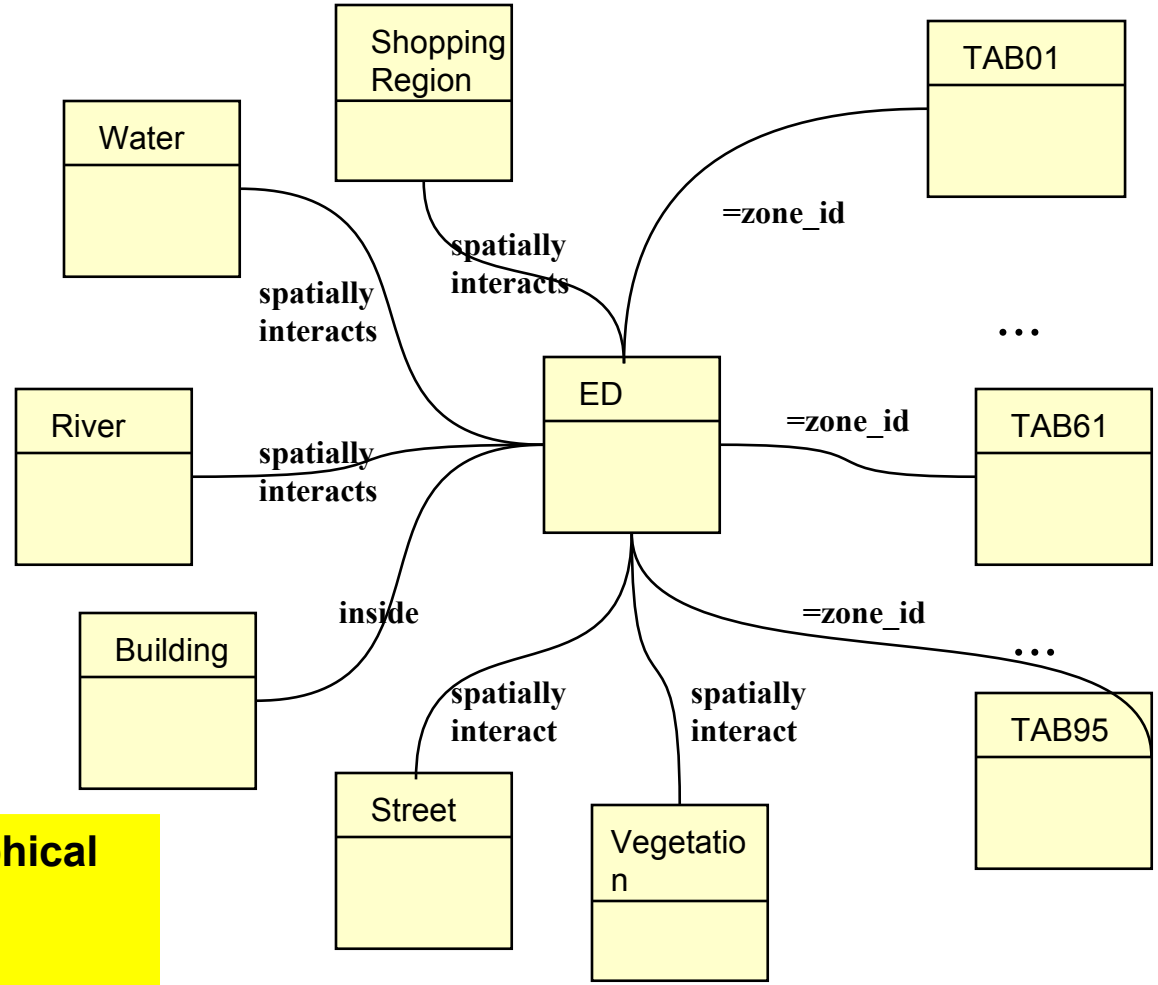
- Health data are provided by the UK Office of National Statistics on annual basis at ward level
- Mortality rate is selected as target variable for explorative study

Mining census data for spatial effects on mortality

- Issue: Mortality rate clearly depends on age and on geographical factors (a ward with a hospital will present a higher number of deaths)
  - Use a weighted sum of mortalities over age categories → standardised mortality ratio (SMR)

In this work, SMR calculation to those aged 0-74 is considered.

Database Schema



Attribute data

95 tables with census data,

~8000 attributes

Spatial Hierarchy

- County
- District
- Wards
- Enumeration district

Geographical Layers

85 tables

Mining census data for spatial effects on mortality

- Data Mining Task: subgroup mining.
- Subgroup mining is used to analyze the dependencies between a selected target variable and a large number of explanatory variables.
- Interesting subgroups with some designated type of deviation, change, or trend pattern are searched.
- In this work, a subgroup is a subset of target objects (e.g., wards with high mortality rate) that is defined by conditions on variable (including those in secondary tables)

Mining census data for spatial effects on mortality

- The results support the assumption that both geography and deprivation are relevant (causal?) for high mortality and their interactions also are important.
- The interaction of these factors is linked to the highest levels of mortality, especially in Greater Manchester and Liverpool.

Mining georeferenced census data for urban accessibility

- The concept of "accessibility" appears initially in the context of geographical science and was progressively introduced in transport planning in the 1960's and 1970's.
- Many different definitions of accessibility and many ways to measure it can be found in the literature.
- In this work authors are interested in urban accessibility, which refers to local (inner city) daily transport opportunities.

A. Appice, M. Ceci, A. Lanza, F.A. Lisi, & D. Malerba (2003). Discovery of spatial association rules in geo-referenced census data: A relational mining approach, *Intelligent Data Analysis*, 7(6):541-566.

Mining georeferenced census data for urban accessibility

- A great effort has been made to define urban accessibility indices, which can be used to assess/compare transportation facilities within different regions of an urban area or between urban regions
- Accessibility is usually measured with respect to key activity locations for individuals (e.g., home, workplace) and evaluates the transportation services provided in these key locations to assess their relative advantages

Mining georeferenced census data for urban accessibility

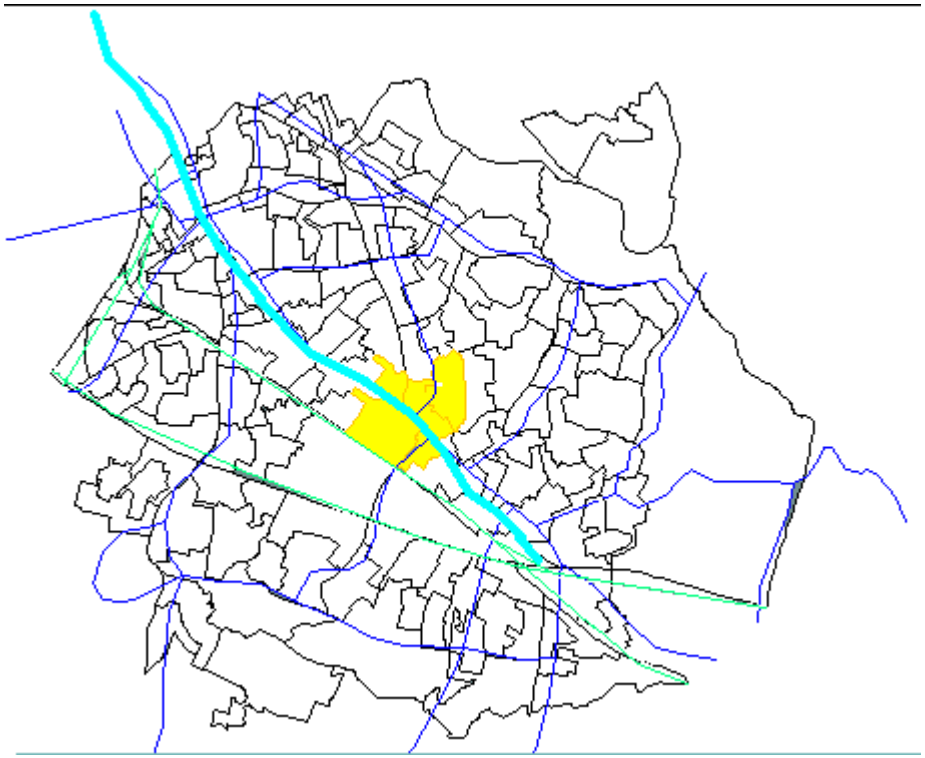
- In this work, authors are interested in the accessibility "to" the Stepping Hill Hospital "from" the actual residence of people living within in the area served by the hospital.
- Since (micro) data on the actual residence of each involved household are not available, we study the accessibility at the ED level.
- This study does not aim to synthesize a new accessibility index, but to discover human interpretable patterns that can also contribute to directing resources for facility improvement in areas with poor transport accessibility.

Mining georeferenced census data for urban accessibility

- Factors:
  - Geographical
  - Socio-economic
  - Transportation
- Data:
  - Census data (UK census 1991)
  - Spatial data (provided by Ordnance Survey)

Mining georeferenced census data for urban accessibility

- Spatial patterns relating 5 Stepping Hill EDs (task relevant objects) with 152 other EDs in a distance of 10Km from SH (reference objects)
- Which reference EDs have access to the task relevant EDs ?
- Use Ordnance Survey data on transport network (roads, railways and bus priority line)



Mining georeferenced census data for urban accessibility

- Main issues of this work:
  - Some background knowledge is available. How to express it? ➔ Use logic formalism.
  - Several runs for tuning two parameters (support & confidence) of the data mining task ➔ to speed-up execution, precompute spatial relations and represent them as facts in a logic formalism
  - Numerical census data cannot be directly handled by the algorithm for mining spatial data ➔ discretize them by means of a contextual discretization algorithm

Mining georeferenced census data for urban accessibility

An example of spatial association rule:

ed_around_stepping_hill(A), can_reach_only_by_road(A,B),
is_a(B,stepping_hill_ED) → no_car(A,[0.228..0.653]) (38.15%, 56.31%)

- This spatial pattern occurs in fifty-eight distinct EDs, thus
- From fifty-eight distinct EDs within a distance of 10Km from Stepping Hill Hospital (ed_around_stepping_hill), it is possible to reach the hospital only by road and the percentage of households with no car is quite high (between 22.8% and 65.3%).
- Social issue: The hospital can be reached by road (perfect accessibility from a graph-theory viewpoint) but … what about those households with no car?

Mining georeferenced census data for urban accessibility

1. 1991 Census data are now obsolete.
2. The crossing of a railway does not necessarily mean that there is a station in an ED. Similar considerations can be made for bus priority lines and roads.
3. Digital maps made available by the Ordnance Survey are devised for cartographic reproduction purposes and not for data analysis. Hence, a road may appear to be 'blocked' in the digital map, because it runs under a bridge.

Calculating economic household indexes

- Problem: having geographical, social and economic information about customers for CRM applications
- Data:
  - Population and Housing Censuses (Spain 1991)
  - Family Expenditure Survey (Spanish National Statistics Institute, published quarterly) for 300 products
- Potential applications:
  - Evaluate which censal sections in the country are more predisposed to expenditure on a given consumer product
  - Longitudinal study of expenditure patterns in a region for marketing analysis
  - Spatial analyis of expenditure patterns

S. Frutos, E. Menasalvas, C. Montes & J. Segovia (2003). Calculating economic indexes per household and censal section from official Spanish databases, *Intelligent Data Analysis*, 7(6): 603-613.

Calculating economic household indexes

- Issue: the two databases cannot be directly crossed
  - Census: no reference to families/households
  - FES: no reference to censal section
- How to estimate quarterly household economic indexes for Spanish Censal Sections? How to build a (linear) forecasting model for estimating the trend of each of the defined indexes?

Calculating economic household indexes

Many steps:

1. Group families surveyed in the FES on the basis of zip code or other proximity indicators
2. Calculate the socio-economic composition of each family group
   - percentage of families whose principal earner is male, percentage of families whose principal earner is female, …
1. Get the average indexes per family group
   - income, expenditure, investment, property, saving, debt indexes
1. Get estimation models for each index
   - Neural-network approach

Calculating economic household indexes

1. Calculate the socio-economic composition of each censal section
2. Get indexes per censal section by applying the learned models
3. Get the temporal evolution of the indexes

Calculating economic household indexes

An example for an expenditure index:

Averaged Electric Power Consumption per home (EPC), in pesetas. Comparison between official and model data. Model includes prediction for period 1997-2000

Mining Official Data

Some important issues:

- Integrating the data
- Communicating the results to the users/decision makers
- Consolidation of knowledge (KDD findings) for action as final step of the process / as the step completing the KDD process
  - This is more difficult for the public sector.
  - In the private sector, the goal is to increase the profit.
  - What is the scope in the public sector?
- Many stakeholders with different goals
- Confidentiality or Privacy-preservation

Mining Official Data

An overseas experience:

- The United States General Accounting Office is applying data mining for **internal** fraud detection
- Fighting fraud, waste and abuse in the use of government credit cards
  - First results in identifying prohibited purchases of goods and services at the Department of Defence (DoD)

Outline of the presentation

- Public Sector Information in the Information Society
- Mining official data
- eGovernance/eGovernment
- Environment
- Health Care

eGovernance

- eGovernance is about the use of ICT to improve the quality and the efficiency of all phases of the life cycle of legislation.
- *Governance* is **not** a synonym of *government*.
- Public institutions and governmental agencies are not the only actors involved in the process of governing society.
  - Other actors are press, political parties and lobbies, general public, non-governmental organizations, …

eGovernance vs. eGovernment

- eGovernment is about the use of ICT to support the work of governmental institutions and agencies
- eGovernance is about the use of ICT to support the guiding or steering of an organization to achieve its goals.
- In the political context, as a special case, eGovernance is about the use of ICT to steer society and promote public interests.

LKBS

- In eGovernance, computer models of legislation and other sources of norms play a central role
- Legal Knowledge-Based Systems (LKBS) are a particular class of computer models of legislation
- They are the evolution of the legal applications of rule-based systems for public administration (legal expert systems, LSE).

LKBS vs. LES

- LKBSs now include the use of all possible sources of legal knowledge
  - Original, authoritative legal texts (legislation, case law)

  In addition to formalized opinion of legal experts
- and all ways of modelling legal knowledge using computers, such as case.based reasoning methods or neural networks

  In addition to rule-based technology

Translating legal texts into a formal logic description

- Very ambitious goal (few prototype available)
- Data Mining + Natural Language Processing technologies can play an important role
- An initial work:

G. Lau, K.H. Law, & G. Wiederhold: Similarity Analysis on Government Regulations, *SigKDD'03.*

Goal:

Propose and validate a framework for regulation management and similarity analysis.

The on-line repository of legal documents created with the help of text mining tools

Translating legal texts into a formal logic description

- Issue: classical text mining techniques don't work
- They ignore the structure of regulations
  - Regulations are organized into deep hierarchies
  - Sections are heavily crossed-referenced
  - Terms are well-defined within regulations

This work uses a document representation based on concepts and key phrases (no bag-of-words) to capture sequencing information on words.

The list of concepts is extracted with the software tool Semio Tagger.

Translating legal texts into a formal logic description

Original Section 4.6.3 from the UFAS

4.6.3 Parking Spaces

… at least 96 in … and an adjacent access aisle...

EXCEPTION: If accessible parking spaces for vans...

Refined Section 4.6.3 in XML format

```
<regElement name="ufas.4.6.3" title="parking spaces">
  <concept name="access aisle" num="3" />
  <indexTerm name="accessible circulation route" num="1" />
  <measurement unit="inch" size="96" quantifier="min" />
  <ref name="ufas.4.5" num="1" />
  …
  <regText> Parking spaces for disabled people ... </regText>
  <exception> If accessible parking spaces for ... </exception>
</regElement>
```

An example of government regulation with the complete set of feature mark-ups that shows *exception*, *measurement*, *ref*, *concept* and *indexTerm* tags in addition to the body text *regText* tag

Translating legal texts into a formal logic description

- Subsequent retrieval is based on a similarity analysis that identifies relevant provisions by utilizing both the hierarchical structure of regulations and referential structures.
- Similarity measure is based on cosine correlation typically used in information retrieval.
- Long run goal: to perform automated analysis of overlaps, completedness and conflicts.

Outline of the presentation

- Public Sector Information in the Information Society
- Mining official data
- eGovernance/eGovernment
- Environment
- Health Care

Environmental management

- A typical national environmental protection agency aims to "*protect public health and to safeguard and improve the natural environment*".
- It sets and enforces national pollution control standards and performs …
- **Environmental monitoring =** periodic or continuous surveillance or testing to determine the level of compliance with statutory requirements and/or pollutant levels in various media or in humans, plants, and animals

  (U.S. EPA Terms, 2000)

Environmental monitoring and protection

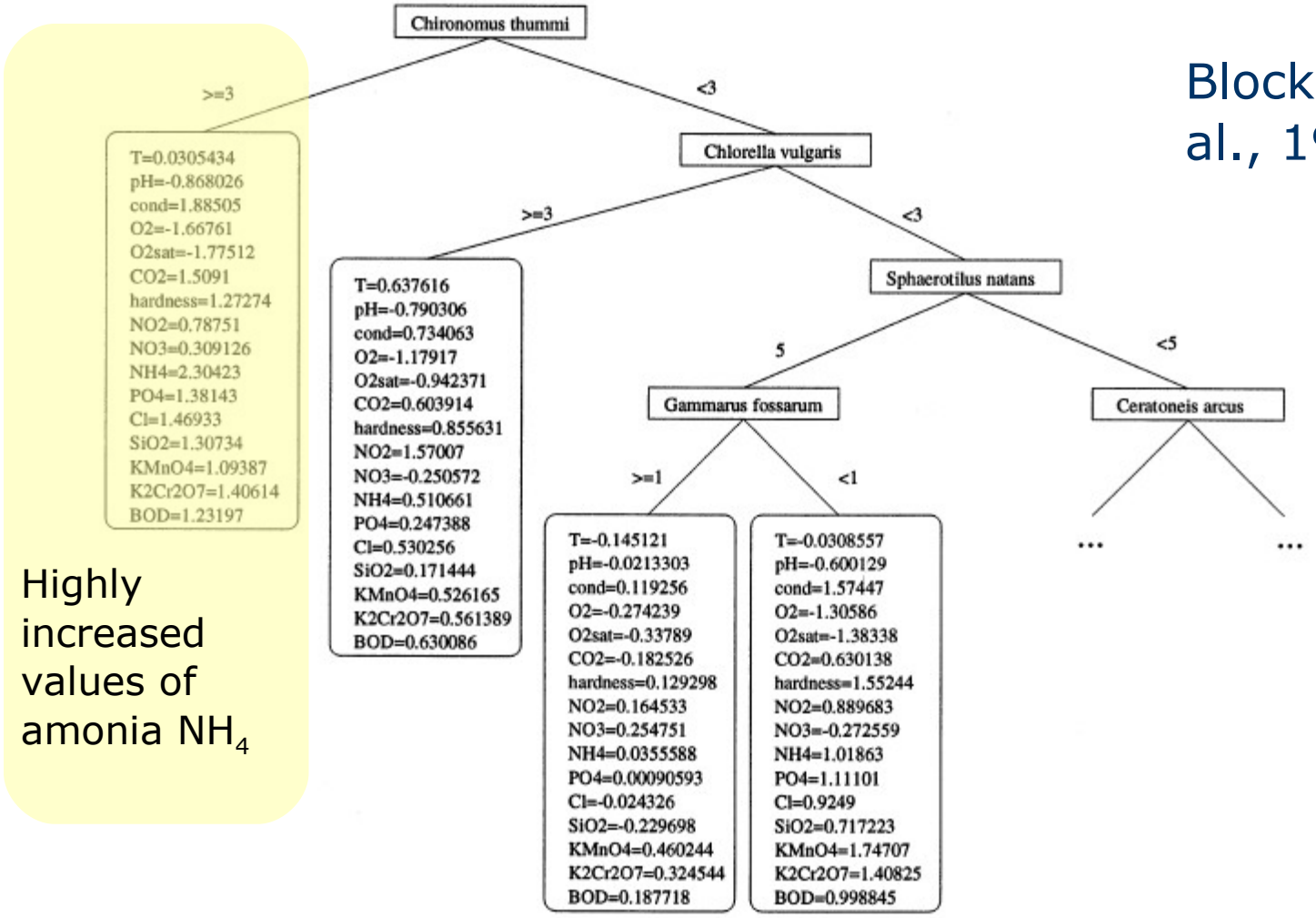
- Problem: interpret and classify samples of river water quality
- Factors:
  - Physical and chemical properties ➔ river water quality at a particular point in time
  - Biota (living organisms) ➔ general picture of water quality over a period of time
- Goal: inferring the chemical properties from the biota
  - Monitoring chemical compounds is expensive
  - In many countries extensive biological monitoring is conducted
- Data:
  - Biological and chemical samples from Slovenian rivers.

S. Dzeroski, J. Grbovic, & D. Demsar. Predicting chemical parameters of river water quality from bioindicator data. *Appl. Intell*. 13, 7–17, 2000.

Environmental monitoring and protection

- Biological data: species/taxa present at the site and their abundance
- Chemical data: measured values of 16 physical and chemical parameters
  - Biological oxigen demand (BOD)
  - Chlorine concentration (Cl)
  - CO_2 concentration
  - …
- Issues:
  - Varying length data records: depending on the site and water quality, the number of taxa present can vary. ➔ Methods for handling structural information
  - Making multiple predictions: most machine learning methods for prediction only deal with one target variable. It might be beneficial to try to predict several interrelated variables simultaneously.

Environmental monitoring and protection



Blockeel et al., 1998

Highly increased values of amonia NH_4

A clustering tree for simultaneous prediction of multiple chemical parameters

Environmental management

- **Ecological modelling =** development of models of the relationships among members of living communities and between those communities
- These models can then be used to support decision making for environmental management
- Modeling topics:
  - Population dynamics of several interacting species
  - Habitat suitability for a given species

Ecological Modeling

- Population dynamics
  - Behaviour of a given community of living organisms (population) over time
  - Factors considered: abiotic (concentrations of nutrients/pollutants,etc.)
  - Modelling formalism used by ecological experts: differential equations
  - Relationships among living communities and their abiotic environment can be highly nonlinear. ➔ neural networks & systems for discovery of differential equations

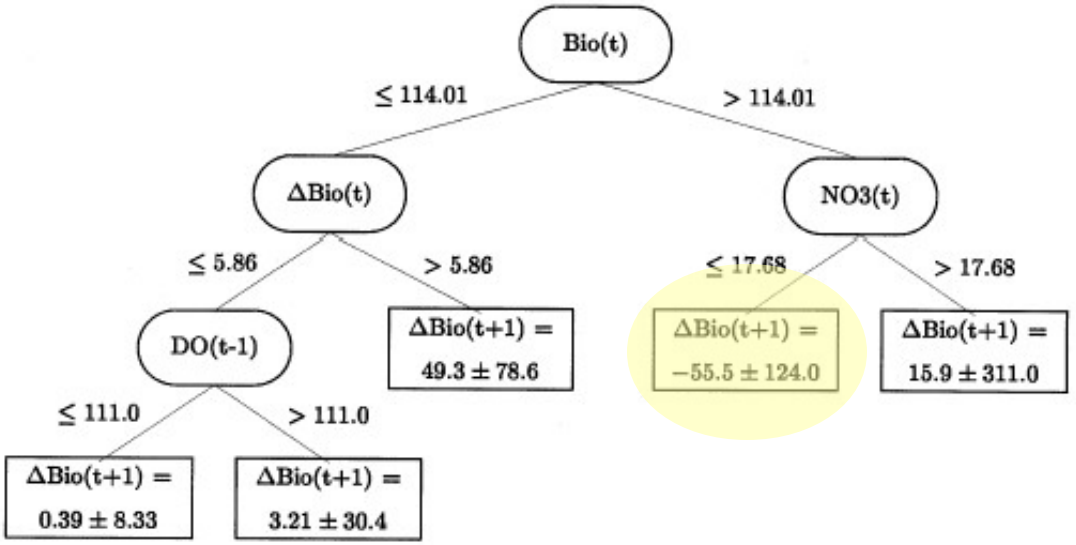
Ecological Modeling

- Problem: modeling algal growth in the Lagoon of Venice
- Factors:
  - Water temperature
  - Pollutants ➔ food (nutrients) for algae
  - Time
- Goal: Would a reduction in the use of fertilizers (phosphorous) reduce the growth of the dominant species of algae?
- Data:
  - Biomass (Bio), water temperature, dissolved nitrogen (NO_3) and phosphorous, dissolved oxygen (DO).

B. Kompare, S., Dzeroski, & V., Krizman. Modelling the growth of algae in the Lagoon of Venice with the artificial intelligence tool GoldHorn. In: Proc. Fourth International Conference on Water Pollution. Computational Mechanics Publications, Southampton, pp. 799–808, 1997.

Ecological Modeling

- Issues:
  - Measurement errors of algal biomass
  - Data on wind is missing (wind might move algae away from the sampling stations).



The regression tree for predicting algal growth shows that nitrogen is the limiting factor (negative answer to the original question).

Outline of the presentation

- Public Sector Information in the Information Society
- Mining official data
- eGovernance/eGovernment
- Environment
- Health Care

Health Care

- As in the case of environmental management, KDD can be widely used to increase scientific knowledge concerning diseases, effect of therapies, etc. from *laboratory data*
- Health care also generates mountains of *administrative data* about patients, hospitals, bed costs, claims, etc.
- **Focus**: KDD applications to such administrative data that provide more cost effective quality health care.

Health Care

- Problem: Preterm Birth Prediction

  Preterm birth, before 37th week of gestation

  Final goal: identify factors that will improve the quality and cost effectiveness of perinatal care
- Factors:
  - demographic ➔ such as age, race, education, religion and marital status
  - clinical
- Goal: predicting preterm or full term delivery
- Data:
  - Duke Univerity's TMR perinatal database.

    71,753 records and approximately 4,000 potential variables per patient.

L. Goodwin et al. Data Mining issues for improved birth outcomes. *Biomed. Sci. Instrumentation*. 34, 291–296, 1997.

Health Care

- Interesting finding: Best results found that seven demographic variables yielded .72 and addition of hundreds of other clinical variables added only .03 to the area under the curve (AUC).
- Demographic variables may offer a small set of low cost variables with predictive accuracy in a racially diverse population.
- Several learning methods tested:

Model	(ROC) Area Under Curve
Neural Net - Demographic Only	0.64
Logistic regression - All variables	0.66
Rule induction - All variables	0.67
Custom classifier software - Demographic variables only	0.72
Custom classifier software - All variables	0.75

Health Care

- Issues:
  - Data quality: many missing values, inconsistent data (clashes)
  - Different cost for false positive and false negative: the lack of needed service could have a negative impact on both patients outcomes and costs.
  - Large volumes of data